A Unified Neurocomputational Model of Prospective and Retrospective Timing

Joost de Jong[1], Aaron R. Voelker[2], Terrence C. Stewart[3], Elkan G. Akyürek[1], Chris Eliasmith*[4], and Hedderik van Rijn*[1]

[1]Department of Experimental Psychology, University of Groningen
[2]Applied Brain Research Inc.
[3]National Research Council Canada
[4]Centre for Theoretical Neuroscience, University of Waterloo

Time is a central dimension against which perception, action, and cognition play out. From anticipating when future events will happen to recalling how long ago previous events occurred, humans and animals are exquisitely sensitive to temporal structure. Empirical evidence seems to suggest that estimating time prospectively (i.e., in passing) is qualitatively different from estimating time in retrospect (i.e., after the event is over). Indeed, computational models that attempt to explain both prospective and retrospective timing assume a fundamental separation of their underlying processes. We, in contrast, propose a new neurocomputational model of timing, the Unified Temporal Coding (UTC) model that unifies prospective and retrospective timing through common principles. The UTC model assumes that both stimulus and timing information are represented inside the same rolling window of input history. As a consequence, the UTC model explains a wide range of phenomena typically covered by specialized models, such as conformity to and violations of the scalar property, one-shot learning of intervals, neural responses underlying timing, timing behavior under normal and distracting conditions, common capacity limits in timing and working memory, and how timing depends on attention. Strikingly, by assuming that prospective and retrospective timing rely on the same principles and are implemented in the same neural network, a simple attentional gain mechanism can resolve the apparently paradoxical effect of cognitive load on prospective and retrospective timing.

Keywords: Timing and Time Perception, Temporal Cognition, Neurocomputational Model, Prospective Timing, Retrospective Timing, Attention

Time features prominently in most of our everyday activities. From waiting for a pot to boil to suddenly realizing that a pleasant conversation ran on for much longer than expected, time is one of the most fundamental dimensions of our mental lives. But despite decades of formal theorizing about 'the sense of time', a clear consensus about underlying cognitive and neural mechanisms is still lacking. Recent modelling efforts with recurrent neural networks (RNNs) have generated excellent fits to complex neural data, but their underlying representational and computational principles are often difficult to probe. Further, most models of timing have focused exclusively on prospective timing (e.g., waiting for the pot to boil), leaving retrospective timing (e.g., recalling the duration of the conversation) in need of a coherent explanation. The few theoretical approaches and computational models that have attempted to explain retrospective timing suggest that prospective and retrospective timing are related,

* Shared senior authorship
Joost de Jong https://orcid.org/0000-0001-8841-5646
We would like to thank Marc Howard, Patrick Simen and Virginie van Wassenhove for their detailed feedback on the manuscript.
Correspondence should be addressed to Joost de Jong, Experimental Psychology, University of Groningen, Grote Kruisstraat 1/2, the Netherlands. joost.de.jong@rug.nl
All code and simulation data are available on https://github.com/dejongejoost/UTC_model.
Chris Eliasmith holds an interest in the patent holder (Applied Brain Research, Inc.) for the LMU.

but essentially distinct processes. In this paper, we propose a new neurocomputational model of timing, the 'Unified Temporal Coding' (UTC) model, with clear underlying representational and computational principles that propose a unified account of prospective and retrospective timing.

This introduction is structured as follows. First, we will outline some basic empirical phenomena related to prospective and retrospective time estimation. We will focus on the differences between both types of timing, and discuss theoretical perspectives related to these differences. Next, we will discuss several classes of models that have attempted to explain these phenomena, and finally introduce our own model, the Unified Temporal Coding (UTC) model.

Time in Passing and Time in Retrospect

Unlike most modern clocks that keep time in a highly precise and accurate fashion, time estimation in humans and non-human animals is modulated by a variety of external and internal factors. Most notably, subjective estimates of time depend on whether time is estimated as it passes or in retrospect. This insight is far from new, as William James (1890) aptly characterized in *The Principles of Psychology*:

> In general, a time filled with varied and interesting experiences seems short in passing, but long as we look back. On the other hand, a tract of time empty of experiences seems long in passing, but in retrospect short (p.320)

James' distinction between 'time in passing' and 'time in retrospect' has become a primary distinction in the time perception literature, where they are commonly termed 'prospective' and 'retrospective' timing (Hicks et al., 1976).

In prospective timing paradigms, subjects are instructed beforehand to pay attention to the duration of individual stimuli or the duration of the task, and are asked to give a temporal estimate after the interval has ended. In retrospective paradigms, subjects are unaware that temporal features are important for the task beforehand, which is only revealed when they are asked to estimate the interval after it is already over. For instance, researchers have employed list memory paradigms to study differences between prospective and retrospective timing (e.g., Poynter, 1983). Subjects are presented with a list of items that they have to remember. In the prospective condition, subjects are given the additional task to track the duration of the memory encoding phase, whereas the retrospective group is kept oblivious about this additional task. After the items have been presented, both groups of subjects report the duration of the encoding phase. In the prospective condition, subjects can form an estimate of elapsed time as the interval unfolds, while subjects in the retrospective condition have to construct an estimate in hindsight. A common finding is that retrospective estimates are both less accurate (lower) and less precise (more variable) than prospective estimates (Block & Zakay, 1997).

A major theoretical issue in the literature is whether prospective and retrospective timing are different in degree or different in kind. Brown (1985) has argued that the same processes underlie timing performance in prospective and retrospective conditions (i.e., similar or identical "in kind"), but that they differ in the degree of attention paid to temporal or stimulus information, respectively. According to this view, time estimates are constructed from encoding temporal cues, such as salient changes (e.g., Poynter & Homa, 1983) or event structure (e.g., Brown & Boltz, 2002). In a prospective condition, the task instructions ensure that sufficient attention is focused on the timing task to ensure reasonable levels of accuracy. In retrospective conditions all attention will be directed towards the main (or a distracting) task, resulting in less frequent and less consistent encoding of temporal cues, resulting in shorter and more variable time estimates. As dual-task conditions can have similar effects on prospective and retrospective timing (e.g., Brown & Stubbs, 1992), this is seen as support for the view that prospective and retrospective timing only differ in degree of temporal processing.

In contrast, Block et al. (2010) argue that prospective and retrospective timing rely on categorically different kinds of processes. In their view, not dissimilar to the view proposed above, prospective timing is based on an internal clock mechanism that needs attention to function properly. Zakay and Block (1995) propose an attentional gate that controls how fast 'ticks' pass to an accumulator. When attention is directed to time, the gate opens, allowing more ticks to pass and leading to longer time estimates. When attention is diverted away from time, the attentional gate narrows, allowing fewer ticks to pass, explaining why prospective time estimates are lower in attention-demanding dual-task conditions. Conversely, Block et al. (2010) propose that retrospective estimates are based on the reconstruction of past events through memory retrieval. Time judgements are constructed by estimating how many contextual changes have happened during the event: More contextual changes lead to longer time estimates (Block & Reed, 1978). When the primary task is more demanding (i.e., higher cognitive load), more attention is focused on incoming stimuli, increasing the number of contextual changes that are encoded and remembered, increasing time estimates. Supporting this dual view of timing, a seminal meta-analysis on the effect of cognitive load on prospective and retrospective timing has shown that prospective estimates decrease under higher cognitive load, whereas retrospective estimates increase under higher cognitive load (Block et al., 2010).

This interaction effect is an important explanatory target for any formal model attempting to jointly explain prospective and retrospective timing. As we will see later, the computational models that have explained the effects of cog-

nitive load on timing typically side with the categorical view advanced by Block et al. (2010) by proposing that cognitive load affects separate processes. We will propose a theoretical alternative that demonstrates that cognitive load could affect a single process (i.e., attention to time), while still capturing the differential effects on prospective and retrospective time estimation.

Timing Phenomena

To demonstrate how our model unifies prospective and retrospective timing, we will consider some target phenomena. First, we introduce behavioral and neural phenomena that have proven robust features of interval timing performance, and which have also been successfully modelled by other theoretical frameworks. Then we discuss some phenomena that directly specify in which situations prospective timing is affected, either by interruptions, cognitive/memory load or inattention. Finally, we will discuss the effect of perceived changes on retrospective timing. When compounding these phenomena, we will demonstrate how prospective and retrospective timing may originate from the common principles.

Timing variability increases over time

Time estimation is remarkably precise. Humans can reliably recognize and reproduce intervals with little variability. Interestingly, however, this variability increases with the target interval, such that the longer the interval to-be-estimated, the larger the variability of those estimates. In fact, many studies have reported that the standard deviation of time estimates scales linearly with time (Lejeune & Wearden, 2006; Wearden & Lejeune, 2008), which is called the scalar property (Gibbon, 1977) [1]. Despite its central role in the timing literature, some deviations have also been observed. For instance, standard deviation has both been shown to increase slower than linear (e.g., Lewis & Miall, 2009) and faster than linear (e.g., Bizo et al., 2006), although it is not clear under which circumstances these distinct violations of the scalar property occur. As such, a major challenge is not just to account for the scalar property, but also to explain why that property may not always hold.

Timing behavior can be learned rapidly

The timing of behavior needs to be flexible. When an interval lasts longer than expected, we need to learn to be more patient in the future, so as to prevent premature responses. Conversely, when an interval ends before we expect it, we need to react more quickly the next time around, in order to not miss out on a window of opportunity. Evidence from humans and non-human animals suggests that this learning can happen impressively quickly, in as little as one or two exposures to a new target interval (Komura et al., 2001; Mello et al., 2015; Simen et al., 2011a). For instance, when humans need to respond as close to the end of an interval (but not after it has ended), they can learn to respond sooner (or later) when the target interval decreases (or increases). This learning happens in as little as one or two exposures to the new interval (Simen et al., 2011a). Few-shot temporal learning clearly contrasts with slower forms of learning (Bueti & Buonomano, 2014), and as such it represents an important benchmark for models of timing.

Complex neural patterns exhibit temporal scaling

Traditionally, it has been hypothesized that the neural mechanisms underlying timing resemble a simple accumulation process. Even though this view has been questioned on empirical and theoretical grounds (see e.g., Kononowicz & Penney, 2016; Kononowicz et al., 2018; van Rijn et al., 2011), it is still a prominent view in the literature (Salet et al., 2022). However, recent studies have uncovered that the neural mechanisms underlying timing might be fairly diverse. Researchers have not just found neurons that steadily increase their firing during an interval (ramping cells; Emmons et al., 2017), but also neurons that decrease their firing (decaying cells; Mita et al., 2009) or fire only at specific moments in time (time-cells; MacDonald et al., 2011). These findings are difficult to align with theories that propose only a single neural mechanism underlying timing performance. Interestingly, this same set of diverse neurons also exhibits temporal scaling: Their firing patterns compress when short intervals are timed and stretch when long intervals are timed (e.g., Emmons et al., 2017; Henke et al., 2021; Shimbo et al., 2021; Wang et al., 2020; Wang et al., 2018; Zhou et al., 2020). In other words, the speed at which their firing pattern unfolds adapts to the target interval. Furthermore, the degree of temporal scaling predicts trial-to-trial fluctuations in time estimation (Wang et al., 2018), suggesting that temporal scaling has an important functional role in timing. Despite the established role of complex neural patterns and temporal scaling in timing behavior, their underlying principles and interconnections are not clear yet.

Interruptions induce delays in timed responses

In realistic contexts, timing may be interrupted, for instance when receiving a call while waiting for the last 30 seconds before draining the pasta. These kinds of interruptions are often studied with gap- and distractor paradigms (Roberts & Church, 1978). Here, subjects are trained to respond after a 'timing' signal (e.g., a tone) has been presented for a certain amount of time. On some trials, a gap in the

[1] The scalar property also pertains to the scaling of the distributions of time estimates. However, in its simplest form, the linear relationship between mean and standard deviation simplifies to Weber's Law for time perception.

timing signal or a salient distractor is presented. Subjects sometimes either ignore the interruption, pause the timing process until the end of the interruption, largely forget how much time has passed before the interruption, or show behavior that is somewhere in between those possibilities (Buhusi & Meck, 2009a). The amount of 'pausing' or 'forgetting' depends on several factors, such as the timing of the interruption (Cabeza de Vaca et al., 1994), the dissimilarity between the distractor and the timing signal (Buhusi, 2012), the length of the target interval (Buhusi & Meck, 2009b), and the novelty of the distractor (Buhusi & Matthews, 2014). Overall, these findings suggest that memories of how much time has passed may be forgotten when timing is interrupted. Nevertheless, it is not yet clear how or why this kind of forgetting happens in the first place.

Working memory load decreases prospective time estimates

The effect of interruptions on timed responses already suggests that timing somehow has limited capacity. More specifically, it seems that timing limitations are related to capacity limitations in working memory (Fortin & Schweickert, 2016). For instance, when performing an N-back task, increasing the number of items that need to be concurrently remembered (i.e., working memory load) does not only affect working memory performance but also decreases prospective time estimates (Polti et al., 2018). This kind of interference is specific to working memory: Processing in working memory interferes with timing, whereas visual search, task switching and long-term memory activation do not (for a review, see Fortin & Schweickert, 2016). A popular interpretation of this effect is that working memory and timing share a common, limited resource (Buhusi & Meck, 2009a). While this theoretical position is often voiced in the literature, it is not clear how such a limited resource is implemented neurally.

