Predictive Processing and the Phenomenology of Time Consciousness

A Hierarchical Extension of Rick Grush's Trajectory Estimation Model

Wanja Wiese

This chapter explores to what extent some core ideas of predictive processing can be applied to the phenomenology of time consciousness. The focus is on the experienced continuity of consciously perceived, temporally extended phenomena (such as enduring processes and successions of events). The main claim is that the hierarchy of representations posited by hierarchical predictive processing models can contribute to a deepened understanding of the continuity of consciousness. Computationally, such models show that sequences of events can be represented as states of a hierarchy of dynamical systems. Phenomenologically, they suggest a more fine-grained analysis of the perceptual contents of the *specious present*, in terms of a hierarchy of temporal wholes. Visual perception of static scenes not only contains perceived objects and regions but also spatial *gist*; similarly, auditory perception of temporal sequences, such as melodies, involves not only perceiving individual notes but also slightly more abstract features (*temporal* gist), which have longer temporal durations (e.g., emotional character or rhythm). Further investigations into these elusive contents of conscious perception may be facilitated by findings regarding its neural underpinnings. Predictive processing models suggest that sensorimotor areas may influence these contents.[1]

Keywords

Auditory perception | Bayesian brain | Consciousness | Event segmentation theory | Ideomotor principle | Phenomenology | Predictive processing | Specious present | Time consciousness | Trajectory estimation model

The aim of this chapter is to try to connect research on predictive processing (PP) with research on the phenomenology of time consciousness. The motivation for this comes, on the one hand, from Grush's work on temporal perception,[2] and, on the other, from Hohwy's work on prediction error minimization.

On the first page of his monograph *The Predictive Mind*, Hohwy suggests that "the idea that the brain minimizes its prediction error […] explains not just that we perceive but *how* we perceive: the idea applies directly to key aspects of the phenomenology of perception." (Hohwy 2013, p. 1). Here, I will attempt to apply this idea to key aspects of the phenomenology of *temporal* perception. Luckily, we can build on existing work by Grush (Grush 2005), who has developed a model of temporal perception he calls the trajectory estimation model (TEM). This draws on control and filtering models, but is pitched at a level of abstraction which makes it compatible with specific PP models. A central point for the purposes of this chapter is that TEM does not posit a hierarchy of representations, whereas the types of PP models considered here do. As I shall argue, extending TEM by drawing on features of hierarchical PP models can help account for key aspects of the phenomenology of temporal perception (which are not addressed by Grush's TEM). I shall call the resulting extension of TEM the hierarchical trajectory estimation model (HiTEM).

The paper is structured as follows. In section 1, I briefly review models of temporal consciousness and highlight two features of conscious temporal perception, which I shall call **endurance** and **continuity**. In section 2, I explain the basic aspects of Grush's TEM and formulate a question, which I call the **interface question** and which is not addressed by TEM. Crucially, providing an answer to this

[1] I am highly grateful to Martin Butz, Jakob Hohwy, Marius Jung, Thomas Metzinger, Mark Miller, Iuliia Pliushch, and Lisa Quadt for providing a number of very useful comments on drafts of this paper. Thanks to Robin Wilson for excellent editorial help.

[2] "Temporal perception" should here be understood as a shorthand for "perception of temporally extended processes or events".

Wiese, W. (2017). Predictive Processing and the Phenomenology of Time Consciousness -
A Hierarchical Extension of Rick Grush's Trajectory Estimation Model.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing*: 26. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573277

question would be necessary to account for **endurance** and **continuity**. To extend TEM, I then explain core features of a computational model by Kiebel and colleagues (Kiebel et al. 2008a) of how the brain represents temporal sequences (section 3). Generalizing from this model, I develop an extension of TEM: HiTEM (section 4). HiTEM provides an answer to the interface question (as I show in section 5). It also suggests how to account for **endurance** and **continuity** (section 6). In section 7, I offer some tentative remarks on what HiTEM says about the contents of consciousness, considering empirical findings on the neural underpinnings of auditory perception.

1 The Phenomenology of Time Consciousness

Experiencing successions of events, as with a series of notes comprising a melody, poses a puzzle. It seems that neither experiencing the different notes simultaneously nor experiencing them in sequence can give rise to the experience of succession. If we experience all notes simultaneously, we experience not a melody but a chord. If we experience first one note, then another, this is a succession of experiences, not an experience of succession (cf. James 1890). So how can we conceive of the experience of successions of events, and of temporally extended processes in general? The two dominant approaches are what Dainton (Dainton 2014) calls *extensional* and *retentional* models, respectively. Interestingly, although these models entail different metaphysical[3] claims about temporal consciousness, they are not necessarily committed to different phenomenological assertions (cf. Dainton 2014, § 3).

According to extensional models, an experience of a succession of events involves a temporally extended experience with proper temporal parts. These correspond to the different temporal parts of the experienced succession of events. For instance, experiencing a succession of two notes involves a *single* experience corresponding to the entire experienced temporal whole (the succession of notes), but the global content of this experience has two temporal parts – one for the first note and one for the second. The notes are experienced as successive, not simultaneous, because the corresponding temporal parts of the total experience are not simultaneous, but successive. In other words, the temporal structure of conscious experience matches the apparent[4] temporal structure of the experienced events (Watzl 2013 calls this the *structural matching thesis*).

According to retentional models, experiencing a succession of events does not always involve a succession of experiences. At least on short timescales, conscious experiences are atomic (cf. Lehmann 2013). Here, "atomic" does not mean that the neural underpinnings are static: This type of atomicity is compatible with the assumption that the neural underpinnings of conscious experiences are always temporally extended (cf. Lee 2014). It just means that the proper temporal parts of a conscious experience cannot be mapped onto different temporal parts of an experienced temporal whole (such as a succession of events). So retentional models reject the assumption that the temporal structure of conscious experience always matches the apparent temporal structure of the experienced events. Still, an experience of a succession has *synchronous* parts which can be mapped onto the different elements of the succession. The parts of the succession are not experienced *as* simultaneous (although they are simultaneously experienced) because the different parts of the experience do not all represent their targets in the same way. As a result, the different events in the succession are represented *as temporally related*. In Husserl's words, events which are just past are represented by *retentions* (cf. Husserl 1991; hence Dainton's label "retentional models").

Disagreements between extensional and retentional models thus mainly concern the metaphysics of our momentary conscious experience (what we are experiencing now, "as present"). According to

[3] Metaphysical claims about consciousness deal, for instance, with the relationship between conscious processes and neural activity, or with properties conscious experiences are deemed to have. Phenomenological claims, by contrast, deal with how consciousness appears from the first-person perspective and with the contents of consciousness.

[4] This qualification is important to allow for temporal illusions in which, for instance, the actual order of a succession of events is misperceived. The apparent temporal structure would then be the temporal structure *as it is experienced*.

extensional models, momentary conscious experiences have different experiences as proper temporal parts. According to retentional models, they don't (see figure 1 for an illustration).

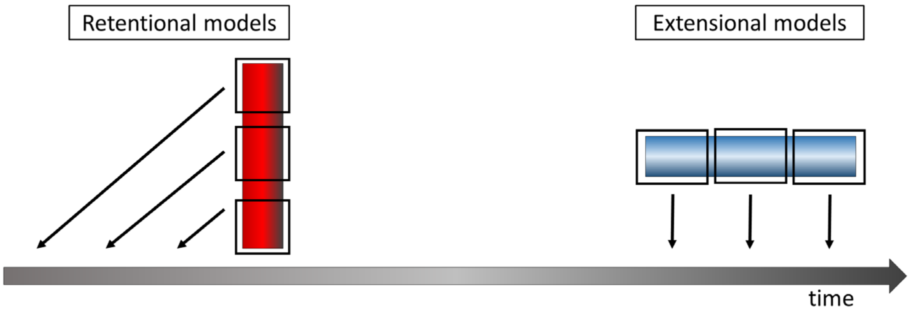

Figure 1: Two conceptions of the specious present: retentional versus extensional models. According to retentional models (the retentional specious present is highlighted in red), different stages of an experienced temporal process are present in consciousness at the same time. According to extensional models (the extensional specious present is highlighted in blue), the conscious experience of a temporal process is itself a temporal process, with proper temporal parts corresponding to the temporal parts of the experienced process. For further details, see the discussion in (Dainton 2014).

What these models agree on is the phenomenological claim that momentary conscious experience constitutes a *specious present* (cf. James 1890). The contents of the specious present comprise an interval extended in time but with parts that are all present (so they are experienced "at the same time", but not *as simultaneous*). James famously afffirmed:

> The unit of composition of our perception of time is a duration, with […] a rearward- and a forward-looking end. It is only as parts of this duration-block that the relation of succession of one end to the other is perceived. We do not first feel one end and then feel the other after it, and from the perception of the succession infer an interval of time between, but we seem to feel the interval of time as a whole, with its two ends embedded in it. (James 1890, pp. 609-610)[5]

What the models disagree about is whether the temporal extension of the specious present itself is explanatorily relevant for the phenomenological claim. The extensionalist claims that the specious present can contain an experience of enduring processes or successions of events, because the specious present itself is a succession of conscious experiences, or an enduring conscious experience. The retentionalist, on the other hand, claims that the specious present can comprise an experience of enduring processes or successions of events, because it has synchronous proper parts which are directed at different times (or represent events which are occurring at different times).

