Synthese (2021) 198:10847–10882
https://doi.org/10.1007/s11229-020-02755-4

Affective experience in the predictive mind: a review and new integrative account

Pablo Fernandez Velasco[1] · **Slawa Loev**[1,2,3]

Received: 22 August 2019 / Accepted: 17 June 2020 / Published online: 29 June 2020

Abstract

This paper aims to offer an account of affective experiences within Predictive Processing, a novel framework that considers the brain to be a dynamical, hierarchical, Bayesian hypothesis-testing mechanism. We begin by outlining a set of common features of affective experiences (or feelings) that a PP-theory should aim to explain: feelings are conscious, they have valence, they motivate behaviour, and they are intentional states with particular and formal objects. We then review existing theories of affective experiences within Predictive Processing and delineate two families of theories: Interoceptive Inference Theories (which state that feelings are determined by interoceptive predictions) and Error Dynamics Theories (which state that feelings are determined by properties of error dynamics). We highlight the strengths and shortcomings of each family of theories and develop a synthesis: the Affective Inference Theory. Affective Inference Theory claims that valence corresponds to the expected rate of prediction error reduction. In turn, the particular object of a feeling is the object predicted to be the most likely cause of expected changes in prediction error rate, and the formal object of a feeling is a predictive model of the expected changes in prediction error rate caused by a given particular object. Finally, our theory shows how affective experiences bias action selection, directing the organism towards allostasis and towards optimal levels of uncertainty in order to minimise prediction error over time.

Pablo Fernandez Velasco and Slawa Loev have contributed equally to this work. The names of the authors are in alphabetical order.

✉ Pablo Fernandez Velasco
p.fernandezvelasco@gmail.com

✉ Slawa Loev
slawa.loev@gmail.com

1 Institut Jean Nicod, Département d'Études Cognitives (DEC), ENS, EHESS, CNRS, PSL University, 29 rue d'Ulm, 75005 Paris, France

2 Munich Center for Neuroscience, Ludwig Maximilian University, Munich, Germany

3 Faculty of Philosophy, Ludwig Maximilian University, Munich, Germany

Keywords Affective experience · Feelings · Emotion · Valence · Predictive processing · Affordances

1 Introduction: affective experiences

What do a tired toddler, a relaxed jazz player, a lascivious llama, a scared cat, a hopeful rebel, a confused voter and a confident nurse have in common? They all are in the grip of an affective experience—assuming that they *feel* their respective states. Affective experiences or feelings[1] are a rich class of phenomena. Nevertheless, we find significant commonalities that affective experiences share among themselves and not with other mental states, distinguishing them as a class. These commonalities constitute the explananda that a good theory of affective experiences needs to account for. In this paper we aim to develop such an account with the help of Predictive Processing (henceforth, PP), a novel framework that aims to provide a unifying vision of the brain, which will be introduced in the next section. In the present section, we will start by outlining the commonalities between feelings. These commonalities run through different strands of philosophical and empirical work on affective states.

First off, feelings are conscious in that they are felt.[2] They are conscious states in the same way perceptual *experiences* are conscious states. We seek to make this clear by referring to feelings also as affective *experiences*.[3] Moreover, feelings are *phenomenally* conscious, i.e. there is "something it is like" to have an affective experience. Feelings, in particular, have an *affective* phenomenal nature. A central aspect of this nature is *phenomenal valence* (Barrett and Bliss-Moreau 2009). Phenomenal valence refers to the aspect of felt positivity (e.g. feeling relaxed) or negativity (e.g. feeling afraid) (Charland 2005; Colombetti 2005) and is often understood in algedonic terms such as (un)pleasantness or evaluative terms such as seeming (dis)value (Carruthers 2017; Teroni 2018). It is worth emphasizing that when we speak of valence here, we mean to refer to valence as a phenomenal property, i.e. to felt or experienced valence.[4]

[1] We use the terms "affective experiences" and "feelings" interchangeably.

[2] It is, thus, not enough to just point to states which have a similar impact as feelings but are not felt (Lacewing 2007, p. 97; see also Winkielman et al. 2005).

[3] It seems largely a matter of definition that feelings are conscious: "That there can be no unconscious feelings is still the position of "commonsense."" (Lacewing 2007, p. 98; see also Clore 1994). However, this is not to say that there are no unconscious emotions or, more generally, unconscious states that are functionally largely analogues to feelings or sub-classes of affective experiences like emotional feelings (Winkielman and Berridge 2004; Lacewing 2007). For perceptual experiences it seems plausible that there are forms of unconscious perception and something similar might be said for affective experiences (say, there might be unconscious affective reactions) (Prinz 2005). However, when making this comparison the aspect of interest is not that perceptual experiences are a form of perception (as is unconscious perception) but that they are a form of experience, and it is a conceptual fact about experiences that they are conscious. Affective experiences should also be distinguished from emotional or affective *processing* which often does occur unconsciously (see e.g. Mathews and MacLeod 2002). Apart from terminological rationale, there is good reason to believe that fine-grained affective experiences require conscious awareness (Pessoa 2005).

[4] This phenomenal quality often but not always correlates with closely associated but ultimately non-phenomenal properties such as *object* valence (Colombetti 2005). It is also worth mentioning that the idea of "unconscious valence" is lately gaining ground (e.g. Berridge and Kringelbach 2015). Unconscious valence has a functional profile similar to phenomenal valence in motivating behaviour. Although we think

Valence is often regarded as the *mark of the affective* (Charland 2005; Barrett 2006), that is, the feature that distinguishes characteristically affective phenomena from other non-affective phenomena.[5]

By way of example, the relaxed jazz player and the hopeful rebel mentioned earlier would feel positive valence. In contrast, the mentioned scared cat and the confused voter would undergo an affective state with negative valence. Phenomenal valence plausibly grounds the evaluative dimension characteristic of affective experiences (e.g. Helm 2009; Bain 2013; Deonna and Teroni 2017). For the relaxed jazz player and hopeful rebel something is going well while there is something amiss for the scared cat and the confused voter.

This evaluative dimension, in turn, is closely associated with another important feature of affective experiences: they are motivational, in that they exert a motivational push and directly motivate behaviour and action (Corns 2014; Scarantino 2014; Carruthers 2017; Kozuch 2018). Generally speaking, the kinds of behaviours that feelings motivate can be described as aversive (cessation, avoidance) or appetitive (continuation, approach) in nature (Corns 2014). And while it is at times not obvious which specific behaviours some feelings mandate, often the motivated behaviours are quite specific. The motivational force of feelings also reaches into the future: they do not only motivate actions in the moment of their occurrence but make certain behaviours more or less likely in the future. For instance, if you stand close to a fire you might experience pain which will not only motivate you to recoil in the moment but will also make it less likely that you will stand that close to a comparable fire in the future (Kozuch 2018).[6] The motivational force and nature of the motivated behaviour is in part determined by the valence of the experienced affective state. This is underlined by the fact that the polarity and intensity of valence usually map onto corresponding dimensions of the motivational push: strongly/weakly (valence intensity) positive/negative (valence polarity) feelings tend to exert a strong/weak motivational push (motivation intensity) towards appetitive/aversive behaviours (motivation polarity) (Kozuch 2018).

Although there is a clear link between valence and the motivation of behaviour, one should be wary of making this link too tight. There are feelings that do not seem to display this tight connection. Affective states such as guilt could appear as negatively valenced as fear (i.e. a highly motivating feeling) but have a weaker link to specific behaviour (Schroeder 2004). We find more striking examples in the context of addiction, where valence and motivation sometimes dissociate (Corns 2014). A heavy smoker regularly has cigarettes that feel unpleasant, but they are highly motivated to smoke despite the affective experience of smoking being negatively valenced.[7] Never-

Footnote 4 continued
there might be some problems with this idea, arguing the point would take us too far afield. Here, our focus are affective *experiences* and we are thus concerned with phenomenal valence. As we will see, however, a virtue of the emerging picture is that it can incorporate the idea of unconscious valence.

[5] This conception of the mark of the affective echoes the conception of the mark (or marks) of the mental, the feature (or set of features) "that set characteristically mental phenomena apart from the characteristically physical phenomena" (Pernu 2017, p. 1).

[6] This also points to the connection between affect and learning, an issue to which we will return in Sect. 4.

[7] Note that it is not only that the (weakly) negative valence does not lead to a weakly negative motivation to avoid cigarettes—we rather observe that there is a strong, positive motivation in place that is poorly explained by the implicated negative valence.

theless, the motivational nature of feelings is something that a good theory of affective states must be able to account for. If valence brings to the fore the evaluative dimension of affectivity, the link between affective states and the motivation of behaviour brings to the fore its regulative dimension.

Finally, feelings are not only phenomenal but also *intentional* states (e.g. Tye 2008; Goldie 2002; Kriegel 2014).[8] They are "about" or are "directed at" something. Feeling pain in one's wrist is about (the bodily events in) one's wrist, being afraid of James' famous bear is about the bear. The individual objects feelings are directed at are traditionally called the intentional or *particular object* of a feeling. Crucially, affective experiences cast their objects in a specific light. The pain is not only about one's wrist but also about its being unpleasant. The fear is not only about the bear but also about its being dangerous. The feeling, plausibly in part via its valence, signals a feeling-specific property exhibited by its particular object. This property specific to a feeling is traditionally called its *formal object* or core relational theme (Kenny 1963; de Sousa 1987; Lazarus 1991; Prinz 2004). To summarise: a feeling represents its particular object as having a feeling-specific property, its formal object.

Note that valence emerges as a property of affective experiences that does not only intuitively distinguish them from other kinds of mental states but also pulls their other features naturally together: First, it (partially) grounds the way in which feelings are evaluative, supplying a general positivity or negativity which is then further specified in the form of feeling-specific formal objects. Second, valence seems (in part) responsible for why affective experiences motivate behaviour and thus can fulfil their regulative function. These considerations make valence appear as an excellent candidate for the mark of the affective.

In this section, we have outlined a set of commonalities of affective experiences, thereby introducing the explananda that a PP account of affective experiences should aim to explain. Namely, we have highlighted that affective experiences are phenomenally conscious and valenced, that they motivate behaviour, and that they are intentional states with a particular and a formal object. After introducing PP in Sect. 2, we will look at different theories that have been proposed within the PP framework to account for affective experiences. We divide these into two families of theories: Interoceptive Inference Theories and Error Dynamics Theories. In Sect. 3 we discuss Interoceptive Inference Theories and argue that while they provide a good mechanism within which to understand affective processes, they face some challenges when it comes to accounting for what we deem the *mark of the affective*, i.e. that which distinguishes affective states from other classes of states and processes (e.g. non-affective bodily sensations). We posit that a computational account of valence might be able to address this challenge in a way that remains compatible with Interoceptive Inference Theories. In Sect. 4, we discuss Error Dynamics Theories and argue that while they manage to account for the mark of the affective through their computational conceptualisation of valence, they are incompatible with one of the central aspects of affective experiences: that affective experiences are phenomenally conscious. In Sect. 5, we provide a synthesis of Interoceptive Inference Theories and Error Dynamics Theories in an effort

[8] There are differing views as to whether or not moods (which we take to be affective experiences) should be considered intentional states. This is an issue that we address in Sect. 5.1.

to account for the commonalities of affective experiences. According to this revised account, valence corresponds to the expected rate of prediction error reduction. This, in turn, provides us with the mark of the affective in PP terms. Section 5.1 elaborates on the intentionality of feelings by building on the notion of regularity within PP we argue that the particular and formal object of a feeling emerge as the inferred hidden causes of variation in predictive success. In Sect. 5.2, we extend the emerging account to explain how affective states motivate and guide behaviour, building on previous PP work on action policy selection. Section 6 recapitulates the central aspects of our account and shows how it provides an explanation for the central elements of affective experience.

2 Predictive processing

The main idea behind the PP framework is that the brain is a dynamical, hierarchical, Bayesian hypothesis-testing mechanism.[9] Based on prior information, the brain actively generates predictions about the world by using an internal model. These predictions go top-down[10] (i.e. from abstract levels down to sensory organs and effectors) and side-ways (i.e. laterally across a given level). In turn, the errors of these predictions (prediction error, henceforth Error) go side-ways and bottom-up and are used to update the system's predictions in a continuous feedback loop. The overall goal of the system is to minimise Error over time (Prediction Error Minimisation, from now on PEM).

Perception is then conceived as the process of continuously explaining away Error through ongoing predictions in order to successfully represent the world, understood as the hidden causes of changes in sensory input. Each layer in the hierarchy tries to predict the input of the layer below, using models developed to capture regularities in the variation of sensory signals. Sensory receptors capture the mismatch between this cascade of predictions and incoming sensory input (i.e. Error) and the Error that cannot be explained away solely by lower layers adapting their predictions travels upward in the hierarchy, through forward error signals (Friston 2003).