Attending to time increases prospective time estimates

Time seems to drag on in boring situations, such as watching a pot boil (Block et al., 1980; Cahoon & Edmonds, 1980). One common explanation posits that in boring situations we 'attend to time', which in turn increases prospective time estimates. This prompts the feeling that we have been waiting for longer than is actually the case. The effects of attention on time perception have also been confirmed in more controlled settings. For instance, when more attention is paid to the timing task in dual-task paradigms, subjective estimates of the interval are longer (Casini & Macar, 1997; Franssen & Vandierendonck, 2002; Macar et al., 1994). Interestingly, for durations up to a minute, self-reported attention to time increased time estimates over and above differences between prospective and retrospective timing instructions (Martinelli & Droit-Volet, 2022). This effect of divided 'attention to time' suggests that keeping track of time demands attention. A related way in which attention increases time estimates is selective attention, for instance when subjects pay attention to a certain region in space. Stimuli in the attended region are perceived to last longer than unattended stimuli (e.g., Enns et al., 1999; Mattes & Ulrich, 1998; Seifried & Ulrich, 2011; Yeshurun & Marom, 2008). In sum, the effect of attention on time estimation is twofold: Divided attention to the timing task increases prospective estimates and selective attention to *stimuli* increases their perceived duration. The concept of 'attention' features prominently in theories of time perception. However, a major challenge for these theories is to implement attention in a neurally plausible way.

Divided attention to time interferes with secondary tasks

Divided attention to the timing task and selective attention to timed stimuli both increase prospective time estimates. However, an important reason to dissociate between their effects on time estimation is that they have opposing effects on stimulus processing. Selective attention both increases prospective time estimates (e.g., Enns et al., 1999; Mattes & Ulrich, 1998; Yeshurun & Marom, 2008) and enhances task performance for attended stimuli. Directing divided attention to the timing task, however, *impairs* secondary task performance (for a review, see Brown, 2006). For instance, when subjects are asked to give priority to the timing task, performance on luminance detection tasks (Casini & Macar, 1997; Macar et al., 1994), visual working memory tasks (Franssen & Vandierendonck, 2002), Stroop interference tasks (Zakay, 1998) deteriorate. It has been suggested keeping track of time requires executive processes (Brown, 2006) specifically those important for memory updating (Ogden et al., 2011). In a classic fMRI study, participants were instructed to divide their attention between a timing task and a color working memory task. When participants attended more to the timing task, not only did their performance on the color working memory task deteriorate, neural responses in brain areas responsible for color perception (V4) were also attenuated (Coull et al., 2004). In contrast, neural responses to selectively attended stimuli are typically enhanced (Treue, 2001). In sum, if 'attention' is invoked to explain variations in time estimation, divided and selective attention need to be dissociated carefully.

Perceived changes increase retrospective estimates

As already suggested by William James, an interval with varied and interesting experiences seems long as we look back. Indeed, when more stimuli are perceived, retrospective time estimates are longer (Block & Reed, 1978; Fountas et al., 2022; Lositsky et al., 2016; Mcclain, 1983; Predebon, 1996). Crucially, the number of *perceived* stimuli is a reliable predictor of retrospective estimates, but

not the number of remembered stimuli (e.g., Block, 1974). The number of perceived stimuli in an interval also lengthens prospective estimates in some situations (Bangert et al., 2019; Faber & Gennari, 2017; Herbst et al., 2012; Kladopoulos et al., 2004; Poynter & Homa, 1983; Roseboom et al., 2019; Waldum & Sahakyan, 2013), but in other situations shortens prospective estimates (Bangert et al., 2020; Liverence & Scholl, 2012; Mcclain, 1983; Poynter & Homa, 1983; Predebon, 1996). While the effect of the number of stimuli seems consistent across prospective and retrospective paradigms (Block & Zakay, 1997), their effects can be dissociated experimentally. For instance, when stimuli are actively processed, increasing the number of stimuli *decreases* prospective estimates, whereas passively viewed stimuli do not have a consistent effect. At the same time, retrospective estimates increase with the number of stimuli regardless of whether these stimuli are processed actively or passively (Mcclain, 1983; Predebon, 1996). A recent study by Bangert et al. (2020) demonstrates that task requirements determine whether event boundaries lengthen or shorten prospective duration estimates. When a naturalistic event boundary happened during a to-be-estimated interval, it shortened prospective estimates. However, temporal proximity between tones was judged as more distant when an event boundary intervened.

Another effect related to perceived changes clearly dissociates prospective and retrospective timing. When a series of items is explicitly segmented, effectively processing more changes in perceptual input, retrospective time estimates increase (Poynter, 1983), whereas prospective estimates are unaffected (Zakay et al., 1994; for a meta-analysis, see Block et al., 2010). In a typical version of this paradigm, participants encode a series of memory items and are instructed to remember some 'high-priority' items at all costs. When these high-priority items are uniformly distributed over the interval (effectively segmenting the input), retrospective estimates are significantly longer than when high-priority items are clustered around the start or end of the interval. Interestingly, segmentation does not affect the estimated number of events, suggesting that segmentation is partly dissociable from the effect of the number of perceived changes (Poynter, 1983). Overall these findings suggest that actively processed changes, and in particular events that segment a stream of input, selectively shape retrospective but not prospective time estimates.

Cognitive load affects prospective and retrospective estimates differently

As referred to earlier, the seminal meta-analysis by Block et al. (2010) on time estimation has found that cognitive load decrease prospective estimates, while it increases retrospective estimates. The effect of cognitive load on prospective and retrospective estimates neatly combines the effects we discussed above. As more attention is paid to the difficult secondary task, less attention is paid to the timing task, decreasing prospective time estimates. Conversely, as the primary task becomes more difficult, more changes are stored in memory, increasing retrospective time estimates. In sum, the differential effects of cognitive load on prospective and retrospective time estimation may provide us with information on how they differ. As we will see in the next section, current state-of-the-art models that attempt to explain this interaction suggest that cognitive load affects different processes for prospective and retrospective time estimation – attention and memory, respectively.

Models of Timing

Despite the central role of timing in everyday activity, its underlying cognitive and neural processes remain an active matter of debate (for a comprehensive review, see Paton & Buonomano, 2018). A wide variety of models have been proposed that show basic timing capabilities, suggesting that many possible mechanisms could keep track of time. Here we will introduce several classes of timing models and discuss how they explain the timing phenomena introduced earlier. We will then zoom in on two models that explain the paradoxical effect of cognitive load on prospective and retrospective time estimation.

Pacemaker-Accumulator Models

The earliest formal models of interval timing were Pacemaker-Accumulator (PA) models (Creelman, 1962; Treisman, 1963; for an extensive review of PA models, see Simen et al., 2013; van Rijn, 2014). These models view timing as the accumulation of 'ticks' that are emitted by a pacemaker. The number of accumulated ticks represents how much time has elapsed since the onset of a single timed event, but also the expectancy of future rewards that follow these events after a predictable interval (Gibbon et al., 1984; Killeen & Fetterman, 1988; Simen et al., 2011a).

There are several ways in which PA-models explain the scalar property, mainly varying in assumptions they make about noise in the pacemaker (e.g., Simen et al., 2013), the memory system that stores temporal information (Gibbon et al., 1984) or in the rate of the pacemaker (Treisman, 1963; Ulrich et al., 2022). Several versions of PA-models have also been proposed that successfully account for violations of the scalar property (e.g., Bizo et al., 2006; Namboodiri et al., 2016). PA-models propose that a steady accumulation of 'ticks' underlies time estimation, which makes it difficult to account for complex neural patterns. However, some PA-models (for an overview, see Simen et al., 2013; also see, Almeida & Ledberg, 2010) propose that shorter or longer intervals are learned by speeding up or slowing down accumulation, which explains the temporal scaling of ramping neurons (Komura et al., 2001). The effect of interruptions

on time estimates can be successfully explained by assuming that time estimation shares some attention and memory resources. During the interrupting event, accumulation (partly) stops and accumulated 'ticks' are gradually forgotten, explaining most of these 'interruption' effects fairly parsimoniously (Buhusi & Meck, 2009a). Similarly, this same set of assumptions can explain the effect of working memory load on time estimates: When working memory resources are taken away from the timing task, prospective time estimates decrease (Fortin & Schweickert, 2016). Conversely, when attention is directed to the timing task, more 'ticks' can be accumulated, increasing prospective estimates (Zakay & Block, 1995). A similar explanation also holds for the effect of attention on time estimates: More ticks are accumulated for attended stimuli. Alternatively, attention to time entails that the current 'tick count' is monitored more consistently. Conversely, when attention is taken away, the count is inspected too late, delaying timed responses (Taatgen et al., 2007; van der Mijn & van Rijn, 2021)[2]. While the effect of 'attention to time' on secondary task performance is not often considered, it is compatible with time estimation sharing a *common* resource with other cognitive processes that underlie secondary task performance (Buhusi & Meck, 2009a). Lastly, PA-models are not ideally suited to explain retrospective time estimation. If we assume a single internal clock, the model would need a clear cue when to start and stop accumulating, which is not available in 'retrospective' scenarios. In principle, PA-models could account for retrospective timing if each event started its own internal clock from which elapsed time would be read out. However, it is not clear whether this solution scales well given the number of events that would need to be timed (and therefore clocks that need to run in parallel) , and if it does, whether it would account for empirical patterns in retrospective timing.

Memory Models

Memory models of timing are a more recent development in the literature and were a clear reaction to the more dominant PA-models (Staddon & Higa, 1999). Instead of proposing an accumulation process, memory models have generally implemented timing as keeping track of the activity of memory traces (French et al., 2014; Grossberg & Schmajuk, 1989; Killeen & Grondin, 2021; Shankar & Howard, 2010, 2012; Staddon & Higa, 1999). These models assume that events create memory traces that decay over time. Elapsed time since an event can be estimated from how much activity is left in memory traces associated with that event, similar to how radioactive decay can be used to date fossils.

Memory models have been successful at explaining both adherence to the scalar property (French et al., 2014; Shankar & Howard, 2010, 2012) as well as violations of the scalar property (Killeen & Grondin, 2021; Staddon & Higa, 1999). While all memory models specifically propose neural decay as central to timing performance, the TILT model (Shankar & Howard, 2010) has also successfully predicted the distribution of neural decay rates in entorihinal cortex (Bright et al., 2020), time-cell activity (MacDonald et al., 2011; Pastalkova et al., 2008), and proposed how these time-cells might exhibit temporal scaling (Liu et al., 2019) as observed in (Shimbo et al., 2021). The effect of interruptions on timing can be captured by memory models quite naturally since recent 'timing' input is gradually forgotten during the interruption (Hopson, 1999). The effect of working memory load has been successfully modelled by the Fading-Gaussian Activation Model of Interval Timing (GAMIT; French et al., 2014), which assumes time estimation competes for attention with concurrent tasks in working memory. GAMIT can also explain the effect of 'attending to time' in dual-task situations, however, it is not clear whether GAMIT also predicts that time estimation degrades secondary task performance (see section GAMIT). GAMIT does not explicitly address the effect of the number of perceived stimuli on retrospective time estimates. Conversely, the Predictive Processing model (Fountas et al., 2022) has shown that the number of perceived events can explain retrospective estimates for eventful scenes (see section Predictive Processing Model). As we will see later, both GAMIT and the Predictive Processing model can explain the differential effects of cognitive load on prospective and retrospective time estimation by assuming that they affect attention and memory processes, respectively.

Recurrent Neural Network Models

In recent years, Recurrent Neural Network (RNN) models have gained prominence in the timing literature (e.g., Buonomano, 2000; Buonomano & Mauk, 1994; Egger et al., 2020; Gavornik et al., 2009; Goudar & Buonomano, 2018; Hardy & Buonomano, 2018; Hardy et al., 2018; Laje & Buonomano, 2013; Pérez & Merchant, 2018; Remington et al., 2018; Shea-Brown et al., 2006; Sohn et al., 2019; Wang et al., 2018; Yamazaki & Tanaka, 2005). While the previously discussed model categories referred to specific mechanisms underlying timing behavior (i.e., accumulation, decay, oscillation), RNN models only constrain the wiring diagram of the neural network to have recurrent connections. As such, several models that were previously discussed are technically RNNs. For instance, the ToPDDM model by Simen et al. (2011a) formalize the implementation of their model as neural network with specific recurrent connections, which allows the model to implement a 'neural clock'. Here, we will mainly talk about RNN models that randomly initialize their recurrent weights (which can be further refined through

[2]This cannot be the whole story, however. Dual-tasking has a large effect on timed motor responses, such as production and reproduction, but also reliable effects on verbal estimates (see Block et al., 2010)

learning mechanisms). When these RNNs are given inputs, complex neural firing patterns ensue from which elapsed time can be read out. Consequently, RNN models demonstrate that any stable, non-repeating trajectory through high-dimensional neural state space can tell time. Indeed, this view suggests that most (if not all) neural circuits have the intrinsic ability to tell time, as opposed to models that assume dedicated timing circuits (Ivry & Schlerf, 2008).