An example of such a retentional model is Grush's trajectory estimation model (TEM) – at least it shares the central intuition of this class of models. According to TEM, the content of the specious present can be described as a trajectory estimate, which contains estimates of what is happening at different times.

Combining this basic idea with theoretical research on predictive processing, I argue that a more fine-grained phenomenological analysis of temporal consciousness can be provided: The content of the specious present is not best conceived as a linear stream of events, but rather as a hierarchy of temporal wholes.[6] I try to show that this view can account for two features of temporal consciousness,

[5] Some people disagree with this description and claim that the contents of consciousness are more like dynamic snapshots (cf. Prosser 2016; this is compatible with the assumption that the neural underpinnings of such snapshots are always temporally extended). I restrict the treatment in this chapter to accounts which are compatible with the specious present view, to avoid making the discussion unnecessarily complicated.

[6] A hint at a similar idea can be found in Thomas Metzinger's *Being No One*: "[C]onvolved holism also reappears in the phenomenology of time experience: Our conscious life emerges from integrated psychological moments, which, however, are themselves integrated into the flow of subjective time." (Metzinger 2004[2003], p. 151).

which are treated as primitive by existing models or left unaddressed. This view can be construed as an extension of TEM, but the central phenomenological analysis (according to which the contents of the specious present consist of a hierarchy of wholes) is compatible with all models that embrace the view that the content of momentary conscious experience comprises an interval (which is common ground between retentional and extensional models). I focus on TEM and not on other retentional, or even extensional, models because TEM is formulated in computational terms, which makes a connection with PP models and further development relatively straightforward.

Conceiving of the contents of the specious present as a hierarchy of temporal wholes can help clarify the following two features of temporal experience:

Continuity $=_{\mathrm{Df}}$ At least sometimes, we experience smooth successions of events (or smooth changes). An example is a series of notes played *legato* by a single instrument (contrast this with a series played *staccato*). Such sequences are experienced as temporal continua (which, strictly speaking, would involve an infinite number of events).

Endurance $=_{\mathrm{Df}}$ At least sometimes, we experience temporally extended events as enduring. An example is an opera singer holding a single note for an extended period (this example is taken from Kelly 2005, p. 208). By contrast, when one is surprised by a sudden bright flash, this punctual event is not experienced as part of an enduring event.

These features are not mutually independent. **Continuity** implies **endurance**: When we experience a temporal continuum, we experience a dynamic event, in which a higher-order event is experienced as enduring through change. This idea is not completely new (see Prosser 2016, especially p. 172) and Zacks' event segmentation theory (EST) is related (cf. Zacks et al. 2007), although there are important differences. I explain it in more detail below, having illustrated the two features and shown that they pose a challenge to Grush's TEM. I draw on a computational PP model by Kiebel et al. to provide a theoretical sketch of how the features can be accounted for and I review some empirical results which enrich the proposal.

2 The Trajectory Estimation Model (TEM)

TEM is an abstract[7] model of how the brain represents consciously experienced, temporally extended sequences at small timescales (on the order of 200 ms,[8] see Grush 2006, p. 444). A core assumption is that, at such timescales, consciously experienced events are represented as related by temporal relationships such as "earlier than" or "simultaneous with", not as events that are occurring "now" or will occur in the future (cf. Grush 2016, p. 8). Consequently, when one event is represented as occurring earlier than another, this does not entail that one is experienced as less real. Furthermore, the content of perception at a time comprises a trajectory – an ordered tuple of events, not just events occurring

[7] The model is abstract in the sense that it does not specify which exact process models are computed by the brain and which exact computational strategies are employed to generate trajectory estimates (cf. Grush 2005, p. S218).

[8] Why does Grush assume that this interval has a length of around 200 ms? The assumption is motivated by research on temporal illusions, especially *postdictive phenomena* (cf. Shimojo 2014), in which a percept of a stimulus is influenced by input received around 100-200 ms after the first stimulus presentation. A classic example is a type of apparent motion in which two stimuli of different colors are used (often called "colored phi", see Kolers and von Grunau 1975). When, say, the brief presentation of a green spot is followed by the presentation of a red spot, this can lead to the percept of a moving spot which abruptly changes its color (from green to red). Clearly, this percept cannot be formed before the second stimulus has been processed. This means that the percept is a function of sensory signals obtained over an interval of time. Since apparent motion is perceived when stimuli are separated by an interval of around 100 ms, this suggests momentary conscious perception reaches "into the past". Similarly, research on an effect called "representational momentum" (see Thornton and Hubbard 2002; Hubbard 2014) suggests that momentary conscious perception also reaches "into the future", i.e., it comprises representations of anticipated events (just about to happen, in the very near future). From such results, Grush concludes that conscious perception presents us with events which are currently happening, events which have just happened, and events which are expected to happen, so conscious perception has "a lag and reach on the order of 100 ms each, for a total temporal magnitude on the order of 200 ms." (Grush 2006, p. 444). Note that this is a claim about conscious perception, not about activity in the brain as such. For the purposes of this paper, nothing hinges on the exact temporal extension of the interval. However, Grush's considerations do make it plausible that the interval covers only a fraction of a second.

simultaneously. In particular, it involves *smoothed*[9] and *predicted* estimates respectively of future and past events, to capture the intuition underlying the posit of *retentions* and *protentions* in Husserl's account of time consciousness (see Grush 2006). Representations of perceived events also comprise *filtered* estimates. These combine current sensory information with prior knowledge about the target (see Grush 2008, p. 152). Formally, Grush describes the trajectory estimate as follows:

> With such tools in place, it is possible to describe a system that combines smoothing, filtering and prediction to maintain an estimate of the trajectory of the modeled domain over the temporal interval *[t - j, t + k]*, by determining, at each time *t*, the following ordered *j* + *k* + 1–tuple: $(\tilde{p}(t-j), \tilde{p}(t-j+1), \ldots, \hat{p}(t), \bar{p}(t+1), \ldots, \bar{p}(t+k))$. (Grush 2005, p. S211)

Here, $\tilde{p}$ denotes a smoothed estimate, $\hat{p}$ a filtered estimate, and $\bar{p}$ a predicted (prospective) estimate. As Grush shows, TEM can account for a variety of perceptual illusions (see Grush 2005; Grush 2006; Grush 2008). Since TEM is a model of conscious *perception*, its scope is explicitly restricted to perceptual representations and, more specifically, to perceptual representations of what is currently happening (within an interval of approximately 200 ms).[10] See figure 2 for an illustration.

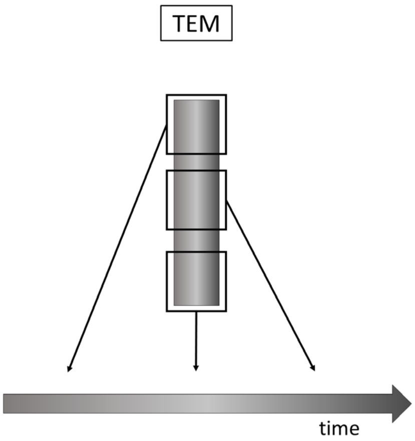

Figure 2: The trajectory estimation model (TEM). Estimates of what is happening at different times are computed simultaneously. Hence, what is experienced at a time is an interval (a succession of events, or a temporally extended process).

Intuitively, it should be plausible that there is difference between *perceiving* events that are currently happening and vividly *remembering* events that happenend in the past, or *imagining* events that may happen in the future. According to Grush, these different conscious experiences involve different types of representation. Remembering and imagining involve *conceptual* representations, while experiencing events that are currently happening involves *perceptual* representations:

[9] "Smoothing" is the technical term for methods in which an estimate at a given time step is generated by taking measurements obtained after that time step into account. An example is the moving average method, in which an estimate at a time is the average of a set of data (obtained before and after that time).

[10] TEM bears an apparent similarity to event segmentation theory (EST). *Event segmentation* refers to the capacity of dividing perceptual streams into meaningful chunks (see Zacks 2008 for a brief introduction). EST posits *event models* to account for this capacity. The apparent similarity to TEM is that an "event model is a representation of 'what is happening now,' which is robust to transient variability in the sensory input." (Zacks et al. 2007, p. 274). Similarly, trajectory estimates in TEM code the contents of the specious present, which is our conscious experience of "what is happening now". There is a huge difference, however, between the temporal grain of events that are relevant to TEM and EST respectively. EST deals with events that have a relatively long duration (several seconds, see Zacks et al. 2001, p. 653; Zacks et al. 2007, p. 274), whereas TEM applies only to events that occur within a fraction of a second. However, some aspects of the information-processing strategy implied by EST may have fruitful connections to the account sketched in this paper. For instance, Zacks et al. "hypothesize that the architecture in EST is implemented simultaneously on a range of timescales, spanning from a few seconds to tens of minutes." (Zacks et al. 2007, p. 276). Similarly, the hierarchical extension of TEM sketched here, HiTEM, posits multiple timescales over which perceptual estimates are computed. Again, however, the relevant temporal grain is much smaller than that referred to by Zacks et al. (and there are some more subtle differences, see section 4 for details).