In a unifying vision, PP can be seen as the result of the free-energy principle, which states that self-organising systems must minimize (variational) free-energy (Friston 2010; for a discussion of pluralistic and unifying versions of PP, see the recent review in Hohwy 2020). Free-energy is the quantity that bounds the evidence for a model of data (i.e. in the case of organisms, free-energy is greater than the surprise in sensory data). In this paper, we will generally discuss things in terms of Error instead of free-energy, because a fundamental assumption of PP is that free-energy corresponds to the amount of prediction error (Friston 2009, p. 293). Expected free energy, then, becomes the tractable bound on long-term Error, and PEM corresponds to the organism minimising expected free-energy (Friston et al. 2015).

[9] PP is a process theory whereas the Bayesian brain is a normative principle, a clarification suggested by an anonymous reviewer.

[10] A recent review introduces the possibility that context-independent structural components constraining bottom-up processing constitute a form of bottom-up predictions (Teufel and Fletcher 2020). It is still early to tell if (and how) such a conception would change the PP framework.

In PP, Error is reduced either by updating predictions or by acting on the world in order to produce or sample the predicted sensory input (Friston et al. 2011; FitzGerald et al. 2012). Against this background, an action happens in order to fulfil an emergent proprioceptive prediction. For example, in order for an agent to open a door, the system predicts (at a high level) the required movement to open the door and a cascade of lower level predictions ensues (e.g. predictions about the movement of the hand, the fingers, etc.) The movement of the agent, then, happens to minimise the Error of the counterfactual proprioceptive predictions corresponding to the opening of the door. Thus, in the standard PP account of action, we move "from predicting the rolling present to predicting the near-future, in the form of the not-yet-actual trajectories of our own limbs and bodies […] Predicting these (non-actual) sensory states actually serves to bring them about." (Clark 2015, p. 112). The idea that action occurs to fulfil (proprioceptive) predictions implies that both action and perception are part of the same process of minimising Error over time (commonly referred to as active inference). One of the upshots of this idea is that it "separates the problems of optimizing action and perception by assuming that action fulfils predictions based on inferred states of the world" (Friston et al. 2016, p. 864).

The system uses Error to constantly update its predictions, but not all Error is equally reliable. For example, the Error coming from stimuli with a high level of noise (e.g. a poorly lit room) is less reliable than the Error coming from stimuli with a low level of noise (e.g. a well-lit room), because the Error of the former is more likely (relatively) to be due to noise than to the inaccuracy of the current prediction. Thus, not all Error has the same weight when it comes to updating hypotheses[11] about the world. The errors coming from sources that are expected to have a high variance are assigned a lower weight compared to the errors coming from sources that are expected to have a low variance. The inverse of variance is called precision (which quantifies the degree of certainty about the signals; see Feldman and Friston 2010 for a Free Energy formulation), and the brain, in addition to first-order predictions, is constantly trying to predict the precision of forthcoming Error. Precision serves then as a form of weighting when the system is deciding the degree to which the current hypothesis should be updated in light of incoming Error—highly precise Error will play a larger role in driving hypothesis updating. Accordingly, the brain is always trying to optimise precision, which is conducive to Error minimisation over time. In the PP literature, this process of precision-optimisation is what defines attention (Hohwy 2012).

PP is particularly compelling in explaining the workings of attention and how attention relates to consciousness. Hohwy suggests that "conscious perception is determined by the prediction or hypothesis with the highest overall posterior probability" (Hohwy 2012, p. 3). The posterior probability quantifies how probable the prediction is given the observed evidence (i.e. the actual sensory input). In other words, the brain is continuously computing and evaluating a plethora of hypotheses about the world,

[11] Although the constructs of prediction, hypothesis and inference are often used interchangeably in the PP literature, it is helpful to briefly clarify their relations and differences. Technically, a hypothesis is a joint set of predictions. Inference is the updating of the prior to the posterior in the light of the prediction and the prediction error. Predictions are expected states based on Error and previous experience while inference is the process of predicting states based on Error and previous experience. That is, inference is the process that yields new predictions.

and the hypothesis that is deemed to be the most probable (i.e. the most likely state of the world) determines what the subject is consciously experiencing. Attention has a strong effect on the content of consciousness because it biases the competition between different hypotheses (by weighting the Error that will drive hypothesis updating), so that whichever is the most probable hypothesis determines conscious experience.

3 Interoceptive inference theories of affective experience

Within PP two families of theories of affective experience have been proposed so far: the first can be called "Interoceptive Inference Theories" (IIT) (e.g. Hohwy 2011; Seth et al. 2012; Seth 2013, 2015; Seth and Friston 2016; Gu et al. 2013; Barrett and Simmons 2015). The second we call "Error Dynamics Theories" (EDT) (Van de Cruys and Wagemans 2011; Van de Cruys 2017; Joffily and Coricelli 2013). We will discuss both in turn.

IIT takes as its point of departure the fact that the central imperative of a biological system is to ensure its survival. Arguably, the most immediate determinant of survival is one's ability to achieve and maintain physiological integrity or homeostasis, i.e. to keep the body's vital parameters within viable bounds (Seth 2015). Consequently, *all* PP is "ultimately in the service of maintaining organismic homeostasis" (Seth 2015, p. 9; see also Barrett 2017). The activity of achieving homeostasis through physiological and behavioural change, is called allostasis. Interoception, the sense of the internal physiological condition of the body, is linked particularly closely to allostasis because it monitors homeostatic change. Against this background, IIT was first developed as a PP-theory of interoception and then taken to naturally extend to affective experiences.

It is worth noting that the idea that interoception is a key determinant of emotion can be traced back to at least Aristotle and the Stoics (Aristotle 1986; Gill 2010) and has been famously developed by James (1884) and Lange (1887) into an influential view that counts prominent recent proponents (Damasio 1994; Damasio and Carvalho 2013; Prinz 2004, 2006; Craig 2003, 2009; Critchley et al. 2004). What IIT brings to the table is the idea that interoception is not a feedforward process (as e.g. in Craig 2003, 2009), but part of the larger PP machinery. The initial idea of IIT was that interoceptive experiences—among which we also find affective experiences according to IIT—result from inference to the causes of interoceptive signals. In Seth's own words:

> On this theory of interoceptive inference […], emotional states (i.e., subjective feeling states) arise from top-down predictive inference of the causes of interoceptive sensory signals […] In direct analogy to exteroceptive PP, emotional content is constitutively specified by the content of top-down interoceptive predictions at a given time (Seth 2015, p. 9).

Note that against this background, *changes* in interoceptive signals will lead to Error that will have to be accounted for by adjusting one's predictions in order to accommodate the Error-generating changes in interoceptive signals. There is something appealing about this idea: it offers a neat account for the intentionality of affective experiences in analogy with exteroceptive experiences. As we have outlined

before: affective experiences are about particular objects. If one becomes afraid of an approaching bear, it is the bear that is the particular object of one's fear.

Here one might initially worry that by appealing to interoceptive signals, IIT risks to be overly "bodycentric", a criticism often levied against non-PP versions of somatic representationalism about affective experiences (as e.g. found in Tye 1995). However, the PP framework in which IIT operates can easily rise to this challenge: It is the *causes of* interoceptive signals that interoceptive predictions represent. These *can* be bodily (or interoceptive) events as in the case of hunger or pain. However, not only bodily events influence interoceptive signals but plausibly also events or things in the world (or the mind), such as (imagined) bears.

One might now wonder: How do non-bodily events get to change interoceptive signals so as to be inferred as their causes in the first place? Low blood glucose levels or physiological damage are clearly integrated within the interoceptive system—but bears don't seem to be. Before we turn to this question of (extrasomatic) distal causation of affective experiences, it is helpful to consider another question that will bring IIT's answer in clear relief. Here, we're after the *mark of the affective* in the predictive mind: we are looking for something that demarcates affective from non-affective experiences within the PP machinery. Outside of PP-theorising, valence seems the most promising candidate for the mark of the affective. Thus, what we are looking for is an account of valence in IIT terms. Now, is there a mark of the affective according to IIT? Or: is there a plausible idea of what valence could be? And if yes, what is it? What distinguishes affective states from non-affective states in IIT, if anything?

Obviously, the mark of the affective cannot simply be that we are engaging in predictive causal inference—since this kind of inference also undergirds e.g. visual experiences that are intuitively non-affective. The remaining candidate is that the inference is *interoceptive* rather than exteroceptive. In fact, this seems to be what proponents of IIT have in mind: "Emotion or affective content […] becomes an attribute of any representation that generates interoceptive predictions" (Seth and Friston 2016, p. 5). This indeed distinguishes interoceptive predictions from e.g. visual predictions. Thus, IIT's mark of the affective is the presence of predictions that are *interoceptive*. So as soon as there are predictions about the causes of *interoceptive* signals, we experience affective experiences, comprising bodily and emotional feelings.[12] Hypotheses that are co-constituted by predictions resulting from interoceptive inference will have a valence.

Now it is true that affective experiences often appear to have a bodily component, just think of bodily pain and orgasms. The problem with this proposal, however, is that not everything that has a bodily component seems to be an affective experience. That is, we seem to have experiences that are plausibly construed as interoceptive and grounded in interoceptive prediction but that are not affective. Examples are feeling

[12] Two points need to be qualified here: (1) We *experience* affect only if the interoceptive predictions in question additionally meet the criteria for supplying the contents of conscious experience. There might be interoceptive predictions that do not meet these criteria, such as those that are computed but discarded because competing predictions are more successful in minimising Error. (2) These interoceptive predictions are normally not all that there is to the affective experience. As affective experiences are usually multimodal, they are co-constituted by predictions from other modalities (e.g. those underlying the visual experience of the bear). It is, however, the interoceptive bit that makes the multimodal experience in question an *affective* experience.

one's heartbeat, stomach at work or breathing in and out. It appears plausible to draw a distinction between such bodily *sensations* and bodily *feelings* such as the mentioned bodily pains and orgasms. One can thus distinguish between cases of "hot" affective interoception (i.e. bodily feelings) and "cold" non-affective interoception (i.e. bodily sensations) (cf. Proust 2015, p. 20 sq.; Gerrans 2015).

The need for such a distinction is suggested by a popular view of bodily pain, which decomposes bodily pain into a sensory-discriminative and an affective-motivational component (Melzack and Wall 1988; Auvray et al. 2010). Moreover, we observe dissociations between these components (e.g. Rubins and Friedman 1948; Berthier et al. 1988; Corder et al. 2019). In pain asymbolia, for instance, subjects show a striking indifference to pain while appearing fully aware of the pain, or at least of its sensory-discriminative aspect (Bain 2014; Klein 2015). On the other hand, pain affect without pain sensation has been documented in humans and rats (Ploner 1999; Uhelski et al. 2012).

The Cotard syndrome generalises this lesson to other affective experiences. It is known for producing the Cotard delusion where patients come to believe that they are dead. Something in the experience of Cotard patients gives rise and explains this delusion: Cotard patients display a general loss of affect, without impairment in their interoceptive awareness (Michal et al. 2014). In other words, Cotard patients' bodily sensations are intact but their conscious life has been purged of affective experiences (Gerrans 2015).

The emerging issue for IIT's initial mark of the affective is that it classifies states as affective that aren't plausibly so. Although interoceptive predictions are a natural component of affective experiences, they also seem to be part of non-affective experiences.[13]

However, recently IIT has been developed in a way that bears promise in accounting for the outlined challenge (see especially Seth and Tsakiris 2018; but also Seth 2015). The problem for the initially sketched version of IIT was that it gave us little guidance to distinguish between *different kinds* of interoceptive inference. Now Seth and Tsakiris' (S&T) development provides exactly this: generally, they distinguish between perceptual inference and active inference. Perceptual inference is inference to the (hidden) causes of sensory signals. It serves Error minimisation by updating predictions based on incoming Error. Active inference, on the other hand, minimises Error by means of issuing actions.[14] Importantly, it can do so in two ways: actions might be directed at selectively sampling information to enhance the models of the predictive system. S&T call this kind of active inference (active) *epistemic inference*.

[13] A way of resolving this issue might be to draw the line between affective and non-affective (interoceptive) experiences by appealing to the implication of deep (i.e. high-level) expectations associated with an individual's concerns or goals, arguing that it is specific to affective experiences (cf. Seth and Friston 2016, p. 5). An initial worry with this idea is that there seem to be affective experiences, such as itches or the enjoyment of one's favourite music, where it is not clear from which high-level expectations they receive support. The suggestion, furthermore, poses the following question: why would the mentioned deep expectations be specific to interoceptive inference? It seems that *all* kinds of inference should be able to be supported by expectations at higher levels of the hierarchy, and, in fact, they regularly are. Thus, the implication of high-level expectations expressive of concerns does not appear like an IIT-friendly mark of the affective. Thanks to an anonymous reviewer for making us think about this point.

[14] Actions themselves are understood as predictions that dictate bodily behaviour, see Sect. 2.