Several RNN models have systematically explored how different sources of neural noise may explain the scalar property and deviations of the scalar property (e.g., Laje & Buonomano, 2013; Pérez & Merchant, 2018). In contrast to PA-models and memory models, RNNs exhibit many dynamic neural patterns: ramping, decaying, oscillating, and more complex patterns (see, e.g., Wang et al., 2018). A variety of RNN models have been developed that exhibit temporal scaling of these complex responses as well (Goudar & Buonomano, 2018; Murray & Escola, 2017; Sohn et al., 2019; Wang et al., 2018). RNN models have not aimed at explaining the effect of interruptions on timing. Interestingly, some researchers have shown that RNNs that were trained to robustly estimate time were insensitive to interruptions (Laje & Buonomano, 2013), which suggests that forgetting may not naturally emerge in these trained RNNs. While the complex dynamics of RNN models resemble those that are found during working memory tasks (Bi & Zhou, 2020; Cueva et al., 2020), a systematic explanation of working memory load effects on time estimation has not been explored. Similarly, RNN models have not incorporated attentional mechanisms, so the effects of selective and divided attention on time estimation are currently not explained. Further, RNN models are typically applied to prospective timing only, leaving open the question of how they explain differences between prospective and retrospective timing.

The flexibility of RNNs allows them to explain a variety of psychological and neural phenomena, but this flexibility may come at the cost of interpretability. Where previous models had a relatively straightforward interpretation of how individual states represent time (e.g., time since onset, history of events), the temporal representations used by RNNs are more elusive. RNNs are trained to perform a host of different timing tasks, after which their behavior can be studied by analysing the dynamics of the network as it performs those tasks (Beiran et al., 2023; Bi & Zhou, 2020; Sohn et al., 2019). However, there are little to no guarantees that the network solves different timing tasks using the same basic mechanisms. In this sense, RNN models generally lack strong theoretical commitments to *common* representational and computational principles underlying temporal processing. In principle, RNN activity could provide the raw materials for decoding a more abstract representation of time (van Wassenhove, 2009), which could conform to common representational principles. However, that still leaves open the question of whether the raw materials that RNNs generate operate according to common principles, or whether they are merely an accidental by-product of its wiring diagram. In sum, while RNN models have generated powerful explanations for a host of neural data, their flexibility complicates a systematic account of the representational and computational principles underlying prospective and retrospective timing. We will now zoom in on two models that have attempted such a unification.

GAMIT

In the memory section, we already mentioned the Fading-Gaussian Activation Model of Interval Timing (GAMIT; French et al., 2014). Since it is one of the few formal models that explain both prospective and retrospective timing, we will explain it in more detail here. GAMIT assumes that prospective and retrospective estimates are made based on decaying neural activity. The model learns a mapping between memory trace activity and objective time: the lower the activation, the more time has passed. GAMIT further assumes that cognitive load affects the rate of decay: when cognitive load is higher than usual, activation decays more quickly, presumably because of interference from distracting concurrent tasks (timing tasks themselves have a special status, and do not affect the rate of decay). When time estimates are made under high levels of cognitive load, activity traces will have decayed more than under typical levels of cognitive load, explaining why retrospective estimates are longer under high cognitive load. The model further proposes that only in prospective conditions, activity traces are sampled by a separate attentional mechanism. This attentional sampling mechanism produces an estimate of how quickly the activity trace decays: if the difference between consecutive activity samples is large, the rate of decay is estimated to be high. As a result, the model estimates that the passage of time is relatively fast. GAMIT assumes that this 'passage of time' estimate adjusts activity-based estimates. For instance, if the estimated rate of decay is faster than typical, activity-based estimates will be adjusted to be shorter, since time seems to be passing more quickly.[3] Crucially, when attention is diverted away from timing, fewer samples are collected, leading to larger differences between consecutive samples and therefore fast passage of time estimates. Activity-based estimates are adjusted to be shorter, explaining why prospective estimates decrease with high cognitive load.

We believe that the hypothesized role of attention in GAMIT may preclude a comprehensive explanation of some

[3]The assumed connection between prospective time estimates and 'passage of time' estimates, however, is more complicated: Passage of time judgements are often not systematically related to prospective time estimates (Wearden, 2015).

important effects. First, GAMIT models divided attention to the timing task, but attentional sampling of the memory trace does not affect the memory trace. Therefore, it is not clear how more attention to time (i.e., more sampling) might cause interference with secondary task performance (see Divided attention to time interferes with secondary tasks), in particular tasks in which such a memory trace might be central to performance, such as working memory tasks (e.g., Franssen & Vandierendonck, 2002). Also, while attentional sampling does not influence neural activity in GAMIT, attention to time does seem to attenuate neural responses related to secondary task performance (Coull et al., 2004). While GAMIT does not explicitly model the effects of selective attention to stimuli on time perception, it seems that the magnitude of neural responses would not play a role in such an explanation, since attention is not assumed to influence the activity trace. In contrast, a large body of evidence suggests that larger neural responses engender longer perceived stimuli (Matthews & Meck, 2016), and selective attention to stimuli amplifies neural responses (Treue, 2001). In sum, GAMIT's assumption that attention does not influence neural activity is inconsistent with several timing phenomena that are related to attention.

Predictive Processing Model

The most recent model attempting to explain the effect of cognitive load on prospective and retrospective time estimation is the Predictive Processing model by Fountas et al. (2022). The core principle of the Predictive Processing model is that time estimates are based on counting the number of surprising events encoded in a sensory processing network. This model captures idiosyncratic biases in prospective timing, where more eventful scenes were judged to last longer than uneventful scenes (Fountas et al., 2022; see also Roseboom et al., 2019; Sherman et al., 2022). The predictive processing model assumes that sensory inputs are processed by a hierarchical Bayesian network. The network continually updates an internal model of the world by generating model-based predictions and comparing these predictions to incoming information. When predictions are violated, the network generates a prediction error. Crucially, when the magnitude of the prediction error crosses a decaying threshold, it is 'surprising' enough and the network encodes the relevant information as an event in episodic memory after which the decaying threshold is reset. Time estimates are generated by reading out the number of surprises in the hierarchical network. When less attention is paid to time in high-load prospective conditions, the dynamic threshold decays slower, resulting in fewer surprising events being encoded in episodic memory and shorter time estimates. In contrast, effects of cognitive load in retrospective conditions are explained by memory retrieval processes, specifically how much effort is put in retrieving events from episodic memory after the interval has ended.[4] The model assumes that in high-load conditions, more effort is put into retrieving events after the interval is over, leading to more retrieved events and therefore longer time estimates.

The Predictive Processing model assumes that as more attention is paid to time, the attention threshold decays faster, leading to more surprising events and longer time estimates. However, by equating time estimates with the number of surprising events, the model may be unable to account for some effects of attention on time estimation. First, while its attentional mechanism can amplify sensory signals, it fails to capture how divided attention to time interferes with secondary task performance. For instance, when attention is directed at time, more events are encoded in episodic memory, but the model does not clarify how more remembered events could lead to worse task performance in secondary tasks, especially working memory tasks (e.g., Franssen & Vandierendonck, 2002; Macar et al., 1994) or luminance detection tasks (Casini & Macar, 1997; Macar et al., 1994). Second, surprising events do not always lead to longer temporal percepts. Several studies have shown that stimuli shown at cued locations are perceived as *longer*, even though stimuli at cued locations were more probable and therefore *less* surprising than stimuli at uncued locations (Enns et al., 1999; Mattes & Ulrich, 1998; Yeshurun & Marom, 2008). Further, as mentioned earlier, the number of stimuli increases prospective estimates, but mainly when stimuli are not actively processed. Instead, when stimuli are actively processed, more stimuli decrease prospective estimates (Mcclain, 1983; Predebon, 1996). It is also not clear how the Predictive Processing model would account for the effect of interruptions (see Interruptions induce delays in timed responses). When a salient distractor (i.e., a 'change' in the timed signal) is introduced in timing tasks, timed responses are delayed in proportion to the dissimilarity to the timed signal (for a review, see Buhusi & Meck, 2009a). In contrast, the Predictive Processing model would predict that more salient changes would lead to *faster* timed responses, given that *more* subjective time is accumulated. Additionally, when the distractor is not familiarised (i.e., more surprising), timed responses are delayed even more (Buhusi & Matthews, 2014). In sum, by equating 'surprising events' to subjective time, the predictive processing model overlooks phenomena in which surprises *compress* subjective time (for a similar critique of this type of explanation, see Phillips, 2012).

The Unified Temporal Coding Model

In this paper, we develop a neurocomputational model of prospective and retrospective timing: the Unified Temporal Coding (UTC) model. The UTC model puts for-

[4] But it is unclear why cognitive load would not affect the encoding of episodes into memory.

ward unifying representational and computational principles underlying prospective and retrospective timing. Here we will provide a conceptual sketch of those principles and how they can account for the timing phenomena we introduced earlier.

At the core of the UTC model is the Legendre Delay Network (LDN; Voelker & Eliasmith, 2018), which is structured to optimally approximate a rolling window of its input history. This neural network maintains and continuously updates a representation of the recent past. To illustrate how this works, consider a sequence of inputs (Figure 2). Our network represents, at any point in time, not only the current input but also its history up to a certain point in the past: it represents a rolling window of input history. Incoming inputs are encoded into the front of the window and past inputs are gradually pushed to the end of the window until they eventually fall outside of it.

It should be stressed that the UTC model does not just process a sequence of stimuli with a certain time-constant (i.e., the one set by the window). In addition to going through a sequence of stimulus representations, it also *represents a sequence of stimuli*. That is, *at any point in time*, the network does not only represent the current input (or a few hundered milliseconds ago, accounting for physical and physiological delays); it represents on a continuous timeline what happened when. This timeline spans the interval between the current timepoint ('now') to some point in the past, defined by the window's size. Crucially, stimulus information within the window is not simply an echo of its initial presentation, and its temporal location inside the window does not depend on its 'strength'. Instead, information continually slides across the window, and as such temporal information is actively constructed, instead of passively receding into the past.

As we will see later, the way that our network actively represents a rolling window explains the variety of complex neural firing patterns underlying timing performance. Further, when the network only needs to remember events that happened very recently, it can shrink the size of the temporal window on-the-fly. This ensures that more recent events are represented with higher fidelity (i.e., smaller error), but at the cost of more distant events that fall outside of the smaller window. Our network accomplishes this by speeding up the dynamics, which explains why complex neural patterns may exhibit temporal scaling (see Changes in window size explain temporal scaling in complex neural patterns).

The UTC model also details how the length of the rolling window can be learned rapidly. A on-shot learning rule (adapted from TDDMs Rivest & Bengio, 2011; Simen et al., 2011a) details how the window needs to be lengthened when the target interval is longer than the window, and shortened when the target interval is shorter than the window. In effect, the UTC model can learn to match the window size to the target interval. These learning mechanisms ensure that all relevant stimulus information during the interval can fit inside the window. Whenever the window shrinks, this ensures that the information still inside the window is represented at higher fidelity. As we will show later, this learning mechanism matches empirical learning rates from behavioral data, and is consistent with rapid adaptation of neural ramping speed (Komura et al., 2001).

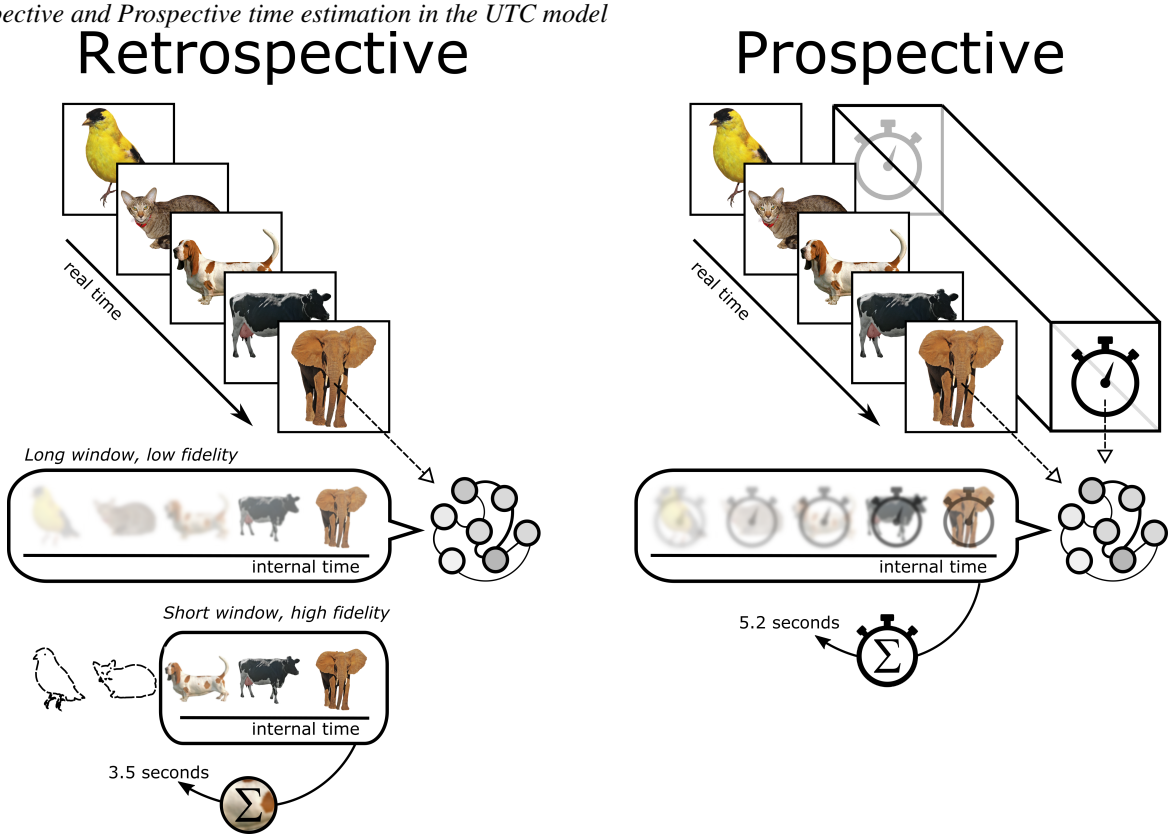
The UTC model assumes that retrospective time estimates are made by summing the overall fidelity of the remembered inputs that are represented inside the temporal window. When more inputs are summed, retrospective time estimates become longer. Similarly, when those same inputs have higher fidelity, retrospective time estimates also increase. This mechanism provides an explanation of why more perceived stimuli increase retrospective time estimates (see Integrating remembered content explains effects of contextual changes).