> There are two ways in which it could be plausibly maintained that contents characterizable only in temporal interval terms play a role in experience. One, which potentially spans a larger interval, might be described as *conceptual* in the sense that it is a matter of interpreting present experience in terms of concepts of processes that span potentially large intervals. Music appreciation would fall into this category. When I recognize something as part of a larger whole (a spatial whole or a temporal whole), then my concept of that whole influences the content grasped via the part. Something along these lines is what appears to be happening with music. On the other hand, there is what might be called a perceptual or phenomenal phenomenon of much brief[er] magnitude. In the music case, the listener is quite able to draw a distinction between some things she is perceiving and some she is not, and notes from a bar that sounded three seconds ago will not typically be misapprehended by the subject as being currently perceived, even though their presence is felt in another, contextual or conceptual sense. (Grush 2006, p. 447)

When I am listening to a piece of music, I can be aware of the temporal context in which the currently sounding notes occur, but, as Grush points out, I do not have the impression that those parts are occurring at the same time as the notes that are currently sounding. It may be debatable whether the term "conceptual" is apt for such representations, but it should at least be plausible that there is a phenomenal difference between perceiving the notes which are sounding *now* and being aware of the notes which sounded a few seconds ago (or that are about to sound). So, for the purposes of this paper, let us stick to Grush's label, and call all non-perceptual conscious representations conceptual. What matters here is that Grush seems to draw a sharp distinction between two types of conscious representation (perceptual versus non-perceptual), and we can use the labels *perceptual* and *conceptual*, respectively, for these types.

I shall argue that a more useful distinction can be drawn by focusing on the timescale at which a representation operates. By this, I mean the temporal extension of the process or event that is represented by a representation. As we shall see in section 3, some PP models posit estimates which track features at different timescales, i.e., features which change more or less quickly (or, conversely, remain invariant for shorter or longer times). In the passage quoted above, Grush already hints at this, when he writes that a representation can be "conceptual in the sense that it is a matter of interpreting present experience in terms of concepts of processes that span potentially large intervals" (Grush 2006, p. 447). A suggestion inspired by work on PP is that events which are currently happening are *always* represented in terms of processes that span intervals of different lengths. Crucially, some of these intervals are shorter than the interval of Grush's TEM, and some are only slightly longer. So when conscious representations are categorized according to the timescale at which they operate, there is no sharp distinction between two types of representation, because there are not only representations operating at very short timescales (Grush's perceptual representations) and representations operating at very long timescales (Grush's conceptual represenations); but there are also *intermediate* representations, which can only arbitrarily be classified as either perceptual or conceptual.

Assuming a sharp distinction between perceptual and conceptual representations would lead to a puzzle when we try to account for **endurance** (and **continuity**). Recall that, according to **endurance**, we sometimes experience temporally extended processes as enduring. We are aware that they have just been present and we are aware that they are still present. If we assume a sharp distinction between conceptual and perceptual representations, some enduring processes would have to be represented by two conscious representations of different types – a conceptual and a perceptual representation. Since these representations are qualitatively different, and since the represented processes are still experienced as identical (it is the *same* process that has occurred and that is still occurring), this raises what I shall call the **interface question**:[11]

> **Interface question:** $=_{\mathrm{Df}}$ How are perceptual representations of sequences integrated with conceptual representations of sequences?

This question is not addressed by TEM (because it is only concerned with perceptual trajectory estimates). Before showing how PP can inspire an extension of TEM, i.e., HiTEM, which avoids the **interface question**, let me illustrate how the question is related to **endurance** and **continuity**, to emphasize its relevance. Hopefully, this will also make the explanatory potential of HiTEM more salient. To a first approximation, a phenomenological formulation of the interface question is: How can I experience past and present events as parts of a single temporal horizon (cf. Husserl 1991, p. 29)? How can I experience recent events as being seamlessly connected to present events?[12] In particular, how can a sound I am perceiving right now (as part of the present) be experienced as the same sound I heard in the recent past?[13] When I perceive an enduring sound, I don't simply experience part of it as present and part of it as past. Noë puts it thus:

> What you experience, rather, is, to a first approximation, the rising of the current sounds out of the past; you hear the current sounds as surging forth from the past. You hear them as a continuation. This is to say, moving on to a better approximation, you hear them as having a certain trajectory or arc, as unfolding in accordance with a definite law or pattern. It is not the past that is present in the current experience; rather, it is the trajectory or arc that is present now, and of course the arc describes the relation of what is now to what has already happened (and to what may still happen). In this way, what is present, strictly speaking, refers to or is directed toward what has happened and what will happen. (Noë 2006, p. 29)[14]

Such phenomenological descriptions cannot be accounted for by TEM, since TEM is just a model of the perceived present (which Grush assumes to have a temporal extension of about 200 ms, at least usually). By contrast, Noë refers to perceived processes which have significantly longer extensions, on the order of seconds.[15] These processes are still experienced as seamlessly connected, as enduring. This is why the **interface question** arises in this context, and why it is beyond the scope of TEM.

Let us now consider some core assumptions underlying PP models. In particular, we shall focus on models of temporal sequence generation and recognition.

3 Hierarchical Models of Sequence Recognition

Stefan Kiebel and colleagues have recently developed computational models of phenomena involving the representation of sequences, including recognition of bird songs (cf. Kiebel et al. 2008a; Yildiz and

[11] Another question is what one could call *the flow question*: What accounts for the experienced temporal flow of events, and for differences in the speed of the flow? A PP-inspired answer to the flow question has been proposed by Hohwy and colleagues (Hohwy et al. 2016).

[12] Again, an example of an event not experienced as seamlessly connected to past events is a sudden, surprising flash.

[13] Note that "experiencing as the same as" is different from "experiencing as being seamlessly connected to". The first description refers to what I am calling **endurance** here, the second to the feature of **continuity**. When an event is experienced as enduring, distinct temporal parts are experienced as belonging to a single event (such as notes experienced as part of a melody). A continuous sequence is experienced when, in addition, no temporal gaps or boundaries between the individual parts are experienced (e.g., when a melody is played *legato*, as opposed to *staccato*). Thanks to Jakob Hohwy for pressing me to clarify this.

[14] This idea, that experienced events have something like a continuous tail which extends into the past, can also be found in Husserl's work: "During the time that a motion is being perceived, a grasping-as-now takes place moment by moment; and in this grasping, the actually present phase of the motion itself becomes constituted. But this now-apprehension is, as it were, the head attached to the comet's tail of retentions relating to the earlier now-points of the motion." (Husserl 1991, p. 32).

[15] The claim that processes on the order of a few seconds are also consciously perceived as integrated wholes is empirically supported by a variety of findings, including speech segmentation, short-term memory tasks, and sensorimotor tasks (for a review, see Pöppel 1997, Pöppel 2009).

Kiebel 2011) and of artificial speech (cf. Kiebel et al. 2009). In these models, sequences (trajectories) are not modeled as successions of events, but as states of "a collection of hierarchical, dynamical systems, where slower environmental changes provide the context for faster changes" (Kiebel et al. 2008a, p. 2). An interesting aspect of such models is that they are compatible with TEM but are more specific — specifying that trajectories are represented hierarchically. Furthermore, they are still neurally plausible, because the cortical hierarchy seems to match the temporal hierarchy entailed by these models (for a review of neuroscientific evidence, see Kiebel et al. 2008b).

These models presuppose the first six features and the ninth of predictive processing (as defined in Wiese and Metzinger 2017). Of these, the **ideomotor principle** and **hierarchical processing** are particularly important. Both are combined with **prediction error minimization**. Let me explain each in turn.

3.1 The Ideomotor Principle

An assumption underlying the models of Kiebel and colleagues is that the perception (recognition) of a sequence is enabled by a model of its generation. The idea applies not only to sequences the subject herself can generate (like the movement of an arm), but also to others (like a falling snowflake perceived by a subject). For sequences that can actually be generated by the subject, like bodily movements, this entails that at least some of the representations which are active when the subject is perceiving the sequence are also active when the subject herself is performing such movements. Following William James (James 1890), we can call this the *ideomotor principle* (see Wiese and Metzinger 2017 and Limanowski 2017, for more details, and Wiese 2016a for a discussion in the context of active inference). Regarding the neural underpinnings of perception, this suggests that areas not ordinally regarded as sensory may influence the contents of conscious perception (more on this in section 7).

3.2 Hierarchical Processing

A further assumption is that a given sequence can be modeled as a hierarchy of dynamical systems, where the output of a dynamical system at a given level functions as a control parameter[16] for the system at the level below. This principle is illustrated in figure 3. The output at the lowest level corresponds to the sensory consequences of the sequence (those signals are received by the perceiving subject); all other states are hidden and have to be inferred.

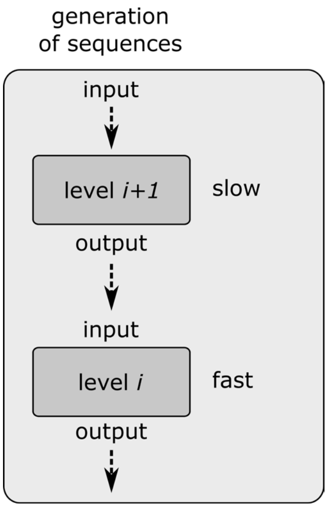

Figure 3: Hierarchical processing: sequences are construed as the states of a hierarchy of coupled dynamical systems. The states of the systems change at different speeds, i.e., they operate on different timescales.