Then there is what they call (active) *instrumental inference*. In instrumental inference, actions are performed to exert predictive control over sensory variables in order to bring them in line with the system's predictions.

These conceptual resources can now be put to work to distinguish between different kinds of *interoceptive* inference. S&T specifically point towards *instrumental interoceptive inference* which, rather than inferring the causes of interoceptive signals (which could be called "perceptual interoceptive inference"), are in the business of model-based *regulation and control* of interoceptive variables (notably, physiological essential variables). In other words, instrumental interoceptive inference directly brings about physiological homeostasis and thus implements the most immediate kind of allostasis. It does so by issuing initially counterfactual interoceptive predictions that result in the engagement of autonomic reflex arcs ("intero-actions") and potentially other allostatic actions which, in turn, bring about the ultimately self-fulfilling predictions.

S&T suggest that instrumental interoceptive inference plausibly gives rise to emotional (and other self-related) experiences:

> [It] is not enough to say that emotional and self-related experiences are the way they are (and are different to, for example, visual experiences) because they emphasise predictions about interoceptive (rather than visual) signals. Instead, it is helpful to consider the nature of predictions associated with interoceptive inference, especially their control-oriented (instrumental) bias. (S&T, 6–7).

In this vein, one might try to draw the initially problematic distinction between hot and cold interoception precisely along the lines of instrumental interoceptive inference and other kinds of interoceptive inference. Importantly, we thereby identify a potential candidate for a refined mark of the affective according to IIT: not interoceptive predictions *simpliciter* but specifically *instrumental* interoceptive predictions make an experience affective. If there are interoceptive predictions geared towards regulating the internal milieu, we experience affect.[15]

Now we are also able to answer the question of extrabodily distal causation of affective experiences: How do non-bodily events get to change interoceptive signals so as to be inferred as their causes in the first place? Think about the fear of the approaching bear again. Spotting a bear will, among other things, recruit instrumental interoceptive inference (your models of a bear encounter mandate that you better prepare your body for the occasion) to issue regulative predictions that lead to changes in interoceptive signals. Perceptual interoceptive and non-interoceptive inference will now infer the bear (together with proximal interoceptive factors recruited by instrumental interoceptive predictions) as the cause of the (changes in) interoceptive signals. This is a story that illustrates the strength of the emerging picture.

However, an initial complication with instrumental interoceptive predictions as the mark of the *affective* seems to be that S&T themselves subsume seemingly non-affective experiences under the products of instrumental interoceptive inference as well. In fact, the explicit target of the article in which they develop the notion of instrumental interoceptive inference is not affective experiences but self-related expe-

[15] For two important qualifications see footnote 12.

riences or "experiences of embodied selfhood" among which S&T count: "experiences of being an embodied organism, experiences of mood and emotion, pre-reflective experiences of selfhood and 'mineness', explicit self-awareness, metacognitive insight, reflective self-awareness, and social aspects of selfhood" (S&T, 7). It is far from obvious that all these experiences are affective. According to S&T, however, something unites these experiences: they are all "grounded in processes of instrumental (control-oriented) interoceptive inference that underpin allostatic regulation of physiological essential variables" (10). It seems, thus, that predictions resulting from instrumental interoceptive inference are not as much intended as a mark of the affective than as a mark of *self-relatedness*.

Apart from this complication which could potentially be resolved (by e.g. arguing that all the mentioned experiences *are*, in fact, affective), we run into other problems when we take control-oriented interoceptive predictions as the mark of the affective. A challenge is posed by the fact that we often have affective experiences where there seems to be plausibly no need for predictive interoceptive control. We find such affective experiences for instance in the context of aesthetics. Consider an example that Jennifer Corns recently advanced against the IIT-congenial idea of valence—what she calls hedonic episodes or (dis)pleasure—as homeostatic utility or disutility:

> A feature is homeostatically controlled only if it is continually monitored and subject to stabilizing processes when errors are detected. It is implausible that every stimulus capable of causing a hedonic episode exhibits features subject to homeostatic control. Implausible, that is, that (dis)pleasure is exhausted by homeostatic (dis)utility. Consider the hedonic episode caused by listening to your favourite piece of music. Is it at all plausible that I have a homeostatic mechanism that continuously monitors my Chopin Cello Sonata in G minor levels? Is it any more plausible that I have a homeostatic mechanism that continuously monitors my music levels? How could such a mechanism have evolved and how could it be realized? What are the detectors constantly monitoring music levels? What are the built-in range of values? Counterexamples proliferate. (Corns 2014, p. 241).

In analogy: what kind of (control-oriented) interoceptive predictions could events such as listening to Chopin, watching a sunset or hearing a joke encompass? It is not easy to think of any. IIT might be able to counter: the interoceptive predictions in question are not to be understood as resulting in control of interoceptive variables *right now* but *later*. Note that the intuitive idea behind instrumental interoceptive inference shifts significantly as soon as the counterfactual interoceptive predictions are not about now but about later. These counterfactual interoceptive predictions cannot straightforwardly be called control-oriented anymore, since they are not issuing interoceptive changes (now) but rather predict (presently counterfactual) interoceptive changes in the future. Consequently, what IIT would need to say is that the organism infers that listening to Chopin now will lead to interoceptive changes later. But then: what are the predicted interoceptive changes associated with Chopin? It is *still* challenging to think of any.[16]

[16] As an anonymous reviewer helpfully suggested, one might try to accommodate music or humour by reference to cultural (hyper-)priors. We are looking forward to the development of a proposal along these

Note further that this reply would complicate matters for IIT in trying to account for the host of bodily sensations that we *do* sometimes have in the very moment of undergoing aesthetic affective experiences, think, for instance, of goosebumps (Chanda and Levitin 2013). Since IIT must claim that the counterfactual interoceptive predictions concern interoceptive changes in the future, the fact that there are interoceptive changes right now seems not to be predicted by IIT. This is for the simple reason that there is (1) apparently no direct track for music (and many other kinds of typically affect-eliciting stimuli) to affect interoceptive signals, and (2) music does not appear to mandate instantaneous regulation of interoceptive signals—what priors might make it seem like a good idea to tinker with essential physiological variables in response to music? Thus, it seems to be an extra-theoretical fact to IIT that in such situations there are bodily events going on.[17]

Even if IIT could come up with stories here, there would still be other serious issues: affective experiences and their valence have polarity and intensity (cf. Kozuch 2018). Sometimes an affective experience is positive and sometimes it is negative. Sometimes an affective experience is more intensely positive (or negative) and sometimes less. It is highly plausible that these dimensions should covary with properties of instrumental interoceptive predictions. But it is not obvious how these dimensions of valence map onto dimensions of (instrumental) interoceptive predictions.

Looking at polarity, a straightforward candidate might come to mind: On the one hand, an affective experience is positive if the predicted interoceptive changes are conducive to homeostasis. On the other hand, an affective experience is negative if the predicted interoceptive changes are unfavourable to homeostasis. A natural way to map this suggestion onto instrumental interoceptive predictions would be to assume that counterfactual interoceptive predictions that predict vital variables to be well within viability range lead to positive affect while counterfactual interoceptive predictions that predict vital variables to approach or exceed viability limit lead to negative affect. Perhaps this is what S&T have in mind when they write that "instrumental interoceptive

Footnote 16 continued
lines. However, we would like to point out a challenge resulting from instances of affective experiences that are similar to e.g. the mentioned aesthetic or humour-related experiences but seemingly unrelated to culture. One is the mentioned sunset. Then there is also the internet phenomenon of "Oddly Satisfying Videos". These clips show events and actions that typically involve the meticulous manipulation of physical objects such as peeling wood (see the same-named subreddit and YouTube channel). Their audience reportedly experience positive affective experiences watching them. By the same token, why is the experience of watching upward flowing water negatively valenced (i.e. why does it feel wrong)? Here is a demonstration: https://youtu.be/NiOAfQZwn0g It appears doubtful to us that these affective experiences are easily explained in terms of instrumental interoceptive predictions or by reference to cultural priors. A strength of the alternative we are going to offer in Sect. 5, is that it derives the positivity and negativity resulting from such expectation-satisfying/violating (perceptual) experiences from underlying PP-principles rather than from priors with case-specific content.

[17] There is another problem with the appeal to future interoception which has to do with the intentionality of affective experiences. If there are no changes in interoceptive signals right now, how come that there is an affective experience that seems to be about this music that we enjoy—something in the here and now? It appears that if there are no immediate changes in interoceptive signals, there will be nothing interoceptive to infer the causes of, and so affective experiences would be about nothing in the here and now. In contrast: affective experiences are more often than not about the here and now (this is, of course, not to say that there are no affective experiences with future or even past-directed temporal orientations, such as hope, pleasant anticipation or regret).

inference requires maintaining physiological essential variables within tight ranges of viability across time" (S&T, 9).

However, this intuitive proposal runs into several problems: for starters, why would listening to Chopin's Sonata lead to favourable counterfactual predictions about vital variables in the first place? The inverse case seems even more implausible: Why would music that one decidedly doesn't like lead to unfavourable counterfactual predictions about vital variables?

Be that as it may, a bigger challenge comes to the fore when we consider that the main function of an instrumental interoceptive prediction is to regulate vital variables so as to fulfil the prediction. Now, if a counterfactual interoceptive prediction predicts vital variables to be well within viability range, then, it seems, it would regulate interoceptive signals so that the vital variables are well within viability range. Put this way, this might seem somewhat puzzling at first glance. Perhaps one can make sense of it if one shifts one's gaze from present- to future-directed counterfactual interoceptive predictions. Imagine you're getting a much-needed neck massage. It feels glorious. Here is how one could account for it: the massage directly effects some favourable interoceptive changes. This, in turn, leads you to predict favourable interoceptive changes in the (near) future—and, thus, to feel glorious.

This picture turns highly implausible, however, when we turn to negative affective experiences: if an instrumental interoceptive prediction predicts vital variables to approach or exceed viability limit, then it would regulate interoceptive signals so that the vital variables approach or exceed viability limit. In other words, counterfactually predicting things to go poorly would make things go poorly. That is, the instrumental interoceptive prediction would be doing the opposite of what it should be doing, namely trying to bring things back on track (i.e. into viable bounds). Against the background of the present suggestion this would paradoxically require the kind of prediction that gives rise to positive rather than negative affect.[18]

These considerations speak against the suggested way of mapping polarity onto instrumental interoceptive predictions. Perhaps there are other, more plausible ways to recover polarity from IIT (see also footnote 20 and 21). Intuitively, it seems that *on top* of the instrumental interoceptive predictions put forward by IIT, there is an evaluation of these changes. We have now seen that it is not straightforward to recover this component from IIT.[19]

[18] Interestingly, this sort of thing is proposed to happen by PP-accounts of conditions like chronic fatigue and depression (Stephan et al. 2016). We agree that this seems like a plausible account of these abnormal affective conditions. However, it does not seem like a good model for *normal* affective states that are our primary focus here. Thanks to an anonymous reviewer for bringing this work to our attention.

[19] It is worth emphasising *what* it is that is missing in IIT. IIT gives us (control-oriented) predictions of interoceptive changes as the mark of the affective. Thus, upon a bear encounter, instrumental interoceptive predictions will issue physiological changes such as an increase in heartbeat rate. These predictions will make the protagonist of our example enter an affective state. Now, when we read about this situation, we have no difficulty judging that our unlucky protagonist will not only be in *some* affective state but in a negative affective state of fear. However, the only thing that we seem to get from IIT as the base of affective experiences are counterfactual predictions about physiological changes. What does not straightforwardly fall out of IIT is the intuitive metric by which we determine that the predicted physiological changes are bad for the wellbeing of the protagonist (see also footnote 20). Such a metric is needed because it is not obvious that an instrumental interoceptive prediction taken by itself is intrinsically positive or negative. Intuitively, the instrumental prediction of physiological changes such as a heartbeat increase can be a bad

If not in the implausible way outlined above, in what way are we to understand this evaluation? In terms of Error? In terms of precision? It is unclear, partly because Error and precision are nothing specific to interoceptive inference. Perhaps it is specifically Error occurring in, or expected precision of, interoceptive inference?

Note that there is a general issue with such "interoception-only" moves (see also footnote 13): giving *interoceptive* inference exclusive rights when it comes to affect production appears in tension with one of IIT's main premises saying that *all* (not only interoceptive) PP activity is geared towards promoting allostasis:

> Interoceptive inference […] should not be considered as a generalisation of predictive coding from exteroceptive modalities such as vision. Instead, perceptual content in all modalities, including modalities such as vision, is a consequence or generalisation of a fundamental imperative towards physiological regulation. Seen this way, all perceptual content is underpinned by inferential mechanisms that have a functional, ontological, and phylogenetic basis in allostasis. (S&T, 11).

While it is plausible that interoceptive inference is somewhat affectively privileged due to its proximity to homeostasis, it appears unwarranted to grant it exclusive affect-production rights.