Prospective estimates of stimulus duration are made in the same way as retrospective estimates. Since the network *continuously* updates input history, the longer a stimulus is presented, the longer the representation of that stimulus in the rolling window. Then, if the network integrates this representation within the rolling window, it gives an accurate estimate of stimulus duration. However, when the stimulus input is interrupted in some way, the representation of stimulus history contains a 'gap', resulting in lower time estimates, effectively delaying timed responses. We will show later that the UTC model explains how the timing of the gap, the similarity of the interrupting distractor and the to-be-estimated interval determine delays in temporal responses (see Forgetting of timing information accounts for the effect of interruptions).

In some situations, we need to prospectively time an 'empty' interval with little to no external stimuli. In this case, we assume that the network receives an internally generated, constant 'timing' input. Crucially, this timing input is represented in the same way as stimulus inputs. As such, there is no difference between prospective and retrospective timing apart from the fact that this input is used to estimate time. Using a constant 'timing' input ensures that time estimates are largely independent of fluctuations or gaps in the stream of stimulus inputs. As we will show later, the more inputs are presented to the network, the more both timing and stimulus inputs will be distorted. We will demonstrate that this pattern of interference is similar to interference found in working memory, which explains why higher working memory load (i.e., more stimulus inputs) interferes with timing performance. Interestingly, this same mechanism also explains why performing a timing task degrades secondary task performance: The 'timing' input interferes with the representation of stimulus inputs as well (see Neural normalization

Figure 1

Retrospective and Prospective time estimation in the UTC model



Note. Left panel: The UTC network receives a series of stimuli and remembers both their content and their temporal position inside a temporal window. Incoming stimuli are encoded at the right of the temporal window. Recent stimuli are continually pushed to the left until they fall outside of the window and are forgotten by the network. The network can control the size of the temporal window based on task demands. Longer windows ensure that longer temporal patterns can be tracked, but at the cost of lower fidelity. Short windows can only track temporal patterns over a brief timescale, but the fidelity of the content is higher. Time estimates are made by adding up how much content the network remembers in the temporal window and mapping that to a unit of time. Right panel: When the UTC model prospectively estimates an interval, it receives an additional constant 'timing' input (see clock in the Figure). This 'timing' input is represented in the same way as the stimulus inputs. The UTC model estimates time by adding up how much of this 'timing' input it remembers.

explains effects of working memory load).

Unlike previous models, the UTC model only incorporates an attentional mechanism with respect to stimulus input. When more selective attention is paid to stimuli, the input is multiplied by an attentional gain factor, consistent with neurophysiological effects of attention (Treue, 2001). As we will see later, this explains why attended stimuli are perceived as longer than unattended ones: Attentional gain increases the vividness of the stimulus input, resulting in longer estimates (see Attentional gain explains effects of selective and divided attention on time estimation). Crucially, to model divided attention, the UTC model assumes that when more attention is paid to timing, less attention is paid to stimulus inputs. Because these stimulus inputs are partially 'ignored' they have less opportunity to interfere with the timing input. This both explains why attending to time increases prospective time estimates (less interference), but also why paying more attention to time interferes with secondary task performance (stimulus inputs are less attended; see Attentional gain explains effects of selective and divided attention on time estimation).

The effects of cognitive load on time perception tend to suggest that prospective and retrospective timing are different kinds of processes. The UTC model, however, suggests a different view. In cognitively demanding tasks, more divided attention needs to be paid to incoming stimuli. These incoming stimuli compete with the timing input, effectively decreasing prospective time estimates. In contrast, when stimuli are attended more in retrospective timing, they lead to more change being encoded in the temporal window, increasing retrospective time estimates. The only difference between prospective and retrospective timing is the 'timing'

input to the network. But precisely *because* stimulus and timing information is processed in the same way, can we account for the interaction effect of cognitive load with a single parameter: The attention paid to stimuli.

Methods

In this section, we will detail the representational and computational principles behind the UTC model. First, we describe the Legendre Delay Network (LDN; Voelker et al., 2019), which a memory network that tracks 'what' happened 'when' over a rolling window. We will describe how complex information is represented by the network as high-dimensional vectors (referred to as Semantic Pointers; Eliasmith, 2013). In the appendix, we also demonstrate how to implement the LDN as a spiking neural network. Finally, we will give a brief overview of the full network architecture of the UTC model.

Legendre Delay Network

How do we represent 'what happened when' up to some arbitrary point in the past? Consider the case where we want to remember the luminance of some light source, from the current moment up to one second ago. One could, in principle, store each new input on the first slot of a memory register, while moving already stored inputs along. The input at the end of the memory register is dropped (after exactly two seconds). At each point in time, the memory register perfectly represents a temporal window of 'what happened when' that spans the current moment up to some point in the past.

There are some obvious problems with this approach, but the most salient one is memory capacity. If we want to perfectly store the history of an input that evolves through *continuous* time, we would need infinite memory capacity. A more scalable solution is to approximate the input's history. Given limited resources, the optimal way to represent the input history up to some point in the past is with the Legendre Polynomials (Voelker, 2019). Similarly to how a signal in the frequency domain can be approximated by a finite combination of sines and cosines, a signal in the time-domain can be approximated by the Legendre polynomials. This representation in the time-domain is optimal in the sense that it minimizes the RMSE between the representation and the history of the input up to some point in the past.

The Legendre Polynomials can be considered a 'temporal basis function', from which one can construct a representation of input history. In our model, we will use a shifted version of the Legendre Polynomials (A) that is defined over the interval between 0 ('now') and θ, where θ is the length of the temporal window (i.e., up to when inputs need to be remembered). We will denote the number of polynomials that are used (i.e., the order of the approximation) as d. Each polynomial adds something unique to the representation in the temporal window. The first dimension represents the mean of the signal, the second one represents the slope, the third one the quadratic component, and so on. We simply need to determine how much 'weight' we should give each polynomial and add them up to form a representation of the input history. We will denote these 'weights' as the d-dimensional vector x: each value in x corresponds to a weight for its associated Legendre polynomial.

Given this optimal method of representing the input history, let us consider how to construct such a representation on-the-fly, that is, specify the algorithm that can be used to generate such a representation. Put differently, at each moment in time, we want to know how to encode new inputs (u) while maintaining and updating our current representation of input history (x). We want our system to represent the history of its input $u(t)$ using a d-dimensional state-vector $x(t)$, where each of the d coefficients applies to a different dimension of our temporal basis function (the Legendre Polynomials). Since we have defined our challenge in continuous time, the most natural way of viewing our system is as a dynamical system:

$$\theta\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t) \tag{1}$$

where θ is the length of the window, $\mathbf{x}(t)$ is a d-dimensional state vector and $\dot{\mathbf{x}}(t)$ is the temporal derivative of $\mathbf{x}(t)$. The input matrix ($\mathbf{B}$) defines how new inputs should be encoded and the dynamics matrix ($\mathbf{A}$) defines how to maintain our current representation of input history (for a detailed derivation, see Voelker (2019)). We can think of the input matrix as mapping the new moment in time into the Legendre polynomial space, in such a way that it is combined with the previous representation of the input history without distorting that history. At the same time, the dynamics matrix maps the current history to the next moment in time, while 'dropping' the oldest point in the memory, since that oldest moment is now longer than θ seconds ago. Performing these mappings over and over means that old information is constantly dropped and new information is constantly added so that, at any moment the vector $\mathbf{x}(t)$ contains exactly θ seconds of historical information.

Note that we have included θ as a variable in the dynamical system that can be adjusted on-the-fly. If we want the system to only remember the last 10 instead of 20 seconds, we may decrease θ, leading to faster encoding and forgetting of information. As changing θ does not influence the dimensionality (i.e., the 'storage space' stays the same), the incoming information can be stored with higher fidelity when θ is reduced. This demonstrates the inherent balance in the system, it can either store information over longer timeframes with lower fidelity, or use the available resources to capture the input at high fidelity over shorter time frames. To illustrate how this system works, Figure 2 shows how the

network represents the last θ seconds of its input history [5].

The shifted Legendre polynomials are timescale-invariant because they are defined over $0 < \frac{\theta'}{\theta} < 1$, where θ' is a time point within the temporal window. Therefore, for any θ, the underlying temporal basis functions will be exactly the same, scaled by θ. To illustrate, given a scaled input, the state-space representation at $\theta' = 10$ms for $\theta = 100$ms is exactly the same as the state-space representation at $\theta' = 10$ seconds for $\theta = 100$ seconds. In sum, the underlying representation of time is the same, regardless of the timescale that we are dealing with. This suggests that the algorithm will apply well across many timescales.

In the appendix (Appendix B), we describe how the algorithm can be implemented as a recurrent neural network in a detailed, biologically plausible neural framework.

Semantic Pointers

To this point, we have described how to implement an algorithm that can represent a rolling window of a 1-dimensional signal (e.g., a network that only tracks the luminance of a single source of light). However, the representations that are the part and parcel of cognition are more complex. How do we move from a 1-dimensional representation to the representation of the letter 'A' on a screen or a complex tone that signals future rewards? A possible solution to this problem is assuming that these symbol-like entities are represented as high-dimensional vectors that we call 'semantic pointers' (Eliasmith, 2013). Semantic pointers are compressed neural representations that provide a consistent representational protocol for supporting a wide variety of biological behaviours, including perception, action, decision-making, and symbolic cognition. Semantic pointers, along with the architecture they are a part of, have been used to build the world's largest functional brain model, Spaun (Eliasmith et al., 2012). This architecture is implemented with the NEF, as it naturally extends to high-dimensional vector representations. In the UTC model, we use the methods of semantic pointers to capture different concepts being represented at different points in time.

The Legendre Delay Network described in the previous section also naturally extends to representing a rolling window of vectors, which can, for instance, be semantic pointers. Instead of presenting a 1-dimensional signal to an LDN, we can present a D-dimensional semantic pointer to D networks that each encode a rolling window of a single dimension of the semantic pointer.[6] Therefore, the collective $D \times d$ state matrix $\mathbf{X}$ represents a rolling window of the history of semantic pointers. To decode the original semantic pointers, we first decode the history of each component of the semantic pointers by multiplying the state vector with the Legendre polynomials ($S\,P_{\text{history}} = \mathbf{X}\mathcal{P}^T$). Then, we compute the column-wise dot-product between the resulting $D \times t$ matrix ($S\,P_{\text{history}}$) and the originally presented semantic pointers (Figure 3).

Semantic pointers allow for complex information processing, from action selection to abstract reasoning (Eliasmith et al., 2012). In our current model, however, we only use semantic pointers to simulate how stimuli are represented in typical cognitive and timing tasks. Nevertheless, this high-dimensional vector representation is a core explanatory principle in the UTC model. As we will see later, by assuming that both 'timing' information and 'stimulus' information are represented by semantic pointers, the UTC model accounts for phenomena where these types of information interact (e.g., stimulus-distractor similarity effects in gap procedures, working memory load effects on prospective timing, and the effect of selective and divided attention on prospective and retrospective timing).