[16] A control parameter is a parameter which can change the phase space of a dynamical system continuously or discontinuously. For instance, it can determine whether the system has a fixed-point or a chaotic attractor, and continuous changes in control parameters can lead to discontinuous changes in phase space (these are called *bifurcations*, cf. Arrowsmith and Place 1998[1992], p. 224).

More formally, the essential aspects can be captured as follows (here, the hierarchy has only two levels; the equations are simplified versions of the ones found in Kiebel et al. 2008a, p. 3):

$$\dot{x}^{(2)} = f^{(2)}(x^{(2)}, \text{“input from level above”}, \text{“slow”}) + \text{noise}^{17} \qquad (1)$$
$$\dot{x}^{(1)} = f^{(1)}(x^{(1)}, \text{“input from level above”}, \text{“fast”}) + \text{noise} \qquad (2)$$

The variables $x^{(1)}$ and $x^{(2)}$ describe the states of two dynamical systems, which unfold according to the differential equations (1) and (2), respectively. These equations each have a parameter governing how quickly the respective system changes (“slow” versus “fast”). Furthermore, how $x^{(1)}$ and $x^{(2)}$ evolve depends on input from the level above (here, the input to the second level is a constant). In the example in (Kiebel et al. 2008a), both dynamical systems are *Lorenz systems*.[18] Lorenz systems can have different types of attractors, depending on their *Rayleigh number*:

> We coupled the fast to the slow system by making the output of the slow system [...] the Rayleigh number of the fast. The Rayleigh number is effectively a control parameter that determines whether the autonomous dynamics supported by the attractor are fixed point, quasi-periodic or chaotic (the famous butterfly shaped attractor). (Kiebel et al. 2008a, p. 3)

The Rayleigh numbers are denoted by “input from the level above” in equations 1 and 2. Coupling two dynamical systems in this way already enables very complex dynamics. For instance, the authors use these equations to simulate birdsongs. Crucially, they simulate not only the generation of a birdsong but also the recognition of the song, exploiting the first principle mentioned above, the ideomotor principle. The principle entails that a recognizing system uses a model of how the song has been generated. Ideally, this model contains the same differential equations that describe how the song has actually been generated and can thus be used as a representation of the song (see Wiese 2016b, sections 3 and 4, for a general description of how such models can be used as representations). Recognition is also based on a third computational principle: prediction error minimization.

3.3 Prediction Error Minimization

Prediction error minimization is here used as a generic term for computational methods in which prediction error terms are minimized. One such method is predictive coding, originally a strategy to compress data (cf. Shi and Sun 1999). The idea is that if we want to transmit data d_1 and d_2 from A (the sender) to B (the receiver), we can reduce the amount of data if we exploit informational relations between d_1 and d_2. For instance, if d_2 is highly predictive of d_1, we can just transmit d_2, and let the receiver infer d_1 (based on d_2). But what does it mean that d_2 is predictive of d_1? A general answer is that d_1 is a mathematical function of d_2. So if we know d_2 and the functional relation $d_1 = f(d_2)$, we can compute d_1. Hence, the amount of data needed to transmit d_1 and d_2 from A to B can be reduced.

In a slightly more realistic setting, there would be more than two pieces of data (e.g., the pixels of an image), so there would be, say, data d_1 and d_2 for which the function relating d_1 and d_2 would not yield a completely accurate estimate of d_1 when applied to d_2. For instance, instead of transmitting the values of all pixels of an image, the sender could transmit only a subset of the pixels, as well as a prediction error which tells the receiver how to correct any errors. Clark (Clark 2013, p. 182) attests:

> In most images, the value of one pixel regularly predicts the value of its nearest neighbors, with differences marking important features such as the boundaries between objects. That means that the code for a rich image can be compressed (for a properly informed receiver) by encoding only the “unexpected” variation: the cases where the actual value departs from the predicted one. What

[17] The “noise” terms capture any unpredictable influences on $x^{(1)}$ and $x^{(2)}$, i.e., they reflect the uncertainty about the respective estimated variables.

[18] A Lorenz system is a set of ordinary differential equations which can have the famous butterfly-shaped attractor.

needs to be transmitted is therefore just the difference (a.k.a. the "prediction error") between the actual current signal and the predicted one.

In other settings, the mapping between different variables will be non-deterministic. This means that there is some uncertainty about our computation of d_1; we can only compute an estimate that is more or less reliable. More formally, we can describe this as follows (where, again, the "noise" terms capture all unpredictable influences):

$$d_1 = f(d_2) + \text{noise} \qquad (3)$$

Given, d_2 and f, we can thus compute an estimate $\hat{d}_1 := f(d_2)$. Depending on the level of noise (or, in general, uncertainty), there will again be a prediction error (because d_1 is not equal to $f(d_2)$). If the sender knows the exact values of d_2 and d_1, the sender can again transmit the prediction error, which will allow the receiver to compute the exact value of d_1.[19]

When it comes to the problem of recognition (perception), things are even worse, because the recognizing system only receives sensory signals. To simplify, say the sensory signals are given by the value of $x^{(1)}$ (this corresponds to d_1). There are two important differences from the situation above. The second value, coded by $x^{(2)}$, cannot be computed from $x^{(1)}$ (at least not given equations (1) and (2)). Furthermore, the recognizing system does not receive a prediction error, but only sensory signals. The solution to this problem is to give the idea a twist. The recognizer does not simply compute an estimate of the value of $x^{(1)}$, but first estimates $x^{(2)}$; this estimate is then used to compute a prediction of $x^{(1)}$ (using equation (2)), and this prediction is compared to the actual signal coded by $x^{(1)}$. A prediction error can then be used to update the estimate of $x^{(2)}$.

This third feature thus exploits the other two features mentioned above, i.e., the ideomotor principle and hierarchical processing. The ideomotor principle entails that recognition of a sequence is based on a model of how the sequence is generated, which enables a prediction of sensory signals. In the simple example given here, we only have two layers, but the same principle can be applied to systems with a more complex hierarchy. Figure 4 illustrates the basic idea (with just two layers).

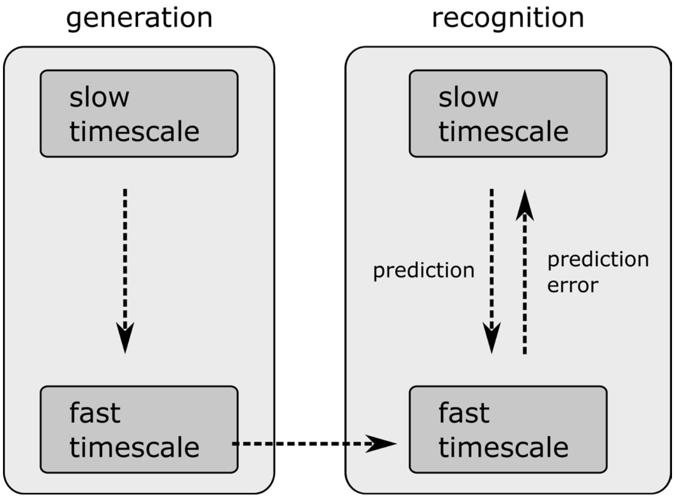

Figure 4: Recognition of a sequence is based on a model of its generation and implemented using prediction error minimization. Processes at higher levels of the hierarchy operate at slower timescales than processes at lower levels. The arrow at the bottom represents sensory signals received by the recognizing system. Processes in the box on the left-hand side are hierarchically coupled dynamical systems (cf. equations (1) and (2)). Processes in the box on the right-hand side model

[19] Here, d_1 is a random variable, because it is a non-deterministic function of d_2. In the example given, it is assumed that the sender knows the exact value of d_2 (which may be a deterministic variable) and has access to a sample, which is modelled as a particular outcome of d_1. This is why, at least in this toy example, the sender can compute the prediction error, although it requires knowledge about the "noise" term, which is by definition unpredictable.

these dynamical systems and therefore enable hierarchical prediction error minimization, which ideally helps keep the model accurate.

4 The Hierarchical Trajectory Estimation Model (HiTEM)

4.1 From TEM to HiTEM

Having described the main aspects of a predictive processing model of sequence perception, we can generalize the model and combine it with Grush's TEM. Recall that the essential part of TEM is a trajectory estimate (which combines smoothing, filtering, and prediction) over the temporal interval $[t - j, t + k]$:

$$T := (\tilde{p}(t - j), \tilde{p}(t - j + 1), \ldots, \hat{p}(t), \bar{p}(t + 1), \ldots, \bar{p}(t + k)). \text{ (cf. Grush 2005, p. S211)} \qquad (4)$$

The challenge now is how to combine TEM with hierarchical models. Two general options are the following:

Localized $=_{\mathrm{Df}}$ The trajectory estimate T coding the perceptual contents of the specious present corresponds to the state of a dynamical system represented at a specific (single) level of the hierarchy.

Distributed $=_{\mathrm{Df}}$ The trajectory estimate T is distributed across at least two levels of the hierarchy.