Eventually, the main problem for recovering polarity from IIT is its vagueness. It simply does not clearly elaborate the elements that would be required to explain the evaluative component. We will return to a suggestion for explaining this component later in Sect. 4.

For now, let us briefly consider how IIT fares when it comes to valence intensity. A possibility is that intensity varies depending on how far in the future the interoceptive changes are predicted. The further away, the less intense, and the more imminent, the more intense. This can't be the whole story, though: the IIT story about Chopin's Sonata, for instance, will plausibly have to locate the interoceptive changes quite far into the future. But don't we often enjoy our favourite music intensely? On the other hand, a relatively faint itch might well be dealt with immediately. The general point is that there seems to be no proportional link between how temporally remote interoceptive changes are and the magnitude of experienced positivity or negativity of an affective experience (Kozuch 2018). A similar problem arises if we take intensity to map onto the magnitude of the predicted interoceptive changes: an IIT proponent would be hard-pressed to argue that the elation of listening to Chopin or watching the sunset is proportional to the magnitude of the expected changes in homeostasis, given that IIT is already overstretched when it comes to simply linking aesthetic experiences with changes in homeostasis regardless of intensity.[20]

Footnote 19 continued
thing in the case of a bear encounter or a good thing in the case of a romantic encounter. In this context, high-level predictions (bear vs. romantic encounter) can plausibly contextualise the counterfactual interoceptive predictions, marking them as positive or negative. Apart from the mentioned issue that high-level predictions are not specific to interoceptive predictions (see footnote 13), there is another problem lurking. We seem to encounter a regress: instrumental interoceptive predictions are plausible reactions *to* a bear encounter—but now we need the contextualising high-level predictions for the instrumental interoceptive predictions to emerge (as negative) in the first place.

[20] Perhaps, then, affect intensity maps in the following way: the more interoceptive changes are predicted (i.e. the more fluctuation in interoceptive variables), the more positive or negative it is. In fact, this might also

To take stock: Earlier we have considered and criticised IIT's initial mark of the affective: the presence of interoceptive predictions. Now we have considered and criticised IIT's refined mark of the affective: the presence of control-oriented or instrumental interoceptive predictions. When trying to defend IIT one might grant the objections and propose still another mark of the affective. To see this one needs to remember the distinction between perceptual inference on the one hand and active inference on the other hand. The latter, in turn, can be subdivided into epistemic inference and instrumental inference. These distinctions can then be applied to different inferential modalities such as interoception. In our discussion we have focused on one kind of active interoceptive inference, namely instrumental interoceptive inference. But perhaps, the IIT proponent might now point out, we should seek the mark of the affective not in instrumental interoceptive inference only but in active interoceptive inference *in general*.[21] That is, the mark of the affective according to IIT is the presence of counterfactual predictions of the control-oriented (instrumental) *or* the explorative (epistemic) type.[22] Put simply, if active interoceptive inference is at work, then there is affect.[23]

Now, in order to assess the merits of this suggestion it is helpful to get a better grasp on what epistemic interoceptive inference could be. What unites both kinds of active interoceptive inference is that they both prescribe action in the form of the engagement of autonomic reflex arcs ("intero-actions") and other allostatic actions. What distinguishes epistemic from instrumental interoceptive inference is that its function lies in gathering information rather than in immediate regulation. That is, in epistemic interoceptive inference interoceptive actions are performed in order to acquire information that will improve an agent's (interoceptive) models, enhancing her predictive and regulatory capabilities. In other words, an agent actively engages in exploring her internal milieu by tinkering with her physiological variables.

On the face of it, this seems like a rather risky endeavour. The idea of (active) epistemic inference is highly plausible in the exteroceptive and motor-domain. We often look or move some way not so much to change (regulate) the world in accordance with our predictions but to acquire better information about the world. However, can we think of any examples of epistemic *interoceptive* inference? It is not easy to think of any. On second thought, changing diets and physical exercise might come to mind

Footnote 20 continued
offer an account of valence polarity: if high/low interoceptive fluctuation is predicted, then the experience is positive/negative. This idea entails that we would have the most intense positive affective experiences if no interoceptive changes are predicted. However, this appears to obviate the need for interoceptive prediction in the first place and it might be thought to render the resulting experience non-affective. It also seems at odds with the observation that positive affective experiences are often accompanied by a myriad of bodily sensations, which would have to map onto a lot of (instantaneous) interoceptive changes. But then, the suggested criterion would predict that the affective experience with many interoceptive changes would have to be negative. Is it then, perhaps, that the negativity of the many instantaneous interoceptive changes is somehow outweighed by the predicted future interoceptive stability? This would appear at odds with how the bodily sensations appear to one in the moment of the positive affective experience—namely as contributing rather than subtracting from its positivity.

[21] We thank an anonymous reviewer for this suggestion.

[22] This mark of the affective is narrower than the first, i.e. interoceptive predictions *simpliciter*, but broader than the second, i.e. instrumental interoceptive predictions.

[23] Note that this proposal can preserve the intuitive distinction between "cold" and "hot" interoception by appealing to the distinction between perceptual and active interoceptive inference.

as potential examples. To a limited degree, this might coincide with some diet- and exercise-related bodily feelings such as feeling satiated or sore. But this does not seem to map very well onto affective experiences more broadly, especially not on the problem cases previously identified. What kind of explorative interoceptive actions are prescribed when we listen to music? It is challenging to think of any.

In general, even if we can come up with a set of plausible examples of epistemic interoceptive inference, such "interoceptive experiments" would be a rather rare phenomenon. And for good reason:

> [I]nstrumental interoceptive inference requires maintaining physiological essential variables within tight ranges of viability across time. This entails precise prior expectations that these variables and their trajectories remain within such stable ranges […] The resulting interoceptive perceptions will therefore be drawn towards stable inferences about self-related variables and their trajectories. (S&T, 9).

This is why Seth and others repeatedly point out that the concept of epistemic inference has only limited application within the context of interoception:

> In this context, exploratory or epistemic interoceptive 'actions' may be less evident because they may be more costly: one does not want to raise one's blood pressure to physiologically dangerous levels just to see whether it can return. (Seth and Friston 2016, p. 5; see also Seth 2015, p. 20; S&T 8–9).

It remains questionable whether the idea of interoceptive experiments can map well onto affective experiences. These considerations seem to be responsible for why S&T do not go into the suggested direction. When discussing affective and self-related experiences they pass by epistemic inference rather quickly to get to what seems to them most essential: instrumental interoceptive inference. In fact, they use the distinction between instrumental and epistemic inference to explain *differences* in the phenomenology between affective and other experiences:

> Based on these distinctions, we propose that instrumental inference undergirds a different phenomenology than epistemic inference related to discovery. Instead of delivering a phenomenology of objecthood, instrumental (control-oriented) interoceptive inference plausibly underlies a phenomenology related to the evaluation of the allostatic consequences of regulatory actions. A non-localised, non-object-based phenomenology associated with both mood and emotion (S&T, 9).

Note that we are after something that *phenomenally unites* affective experiences—as phenomenal valence seems to do. S&T, however, propose that epistemic interoceptive inference is phenomenally distinct from instrumental interoceptive inference because they are different kinds of active inference. Against this background, active interoceptive inference (compared to instrumental interoceptive inference) appears rather unsuitable as the unifying mark of the affective.[24]

[24] We have mostly focused on *single* aspects of the IIT machinery, seeking to identify a plausible mark of the affective among them that should give us something resembling the valence of affective experiences.

In closing, we would like to emphasise that the projects that IIT sets for itself are not undermined by its struggle to provide a mark of the affective. In the end, IIT successfully elucidates the nature of its targets, such as interoception and self-related experiences, and provides a convincing account of the regulative role of affect in the mechanism of instrumental interoceptive inference. The identified shortcoming is in part due to something that might seem like a strength of IIT, which lies in "eliminating any bright line separating emotion (or perception) from cognition." (S&T, 3). The laudable enterprise of blurring the line between emotion (or more generally, affect), perception and cognition complicates the identification of an aspect that firmly distinguishes the phenomena from each other. However, the intuitive idea that affect, perception and cognition work in close tandem is not per se in conflict with the equally intuitive idea that there are features unique to affective experiences and that distinguish them from perception and cognition. That we cannot straightforwardly extract *some* of the characteristic features of affective experiences (such as valence) is, in fact, something explicitly acknowledged by S&T: "it remains an open question as to how interoceptive experiences [which IIT equates with affective experiences] map onto the computational machinery of interoceptive inference" (S&T, 3).

Our aim, then, is not to argue against IIT, but to complement it with the mark of the affective. In the next section, we will look for the mark of the affective in Error Dynamics Theories of affective experience. We will suggest a much needed amendment to EDT at the end of the section, and in Sect. 5 we will develop a revised account that builds on the strengths of both, IIT and EDT, to explain the features of affective experience outlined in Sect. 1.

4 Error dynamics theories of affective experience

The central tenet of error dynamics theories (EDT) is that positive (negative) valence is equivalent to a positive (negative) *rate of Error reduction* (a general property of Error dynamics) (Van de Cruys 2017) or, in an analogous free-energy formulation, to a negative (positive) rate of change of free-energy over time (Joffily and Coricelli 2013).

According to PP, the central goal of an organism is to minimise Error over time, and Van de Cruys argues that valence plays a key role in the attainment of this goal. This appears intuitive if we consider that affective states usually express our concern and care for things. Now, in PP the main concern of an organism is to reduce Error. The idea

Footnote 24 continued
IIT might now rejoin: valence is not to be found in single IIT aspects—but in the interplay of several of those. For instance: active interoceptive inferences will not only yield predictions that are either epistemic *or* instrumental but predictions that will be *both* at the same time, expressing a balance between epistemic and instrumental components. Furthermore, these dual-natured counterfactual interoceptive predictions will be contextualised by high-level predictions (see footnotes 13 and 20). We admit that evaluating such a proposal introduces a level of exegetical complexity that we hesitate to take up, leaving it to the proponents of IIT to spell out against the background of the raised issues. To us it is not obvious at first look that this (or a similar) mixed proposal can successfully overcome the raised issues and provide a graded polarised metric which plausibly maps on affective experiences. Also, *mixing* several components together would introduce the further challenge to explain how varying combinations of relatively heterogenous components map onto valence, a property that appears as a rather unitary fundamental component of affective experiences.

is then that affective states are bound up with how we fare in reducing Error. But how? Van de Cruys' answer is that affective experiences are bound up with Error reduction via a constitutive and unique component of affective experiences: their valence. And this is so because according to Van de Cruys, valence in PP terms just *is* the rate of Error reduction.

From now on we will simply refer to this rate as *Rate*.[25] Rate is, accordingly to EDT, equivalent to valence—positive when valence is positive, and negative when valence is negative. Intuitively, Rate (i.e. valence) reflects the speed or pace with which the organism makes progress in reducing Error. It is this Rate which Van de Cruys proposes to be reflected in the valence of affective experience: If Rate is positive/negative, then affect is positive/negative. Thus, EDT substantiates the intuitive idea that valence is a (nonconceptual) representation of (dis)value to the organism (Levy and Glimcher 2012; Carruthers 2017), an idea dating at least back to Meinong (1917) and at the heart of perceptualist theories of emotions (e.g. Tappolet 2000; Döring 2007). According to EDT, valence is in fact a representation of (dis)value: it represents Rate, the pace at which the agent is making progress in reducing Error over time. While IIT (with its focus on allostasis) seems to emphasise the regulative aspect of affective experience, EDT (with its focus on valence) seems to emphasise their evaluative aspect.

An initial worry with EDT's idea follows from the principle of parsimony: before looking at dynamic variables such as Rate, one should try to use simpler variables to explain valence, such as instantaneous prediction error (i.e. Error). In fact, the initial approach of EDT was to simply consider Error as always "negative in valence" (Van de Cruys and Wagemans 2011). This approach was later abandoned in favour of considering a negative Rate as always negative in valence (Van de Cruys 2017). To justify moving away from his earlier view of equating valence and Error to his more recent view of equating valence and Rate, Van de Cruys builds on the idea that the role of predictive models ensures that PEM leads to homeostasis (based on work in Pezzulo et al. 2015):

> Once homeostasis, rather than being reactive, relies on predictive models, errors often do not have direct effect on homeostasis (or fitness). It then becomes equally important to monitor prediction error dynamics, as it is to monitor the errors as such. Mere presence of instantaneous prediction error does not seem to be an adequate basis of emotional valence. Positive affect might still occur for a large instantaneous error as long as this error is (or has been) in the process of being reduced. (Van de Cruys 2017, p. 8).