The UTC Model - Network Architecture

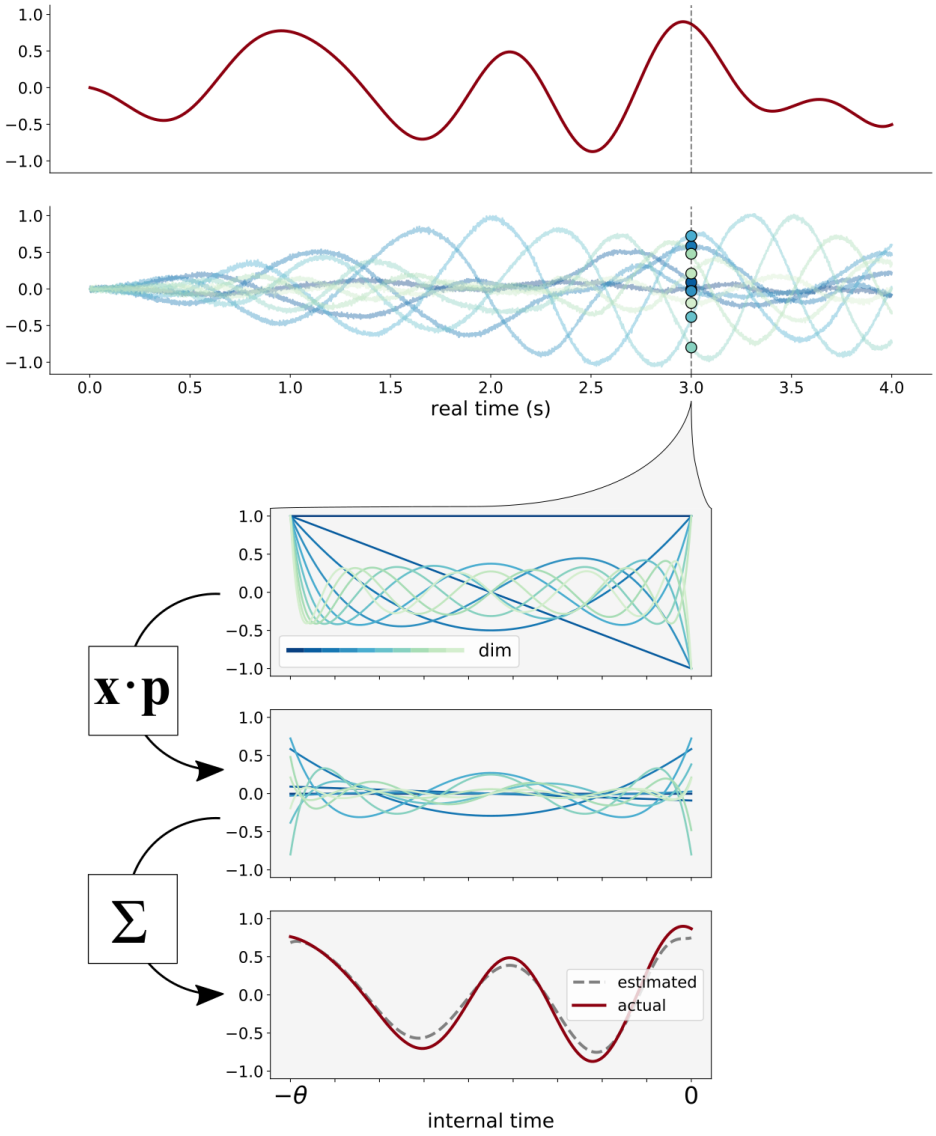
The sections above describe the mechanisms relevant to the implementation of the UTC model, to which we will now turn. The basic premise of the UTC model is that both stimulus and timing information are represented, encoded and read out in the exact *same* way. That is, both types of information are represented as semantic pointers, encoded by the LDN, and read out by integrating the representations in the temporal window of the LDN. The only difference is that stimulus inputs are waxing and waning, whereas the timing input is assumed to be relatively constant. Crucially, we propose that this is also the only difference between prospective and retrospective timing. Only when the timing task is known beforehand (prospective timing) can we have a constant timing input, otherwise the network can only reconstruct a temporal estimate based on stimulus information. The network architecture is presented in Figure 4, alongside its fixed and free parameters in Table 1, and activity traces for a typical trial in a dual-tasking timing experiment in Figure 5.

The inputs to the network are D-dimensional (64-dimensional in these simulations) vectors and come from two external sources: stimulus information (**s**) and temporal information (**t**). Stimulus information encodes external stimuli that are encoded for the task at hand (Figure 5). Temporal information feeds a constant, step-like input to the network, with its onset matched to the to-be-timed interval (Figure 5). In some experimental paradigms, this interval is filled (i.e., the stimulus stays on the screen for the duration of the interval), while in others the interval is empty (i.e., onset and

[5] A more intuitive way to understand the LDN is to see it operate in real-time (see this video-version of Figure 2 https://youtu.be/2jNp6Sf_Vsc)

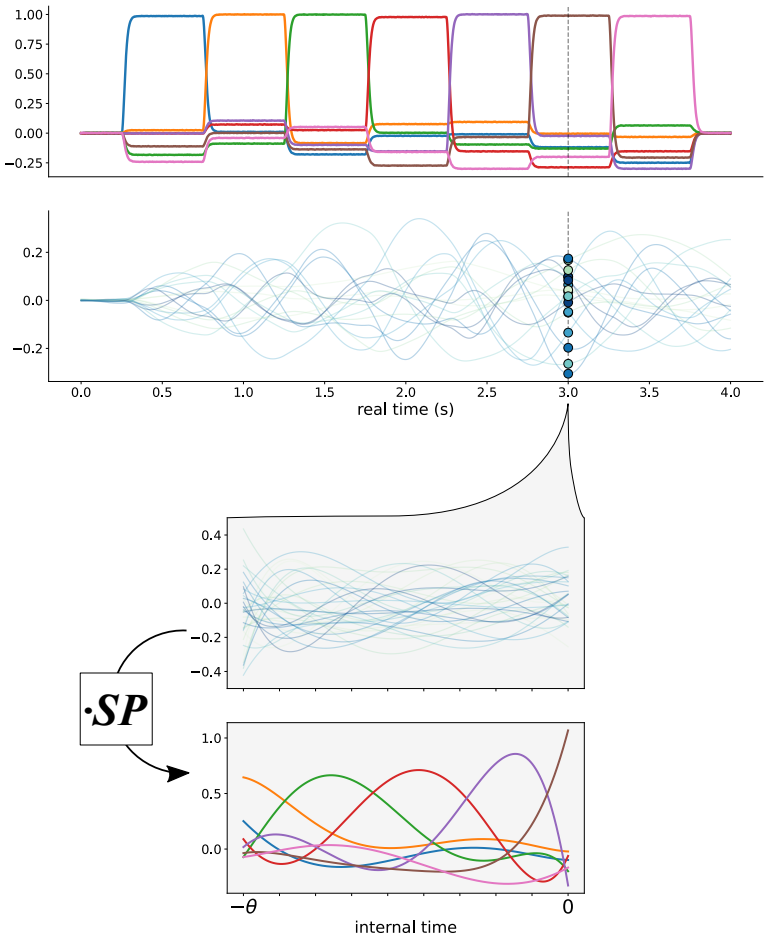
[6] Our implementation of separate LDNs representing a single dimension of the Semantic Pointer is just one possible implementation. Alternatively, a single recurrent neural network could have neurons sensitive to certain directions in the $D \times d$ space.

Figure 2
The LDN can solve the delay challenge



Note. The input u (top row) is fed to the system, which continually updates the state vector $\mathbf{x}$ containing coefficients (second row) on the temporal basis functions (the Legendre Polynomials; third row). At any point in time, the network has an instantaneous representation of the last θ seconds (here 2s) of its input history. For instance, when we take the coefficients in $\mathbf{x}$ at 3s (dots in the second row), multiply them with the Legendre polynomials $\mathbf{p}$ and take their sum, we end up with a fair representation of the input history between 0 ('now') and θ seconds ago.

Figure 3
Network example with Semantic Pointers



Note. The input (top row) is a series of D-dimensional vectors, Semantic Pointers. We plot the dot product between the input state and the ideal vectors. The network ($d = 6$)continually updates the state vector $\mathbf{X}$ containing coefficients (second row) on the temporal basis functions for each of the separate components of the input vector. At any point in time, the network has an instantaneous representation of the last θ seconds (here 2s) of its input history. When we take the coefficients in $\mathbf{x}$ at 3s (dots in the second row) and decode the history of each input component ($S\,P_{\text{history}} = \mathbf{X}\mathcal{P}^T$; third row), we can compute the dot product with the set of original vectors to decode the history of original Semantic Pointers.

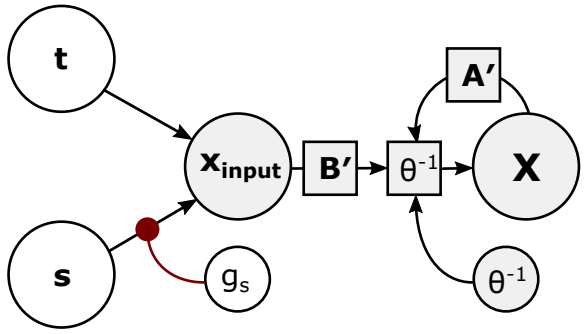
offset are defined by brief stimuli, with nothing in between). A consistent finding in the literature is that 'filled' durations are perceived as longer than 'empty' durations (Wearden & Ogden, 2021). For simplicity, we assume that in both scenarios, the network is fed a constant input, whether it is defined by a filled stimulus or whether it is self-sustained activity. This self-sustained activity can be readily implemented in the NEF (see for instance, Bekolay, Laubach, et al., 2014).

While our network clearly separates stimulus and temporal information into separate input channels, this is only done for clarity. Both sources of information are summed together in the next step, meaning that not their source but their content determines how they drive behavior. This assumption is in stark contrast to PA-models, which

Table 1
Network architecture parameters

	name	description	value
inputs	D	Dimensionality of semantic pointers.	64
	g_s	Attentional gain on stimulus input **s**	Varies between conditions and experiments (default = 1)
X	θ	Window length.	Matched to relevant timescale
	θ^{-1}	Speed of integration and forgetting.	Inverse of θ
	d	Dimensionality of LDN. Controls precision of represented history.	Fixed within experiments. Varies between (sets of) experiments.
	N	Number of neurons per LDN dimension	200 (unless otherwise indicated)
	max rate	Maximum firing rates of neurons in LDN.	Matched to maximum firing rates in modelled neural data
	$\tau_{\theta^{-1}}$	Synaptic time-constant for $\theta^{-1} \rightarrow \mathbf{X}$	0.005
	$\tau_{\text{recurrent}}$	Synaptic time-constant for $\mathbf{X} \rightarrow \mathbf{X}$	0.1

Figure 4
Network Architecture of the UTC model



Note. The network receives two external inputs: a temporal vector **t**, and a stimulus vector **s**. The stimulus vector is multiplied by an attentional gain factor (g_s). The input vectors are added in the neural population $\mathbf{x_{input}}$. Each dimension of $\mathbf{x_{input}}$ is fed into a separate LDN, collectively referred to as the neural population **X**. The input matrix $\mathbf{B}'$ encodes inputs into the window represented by **X**, and $\mathbf{A}'$ pushes past inputs towards the end of the window until they are eventually forgotten. The length of the temporal window (θ) is adapted by controlling the rate of encoding and forgetting (θ^{-1}), which is a result of simply multiplying the input matrix $\mathbf{B}'$ and recurrent matrix $\mathbf{A}'$ by θ^{-1}.

assume that, regardless of stimulus content, timing behavior is driven by the reading of an accumulated 'clock' reading. While specific kinds of stimulus content may trigger the onset of accumulation, stimulus content plays no role in subsequent timing processes. The UTC model, on the other hand, proposes that timing behavior critically depends on an integrated representation of exactly the stimulus content it is supposed to track.

The neural population $\mathbf{x_{input}}$ combines **t** and **s** by adding the two vectors together. Temporal information (**t**) is always a constant, unit-length input, ensuring stable timing behavior. However, we assume that the model can control how much it attends to stimulus inputs (**s**) by multiplying that vector by an attentional gain factor (g_s)[7]. When g_s is high, stimulus information will be better decodable from $\mathbf{x_{input}}$ (Figure 5). This attentional gain factor g_s, as we will see later, captures the effects of selective attention on time perception, models the effects of divided attention by controlling the degree of mutual interference between temporal (**t**) and stimulus information (**s**), and explains differential effects of cognitive load on prospective and retrospective timing.