The simplest version of **localized** would be a two-layer hierarchy in which sensory signals are found at the bottom layer, and the trajectory estimate is located at the second layer. A slightly more complex version would involve more than two layers, but the trajectory estimate coding the perceptual contents of the specious present would still be found at a single level. Note that the neural activity coding the value of T could still be parallel distributed processing, but not over different levels of the processing hierarchy. What **localized** entails is that, given a hierarchical model like the one described in (Kiebel et al. 2008a), which specifies a hierarchy of dynamical systems, there is exactly one level of the hierarchy such that T corresponds to the state of the dynamical system at that level.

As an illustration, consider the following statement by Andy Clark (without implying that Clark would endorse **localized**): "Just as the higher levels in a shape-recognition network respond preferentially to invariant shape properties (such as squareness or circularity), so we should expect to find higher-level networks that model driving sensory inputs (as filtered via all the intervening levels of prediction) in terms of tomatoes, cats, and so forth." (Clark 2012, p. 762). One (though not the only) way to interpret this is that most levels of the PP hierarchy process information unconsciously but at one level *it all comes together* (as in the Cartesian Theater, cf. Dennett and Kinsbourne 1992, p. 183), and this is where information is processed consciously. Again, I would not interpret Clark in this way, but the quotation is at least suggestive, and it is not obviously incoherent to claim that the contents of consciousness are coded at a single level of the hierarchy. This means that localized cannot be dismissed without further argument.

Distributed, by contrast, entails that the description of the trajectory estimate in equation (4) may not map neatly to the estimates over which computations are carried out in the predictive processing hierarchy. So if the states of hierarchically nested dynamical systems can be described by variables x_1, x_2, x_3, …, it is not the case that T corresponds to the value of exactly one x_i. Instead, T corresponds to the states of at least two dynamical systems in the hierarchy. This means that a more detailed description of T could look like this:

$$\begin{pmatrix} T^{(2)} \\ T^{(1)} \end{pmatrix} := \begin{pmatrix} \tilde{p}^{(2)}(t-j) & \tilde{p}^{(2)}(t-j+1) & \ldots & \hat{p}^{(2)}(t) & \bar{p}^{(2)}(t+1) \ldots \bar{p}^{(2)}(t+k) \\ \tilde{p}^{(1)}(t-j) & \tilde{p}^{(1)}(t-j+1) & \ldots & \hat{p}^{(1)}(t) & \bar{p}^{(1)}(t+1) \ldots \bar{p}^{(1)}(t+k) \end{pmatrix} \qquad (5)$$

Note that this hierarchical trajectory estimate is just a "doubled" version of Grush's trajectory estimate and hence does not differ significantly from it. In particular, it does not yet capture the essential part of the hierarchical architecture – that the timescales on which the different levels operate are different. To make this formally explicit, let me adopt a notational convention proposed by Grush:

> I will let $\hat{p}$ stand for a perceptual representation (p, without a hat, will stand for the domain that is being represented), and will indicate the time that the representation represents and the time that the representation is produced by two subscripts separated by a slash, so that $\hat{p}_{a/d}$ is notation for a perceptual representation produced at time d that represents what is (/was/will be) happening at time a. This notation can be generalized to intervals: $\hat{p}_{[a,c]/[d,f]}$ will stand for a perceptual representation of what happened over the interval *[a,c]* that is produced over the interval $[d, f]$. (Grush 2008, p. 151)

For the discussion at hand, the represented time is more important than the time of representing (note that in TEM, the different elements of the trajectory estimate are produced at the same time). For this reason, I will drop the reference to the latter. As in equation (5), any reference to times will be to the represented time. So $\hat{p}_{[a,c]}$ is an estimate of what is happening over the interval $[a, c]$. If we let $t_{-(j+1)} < t_{-j} < t_{1-j} < t_{2-j} < \cdots < t_{-1} < t_0 < t_1 < t_2 < \cdots < t_{k-1} < t_k < t_{k+1}$, we can define a more interesting distributed version as follows (T_d for "distributed trajectory estimate"):

$$T_d := \begin{pmatrix} T_d^{(2)} \\ T_d^{(1)} \end{pmatrix} := \begin{pmatrix} \tilde{p}^{(2)}_{[t_{-(j+1)},t_{2-j}]} & \tilde{p}^{(2)}_{[t_{2-j},t_{5-j}]} & \cdots & \hat{p}^{(2)}_{[t_{-1},t_2]} & \bar{p}^{(2)}_{[t_2,t_5]} & \cdots & \bar{p}^{(2)}_{[t_{k-2},t_{k+1}]} \\ \tilde{p}^{(1)}_{[t_{-j},t_{1-j}]} & \tilde{p}^{(1)}_{[t_{1-j},t_{2-j}]} & \cdots & \hat{p}^{(1)}_{[t_0,t_1]} & \bar{p}^{(1)}_{[t_1,t_2]} & \cdots & \bar{p}^{(1)}_{[t_{k-1},t_k]} \end{pmatrix} \qquad (6)$$

The essential difference between this estimate and the estimate in equation (5) is that the represented times are different on the two levels: In equation (6), the intervals on level two are longer than the intervals on level one. So the events represented at the second level have a longer duration than the events on the first level. An informal illustration of this idea can be found in figure 5:

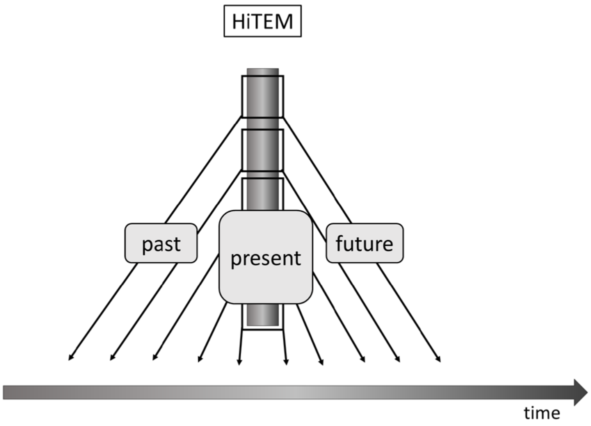

Figure 5: A hierarchical extension of TEM (HiTEM). The core feature of HiTEM is that it posits a hierarchy of temporal wholes. A phenomenological prediction is that slowly-changing features (which remain invariant for more than 200 ms) can also contribute to the perceptual contents of the specious present. Most of these representations do not represent

individual events in time, but represent features which remain invariant over shorter or longer timescales. Some of these representations only refer to the present, but there are at least some which represent features we experience as past, present, *and* future – or, rather, we experience some of them as rising from the recent past, and as continuing into the near future, because they remain invariant over an interval which is slightly longer than the interval of what we experience as happening *now*.

4.2 HiTEM and Event Segmentation Theory

Here, we encounter a similarity to event segmentation theory (EST), which has been developed by Jeffrey Zacks and colleagues. According to EST, the brain constructs *event models* at different temporal scales. Crucially, a given event model can remain active even in the presence of changing perceptual input:

> For example, an event model for a tooth-brushing might include information about the water cup and toothbrush and their locations, the goal of cleaning teeth, and the person doing the brushing. For event models to be useful, they need to be stable in the face of moment-to-moment fluctuations in sensory input – the cup needs to remain in the model even if it is temporarily occluded. (Tversky and Zacks 2013, pp. 89 f.)

Similarly, since the elements of $T_d^{(2)}$ represent features which remain invariant over longer intervals than those represented by $T_d^{(1)}$, the estimate comprised by $T_d^{(2)}$ will often be stable in the face of changes in the estimates contained in $T_d^{(1)}$. Furthermore, T_d contains a hierarchy of representations (restricted to two levels here for the sake of simplicity). This is in line with EST's assumption that events are segmented hierarchically. The represented temporal boundary of an event often corresponds to the represented achievement of a goal; since many tasks can be divided into subtasks (with sub-goals), a given event is typically represented as consisting of a sequence of smaller events (cf. Tversky and Zacks 2013, p. 87).

At this point, it will be helpful to point out two marked differences between the hierarchy of event models posited by EST and the hierarchy of trajectory estimates posited by HiTEM:

1. Event models operate at significantly longer timescales than the estimates posited by TEM and HiTEM.
2. The notion of an event used by Zacks et al. is more restricted than the notion that is relevant to the contents of the specious present. This has some subtle implications which become salient when it comes to accounting for **endurance** and **continuity**. In particular, events need not be represented as having determinate boundaries, according to HiTEM.

Let us call this second feature **fuzzy boundary**:

> **Fuzzy boundary:** $=_{\mathrm{Df}}$ At least at very short timescales (on the order of 200 ms), some events are not represented as having determinate boundaries, and some processes are not represented as having a determinate beginning or ending.

To clarify the second point, let us consider in more detail what the elements of T_d represent. In which sense are they representations of events? Arguably, the most useful and conservative interpretation is as follows. The elements of T_d do not represent events (in the sense of EST) but property instantiations. At least according to certain conceptions of events, an event is just a property instantiation at a given time and place (or the exemplification of a property by an object at a time, see Kim 1966, p. 231). The crucial point is that a property represented at the second level and a property represented at the first level can correspond to a single event. So the estimate

$$\begin{pmatrix} \hat{p}^{(2)}_{[t_{-1},t_2]} \\ \hat{p}^{(1)}_{[t_0,t_1]} \end{pmatrix}$$

can represent a single event. Let us call such an event representation a dynamic event representation (DER).