Although Van de Cruys does not elaborate further on his reasons for reconceiving valence, it should be noted that a further problem with simply equating Error with valence is that there is always Error involved in any processing, which—if we wanted to simply equate positive amounts of Error with negative valence—would mean that valence is always negative, and that some situations just feel less negative than others. A possible solution to this problem would be to postulate a threshold for (the amount of) Error that is acceptable and that draws the line between negative and positive

[25] Rate is short for instantaneous rate of decrease of prediction error, which is the negative of the instantaneous rate of change of prediction error.

valence. This move, however, would require positing a further element to the theory, the threshold itself, which would deprive the idea of its parsimony (i.e. the initial motivation for considering Error *simpliciter* as a candidate for valence).

Central support for the idea that valence is dependent on error dynamics comes from Joffily and Coricelli's computational account. In their 2013 paper, Joffily and Coricelli offer a computational model to support EDT, in which they compare two agents in a non-stationary environment—one agent using valence and the other explicitly estimating the volatility of the environment. The agent using valence successfully replicates the behaviour of the other agent and achieves this by representing fewer hidden states and fewer parameters (Joffily and Coricelli 2013).

Moreover, EDT's notion of valence as dependent on error dynamics is in line with Frijda's view that "pleasure is the positive outcome of constantly monitoring one's functioning" (Frijda 2007, p. 82), with the conception of affect as a "neurophysiologic barometer of the individual's relationship to an environment at a given point in time" (Duncan and Barrett 2007, p. 1186) and with the control-theoretical approach to emotions of Carver and Scheier, who understand affective experience as an expression of the mismatch between the actual and expected dynamics of a given task (Carver and Scheier 1990, 2001; see also Proust 2015). Finally, EDT also finds support when one interprets various emotion studies in light of PP. For instance, Batson and colleagues argue that a transition from a less valued state (in PP terms, large levels of Error) to a more valued state (low levels of Error) corresponds to positive valence (Batson et al. 1992), which is analogous to EDT's claim that positive valence corresponds to an increase in Rate. Although the support for EDT comes from computational models rather than from neuroscientific studies, it provides a computational story that puts flesh on previous conceptions of valence. In return, EDT finds support from the mentioned work on the evaluative aspect of affective states.

To distance itself from IIT, EDT points out that affect by itself is transmodal and distinct from interoception.[26] Addressing IIT, Van de Cruys claims that valence, the basic building block of affective states, doesn't originate in the inference of the causes of physiological states, but in Error dynamics (i.e. Rate), and because these dynamics are ubiquitous, he claims that feelings "can emerge from any processing, not just that about the body" (Van de Cruys 2017, p. 2). While it might seem at first that this makes IIT and EDT irreconcilable approaches, a closer look at more recent versions of IIT shows this not to be the case. IIT is not committed to the idea that affective experiences can only emerge from the inference of the causes of physiological states. While EDT is a good candidate for accounting for the mark of the affective (with its characterisation of valence), it can very well do so in a way that is compatible with IIT. Furthermore, by providing a computational equivalent of valence, EDT (if correct) might be able to complement IIT. We propose that the resulting synthesis (to be developed in Sects. 5) provides a plausible account for the elements of affective experience expounded in Sect. 1.

[26] The early versions of EDT emerged as accounts of aesthetic experience (Van de Cruys and Wagemans 2011), just like early versions of IIT emerged as accounts of interoception. As we will see, this move "away from the body" puts EDT in a better position to address the challenge of the aesthetic cases presented in the previous section, such as listening to Chopin.

An interesting question concerns the role that valence plays in the overall goal of PEM. In our view, valence informs the system in order to optimise PEM over time. On the one hand, feelings direct the organism towards homeostasis (a point that is central to IIT proponents, but that Van de Cruys emphasises as well). On the other, feelings direct the organism towards desirable levels of uncertainty (a point that Joffily and Coricelli emphasise). If we grant that evolutionary pressure has made sure that allostasis and PEM are two sides of the same coin (something that IIT and EDT agree on), then valence (Rate) can be used to maintain the policies that minimise Error over time. A key difference between IIT and EDT is that EDT claims that affective experiences play a central role in achieving homeostasis, while IIT claims that affective experiences *are* homeostasis,[27] with the problematic consequences that we discussed in the previous section.

What EDT brings to the table is that Rate also serves the function of optimising levels of uncertainty. Too much uncertainty (e.g. bear scenario) is dangerous for short-term PEM and requires excessive cognitive resources, too little uncertainty is counterproductive for long-term PEM. The reason to avoid scenarios with too little uncertainty is that they would hamper learning, because the organism wouldn't be able to appropriately update its models of the world. As Kiverstein and colleagues put it, "being sensitive to error dynamics guarantees that the agent avoids wasting time in places where regularities are either already learned or too complex given the agent's skill level" (Kiverstein et al. 2019, p. 2864).

Here the learning rate is an important element. In PP, the learning rate is the pace at which Error changes hypotheses. According to PP, the brain constructs and updates "a vast hierarchy of expectations that overall help regulate the learning rate and thereby optimize perceptual inference for a world that delivers changeable sensory input" (Hohwy 2017, p. 77). Regarding the learning rate, Joffily and Coricelli highlight the following:

> An important function of emotional valence turns out to regulate the learning rate of the causes of sensory inputs. When sensations increasingly violate the agent's expectations, valence is negative and increases the learning rate. Conversely, when sensations increasingly fulfil the agent's expectations, valence is positive and decreases the learning rate (Joffily and Coricelli 2013, p. 1).

A necessary addition is that some instances in which the Rate (or valence) decreases over time but stays (slightly) positive (e.g. a highly predictable but changing environment) might mean that a higher Rate is waiting elsewhere and that the present activity should cease (i.e. a low but positive Rate can lead the system to infer that uncertainty will be optimised by chancing activities), and a highly negative Rate might mean that the situation is too complex to learn predictable patterns (e.g. a novice playing the harder levels of Tetris) or, even more important, that the situation is dangerous (e.g. bear scenario) and that the agent should change her behaviour accordingly (not only out of learning rate considerations). This informs the above-mentioned idea that feelings contribute to directing the organism towards optimal levels of uncertainty in

[27] More precisely, they are "the content of the joint set of predictions geared towards allostasis" (Seth and Tsakiris 2018, p. 6).

order to fulfil PEM over time. This idea is in line with e.g. what Kidd and colleagues have termed the 'Goldilocks Effect': "infants actively seek to maintain an intermediate level of information absorption, avoiding allocating cognitive resources to either overly predictable or overly surprising events" (Kidd et al. 2012, p. 6). In PP-terms, maintaining optimal levels of uncertainty is conducive to PEM because it helps the agent to optimise the allocation of cognitive resources, and, as we already mentioned, avoid overly high levels of uncertainty.

However, there are many situations in which the organism does not seem to avoid low levels of uncertainty, routines and habits being a clear example.[28] If one cleans the house every Saturday it seems unnatural to say that the goal is to optimise uncertainty. Rather, it seems like the goal is to have a clean house over time, even if this goes against the grain of considerations about uncertainty optimisation. Here, the story to follow is not uncertainty optimisation, but the standard PP story, PEM over time. The point is not that uncertainty optimisation substitutes PEM. Rather, PEM subsumes uncertainty optimisation. In other words, uncertainty is optimised *to* minimise Error over time. Our point is that in an EDT formulation, valence is not simply reduced to an engage/disengage signal (of the sort proposed by Prinz 2010). Valence also informs the organism about potential learning opportunities, and as a consequence, the organism might engage in negatively valenced processes if they improve learning in a way that is conducive to PEM. Just as well, the organism will still engage in low-uncertainty processes such as habits and routines, as long as low uncertainty is conducive to PEM over time. Furthermore, conceiving of valence as Rate in the way we have just presented can address the challenge of why there is positive valence associated with activities that do not seem clearly linked to allostasis: some affectively charged activities might be best explained as optimising uncertainty in ways conducive to PEM over time, rather than as instances of allostasis-directed behaviour. By biological necessity, PEM and allostasis need to converge over time. However, there might be occasions in which the link between affect and allostasis is not straightforward, because in computational terms, the goal of the organism is PEM, not allostasis. Generally speaking, by casting valence in terms of Error dynamics, which are more germane to PEM, EDT is better positioned to take care of situations such as the affective responses of listening to Chopin.

Nevertheless, there is a shortcoming in the version of EDT defended by Van de Cruys: it does not account for a central element of affective experiences, namely, that feelings are conscious phenomenal states. Van de Cruys does not simply think that valence determines affective experiences, but also that, since valence is usually understood as a phenomenal aspect of affective experience, changes in Error are consciously experienced as valence. A problem with this is that affective feelings are conscious and, according to the standard view of PP, Error is not part of conscious experience. By extension, changes in Error *qua* Error cannot be part of conscious experience. Within the PP framework, the content of conscious experience are predictions. In Hohwy's words, "Conscious perception is determined by the hypotheses about the world that best predicts input and thereby gets the highest posterior probability. More specifically, since the inversion of the generative model is implicit, what is conscious is

[28] We thank an anonymous reviewer for flagging up this point.

the interconnected set of currently best performing predictions down throughout the perceptual hierarchy" (Hohwy 2012, p. 201). Of course, Error is constantly used to update predictions but in the end what subjects consciously experience is not Error but predictions.

To defend the idea that predictions determine the content of conscious experience, Hohwy follows an abductive argument. Within the PP framework, the conception of conscious content as predictions is the best explanation for a series of explananda arising from empirical studies in cognitive psychology, psychophysics and psychopathology, as well as from a conceptual analysis of our folk understanding of conscious experience. In particular, Hohwy argues that the concept of conscious content tracks characteristics of conscious perception such as "binding, penetrability, reality testing, illusions, inextricably rich experience, and first person perspective" (Hohwy 2013, p. 202). A good example of the concept of conscious content as predictions at work is provided by Hohwy's explanation of binocular rivalry (Hohwy et al. 2008).[29]

That predictions constitute the content of conscious experience is the standard view in PP. For instance, Barret writes that "once prediction error is minimised, a prediction becomes a perception or an experience" (Barrett 2017, p. 12). This view extends beyond exteroception. Hohwy argues that this view also provides a good explanation for the role of expectations for bodily sensation (Hohwy 2011, p. 271), namely, that innocuous stimuli that are *expected* to be painful tend to be *experienced* as painful (Brown et al. 2008). Importantly, IIT (in contrast to EDT) is very clear in upholding the view that predictions determine the content of conscious experience: "An important challenge in this context is to identify which aspects of inference support specifically *conscious* emotional experience, with predictions (rather than prediction errors) being the preferred vehicle" (Seth and Friston 2016, p. 5).

This view poses a challenge for EDT. The resulting argument goes like this:

(P1) Only predictions constitute the content of conscious experience.

(P2) Valence is a content of conscious experience.

(P3) *Rate* is not a prediction.

(C) *Rate* is not equivalent to valence.

A viable solution to this issue is to extend EDT by casting valence in terms of predictions. Here, we propose that a given set of predictions can include predictions of Error dynamics. That is, the predictions that the subject experiences can include predictions of the (expected) Rate. Therefore, if, as EDT claims, feelings co-vary with properties of Error dynamics, what we experience as these feelings cannot come directly in the form of Error dynamics but must come in the form of the system's

[29] In the most common version of binocular rivalry, one eye is presented with a picture of a house and the other eye is presented with a picture of a face, and conscious experience switches between the two percepts. Hohwy's PP explanation is that the likelihood of the combined face-house blend hypothesis cannot overcome "the exceedingly low probability that a face and a house could co-exist in the same spatiotemporal location […] so the hypothesis that is selected, and which determines perception, is either the face or the house hypothesis." (Hohwy 2013, pp. 21–22).

predictions about Error dynamics. The resulting view is that the valence of affective experiences is determined by the prediction of the Rate, *Expected Rate* from now on. This means that positive (negative) phenomenal valence corresponds to the Expected Rate being positive (negative).[30]

It might seem that adding "another layer" of computations (i.e. from Rate to Expected Rate) might bite the bullet at the expense of parsimony. However, the opposite turns out to be the case. Taking a formal account of volatility and learning rate under Bayesian inference as a starting point, Perrykkad and Hohwy argue that the rate at which a prediction error is minimised is partially determined by the precision ascribed to that prediction, so that if an organism holds expectations about precision (a central tenet of PP) it then holds implicit *expectations* about Rate:

> Technically, the rate at which a prediction error is minimised is partially determined by the precision (or confidence) ascribed to that prediction. Intuitively, we update our beliefs faster and more dramatically when we believe our data to be more reliable and when we suspect the world is frequently liable to change […] This means that those creatures whose internal models allow them to make predictions about this precision implicitly hold (sub-personal) beliefs about the expected rate of error-minimisation, conditioned on the data they choose to sample. Given the inverse relationship between precision and ambiguity, we can associate beliefs about the ambiguity expected under a given policy with beliefs" (Perrykkad and Hohwy 2020, p. 3).[31]

Perhaps some creatures hold explicit beliefs about Expected Rate, but as long as the organism is calculating expected precision, it is, by extension, implicitly estimating Expected Rate—making Expected Rate more parsimonious than Rate *simpliciter*. This elaboration of EDT makes Expected Rate the best candidate for valence and, more importantly, for the mark of the affective, which is a necessary element to account for affective experiences.