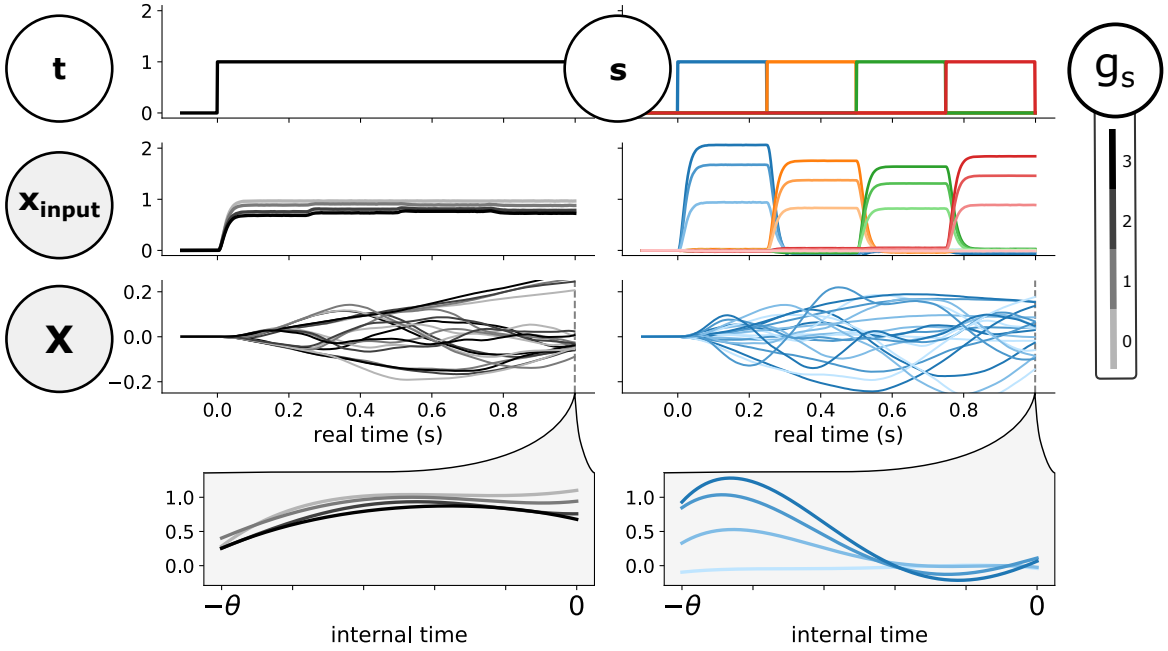
The D-dimensional vector in $\mathbf{x_{input}}$ is fed to D LDNs, collectively denoted as **X**. In effect, **X** is a D-by-d matrix, containing the coefficients on our temporal basis function for each dimension of the D-dimensional input vector $\mathbf{x_{input}}$. To encode the new information into the temporal window, each input dimension is multiplied by B' and by θ^{-1}, which controls how quickly new information is encoded into the window on-the-fly. Information already inside the temporal window, represented by **X**, is gradually pushed outside of the window through multiplication with the recurrent matrix, defined by $\mathbf{A}'$. Again, the speed at which information is pushed outside of the window (i.e., forgotten) depends on window size (θ) and is controlled by setting θ^{-1}. In section , we propose a one-shot learning rule that can learn to match the window size to the appropriate timescale of the task. Given the speed and accuracy of this learning rule, and to simplify our modeling, we simply fitted θ to match the correct timescale of the timing task. For instance, when a 10-second interval needs to be produced, θ was set to 10 seconds.

At each moment in time, the LDNs (**X**) represent a rolling window of its input ($\mathbf{x_{input}}$). This representation would allow for complex judgements about temporal patterns, however, in this paper we are only concerned with one-dimensional judgements about single intervals (e.g., 'how long did the task last?', or 'press this button after 2 seconds'). We propose that these one-dimensional time estimates (t_s) are made by integrating the absolute represented value for

[7]Note that, in the UTC model, attentional and recurrent gain are different. *Recurrent* gain (θ^{-1}) multiplies the recurrent and input matrices of the LDN network. As a result, it scales the window size. *Attentional* gain, on the other hand, only multiplies the stimulus vector. As a result it increases the decodability of the stimuli.

Figure 5

Activity traces during dual-task timing



Note. On the left, the temporal information (**t**), on the right, stimulus information (**s**). First row: input to the memory population ($\mathbf{wm_i}$). Second row: temporal (left) and stimulus information represented in $\mathbf{x}_{\text{input}}$. When g_s is increased (darker colors), meaning that more attention is paid to incoming stimuli, stimulus information is represented more clearly and interferes with the temporal information. Third row: example traces from the LDN populations that are sensitive to either temporal or stimulus information. Bottom: 'internal time', the decoded temporal window at t_{real}=1s. Final row: the network represents the last second of its inputs. The network simultaneously represents temporal (left) and stimulus information (right; first presented stimulus in blue). For higher values of g_s, the stimulus information is represented better to the detriment of temporal information.

each semantic pointer over the entire window and summing up those integrated values for all semantic pointers that are represented in the window:

$$t_s = \sum_{i=1}^{N} \int_{-\theta}^{0} |\mathbf{X}(t)\mathcal{P}^T \cdot \text{SP}_i| dt \tag{2}$$

where $\cdot$ is the dot product, $\mathbf{X}$ is the $d \times D$ matrix of coefficients represented by the network at time t, $\mathcal{P}$ are the Legendre Polynomials, SP_i is the i^{th} semantic pointer in the vocabulary (i.e., the set of possible semantic pointers that we feed into the network) and N is the number of semantic pointers in the vocabulary (see Figure 3). Here, the difference between prospective and retrospective estimates is that prospective estimates are only based on the 'temporal' semantic pointer (**t**), while retrospective estimates are based on semantic pointers presented by the stimulus input (**s**).

All code and simulation data are available on https://github.com/dejongejoost/UTC_model.

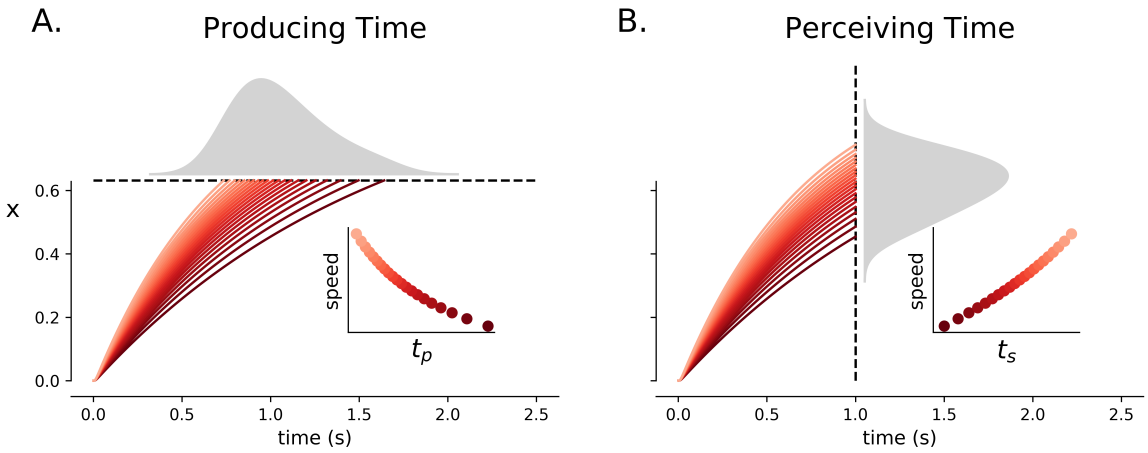
Results

In order to prospectively produce an interval, we present the network with a constant input that is integrated until it reaches a fixed threshold at which time it produces a response. Different intervals are produced by adjusting the window size to match the desired interval. The desired interval is provided in the experimental instructions. The window size, θ, is adjusted by controlling the recurrent gain θ^{-1}, which corresponds to controlling the speed at which the input is integrated. Increasing the recurrent gain will result in a shorter temporal window and will therefore produce a response after a shorter interval, while decreasing the recurrent gain will result in a longer temporal window and produce a response after a longer interval (Figure 6). Our approach is highly similar to some pacemaker-accumulator models, in particular TDDMs (Simen et al., 2013), where the speed of integration (i.e., drift-rate) is adjusted to produce different intervals. Further, when we implement a first-order LDN, this network represents the mean of the temporal window, corresponding to a leaky integrator with a time constant that equals window size (θ).

In contrast to producing an interval, the exact

Figure 6

Prospectively producing and perceiving an interval of one second



Note. (A) When the network produces time intervals, the network integrates a constant input until a threshold (dotted line) is reached, upon which a motor response is made. If the speed of integration is faster (slower), the produced interval (t_p) will be shorter (longer). (B) When the network perceives an interval, a constant input is integrated until the end of the interval (dotted line). The state of the network at the end of the interval serves as a measure of time. If the speed of integration is faster (slower), the perceived interval (t_s) is longer (shorter).

timescale for integrating inputs is not known in perceptual timing tasks, since the target interval is not known at stimulus onset. Therefore, adjusting the recurrent gain of the network on a trial-to-trial basis alone is not an effective timing strategy. Nevertheless, it is evident that subjects are sensitive to the distribution of intervals that have to be timed in a given context (e.g., de Jong et al., 2021; for reviews, see Shi et al., 2013; van Rijn, 2016). In the current model, adapting window size based on the estimated mean of the distribution would improve performance compared to using the same window size for each temporal context. For instance, when an interval of 1 second has to be estimated, the network will not be able to tell time if the window size is 0.1 seconds: the state will have evolved to its maximum value too soon and any differences between intervals thereafter are lost. In contrast, a temporal window of 10 seconds will also render differences in the state around the target duration too small, since the state evolves too slowly. There is some neurophysiological evidence to support the claim that the speed of neural dynamics during the perception of an interval is adapted to the expected range of intervals. For instance, when a short (long) interval is expected, neural trajectories in dorsomedial prefrontal cortex (DMPFC) move faster (slower) (Sohn et al., 2019). Some evidence suggests that these speed adjustments may happen on a trial-to-trial basis; when a previous interval is short (long), neural dynamics move faster (slower) (Damsma, Schlichting, & van Rijn, 2021).

To account for time perception, we assume that the window size remains constant for different target intervals and is normally distributed around the target interval. On each trial, we present the network with a constant input that is terminated after the target interval has elapsed (Figure 6). It is clear that for a fixed window size, the neural state will represent different values for different target intervals. In Figure 6, a 1-second interval is repeatedly estimated, and the recurrent gain varies normally around 1.

Different sources of noise account for different forms of Timing Variability

The scalar property of time (Gibbon, 1977), i.e., the linear scaling of the standard deviation of time estimates with its mean, has long been assumed to be a lawful property of timing. The most straightforward way of testing the scalar property is assessing whether the coefficient of variation (CV; $\frac{\sigma}{\mu}$) of time estimates is constant over a range of different target intervals. Despite much evidence suggesting that the scalar property generally holds (Lejeune & Wearden, 2006; Wearden & Lejeune, 2008), several theoretically interesting exceptions have emerged. For instance, in a set of experiments with target intervals ranging from 68 milliseconds to 16.7 minutes, Lewis and Miall (2009) found a consistently decreasing CV. In contrast, other researchers have found that the CV first decreases, and then increases for longer intervals (e.g., Bangert et al., 2011; Bizo et al., 2006; Getty, 1975; Gibbon et al., 1997; Grondin, 2014; Matthews & Grondin, 2012).

Explaining different forms of scalar variability prompted us to look at different sources of noise in our network. PA-models of timing have demonstrated that, depending on which component is affected by noise, this may either reproduce the scalar property, show a decreasing coefficient

of variation, or an increasing coefficient of variation (for an extensive review, see Simen et al., 2013). The brain is a noisy system, so it is likely that many components in our network (e.g., input, window size, individual neurons) will be affected by noise to some degree, both within a single trial and between trials. The UTC model does not make strong assumptions about different sources of noise, since these may vary between subjects, tasks and even over the course of learning. Nevertheless, we will analyse two sources of noise that are theoretically most relevant in our network: within-trial noise in the input and between-trial noise in recurrent gain. Within-trial noise is central to recent explanations of the scalar property in PA-models (Simen et al., 2013). More specifically, a constant CV is produced when within-trial noise in diffusion is scaled by the square root of the drift rate (which is a result of balancing excitatory and inhibitory inputs to the accumulator). However, some of these PA-models also assume that pacemaker speed is adjusted on a trial-to-trial basis (Simen et al., 2011a), which is corroborated by neurophysiological evidence (Wang et al., 2020). Surprisingly, this trial-to-trial variability does not feature explicitly in their explanation of the scalar property.

In order to examine the role of noise in interval production and perception, we take the mathematical implementation of our network and perturb the input and recurrent gain (θ^{-1}) with noise. Within individual trials, the constant input to the network is normally distributed. That is, the input is 0 before the start of the interval and $\mathcal{N}(\mu = 1, \sigma = \sigma_{\text{input}})$ for the duration of the interval, such that on each timestep of the simulation a random sample is taken from this normal distribution. We consider a scenario in which σ_{input} is constant, and a scenario in which σ_{input} scales with $\sqrt{\theta}$. On a between-trial level, a value drawn from a Normal distribution, $\mathcal{N}(\mu = 0, \sigma = \sigma_{\text{recurrent gain}})$, is added as a constant to θ^{-1} throughout each trial.

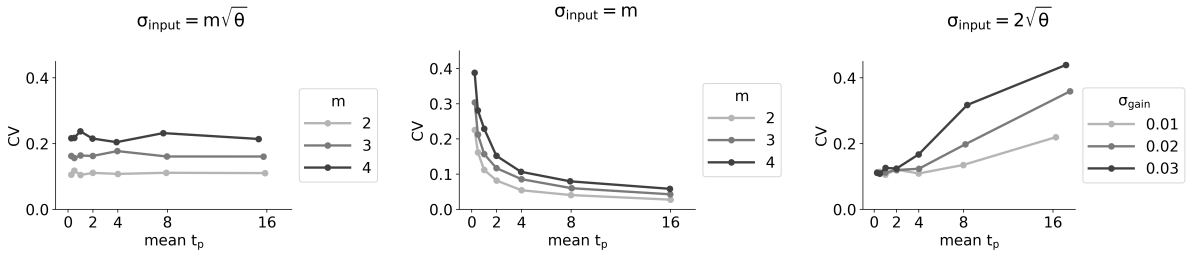
For producing intervals, we follow the rationale described above: different intervals are produced by adjusting the mean recurrent gain. We simulated 250 trials per target interval for different levels of within-trial noise in the input and between-trial variability in the recurrent gain. Our findings suggest that different sources of noise will produce different forms of scalar variability (Figure 7). In particular, we found a constant CV when σ_{input} scaled by $\sqrt{\theta}$. That is, the UTC model can explain adherence to the scalar property by assuming that noise in the input somehow scales with $\sqrt{\theta}$. When σ_{input} is constant, we observe a decreasing coefficient of variation across different levels of noise, clearly violating the scalar property. These observed patterns are in line with findings of decreasing CVs over time (e.g., Damsma, Schlichting, van Rijn, & Roseboom, 2021; Lewis & Miall, 2009). When between-trial noise in recurrent gain is added (while assuming $\sigma_{\text{input}} = m\sqrt{\theta}$), CV increases over the tested range of intervals. These findings may therefore explain some violations of the scalar property where the CV increases for longer target intervals (e.g., Bangert et al., 2011; Bizo et al., 2006; Getty, 1975; Gibbon et al., 1997; Grondin, 2014; Matthews & Grondin, 2012). In sum, input noise and between-trial recurrent gain variability can account for decreases and increases in CV, respectively.