Usually, dynamic events are posited to accommodate the intuition that events can involve change (in fact, some authors would claim that something which does not contain any change *cannot* be an event, see Casati and Varzi 2015, § 2.2). A DER represents change in a minimal form: There is at least one property which is not instantiated during the entire interval in which the event unfolds. A stronger form of change would be a succession of property instantiations. One could object that such a succession would correspond to a succession of events (in which case a DER would just be a representation of a succession of events – just as a higher-order event model in EST models an event which is just a succession of lower-order events). But a DER does not necessarily involve such successions. The key theoretical advantage of the concept is that the times during which the different properties are represented as being instantiated can overlap. This opens the interesting possibility that represented events can overlap in time as well.[20]

Consider the following two estimates:

$$\hat{e}_1 := \begin{pmatrix} \hat{p}^{(2)}_{[t_{-1},t_2]} \\ \hat{p}^{(1)}_{[t_0,t_1]} \end{pmatrix}, \hat{e}_2 := \begin{pmatrix} \hat{p}^{(2)}_{[t_{-1},t_2]} \\ \hat{p}^{(1)}_{[t_1,t_2]} \end{pmatrix}.$$

If these estimates are elements of the same trajectory estimate, they can share a part, $\hat{p}^{(2)}_{[t_{-1},t_2]}$. This does not make sense in all cases: If two red balls bounce at the same time at two different locations, there is a sense in which these events overlap (because they share the property of redness), but the two bouncing events are clearly distinct (the property of redness is exemplified by two distinct objects). But there are cases in which it can make sense to represent distinct events as sharing a single property instantiation.

Take the example of music. When a single instrument like a flute plays a sequence of notes *legato* (as opposed to *staccato*), some properties remain invariant (at least during short intervals), for instance the timbre of the notes. Hence, we can describe the melody as a sequence of overlapping property instantiations: Properties like pitch may change more quickly than properties like loudness or timbre. This means we can have an interval $[t_1, t_2]$ during which, say, two different pitches are instantiated (e.g., the first during the first half of the interval, the second during the second half), and a single timbre is instantiated (during the entire interval). Since the timbre is in fact the same timbre during the interval (after all, the notes are produced by the same instrument), it makes sense to use a single representation for this property instantiation. The key difference with the case of the two bouncing balls is that there is a common cause underlying the sensory signals.[21] Furthermore, using a DER not only makes computational sense (because it is more efficient than representing the same property twice); it may also account for the experienced continuity (e.g., in music perception). Let us explore this by first considering how this proposal can deal with the interface question.

[20] A perhaps controversial implication is that some events are not represented as having a determinate beginning and ending. But wouldn't it be important to know the exact time at which an event starts (and ends, respectively)? If we think of *intentional binding* (cf. Haggard et al. 2002), in which the interval between two causally connected events is systematically underestimated, it seems there are cases in which the exact timing of events does not matter. Instead, it seems more important to represent temporally distinct events as parts of a temporal whole (perhaps as having a common cause). So representing events as overlapping (with indeterminate temporal boundaries) may be a way of prioritizing the temporal whole over its parts.

[21] At a deeper level, there can of course still be a common course, for instance, if both balls are thrown by a juggler (thanks to Jakob Hohwy for this example). This would be comparable to a case in which two instruments (e.g., a flute and a drum) were being played by the same person.

4.3 Answering the Interface Question

Recall that we are still considering how to combine TEM with hierarchical predictive processing models, and the crucial challenge is whether we should favor a **localized** or a **distributed** option. In this section, I argue that a hierarchically distributed extension of TEM (involving something like T_d, i.e., an estimate with at least two levels, in which properties at different levels of temporal granularity are represented) provides an answer to the **interface question**. Recall the formulation of the latter:

> **Interface question** $=_{\mathrm{Df}}$ How are perceptual representations of trajectories integrated with conceptual representations of trajectories?

HiTEM answers the question thus:

- There is no compelling reason to assume a sharp boundary between perceptual (concrete, perspective-dependent) and conceptual (abstract, perspective-invariant) representations.
- In PP, the continuum between perceptual and conceptual representations is typically assumed to be distributed over the hierarchy.

This suggests that clearly perceptual and clearly conceptual representations are found at different levels of the hierarchy. Furthermore, if there are neural representations that are neither purely perceptual nor purely conceptual, these could function as mediators between (conceptual) representations of remembered events and (perceptual) representations of currently occurring events.[22] A theoretical advantage of the distributed option is that the neural vehicles of perceptual and conceptual trajectory estimates can overlap spatially (i.e., by sharing parts), so mediating estimates would not have to be posited as additional representations, but would partly determine both perceptual and conceptual experiences of temporal wholes (I explore this idea in a much wider context in Wiese 2017).

Let me make more explicit what a mediating representation would be in this context. According to TEM, all elements of the trajectory estimates represent events that are occurring within the interval of the specious present (which has, according to Grush, a duration of about 200 ms). All these events are experienced as currently happening; they have features which are represented as being instantiated during this interval. By contrast, a mediating representation in HiTEM represents features as being instantiated during an interval which is longer than that identified by Grush: It represents features as having been present in the recent past, as being present now, and as continuing into the near future. Crucially, such features can be bound to features which are represented as changing more quickly. The result is a dynamic event representation (DER), which corresponds to the experience of an event as present, but also as having been present in the recent past. On the other hand, such mediating features can also be bound to features which are represented as changing more slowly (or features pertaining to events which are represented as past). The result is again a DER, but more akin to what Grush describes as a conceptual (as opposed to a perceptual) representation. Since there is no sharp boundary between purely perceptual and purely conceptual representations, instances of these two types of representation can be integrated by mediating representations.

5 How Can We Account for Continuity and Endurance?

Let us next consider to what extent mediating representations (operating at intermediate timescales, between clearly perceptual and clearly conceptual repsentations) can help account for **continuity** and **endurance**. Let me repeat the definitions of these two features of temporal consciousness:

[22] Just as there can be representations that mediate between purely perceptual (descriptive) representations and (prescriptive) goal representations (cf. Wiese 2014).

Continuity $=_{\mathrm{Df}}$ At least sometimes, we experience smooth successions of events (or smooth changes). An example is a series of notes played *legato* by a single instrument (in contrast with a series played *staccato*). Such sequences are experienced as temporal continua (which, strictly speaking, would involve an infinite number of events).

Endurance $=_{\mathrm{Df}}$ At least sometimes, we experience temporally extended events as enduring. An example is an opera singer holding a single note for an extended period (this example is taken from Kelly 2005, p. 208). By contrast, when one is surprised by a sudden bright flash, this punctual event is not experienced as part of an enduring event.

The answer given to the **interface question** already suggests how to account for **endurance**: when an event is represented by a conscious DER, some of its properties are represented as remaining the same while others are changing, which corresponds to the experience of an event as present, but also as having been present in the recent past; in other words, a conscious DER represents an event as *enduring*. Not all its features are however experienced as having been present in the past, and this is why it can be so difficult to describe our experience of enduring events phenomenologically. An example from Kelly provides an excellent illustration:

> There you are at the opera house. The soprano has just hit her high note – a glass-shattering high C that fills the hall – and she holds it. She holds it. She holds it. She holds it. She holds it. She holds the note for such a long time that after a while a funny thing happens: You no longer seem only to hear it, the note as it is currently sounding, that glass-shattering high C that is loud and high and pure. In addition, you also seem to hear [...] something about its temporal extent. (Kelly 2005, p. 208)

This “something” is, according to HiTEM, a slightly more abstract feature of the note, which is represented as being invariant for more than 200 ms (perhaps even a few seconds). One’s conscious experience will certainly have other, additional aspects which characterize what it is like to hear such an enduring high C (for instance, a feeling of tension or stress). But at least some aspects correspond to perceptual (or quasi-perceptual) features of the note which change slowly.[23]

Such features are, according to HiTEM, always experienced, but they are not always very salient. For instance, to most people hearing a melody, it seems obvious that what they are perceiving is not just one note after the other; but to describe what exactly it is that makes the difference might seem more difficult. HiTEM suggests that the additional experienced features are slightly more abstract (more gist-like) than features such as pitch or loudness (and hence more difficult to describe). Crucially, the additional features contribute to the perception of each individual note; since these features are shared by all of them, temporally separated notes can be experienced as a temporal whole, as flowing into each other. This accounts for **continuity**.

Let us compare the proposal again with EST. A hierarchy of event representations in EST would not necessarily involve a representation of a continuous flow (the event models in EST seem to be more abstract, purely conceptual representations of events). The temporal boundaries of events are assumed to be determinate in EST, so even if a tooth-brushing event is represented as a succession of shorter events (brushing the first tooth, brushing the second tooth, …), this would still only be a succession of events: First is A, then B, and both jointly constitute an event C.

[23] As Kiebel et al. 2008a point out (with respect to their model of birdsong), such features can also provide information about the creature which generated the temporal sequence: “Birdsong contains information that other birds use for decoding information about the singing (usually male) bird. It is unclear which features birds use to extract this information; however, whatever these features are, they are embedded in the song, at different time-scales. For example, at a long time-scale, another bird might simply measure how long a bird has been singing, which might belie the bird’s fitness. At short time-scales, the amplitude and frequency spectrum of the song might reflect the bird’s strength and size.” (Kiebel et al. 2008a, p. 2). Thanks to Jakob Hohwy for suggesting this citation.