5 The affective inference theory of affective experience

In the remainder of this article we will build around the notion of Expected Rate to propose a revised PP-theory of affective experiences, the Affective Inference Theory (AIT), that cannot only accommodate the evaluative but also the motivational and intentional character of feelings. AIT's central tenet is that the valence of affective experiences corresponds to Expected Rate, so that positive (negative) feelings correspond to Expected Rate being positive (negative). According to AIT, Expected Rate is the mark of the affective. AIT integrates insights of IIT and EDT to develop an account that is compatible with the core of both theories while trying to rise to their challenges. Importantly, Expected Rate (unlike EDT's Rate) is a plausible candidate for *phenomenal* valence, and consequently, it provides a plausible mark of the affective in computational terms (unlike IIT's interoceptive inference). Moreover, AIT's

[30] On the other hand, (instantaneous) Rate might be best understood as unconscious valence (e.g. Berridge and Kringelbach 2015) (see footnote 4).

[31] Thanks to an anonymous reviewer for directing us towards this work.

conception of valence can account for both the polarity and the intensity of valence: the polarity of valence corresponds to the polarity of Expected Rate (i.e. the polarity is negative if Expected Rate is negative and positive if Expected Rate is positive), and the intensity of valence corresponds to the magnitude of Expected Rate (i.e. the intensity of valence is smaller the closer Expected Rate is to zero).

A good example of AIT at work are affective experiences when e.g. listening to music (see Sect. 3). A major current in aesthetics holds that affective experiences when listening to music arise from the composer's manipulation of expectations (Meyer 1956; for empirical support see Huron 2006). AIT can provide a computational PP account of affective experience in line with the view of music as expectation manipulation. Music has constantly evolving patterns that result in unique error dynamics. Say we have finally "gotten used" to a pattern, the music changes, and the organism predicts negative Expected Rate (negative valence); this negative valence then continues until the organism becomes more confident in its capabilities to predict the new pattern, resulting in a positive Expected Rate (positive valence), until the music changes again, and so on and so forth. Importantly, AIT offers a model for how Expected Rate can operate at several timescales at once.[32] If the organism expects to be about to generate a good model of the underlying structure of a piece of music (phenomenologically, the anticipation of insight), this will result in positive Expected Rate (i.e. positive valence directed at an abstract, longer term level) even when the Expected Rate of the immediately incoming auditory stimuli is negative (i.e. negative valence directed at a concrete, short-term level). This tension between short-term and long-term Expected Rate underlies much of the phenomenology of experience of music, which is laden with emotional tension, fulfilment, disruption and anticipation.[33]

5.1 AIT: intentionality

With AIT, we can now explain two of the characteristics of feelings outlined in Sect. 1: that feelings are conscious, and that feelings have valence. Now we turn to another aspect of affective experiences: intentionality. As we mentioned in Sect. 1, phenomenal valence is the best candidate for the mark of the affective, partly because it grounds the other features of affective experiences. We can build on AIT's notion of phenomenal valence to provide an account of the intentional and motivational dimension of affective experiences. First, let us focus on intentionality. Felt valence emerges as a nonconceptual representation of value fleshed out in terms of Expected Rate, an agent's

[32] The temporality issue was particular to EDT rather than to IIT. Because on IIT predictions are not only about the states of hidden causes, but about their trajectories over time. Thus, both AIT and IIT are well-placed to explain the phenomenological insight that the here-and-now is always intertwined with the past and the future. We thank an anonymous reviewer for bringing this to our attention.

[33] This is in line with the early EDT idea that "prediction errors at the level of style (perceptual ones) sometimes can be resolved on the level of meaning" (Van de Cruys and Wagemans 2011, p. 1053). However, Van de Cruys and Wagemans fail to flesh out this idea, claiming at once that "a central assumption of our theory is that prediction errors are always to some extent emotional, more specifically negative in valence" (p. 1047) and that "this playful and safe as-if context of art, where our guards can be lowered and our actions suspended, allows for the usually negative prediction errors to be enjoyed" (p. 1041), without explaining how negative prediction error suddenly turns into positive phenomenal valence in the context of aesthetic experience.

expected success in predicting the world (Carruthers 2017; Van de Cruys 2017). This phenomenal affective dimension constitutes the *primary affective intentionality* characteristic to feelings.[34]

When we say that valence can be computationally captured as Expected Rate, it is important for the system to also know what (and in what way) is causing (or modulating) the changes in Expected Rate. In other words, it is important that the feeling in question captures *what* is most likely causing the changes (which we equate with its particular object) and the most likely reason *why* it is causing these changes (which we equate with its formal object). Remember that a feeling represents its particular object as having a feeling-specific property, i.e. its formal object. This is so because representing the connection between the valence and its modulator is crucial for the system to effectively deal with the sources of perturbation in predictive success. One way in which a predictive system can naturally establish these connections is by tracking and predicting regularities, resulting in an experienced connection of the features exhibiting the regularity. Wiese's recent regularity account of PP illuminates this notion: "When the brain tracks a regularity that is predictive of different features (or of different objects or events), there will be an experienced connection between those features (or the respective objects or events). We can then say that the regularity connects those features (or objects or events)" (Wiese 2018, p. 259).

The idea behind the regularity account is that some features are strongly predictive of other features based on a robust observed correlation between these two sets of features. A good example of regularity is binding. Binding is a term usually employed to refer to the binding of multiple intra- and cross-modal features to a single object, such as the binding of the visual information about the shape and the colour of a bouncing ball (intra-modal) or the binding of the ball's visual features and the auditory features of it bouncing (cross-modal) (see e.g. Hommel 2004). Now, a PP system does not have to first process stimuli separately and then bind them. Instead, it uses causal inference embodied in the parameters of its generative model to assume bound attributes and then "predicts them down through the cortical hierarchy. If they are actually bound in the states of the world, then this will minimise prediction error, and they will be experienced as such" (Hohwy 2012, p. 6).

In short, binding in PP terms is simply the exploitation of statistical regularities in the form of Bayesian priors to predict a strong connection between two or more variables. In the case of the intramodal binding of the shape and colour of a bouncing ball, these two features are predicted to have a strong, localised connection, and this regularity results in a high degree of experienced localised integration. The shape and colour of the ball are experienced as the same object, localised precisely in space. In the case of the cross-modal binding, the visual and auditory features of the bouncing ball are related, but not as strongly as its shape and colour, and although both stem from the ball, they occupy different regions in space. This results in the visual and auditory features being experienced in synchrony and as related to the same object,

[34] Insofar, primary affective intentionality is a kind of phenomenal intentionality (Horgan and Tienson 2002). Phenomenal intentionality is the thesis that the intentional content of a state is fixed by the phenomenology or "what it is like"-ness of the state. Thus, what a feeling is about or directed at is (partly) specified by the phenomenology of the feeling.

so that this cross-modal aspect of experience has a high degree of integration, but a lower one than the intra-modal aspect of the experience.

Thus, if there is an observed regularity (e.g. causality or correlation) between an object and changes in Expected Rate, this object will be experienced as the intentional object of the feeling with a certain feeling-specific property that explains these changes. Based on a tracked regularity, the particular and formal object of a feeling emerge as the *inferred* hidden causes of a variation in predictive success.[35] What follows is that formal and particular objects and valence are all integrated into the phenomenal whole of a specific feeling. When it comes to feelings, the connection between their objects and their affective phenomenal dimensions exhibit statistical regularities that are useful for prediction, but weaker than the statistical regularities underlying sensory binding. This is because the features between which these regularities are observed are of an abstract, transmodal nature, involving, for instance, multi-modal perception of objects (which can be whole situations) together with their interoceptive consequences (if any). This results in an affective experience that is often concurrently coded in many modalities (exteroception, interoception) and, at the same time, has a substantial degree of integration, but a lower one than typical experiences of sensory binding (cf. Frijda 2005).

The particular object of a feeling is what is predicted to be the most likely cause of the expected change in Rate.[36] In the example of a person walking through the forest and becoming afraid when spotting a bear, it is the bear that is predicted to cause a change in Rate and it is therefore the bear that is experienced as the particular object of the ensuing feeling, i.e. the object towards which the feeling is directed. In turn, the formal object, i.e. danger, corresponds to the Expected-Rate-relevant properties of said particular object, bound up with other (perceptually) predicted properties such as the size of the bear. In this context, a formal object works as a predictive model of the expected changes in Error dynamics caused by a particular object. The system has priors about the presence of certain relevant properties given such and such input (e.g. given the valence, particular object and context of a situation).

Given a certain input, the system predicts the presence of a formal object (as the property of a particular object), similarly to how the system predicts the presence of the colour red (as the sensory properties of a certain object), given specific events on sensory sheets. This analogy raises the following question: are the properties that are formal objects of feelings different from sensory properties that bottom out in sensory sheets? Some formal objects such as bodily damage or comfort in the case of pain or affective touch respectively seem to be very similar in that they have dedicated sensory pathways (Auvray et al. 2010; McGlone et al. 2014). Other formal objects such as danger or funniness are different in that they are more abstract in a specific way; we do not have dedicated sensory receptors upon which (precursors of) such properties could impinge.[37] In any case, formal objects are inferred high-level properties that is

[35] To conceive of affective experiences as involving a similar kind of inference is reminiscent of Barrett's pre-PP theory of constructed emotions (e.g. Barrett 2014). Barrett has integrated her theory with PP since then, siding with IIT (e.g. Barrett and Simmons 2015; Barrett 2017).

[36] Note that an expected change in Rate is equivalent to a change in Expected Rate.

[37] Here as well there seems to be somewhat of a gradient. One could make the case that we have something like dedicated fast-track threat detection pathways (for a critical review see Pessoa and Adolphs 2010).

some cases bottom out in sensory sheets (e.g. bodily damage) and in others don't (e.g. funniness). Going back to the example of the bear encounter, the bear is expected to be dangerous, and danger is the formal object of this experience. Unlike the colour of the bear's fur, danger is not part of the sensory content of the experience. Rather, danger is a property that is affectively framing the experience itself. Inferring the hidden causes of changes in predictive success in the form of particular and formal objects establishes *secondary affective intentionality* to feelings. Due to the ability of a PP system to establish experienced connections based on tracked regularities, this inference-based intentionality will be phenomenal in nature as well.[38]

The emerging picture allows for a natural way to understand moods such as depression, anxiety or elatedness which are seemingly objectless or, alternatively, directed at the whole world. Note that felt moods are peculiar when it comes to their particular objects while apparently having formal objects similar to directed emotions such as sadness, fear or joy (cf. Mendelovici 2014). Against the background of AIT, moods can be understood as predictions of error dynamics whose specific cause and, thus, particular object cannot be inferred. This can happen for a variety of reasons. The system might simply *fail* to inferentially identify a particular cause. This may be due to the inaccessibility of the modulators of Rate when they, for instance, come in the form of unobservable neurochemical fluctuations. Sometimes, however, there is simply nothing *in particular* that is responsible for a change in Rate. Instead there might be a whole series of events (and not one in particular) where the subject performed well or poorly at reducing Error, leading accordingly to a change in Expected Rate. In such cases, the mood can be thought to directly (and not through a particular object) represent how the subject expects to fare at reducing Error in the world. As a consequence, the subject will experience the world as generally deprived of opportunities (to reduce Error), full of threats to opportunities (to reduce Error) or as full of opportunities (to reduce Error) (cf. de Haan et al. 2013). In other words, moods are affective experiences that do not lack intentionality altogether but (parts of) the inference-based secondary affective intentionality. This seems to us as a natural conceptualization of depression, anxiety or elatedness.[39]

So far, we have an emergent theory in which the PP system generates predictions about Rate (valence), determining a feeling that, through the prediction of regularities, gets bound to a particular object and a formal object (the inferred cause of expected

Footnote 37 continued
Such a case for dedicated sensory pathways is arguably easier to make for feelings such as fear than for feelings such as funniness, (see also Fulkerson 2019).

[38] Of course, also phenomenal valence, i.e. the primary affective (phenomenal) intentionality bit, is a result of inference. The "inference-based" in secondary affective intentionality aims to emphasise that the causes of changes in Expected Rate need to be inferred in turn and that the success of this process is not guaranteed. Consequently, an affective experience cannot fail to have valence (otherwise it simply wouldn't be an affective experience) while the valence of an affective experience can fail to bind to specific inferred causes (making it lack at least phenomenal object-directed intentionality). This point will become clearer in the following paragraph on moods that have primary affective intentionality but lack (part of) secondary affective intentionality.

[39] Similar things could be said for other feelings that appear relatively undirected, such as feeling tired, relaxed or lascivious. Tentative sketches: In the case of tiredness/relaxation one feels like one's capability to go on with one's bodily activity (of Error reduction) is decreasing or smoothly increasing. In the case of feeling lascivious the world appears as populated by mating opportunities, which against a certain model results in PEM.

changes in Rate) into a global phenomenal whole. Because this process happens in a feedback loop, feelings get constantly updated. The system takes into account input from different modalities and at different levels of the predictive hierarchy to modify its predictions, thus making feelings context-sensitive.