While a constant CV is a clear sign of timescale invariance, there are more demonstrations. For example, the entire distribution of timing behavior often scales with the target time (for a review, see Wearden & Lejeune, 2008). This 'superimposition' property of timing behavior is most readily assessed by plotting responses on a relative timescale by dividing the response times by their mean (in the case of temporal (re)production). If timescale invariance holds, these distributions should overlap perfectly. Here, we model an experiment by (Simen et al., 2016) who found overlapping normalized response time distributions for a range of target times (2.2s, 5.1s and 11.3s). When we simulate response times (N=1.000 per duration) from the UTC model under the assumption that $\sigma_{input} = 2\sqrt{\theta}$, we found good superimposition of relative response times, with an overall CV of 0.2, matching behavioral data well (see Figure 8). This suggests that the UTC can approximate true timescale invariance. It should be noted, however, that other models have closed-form solutions that guarantee timescale invariance (e.g., TD-DMs, TILT, etc.), and as such they have an edge over models that only approximate it.

For perceiving time intervals, we follow the rationale outlined above. The mean window size is manually matched to the target interval, assuming that the target interval does not vary much from trial to trial. In effect, this modelling setup resembles classic psychophysical timing experiments, where the criterion interval only varies between blocks, not within blocks (Getty, 1975). This procedure ensures that the effects of the distribution of intervals within a block are minimized, resulting in a cleaner estimation of scalar variability. As we have seen before (Figure 6), the network representation at the end of the interval depends on window size. In other words, the network represents the duration of the interval relative to window size. Therefore, in order to generate a response that is only based on the perceived interval, we multiply the value represented by the network by the mean window size (θ) and the inverse of the fixed threshold ($\frac{1}{threshold}$) to obtain the estimated interval (t_s). This correction effectively rescales the 'relative' time in the window back to 'objective' time. For time perception, we obtain similar results to time production (Figure 9). When the only source of noise is in the input, the CV decreases, while trial-to-trial variability in the recurrent gain causes an increasing CV.

Figure 7

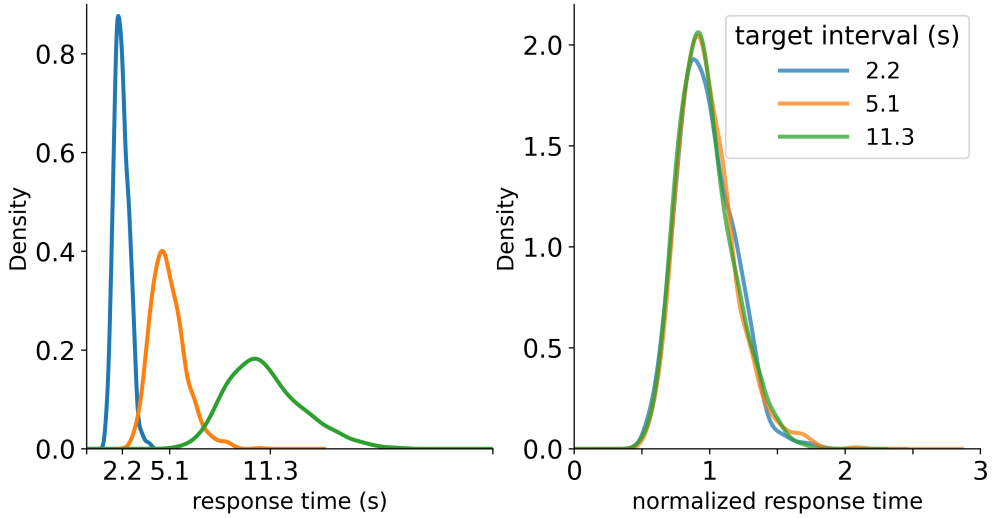
Coefficient of variation (CV) for time production under different assumptions about within- and between-trial noise



Note. Coefficient of variation is plotted across a range of produced target intervals. In the left panel, the within-trial noise in the input signal (σ_{input}) is assumed to scale with $\sqrt{\theta}$, with m determining the overall level of noise. This scaling produces approximately flat CV over mean produced interval, with m scaling the overall level of noise. In the middle panel, σ_{input} is assumed to be constant (m), which produces a decreasing CV. In the right panel, σ_{input} is assumed to scale with $2\sqrt{\theta}$, but between-trial noise in the recurrent gain (σ_{gain}) is varied. This produces an increasing CV over the target interval.

Figure 8

Superimposing normalized response time distributions.



Note. To demonstrate true timescale-invariance in the UTC model, we model the time production experiment by Simen et al. (2016). The model produces intervals of 2.2, 5.1 or 11.3 seconds. Here, we assume that σ_{input} scales with $2\sqrt{\theta}$, which produces a CV of around 0.2. In the left panel, the response time density functions center around the target interval and becomes progressively wider with target interval. In the right panel, response times are normalized by θ. The overlap between normalized response time distributions suggests timescale-invariance.

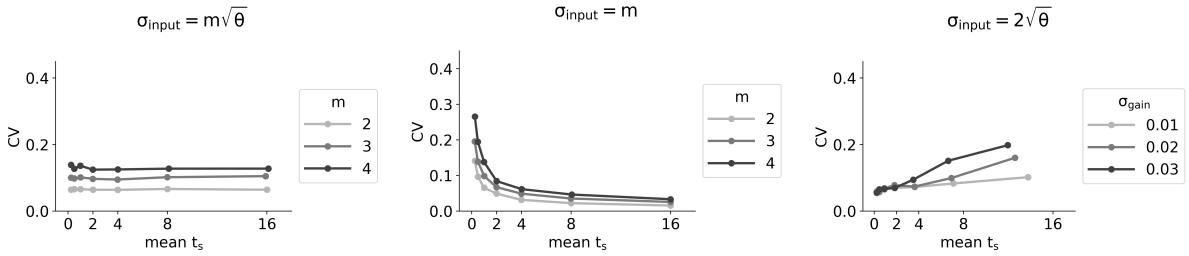
One-shot learning of window size explains rapid temporal learning

Humans and non-human animals can quickly adapt the timing of their behavior to changing temporal contingencies. Accurate timing of new target intervals can be accomplished in as little as one or two trials (Komura et al., 2001; Mello et al., 2015; Simen et al., 2011a). These findings put a lower bound on the learning rate that a model should exhibit. This is especially important for the UTC model, which assumes that, in order to accurately produce or perceive an interval, the window size (θ) is matched to the target interval. How can window size be learned so rapidly? We have taken inspiration from one-shot learning rules developed by TDDM models, which can adapt neural ramping speed to new intervals after a single exposure. We adapted these learning rules to the UTC, so that θ can be learned.

The learning rules consist of an 'early-timer rule' and a 'late-timer rule'. Consider the scenario where the model needs to respond as close as possible to a target interval (but not after it has already ended). When the UTC model responds too early, it should increase θ, so that on the

Figure 9

Coefficient of variation (CV) for time perception under different assumptions about within- and between-trial noise



Note. Coefficient of variation is plotted across a range of perceived target intervals. Similar results to time production are obtained for different assumptions about noise.

next trial, it will respond later. The 'early-timer rule' employed by TDDMs specify how the rate of neural integration (which is akin to θ^{-1} in the UTC model) should be decreased in real-time, starting from the response until the end of the interval. It turns out that this learning rule can be applied to the UTC with little modification to explain rapid learning of longer θ (see Figure 10). From the moment the model responds until the end of the interval, the recurrent gain is decreased θ^{-1} at a rate of $(\theta^{-1})^2$. The 'late-timer rule' details how θ should be decreased when the model responds too late. When the model responds too late, θ needs to be decreased by the relative distance that still needs to be traversed by x until the threshold. Again, this decrease in θ can be implemented through the 'late-timer rule' employed by TDDMs, without much modification (although it should be noted that the accuracy of the 'late-timer rule' depends on d). In effect, the recurrent gain (θ^{-1}) needs to be increased by $\theta^{-1} * \frac{threshold-x}{x}$. Intuitively, when x is at the threshold exactly when the interval ends, there is no update. If x is only halfway there, the θ^{-1} should be doubled. The two learning rules work in concert to rapidly adapt θ to accurately produce intervals (see trial-by-trial Figure 11). In turn, learning the window size in UTC mirrors adaptive 'temporal scaling' of neural responses to target intervals, which we discuss next.

Changes in window size explain temporal scaling in complex neural patterns

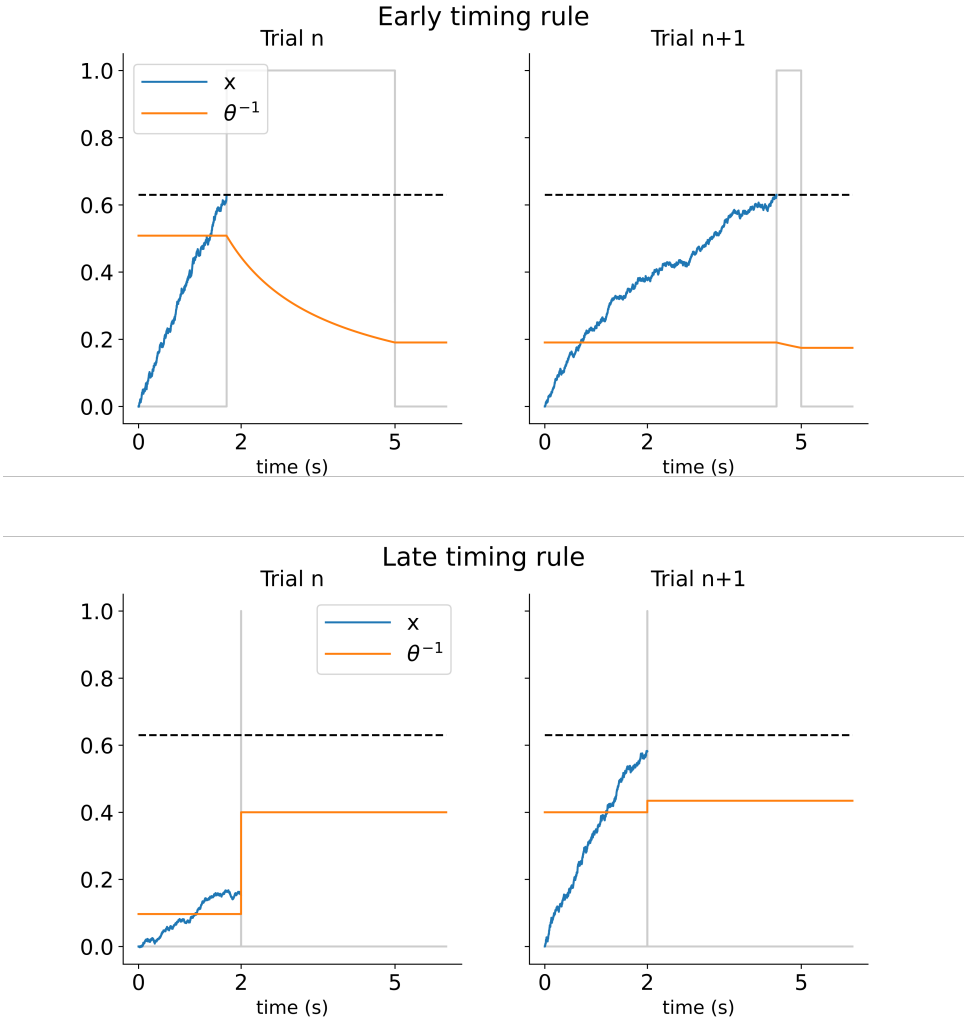
As discussed earlier, the neural firing patterns during timing performance are diverse (e.g., ramping, decaying, time-cell activity). Further, these same complex responses compress and stretch as shorter and longer intervals are timed, respectively. How does our network explain both of these features? First, instead of having a single node that represents a dimension in the network, each spiking neuron in our recurrent network encodes a particular combination of dimensions (see Appendix B. For instance, a single neuron may be sensitive to positive values of the first dimension (meaning an increase in firing rate when the mean of the signal in the temporal window becomes more positive), while at the same time being sensitive to negative values of the second dimension (meaning a decrease in firing rate when the slope of the signal in the temporal window becomes more negative). This heterogeneity of tuning in our spiking neurons systematically captures heterogeneity observed in electrophysiological experiments. A separate population of neurons controls the window size, where individual neurons may encode increases or decreases in window size, therefore speeding up or slowing down neural dynamics. In this section, we demonstrate that the network can jointly capture behavioral and neural data from a temporal production task (Wang et al., 2018) and a perceptual timing task (Gouvêa et al., 2015).