By contrast, a DER would represent a succession of events that do not have determinate temporal boundaries as follows.

$$\begin{pmatrix} & \hat{p}^{(2)}_{[t_{-2},t_2]} & \\ \hat{p}^{(1)}_{[t_{-1},t_0]} & & \hat{p}^{(1)}_{[t_0,t_1]} \end{pmatrix}$$

Here, the entire matrix represents, say, a succession of notes, but $\hat{p}^{(1)}_{[t_{-1},t_0]}$ & $\hat{p}^{(2)}_{[t_{-2},t_2]}$ jointly constitute a single representation of a note (the first note in the succession), and $\hat{p}^{(1)}_{[t_0,t_1]}$ & $\hat{p}^{(2)}_{[t_{-2},t_2]}$ likewise (the second note in the succession). On the one hand, the first note is represented as occurring before the second, because $\hat{p}^{(1)}_{[t_{-1},t_0]}$ and $\hat{p}^{(1)}_{[t_0,t_1]}$ represent properties (say, pitch) as being instantiated during distinct intervals ($[t_{-1},t_0]$ and $[t_0,t_1]$, respectively). It is not true, however, that the first note is represented as occurring completely before the second, because the other property associated with the notes (say, timbre), which is represented by $\hat{p}^{(2)}_{[t_{-2},t_2]}$, is represented as being instantiated during a longer interval. Hence, the notes are represented as being distinct, but overlapping (where the overlapping part is not just a further note). This is why the entire representation is not just a representation of two events, or of a succession of events, but of a continuous succession, where one event flows smoothly[24] into the next.

6 What Are the Contents of Mediating Representations?

Recall that Grush draws a rather sharp boundary between perceptual and conceptual representations. By constrast, assuming that the contents of the specious present are coded by a hierarchy of representations, it is already suggestive to believe that there are mediating representations. If they contribute to the contents of consciousness, however, it will be relevant to determine their contents. I alluded to the example of auditory perception and suggested that examples of mediating representations could include representations of timbre or rhythm. To explore more options, and to make first steps towards finding neural evidence for such representations, let us consider results from empirical research on auditory processing in the brain.

As Lima et al. (Lima et al. 2016) point out in a recent review, neural processing of auditory information is distributed over anatomically and functionally different streams, which can broadly be divided into an anteroventral "what" pathway and a posterodorsal "how/where" pathway (cf. Lima et al. 2016, p. 530). Interestingly, whereas the hierarchy in the "what" pathway seems to provide more and more abstract re-representations of semantic information, the "how/where" pathway seems to provide sensorimotor representations, involving also supplementary motor areas (SMA) and pre-supplementary motor areas (pre-SMA). Furthermore, SMA and pre-SMA not only play a role in speech perception, but also in music perception and auditory imagery (cf. Lima et al. 2016, p. 532). The authors hypothesize that these "regions mediate spontaneous motor responses to sound, and support a more controlled generation of sensory predictions based on previous sensorimotor experience, predictions that can be flexibly exploited to enable imagery and optimize a variety of perceptual processes." (p. 539). This hypothesis suggests that SMA and pre-SMA contain the kind of sensorimotor representations which are posited by the ideomotor principle and which are required if indeed the perception of the (auditory) sequence is based on a model of its generation.

So, given that activity in these regions correlates not only with motor or cognitive processes (for evidence, see the references cited in Lima et al. 2016, p. 534), we can speculate that these regions harbor mediating representations, which are not purely perceptual (becaue they are also relevant for motor tasks) but still correlate with consciously experienced perceptual contents (which may be gist-like). The evidence presented by Lima et al. seems to be consistent with this hypothesis, but more work will

[24] Note that this also involves computationally smoothed estimates, but this computational technique does not account for the smoothness of the flow (because trajectory estimates in TEM involve smoothed estimates as well, without thereby accounting for the experienced smoothness).

have to be done to find stronger support for it (for instance, it is not clear whether activity in SMA and pre-SMA correlates with *conscious* perception; cf. Repp 2001). Bearing this in mind, let us briefly consider with which perceptual contents activity in these regions has been associated. This will at least illustrate what the contents of representations at higher levels in a hierarchical trajectory estimate could be.

According to Lima et al.'s review, SMA and pre-SMA become specifically activated by non-verbal vocal emotional cues (cf. Lima et al. 2016, Box 2 on p. 532, and the evidence cited therein). It is plausible that such contents are part of what determines the perceptual character of conscious music perception, and at the same time part of what makes such conscious experiences difficult to describe. In general, the way in which these areas contribute to auditory perception is complex, as Lima et al. point out:

> There is no consensus position on the roles of SMA and pre-SMA responses in auditory processing and imagery. When such responses are discussed, they have been linked to a variety of processes. Timing functions have been suggested for perceptual tasks requiring evaluations of temporal aspects of auditory stimuli […], or for stimuli varying in the sequential predictability and rhythmic regularity that they afford […]. SMA and pre-SMA, together with the cerebellum and the basal ganglia, have in fact been considered to form the substrates for a 'temporal processing' network […]. (Lima et al. 2016, p. 535)

Rhythmic regularities are among the features which are especially relevant in this context, because they change more slowly than such features as loudness or pitch. But the general picture is even more complex. In one study cited by Lima et al. (Raij and Riekki 2012), activity in pre-SMA was stronger for voluntarily generated imagery than for auditory hallucinations, suggesting a role in coding voluntary imagery (Lima et al. 2016, p. 532). Futhermore, activity in SMA seems to be correlated with perceived vividness of auditory imagery (p. 534).

Such findings are consistent with the claim that the perceptual contents of the specious present involve more than just successions of events. Instead, individual events (such as the sounding of a single note) are experienced in the context of larger temporal wholes, which may be marked by an affective character, a rhythmic regularity, volitional aspects (like an "urge to move", cf. Lima et al. 2016, p. 537; see also Grahn and McAuley 2009), or the experienced vividness of imagery.

We can distinguish between two types of hierarchy of temporal wholes here, *nested* and *non-nested*. The elements of a nested hierarchy stand to each other in part-whole relations (just as brushing the first tooth may be part of a larger tooth-brushing event, which has a longer temporal duration). The elements of a non-nested temporal hierarchy are only hierarchically ordered by the relation "has a longer duration than". For instance, the emotional response accompanying hearing a short melody could have a longer temporal extension than the melody itself, but it is not experienced as part of the melody (and neither is the melody experienced as part of the emotional response). A functional difference between these two types of hierarchy might be that one could selectively attend to the elements of a non-nested temporal hierarchy (only to the melody, or only to the emotional response), yet not always be able to do so for a nested temporal hierarchy (e.g., it may be impossible to attend only to the rhythm of a melody, without thereby also attending to the sounds of which the melody is composed).

7 To What Extent Are Mediating Representations Predictive of Perceptual Contents?

So far, I have only suggested that regularities tracked at different temporal (and spatial) grains may determine the contents of our conscious perception of temporal processes and successions of events. This idea sits well with hierarchical PP models, and I gave examples in the previous section, but I have not yet addressed the question as to whether features tracked at different levels of the hierarchy can plausibly be assumed to be predictive of each other. Despite the differences between TEM (and HiTEM) and EST to which I alluded, we here encounter an interesting parallel: EST entails that predictions are derived from event models, and when there is an increase in prediction error an event boundary is inferred, and the event model is updated (cf. Reynolds et al. 2007, p. 616). The fact that a single event model can be predictive of a stream of perceptual input is exploited here, and this idea can of course be generalized to hierarchical models (cf. Butz 2016).

Applying this to conscious auditory perception of melodies, can we identify predictive relationships between the contents mentioned in the previous section? More specifically, to what extent are mediating representations (neither purely perceptual nor purely conceptual) predictive of perceptual representations? First of all, representations of rhythm or meter are predictive of the *timing* of individual notes. Furthermore, emotional responses can be predictive of the *key* in which a melody is played, and the key can be predictive of intervals in a melody. An urge to move may be an even higher-level representation, which is not predictive of a particular rhythm but perhaps of a certain class, e.g., rhythms familiar to the subject or with a clearly perceivable meter or beat. So it is at least plausible to assume that the contents experienced in temporal perception (like music perception), are not only ordered (or nested) in a temporal hierarchy but are also predictive of each other. Therefore, it should be possible to model them in the way suggested by the hierarchical predictive processing models mentioned above (section 3).[25]

8 Conclusion

This chapter has focused on two features of temporal consciousness, which I called **endurance** and **continuity**:

Continuity $=_{\text{Df}}$ At least sometimes, we experience smooth successions of events (or smooth changes).

Endurance $=_{\text{Df}}$ At least sometimes, we experience temporally extended events as enduring.

Rick Grush's trajectory estimation model (TEM), a compelling model of conscious temporal perception, cannot account for these features, but I have tried to show that the model can be extended by drawing on features of hierarchical predictive processing models. Such models posit representations operating at various timescales. As a result, sequences are not just represented as successions of events but as hierarchical wholes. This accounts for **endurance** if the proposal in this chapter is on the right track. A key feature, which I call **fuzzy boundary**, is that events need not be represented as having determinate temporal boundaries. This may account for **continuity**.