5.2 AIT: motivation

A final elaboration is the connection between feelings and behaviour.[40] As we mentioned earlier, in the PP framework, action happens to fulfil proprioceptive predictions. Our claim here is that feelings bias action selection, so that specific behaviours become more likely. This biasing happens at different levels and timescales. It is likely that not only immediate actions will become more probable because of feelings but also certain actions in the future or distal predictions such as goals (see Pezzulo and Cisek 2016). In PP we can think of goals as high-level predictions that influence action in the following way: "We imagine a future goal state as actual, then use Bayesian inference to find the set of intermediate states (which can now themselves be whole actions) that get us there" (Clark 2015, p. 120). It is also important to emphasise that the dynamics involved here are non-linear. Because all of this happens in a feedback loop, feelings will modify the likelihood of actions, but actions will also modify feelings in turn, by modifying the error dynamics and with them, the predictions (e.g. valence) determining the feelings. In general, engaging in actions spans predictive trajectories in whose course feelings can arise in a regulative fashion (cf. Proust 2013). What this means is that in performing the component steps that implement an action, error dynamics unfold, leading to possible alterations in Expected Rate, giving, in turn, rise to regulative feelings in the form of predictions about error dynamics (i.e. Expected Rate).

This conception of the influence of feelings on behaviour resonates with the concept of action tendencies, the idea that a given emotion will make a set of actions more likely than others (Frijda 1986). To fully address the complexity of the feeling-behaviour relation, we need to revisit the notion of formal objects introduced earlier. We equated formal objects with the hypothesised Rate-relevant properties of the particular object. These hypothesised properties (e.g. the danger of a bear in the case of fear) work as a partial model of the particular object that can help the agent direct her actions towards the desirable values of Rate (i.e. the values that are expected to be conducive to PEM over time). Formal objects should therefore not be viewed simply as passive descriptions of the particular object, but also as calls for action in the form of action tendencies. The property "danger" is not simply the likelihood of the bear deciding to attack, but also the affordance to escape, or to raise one's arms trying to look big (cf. Proust 2015). The prediction of the property "danger" is thus a "pushmi-pullyu representation", both descriptive and directive (Millikan 1995). This makes sense when we see that formal objects are high-level hypotheses (i.e. a joint set of predictions), and that in PP, high-level hypotheses are "intrinsically affordance-laden: they represent both how the world is and how we might act in that very world" (Clark 2015, p. 187). Given a particular object, a context, and a certain valence, the formal object is a way of

[40] Here, we use the term action and behaviour interchangeably, reflecting the usage in the PP literature.

using prior knowledge to predict how a situation will unfold (descriptive properties) and to regulate behaviour (directive properties) in order to direct uncertainty to the desired levels.

The notion that formal objects are not only descriptive but also directive, links with the work developed in PP around action policies. In the paper we mentioned earlier, Perrykkad and Hohwy give a good definition of action policies: "A policy is a set of possible actions (or individual control states) that have been grouped together by the individual for its history of success as a strategy to reduce prediction error when faced with situations with learned commonalities, which cue success for that policy." In the same breath, they also specify that expectations for "rate of prediction error minimisation will impact on policy selection" (Perrykkad and Hohwy 2020, p. 3). Accordingly, based on previous dynamic patterns of Rate, formal objects arise to predict how Rate is going to unfold (descriptive) and (based on Expected Rate) to regulate actions to minimise Error over time (directive). Here, we are just extending to feelings the PP idea that "agents are assumed to infer 'what is the current state of the world' and 'what are the best actions to pursue' based on the same generative model of the environment" (Schwartenbeck et al. 2019, p. 3). According to AIT, when it comes to *inferring* the dynamics of Error (i.e. Expected Rate) and what are the best actions to pursue (based on Expected Rate), then we are talking about formal objects that are at once *affective* models and *affective* action policies.

Here, it is worth revisiting some of the discussion about IIT in Sect. 3. We noted that, for IIT, the central function of instrumental interoceptive inference is the model-based *regulation* of interoceptive variables. According to AIT, this function is fulfilled through the modelling of Expected Rate (the evaluative aspect of affective experience), which, as we have just seen, involves action policies (the regulative aspect of affective experience). This is possible because, as we saw in Sect. 4, the organism employs Expected Rate not only to optimise uncertainty, but also to direct the organism towards allostasis. Importantly, the regulation does not only happen through proprioceptive actions but also through intero-actions (autonomic reflex arcs).

Pezzulo and colleagues provide a useful account to further specify the path from feelings to behaviour via formal objects (and precision). They argue that control (goal selection) and motivation (goal prioritization) are two intertwined parts of the same process of active inference. So much so, that they prefer to talk in terms of controlled motivation, or motivated control. A deep hierarchy generates goals (action policies, in the terms of our present discussion), which influence the value of different actions (i.e. motivation), but the maintenance and propagation of these goals are dependent on expected precision, which is again dependent on motivation (Pezzulo et al. 2018). As for formal objects *qua* action policies, this is consistent with our idea that formal objects work as a partial model of the particular object that can help the agent direct her actions towards the desirable values of Rate—a given action policy (formal object) increases the expected precision of the actions that are inferred to fulfil said action policy, making those actions more likely. Of course, feelings emerge in a context-sensitive fashion, and the same is true of the resulting actions. To say that a feeling (e.g. fear) has a formal object (e.g. danger) that directs action in particular ways (e.g. running), is not the same as saying that a feeling corresponds to a rigid set of actions (e.g. depending on the context, an agent might decide to flee or to fight, and will implement either strategy

in an adaptive way). This context-sensitivity is something that is acknowledged not only in discussions of affective experience, but also in the account of action policies expounded by Pezzulo and colleagues, who argue that motivational control is a way to deal with complex hierarchies of contextual constraints (Pezzulo et al. 2018).

On the one hand, the idea of motivated control provides a PP mechanism for action tendencies and for the conception of feelings as forms of affordance-sensing (Proust 2015). On the other hand, the AIT regularity account of formal objects provides a plausible account for how goals become "affectively meaningful" (Pezzulo et al. 2018, p. 296). What is interesting in the account of Pezzulo and colleagues is that it shows how Expected Rate can then influence the formal object and the resulting action in a dynamic way. In their own words, "this view may help understand the multifarious phenomenology of goal processing, such as the positive emotions associated with progress towards the goal (anticipation, enthusiasm) and the negative emotions associated with failures (disappointment, regret), in terms of increased (or decreased) confidence that the selected policy will achieve the desired goals" (Pezzulo et al. 2018, p. 304). Think now of formal objects *qua* models: If there are unexpected changes in Rate, this will decrease the precision of the formal object as a model of Rate, so that either the model will need to adapt (again, in an evaluative-regulative fashion, as with increasing frustration about a difficult task) or it will be substituted by a different model (e.g. going from anticipation to disappointment).

In a recent paper, Kiverstein and colleagues suggest that valence lies directly in the affordances of the environment, so that a nearby apple will have positive valence if it looks tasty and negative valence if it looks rotten (Kiverstein et al. 2019). In our view, valence lies in expected error dynamics, namely Expected Rate, and constitutes the primary intentionality of the affective experience. Valence only gets tied to environmental objects (i.e. particular objects) and to affordances (i.e. the directive aspect of formal objects) derivatively through the prediction of regularities. This constitutes the secondary intentionality of the experience. AIT's regularity account, together with the conception of formal objects as both descriptive and directive models, shows how affordances and environmental objects can gain their apparent valence, which is better understood as object valence derived from phenomenal valence (see also footnote 4). In consonance with a recent computational model of valence and policy selection, we think that it is better to think of affordances as "affectively charged" (Hesp et al. 2019).[41]

The idea that affect is linked to affordances is particularly compelling when we consider cases of prototypical negatively valenced feelings, as in the example of encountering a bear. In said example, the formal object "danger" is not simply a description of the bear, but a display of affordances that are experienced as directive pushes for action, such as raising one's arms, or turning around and run. However, it does seem that not all feelings are clearly linked up to specific affordances in this

[41] An anonymous reviewer brought to our attention this manuscript, which is congenial to the views expressed here. The computational model of Hesp and colleagues can be seen as an extension of the principles behind EDT (Joffily and Coricelli 2013) to the inference of future states of subjective fit (i.e. affective inference) and consequent action selection. Whereas our work is mainly philosophical in nature, their work is mainly computational in nature, but we see this convergence as an auspicious sign.

way. Take for instance cases of positive affective experiences: What is the affordance of pleasant relaxation or simple joy or happiness?

Some formal objects are best understood as affordance-laden hypotheses while others, perhaps, are not. Here the PP framework offers a natural way to understand the gradation between descriptive and directive predictions. It is useful to think of the formal objects of feelings associated with a lot of action-relevant content as *bona fide* affordances (as e.g. in fear), but there is a gradation to the amount of action-relevant content a prediction can have. In other words, some feelings qua formal objects are more directive than others. Additionally, we need to keep in mind that some affective experiences will be immediately and obviously directive (e.g. running away) while others will be immediate but much more subtle (e.g. recalling pleasant memories). Yet others will only raise the probabilities of situationally removed future behaviours (e.g. seeking out more works by a given artist in the future).

6 Conclusion

We started this paper by outlining the characteristics of affective experiences which a PP theory should aim to account for. Affective experiences are phenomenally conscious. A central aspect of this phenomenal nature is *phenomenal valence*, which we argued is the most promising candidate for the mark of the affective. Furthermore, affective experiences motivate and guide behaviour and are intentional, that is they are about something. Traditionally, their intentionality is thought to comprise a specific object they are directed at, say a body part or a bear, and a feeling-specific property assigned to this object by the affective experience. These are respectively called the particular and formal object of the feeling.

Secondly, we reviewed the existing theories of affective experience in the PP framework and divided them into two families: IIT and EDT. We showed the virtues and shortcomings of the two families of theories in accounting for some of the outlined features. Regarding IIT, we argued that while it provides an excellent account of the role of affect in instrumental interoceptive inference and allostasis, it does not provide an adequate mark of the affective.[42] EDT, on the other hand, offers promising computational accounts of valence, and by extension of the mark of the affective. However, EDT disregards a basic premise in PP: that it is predictions, and not error, that determine conscious experience. If we want to account for feelings as phenom-

[42] In the end, AIT is a synthesis and it is therefore potentially compatible with revised versions of both IIT and EDT. A recent paper by Tschantz, Seth and Buckley propose an action-oriented model of goal-oriented and epistemic behaviour based on expected free-energy (Tschantz et al. 2020). A version of IIT in which valence were cast in terms of expected free-energy (in the style of Joffily and Coricelli 2013)—so that, in the analogous terms of prediction, valence could be understood as expected Rate—would be very much in line with AIT (thanks to an anonymous reviewer for suggesting this possibility). Allostasis is a biological imperative, so following the free-energy principle, free-energy minimisation will lead to allostasis. Feelings will guide organisms to minimising Error and, by extension, to allostasis, so there will be a tight connection between affect and interoception in the lines proposed by IIT. The crucial difference between AIT and IIT is that AIT is committed to valence being cast in terms of predictions (of Rate), not directly in terms of interoception.

enally conscious affective states, we need a theory that casts feelings as predictions, and not errors.

Finally, we synthesised the discussed theories into AIT and showed how this revised theory solves the issues of the previous theories and provides a satisfactory account of the characteristic features of affective experiences. According to AIT, valence corresponds to an agent's prediction of her own success in modelling the world (Expected Rate). Through the tracking and prediction of regularities, valence gets experientially bound to a particular object (the inferred cause of expected changes in Rate) and a formal object (a model of expected changes in Rate). We claim that in most cases formal objects are better understood not as passive properties of the particular object, but as pushmi-pullyu representations that are both descriptive and directive. Affective experiences favour certain behaviours, and the function of formal objects is to model the particular object and its link to Expected Rate so that action can emerge in a regulative fashion—enabling feelings to fulfil their role of guiding the organism to allostasis and to optimal levels of uncertainty in order to minimise prediction error over time.

Acknowledgements We would like to thank José Araya, Valérian Chambon, Andy Clark, Marco Inchingolo, Solène Le Bars, Elisabeth Pacherie, Agostino Pinnapintor, George Neish, Takuya Niikawa and Nura Sidarus for their helpful comments on previous drafts of this paper. We would also like to thank two anonymous reviewers for their insightful and constructive comments and suggestions, which helped us to significantly improve the paper. This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement Number 675415; and the Agence Nationale de la Recherche under Grant Agreement Numbers ANR-17-EURE-0017 (FrontCog) and ANR-10-IDEX-0001-02 (PSL).