Since this extension of Grush's TEM, which I call HiTEM (hierarchical trajectory estimation model), draws on features of existing computational PP models, it is at least theoretically supported. Empirically, more work needs to be done to find direct support for the model, but current evidence on neural underpinnings of auditory perception is at least consistent with HiTEM. In particular, empirical results may also enrich phenomenological descriptions of temporal consciousness: They will allow us to say in more detail what exactly we experience when we consciously perceive temporally extended processes or successions of events.

[25] With respect to auditory perception, an excellent overview and a model can be found in (Winkler and Schröger 2015).

References

Arrowsmith, D. & Place, C. M. (1998[1992]). *Dynamical systems: Differential equations, maps, and chaotic behaviour.* London: Chapman & Hall / CRC Press.

Butz, M. V. (2016). Toward a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology, 7.* https://dx.doi.org/10.3389/fpsyg.2016.00925.

Casati, R. & Varzi, A. (2015). Events. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/win2015/entries/events/.

Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind, 121* (482), 753–771. https://dx.doi.org/10.1093/mind/fzs106.

—— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36* (3), 181–204. https://dx.doi.org/10.1017/S0140525X12000477.

Dainton, B. (2014). Temporal consciousness. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/spr2014/entries/consciousness-temporal/.

Dennett, D. C. & Kinsbourne, M. (1992). Time and the observer. *Behavioral and Brain Sciences, 15* (2), 183–201.

Grahn, J. A. & McAuley, J. D. (2009). Neural bases of individual differences in beat perception. *Neuroimage, 47* (4), 1894–1903. https://dx.doi.org/10.1016/j.neuroimage.2009.04.039.

Grush, R. (2005). Internal models and the construction of time: Generalizing from *state* estimation to *trajectory* estimation to address temporal features of perception, including temporal illusions. *Journal of Neural Engineering, 2* (3), S209–S218. https://dx.doi.org/10.1088/1741-2560/2/3/S05.

——(2006). How to, and how not to, bridge computational cognitive neuroscience and Husserlian phenomenology of time consciousness. *Synthese, 153* (3), 417–450.

——(2008). Temporal representation and dynamics. *New Ideas in Psychology, 26* (2), 146–157. https://dx.doi.org/10.1016/j.newideapsych.2007.07.017.

——(2016). On the temporal character of temporal experience, its scale non-invariance, and its small scale structure. *Manuscript.* https://dx.doi.org/10.21224/P4WC73.

Haggard, P., Clark, S. & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5* (4), 382–385.

Hohwy, J. (2013). *The predictive mind.* Oxford: Oxford University Press.

Hohwy, J., Paton, B. & Palmer, C. (2016). Distrusting the present. *Phenomenology and the Cognitive Sciences, 15* (3), 315–335. https://dx.doi.org/10.1007/s11097-015-9439-6.

Hubbard, T. L. (2014). Forms of momentum across space: Representational, operational, and attentional. *Psychonomic Bulletin & Review, 21* (6), 1371–1403. https://dx.doi.org/10.3758/s13423-014-0624-3.

Husserl, E. (1991). *On the phenomenology of the consciousness of internal time (1893-1917).* Dordrecht, Boston, London: Kluwer Academic Publishers.

James, W. (1890). *The principles of psychology.* New York: Henry Holt.

Kelly, S. D. (2005). Temporal awareness. In D. W. Smith & A. L. Thomasson (Eds.) *Phenomenology and philosophy of mind* (pp. 222–234). Oxford: Oxford University Press.

Kiebel, S. J., Daunizeau, J., Friston, K. J. & Sporns, O. (2008a). A hierarchy of time-scales and the brain. *PLoS Computational Biology, 4* (11), e1000209. https://dx.doi.org/10.1371/journal.pcbi.1000209.

——(2008b). Supporting information. *PLoS Computational Biology, 4* (11), e1000209. https://dx.doi.org/10.1371/journal.pcbi.1000209.s001.

Kiebel, S. J., von Kriegstein, K., Daunizeau, J. & Friston, K. J. (2009). Recognizing sequences of sequences. *PLoS Comput Biol, 5* (8), e1000464. https://dx.doi.org/10.1371/journal.pcbi.1000464.

Kim, J. (1966). On the psycho-physical identity theory. *American Philosophical Quarterly, 3* (3), 227–235.

Kolers, P. A. & von Grunau, M. (1975). Visual construction of color is digital. *Science, 187* (4178), 757-759. https://dx.doi.org/10.1126/science.1114322.

Lee, G. (2014). Temporal experience and the temporal structure of experience. *Philosopher's Imprint, 14* (3), 1–21. www.philosophersimprint.org/014003/.

Lehmann, D. (2013). Consciousness: Microstates of the brain's electric field as atoms of thought and emotion. In A. Pereira Jr. & D. Lehmann (Eds.) *The unity of mind, brain and world: Current perspectives on a science of consciousness.* (pp. 191–218). Cambridge: Cambridge University Press.

Lima, C. F., Krishnan, S. & Scott, S. K. (2016). Roles of supplementary motor areas in auditory processing and auditory imagery. *Trends in Neurosciences, 39* (8), 527-542. https://dx.doi.org/10.1016/j.tins.2016.06.003.

Limanowski, J. (2017). (Dis-)attending to the body. Action and self-experience in the active inference framework.

In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Metzinger, T. (2004[2003]). *Being no one: The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

——(2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Noë, A. (2006). Experience of the world in time. *Analysis, 66* (289), 26–32. https://dx.doi.org/10.1111/j.1467-8284.2006.00584.x.

Prosser, S. (2016). *Experiencing time.* Oxford: Oxford University Press.

Pöppel, E. (1997). A hierarchical model of temporal perception. *Trends in Cognitive Sciences, 1* (2), 56–61. https://dx.doi.org/10.1016/S1364-6613(97)01008-5.

——(2009). Pre-semantically defined temporal windows for cognitive processing. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364* (1525), 1887–1896. The Royal Society. https://dx.doi.org/10.1098/rstb.2009.0015.

Raij, T. T. & Riekki, T. J. J. (2012). Poor supplementary motor area activation differentiates auditory verbal hallucination from imagining the hallucination. *NeuroImage: Clinical, 1* (1), 75–80. https://dx.doi.org/10.1016/j.nicl.2012.09.007.

Repp, B. H. (2001). Phase correction, phase resetting, and phase shifts after subliminal timing perturbations in sensorimotor synchronization. *Journal of Experimental Psychology: Human Perception and Performance, 27* (3), 600–621.

Reynolds, J. R., Zacks, J. M. & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science, 31* (4), 613–643. https://dx.doi.org/10.1080/15326900701399913.

Shi, Y. Q. & Sun, H. (1999). *Image and video compression for multimedia engineering: Fundamentals, algorithms, and standards.* Boca Raton, FL: CRC Press.

Shimojo, S. (2014). Postdiction: Its implications on visual awareness, hindsight, and sense of agency. *Frontiers in Psychology, 5* (196). https://dx.doi.org/10.3389/fpsyg.2014.00196.

Thornton, I. M. & Hubbard, T. L. (2002). Representational momentum: New findings, new directions. *Visual Cognition, 9* (1-2), 1–7. https://dx.doi.org/10.1080/13506280143000430.

Tversky, B. & Zacks, J. M. (2013). Event perception. In D. Reisberg (Ed.) *Oxford handbook of cognitive psychology* (pp. 83–94). New York: Oxford University Press.

Watzl, S. (2013). Silencing the experience of change. *Philosophical Studies, 165*, 1009–1032. https://dx.doi.org/10.1007/s11098-012-0005-6.

Wiese, W. (2014). Jakob Hohwy: The predictive mind. *Minds and Machines, 24* (2), 233–237. https://dx.doi.org/10.1007/s11023-014-9338-6.

—— (2016a). Action is enabled by systematic misrepresentations. *Erkenntnis*. https://dx.doi.org/10.1007/s10670-016-9867-x. http://rdcu.be/nZs0.

——(2016b). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 1-22. https://dx.doi.org/10.1007/s11097-016-9472-0.

——(in press). *Experienced wholeness. Integrating insights from Gestalt theory, cognitive neuroscience, and predictive processing.* Cambridge, MA: MIT Press.

Wiese, W. & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing.* Frankfurt am Main: MIND Group.

Winkler, I. & Schröger, E. (2015). Auditory perceptual objects as generative models: Setting the stage for communication by sound. *Brain and Language, 148*, 1–22. https://dx.doi.org/10.1016/j.bandl.2015.05.003.

Yildiz, I. B. & Kiebel, S. J. (2011). A hierarchical neuronal model for generation and online recognition of birdsongs. *PLoS Computational Biology, 7* (12), e1002303. https://dx.doi.org/10.1371/journal.pcbi.1002303.

Zacks, J. M. (2008). Event perception. *Scholarpedia, 3* (10), 3837.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L. & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience, 4* (6), 651–655.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S. & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin, 133* (2), 273. https://dx.doi.org/10.1037/0033-2909.133.2.273.