References

Aristotle. (1986). *De anima (on the soul)* (H. Lawson-Tancred, Trans.). Harmondsworth: Penguin Books.

Auvray, M., Myin, E., & Spence, C. (2010). The sensory-discriminative and affective-motivational aspects of pain. *Neuroscience and Biobehavioral Reviews, 34*(2), 214–223.

Bain, D. (2013). What makes pains unpleasant? *Philosophical Studies, 166*(1), 69–89.

Bain, D. (2014). Pains that don't hurt. *Australasian Journal of Philosophy, 92*(2), 305–320.

Barrett, L. F. (2006). Valence is a basic building block of emotional life. *Journal of Research in Personality, 40*(1), 35–55.

Barrett, L. F. (2014). The conceptual act theory: A precis. *Emotion Review, 6*(4), 292–297.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience, 12*(1), 1–23.

Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology, 41,* 167–218.

Barrett, L. F., & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience, 16*(7), 419–429.

Batson, C. D., Shaw, L. L., & Oleson, K. C. (1992). Differentiating affect, mood, and emotion: toward functionally based conceptual distinctions. *Emotion, 13,* 294–326.

Berridge, K. C., & Kringelbach, M. L. (2015). Pleasure systems in the brain. *Neuron, 86*(3), 646–664.

Berthier, M., Starkstein, S., & Leiguarda, R. (1988). Asymbolia for pain: A sensory-limbic disconnection syndrome. *Annals of Neurology, 24*(1), 41–49.

Brown, C. A., Seymour, B., Boyle, Y., El-Deredy, W., & Jones, A. K. (2008). Modulation of pain ratings by expectation and uncertainty: Behavioral characteristics and anticipatory neural correlates. *Pain, 135*(3), 240–250.

Carruthers, P. (2017). Valence and value. *Philosophy and Phenomenological Research, 97*(3), 658–680.

Carver, C. S., & Scheier, M. F. (1990). Origins and functions of positive and negative affect: a control-process view. *Psychological review*, *97*(1), 19.
Carver, C. S., & Scheier, M. F. (2001). On the self-regulation of behavior. Cambridge University Press.
Chanda, M. L., & Levitin, D. J. (2013). The neurochemistry of music. *Trends in Cognitive Sciences, 17*(4), 179–193.
Charland, L. (2005). The heat of emotion: Valence and the demarcation problem. *Journal of Consciousness Studies, 12*(8–10), 8–10.
Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford: Oxford University Press.
Clore, G. L. (1994). Why emotions are never unconscious. In P. Ekman & R. J. Davidson (Eds.), *The nature of emotion: Fundamental questions* (pp. 285–290). New York: Oxford University Press.
Colombetti, G. (2005). Appraising valence. *Journal of Consciousness Studies, 12*(8–10), 8–10.
Corder, G., Ahanonu, B., Grewe, B. F., Wang, D., Schnitzer, M. J., & Scherrer, G. (2019). An amygdalar neural ensemble that encodes the unpleasantness of pain. *Science*, *363*(6424), 276–281.
Corns, J. (2014). Unpleasantness, motivational oomph, and painfulness. *Mind and Language, 29*(2), 238–254.
Craig, A. D. (2003). Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology, 13*(4), 500–505.
Craig, A. D. (2009). How do you feel—Now? The anterior insula and human awareness. *Nature Reviews Neuroscience, 10*(1), 59–70.
Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A., & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience, 7*(2), 189.
Damasio, A. R. (1994). *Descartes' error: Emotion, rationality and the human brain*.
Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience, 14*(2), 143–152.
De Haan, S., Rietveld, E., Stokhof, M., & Denys, D. (2013). The phenomenology of deep brain stimulation-induced changes in OCD: An enactive affordance-based model. *Frontiers in Human Neuroscience, 7,* 653.
De Sousa, R. (1987). *The rationality of emotion*. Cambridge: MIT Press.
Deonna, J. A., & Teroni, F. (2017). Getting bodily feelings into emotional experience in the right way. *Emotion Review, 9*(1), 55–63.
Döring, S. A. (2007). Seeing what to do: Affective perception and rational motivation. *Dialectica, 61*(3), 363–394.
Duncan, S., & Barrett, L. F. (2007). Affect is a form of cognition: A neurobiological analysis. *Cognition and Emotion, 21*(6), 1184–1211.
Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4,* 215.
FitzGerald, T. H., Friston, K. J., & Dolan, R. J. (2012). Action-specific value signals in reward-related regions of the human brain. *Journal of Neuroscience, 32*(46), 16417–16423.
Frijda, N. H. (1986). *The emotions*. Cambridge: Cambridge University Press.
Frijda, N. H. (2005). Emotion experience. *Cognition and Emotion, 19*(4), 473–497.
Frijda, N. H. (2007). *The laws of emotion*. OCLC: 938467399. Mahwah: Lawrence Erlbaum Associates.
Friston, K. (2003). Learning and inference in the brain. *Neural Networks, 16*(9), 1325–1352.
Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews, 68,* 862–879.
Friston, K., Mattout, J., & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics, 104*(1–2), 137–160.
Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience, 6*(4), 187–214.
Fulkerson, M. (2019). Emotional perception. *Australasian Journal of Philosophy, 98,* 16–30.
Gerrans, P. (2015). All the self we need. *Open MIND*.
Gill, C. (2010). *Naturalistic psychology in galen and stoicism*. Oxford: Oxford University Press.

Goldie, P. (2002). Emotions, feelings and intentionality. *Phenomenology and the Cognitive Sciences, 1*(3), 235–254.
Gu, X., Hof, P. R., Friston, K. J., & Fan, J. (2013). Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology, 521*(15), 3371–3388.
Helm, B. W. (2009). Emotions as evaluative feelings. *Emotion Review, 1*(3), 248–255.
Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K., & Ramstead, M. (2019). Deeply felt affect: The emergence of valence in deep active inference. https://doi.org/10.31234/osf.io/62pfd.
Hohwy, J. (2011). Phenomenal variability and introspective reliability. *Mind and Language, 26*(3), 261–286.
Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3,* 96.
Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.
Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition, 47,* 75–85.
Hohwy, J. (2020). New directions in predictive processing. *Mind and Language, 35*(2), 209–223.
Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition, 108*(3), 687–701.
Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences, 8*(11), 494–500.
Horgan, T., & Tienson, J. (2002). The intentionality of phenomenology and the phenomenology of intentionality. In D. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 520–533). Oxford: Oxford University Press.
James, W. (1884). What is an emotion? *Mind, 9*(34), 188–205.
Joffily, M., & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology, 9*(6), e1003094.
Kenny, A. (1963). *Action, emotion and will*. London: Routledge & Kegan Paul.
Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE, 7*(5), e36399.
Kiverstein, J., Miller, M., & Rietveld, E. (2019). The feeling of grip: Novelty, error dynamics, and the predictive brain. *Synthese, 196*(7), 2847–2869.
Klein, C. (2015). What pain asymbolia really shows. *Mind, 124*(494), 493–516.
Kozuch, B. (2018). No pain, no gain (in Darwinian fitness): A representational account of affective experience. *Erkenntnis, 85,* 693–714.
Kriegel, U. (2014). Towards a new feeling theory of emotion: Towards a new feeling theory of emotion. *European Journal of Philosophy, 22*(3), 420–442.
Lacewing, M. (2007). Do unconscious emotions involve unconscious feelings? *Philosophical Psychology, 20*(1), 81–104.
Lange, C. G. (1887). *Uber Gemütsbewegungen. Ihr Wesen Und Ihr Einfluß Auf Körperliche, Besonders Auf Krankhafte Lebenserscheinungen. Ein Medizinisch-Psychologische Studie.* T. Thomas.
Lazarus, R. S. (1991). *Emotion and adaptation*. Oxford: Oxford University Press.
Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology, 22*(6), 1027–1038.
Mathews, A., & MacLeod, C. (2002). Induced processing biases have causal effects on anxiety. *Cognition and Emotion, 16*(3), 331–354.
McGlone, F., Wessberg, J., & Olausson, H. (2014). Discriminative and affective touch: Sensing and feeling. *Neuron, 82*(4), 737–755.
Meinong, A. (1917). *Uber emotionale Präsentation*. Vienna: A. Hölder.
Melzack, R., & Wall, P. D. (1988). *The challenge of pain*. London: Penguin.
Mendelovici, A. (2014). Pure intentionalism about moods and emotions. In U. Kriegel (Ed.), *Current controversies in philosophy of mind* (pp. 135–157). New York: Routledge.
Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.
Michal, M., Reuchlein, B., Adler, J., Reiner, I., Beutel, M. E., Vögele, C., et al. (2014). Striking discrepancy of anomalous body experiences with normal interoceptive accuracy in depersonalization-derealization disorder. *PLoS ONE, 9*(2), e89823.
Millikan, R. G. (1995). Pushmi-Pullyu representations. *Philosophical Perspectives, 9,* 185–200.
Pernu, T. K. (2017). The five marks of the mental. *Frontiers in Psychology, 8,* 1084.
Perrykkad, K., & Hohwy, J. (2020). Fidgeting as self-evidencing: A predictive processing account of non-goal-directed action. *New Ideas in Psychology, 56,* 100750.

Pessoa, L. (2005). To what extent are emotional visual stimuli processed without attention and awareness? *Current Opinion in Neurobiology, 15*(2), 188–196.
Pessoa, L., & Adolphs, R. (2010). Emotion processing and the amygdala: From a 'low road' to 'many roads' of evaluating biological significance. *Nature Reviews Neuroscience, 11,* 773–782.
Pezzulo, G., & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences, 20*(6), 414–424.
Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology, 134,* 17–35.
Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences, 22*(4), 294–306.
Ploner, M., Freund, H. J., & Schnitzler, A. (1999). Pain affect without pain sensation in a patient with a postcentral lesion. *Pain*, *81*(1–2), 211–214.
Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford: Oxford University Press.
Prinz, J. (2005). Are emotions feelings? *Journal of Consciousness Studies, 12*(8–9), 9–25.
Prinz, J. J. (2006). Is emotion a form of perception? *Canadian Journal of Philosophy, 36*(sup1), 137–160.
Prinz, J. (2010). For valence. *Emotion Review, 2*(1), 5–13.
Proust, J. (2013). *The philosophy of metacognition: Mental agency and self-awareness*. Oxford: Oxford University Press.
Proust, J. (2015). The representational structure of feelings. In T. Metzinger & J. M. Windt (Eds.), *Open MIND*. Frankfurt am Main: MIND Group.
Rubins, J. L., & Friedman, E. D. (1948). Asymbolia for pain. *Archives of Neurology & Psychiatry, 60*(6), 554–573.
Scarantino, A. (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.
Schroeder, T. (2004). Three faces of desire. Oxford University Press.
Schwartenbeck, P., Passecker, J., Hauser, T. U., FitzGerald, T. H., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *Elife*, *8*, e41703.
Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17*(11), 565–573.
Seth, A. K. (2015). The cybernetic bayesian brain. In T. Metzinger & J. M. Windt (Eds.), *Open MIND*. Frankfurt am Main: MIND Group.
Seth, A. K., & Friston, K. J. (2016). Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1708), 20160007.
Seth, A. K., Suzuki, K., & Critchley, H. D. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology, 2,* 395.
Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences, 22*(11), 969–981.
Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., Gard, T., et al. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience, 10,* 550.
Tappolet, C. (2000). *Emotions et Valeurs*. FeniXX.
Teroni, F. (2018). Emotionally charged—The puzzle of affective valence. In C. Tappolet, F. Teroni, & A. Konzelmann Ziv (Eds.), *Shadows of the soul: Philosophical perspectives on negative emotions* (pp. 10–19). New York: Routledge.
Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system. *Nature Reviews Neuroscience, 21,* 231–242.
Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). Learning action-oriented models through active inference. *PLoS Computational Biology, 16*(4), e1007805.
Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind*. Cambridge: MIT Press.
Tye, M. (2008). The experience of emotion: An intentionalist theory. *Revue Internationale de Philosophie, 243*(1), 25–50.
Uhelski, M. L., Davis, M. A., & Fuchs, P. N. (2012). Pain affect in the absence of pain sensation: Evidence of asomaesthesia after somatosensory cortex lesions in the rat. *Pain, 153*(4), 885–892.
Van De Cruys, S. (2017). Affective value in the predictive mind. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Van de Cruys, S., & Wagemans, J. (2011). Putting reward in art: A tentative prediction error account of visual art. *i-Perception, 2*(9), 1035–1062.

Wiese, W. (2018). *Experienced wholeness: Integrating insights from gestalt theory, cognitive neuroscience, and predictive processing*. Cambridge: MIT Press.

Winkielman, P., & Berridge, K. C. (2004). Unconscious emotion. *Current Directions in Psychological Science, 13*(3), 120–123.

Winkielman, P., Berridge, K. C., & Wilbarger, J. L. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality and Social Psychology Bulletin, 31*(1), 121–135.