Temporal Production (Wang et al., 2018). In this section, we model a study by Wang et al. (2018) who found that during a temporal production task neural firing patterns in several brain areas are highly heterogeneous and whose activity exhibited temporal scaling. Wang et al. (2018) recorded single-cell activity from multiple brain areas that are believed to be crucial for temporal production: the medial frontal cortex (MFC), caudate (striatum) and thalamic neurons that projected to MFC. In this task, monkeys were presented with a colored cue at the start of each trial, indicating whether they had to produce an interval of 800ms (red) or 1500ms (blue). Then, after some delay, a 'Set' stimulus was presented, marking the onset of the target interval. Monkeys were required to produce a motor response after the cued interval had passed and were rewarded if their produced interval was close enough to the target. Wang and colleagues found that neural firing patterns during the interval were highly heterogeneous, containing neurons with ramping, decaying, oscillating and more complex temporal profiles.

The authors systematically assessed several classic model alternatives, such as oscillatory models, ramping activity with a flexible threshold, flexible ramping speed or both a flexible threshold and speed, however, these models were unable to capture the heterogeneity of the observed neural responses. In contrast, the best-fitting model was one in

Figure 10

Early- and late timing rules employed by the UTC model



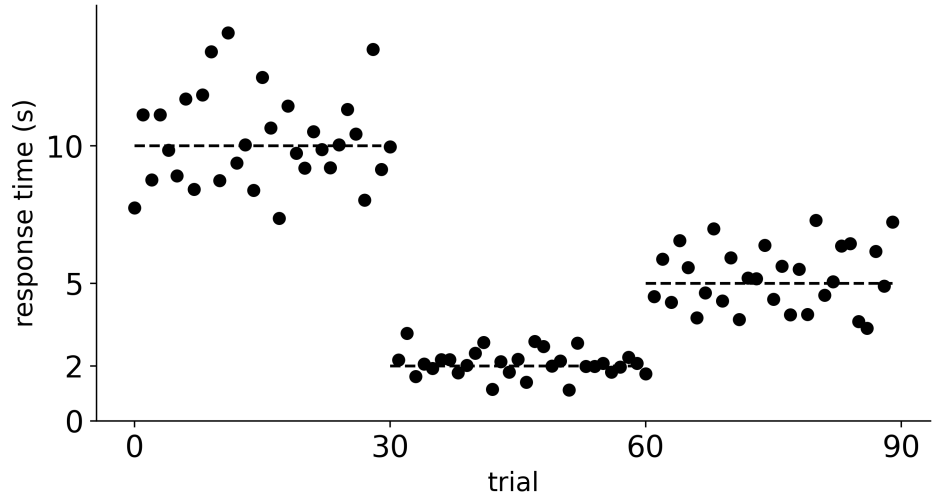
Note. In the top panel, the UTC model produces an interval of approximately 2s on trial n, while the actual target interval is 5s. The early timing rule decreases θ^{-1} continually from the moment of responding (when x crosses the threshold) until the target interval ends. On trial $n+1$, the window size (θ) more closely matches the target interval of 5 seconds (i.e., $\theta^{-1} \approx \frac{1}{5}$). In the bottom panel, the UTC model has not reached the response threshold yet when the 2s interval is already over on trial n. The late timer rule increases θ^{-1}, so that on trial $n+1$, the window size (θ) more closely matches the target interval of 2 seconds (i.e., $\theta^{-1} \approx \frac{1}{2}$).

which the firing patterns were fit with a single polynomial for each neuron that stretched or compressed along the time-axis depending on the length of the produced interval. The degree of scaling on individual trials was highly predictive of behavior. Interestingly, neural responses in thalamus exhibited significantly less temporal scaling than MFC or caudate. Instead, a major portion of the variance in thalamic activity was explained by a component where activity remained constant throughout the interval, but the mean activity scaled with the produced interval. These findings imply that the thalamus may be involved in controlling the speed of neural dynamics in brain areas MFC and caudate. Wang and colleagues were able to model this division of labour with a recurrent neural network, where the recurrent units received a constant input throughout the interval that scaled with the desired interval. This constant input, which is thought to reflect thalamic input to MFC, effectively controlled the speed of the neural dynamics, which in turn allowed the network to produce the desired intervals.

The UTC model resembles the speed-control mechanism uncovered by Wang et al. (2018). We modelled their experiment by setting up the spiking implementation of the LDN (d=3, N=600). As described before, we assumed that the network receives a constant input throughout the pro-

Figure 11

One-shot learning rules allow the UTC model to rapidly learn new target intervals



Note. We initialize the UTC model to produce 10s intervals by setting $\theta = 10$. Then, the target interval (dashed line) changes to 2s and the model rapidly adapts to this shorter target interval through the late timer rule. After several trials, the target interval increases to 5s. The early timer rule allows the model to rapidly adapt θ to match this new target interval.

duced interval. The interval is terminated when the readout of the network reaches a certain threshold. This allowed us to produce target intervals (T) of 0.8s or 1.5s by setting the temporal window to those values. In other words, we adjusted the speed ($\frac{1}{\theta}$) to produce different intervals. In the UTC network architecture, this is achieved by jointly multiplying the input and recurrent activity by $\frac{1}{\theta}$ so as to control the rate of encoding and forgetting, respectively. The speed was normally distributed across trials $\mathcal{N}(\mu = \frac{1}{\theta}, \sigma = 0.1)$. This naturally produced more variable response times for the longer intervals ($\sigma_{0.8}$ = 64ms, $\sigma_{1.5}$ = 164ms), mirroring behavioral results from Wang et al. (2018).

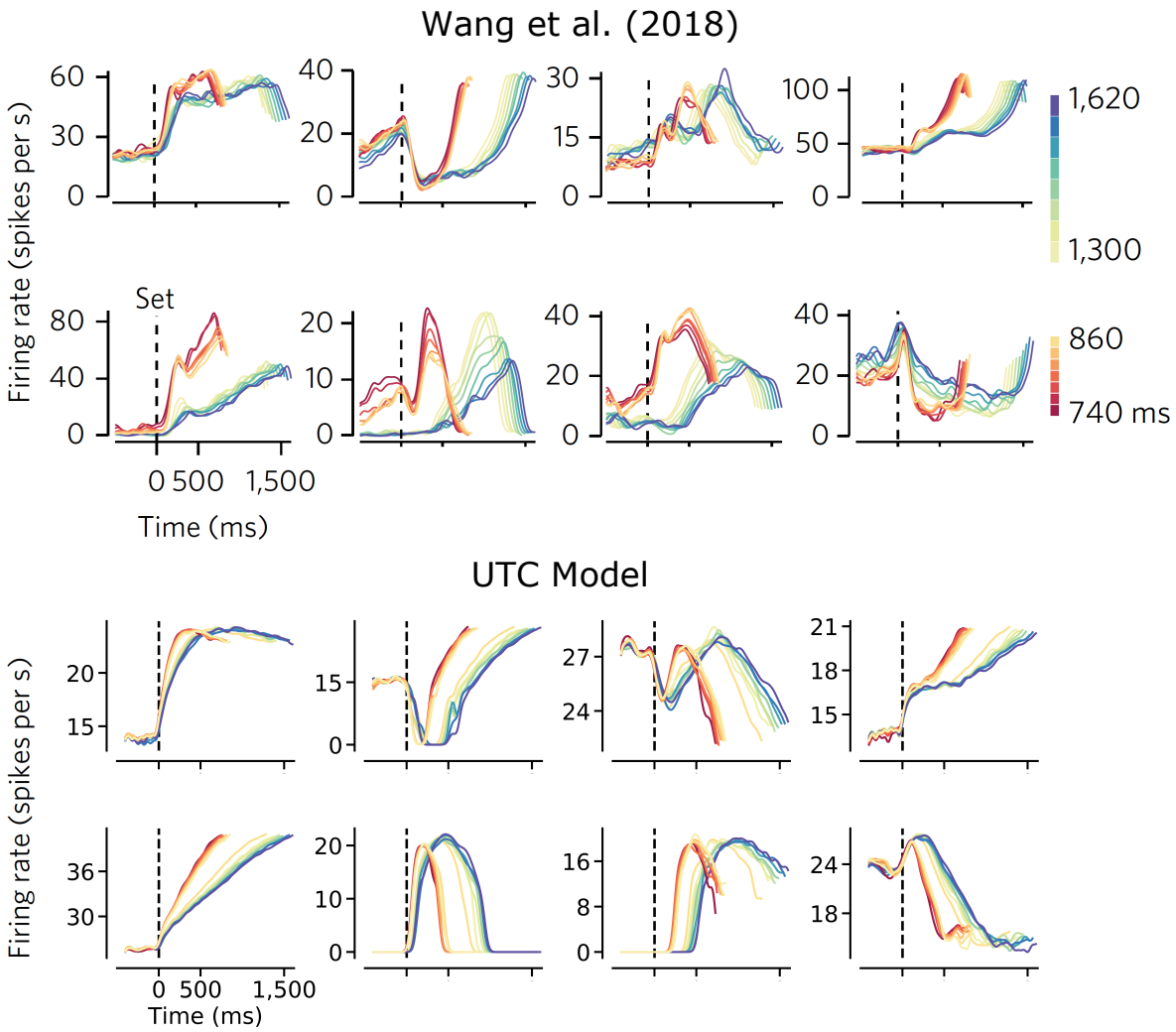
We simulated the model 50 trials for each target interval. In Figure 12, we show the firing patterns of several representative neurons, which are highly heterogeneous, similar to those found by Wang et al. (2018). This heterogeneity is due to individual neurons being sensitive to different combinations of the underlying state vector **x**. We performed no hand-tuning of neurons to generate these responses. Instead, we randomly chose combinations of dimensions over the unit hypersphere, which allows each point in the state vector to be equally likely to be represented by the neurons (see Appendix B). As a result, some neurons exhibit upward ramping, since they are mainly sensitive to positive changes in the first dimension. Other neurons have 'bell-shaped' firing patterns, since they are predominantly sensitive to negative values in the second dimension. Further, when the trials are binned according to response time, the neural firing patterns exhibit scaling along the time axis: firing patterns are 'stretched' for longer intervals and 'compressed' for shorter ones. In sum, adjusting window size accounts for both behavior and the temporal scaling of heterogeneous response profiles in MFC and caudate.

The connection between the polynomials used by Wang et al. (2018) and the Legendre polynomials in our network is readily made. In particular, the Legendre basis *is* a polynomial basis; one which is optimal for minimizing representational error. The main difference is that the UTC model provides a process account of how a polynomial basis is continuously updated to represent time, instead of fitting a polynomial basis to observed neural data. One consequence of this is that our mechanistic account of heterogeneity suggests a general way to model neural timing data, which can in turn inform the optimal dimensionality of our network for a given task. Future work should quantify how well the UTC model can account for the heterogeneity of neural patterns observed in timing tasks, from simple ramping neurons to complex oscillatory responses.

Time Perception (Gouvêa et al., 2015). In this section, we model an experiment by Gouvêa et al. (2015), who found that the speed of neural dynamics in the dorsal striatum explained sensory interval timing on a trial-to-trial basis. Rats were trained to perform a duration categorization task: they had to judge whether auditory intervals were longer or shorter than the mean interval of 1.5 seconds. The dorsal striatum was found to be crucial for timing performance, since performance dropped significantly when it was pharmacologically inactivated. Individual neurons in the striatum were sensitive to different intervals: during the presentation of the longest interval, some neurons decreased

Figure 12

The UTC captures heterogeneity and temporal scaling of neural responses as observed in Wang et al. (2018)



Note. The UTC model qualitatively fits the heterogeneity found in MFC neural responses, with ramping neurons, decaying neurons, oscillating neurons, and neurons with activity bumps. Further, these responses scaled along the temporal axis according to the produced interval, where red hues represent short intervals and blue hues represent long intervals. Adapted from Wang et al. (2018) https://creativecommons.org/licenses/by/4.0/.

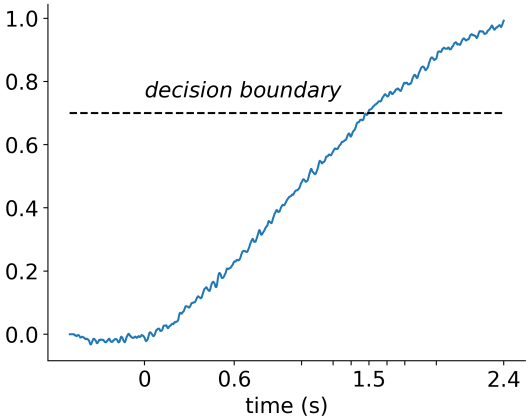
their firing rates over time, some had peak firing rates somewhere in the middle of the interval, while others increased their firing rates over time. Crucially, Gouvêa et al. (2015) found that the speed at which the neural firing rates changed predicted behavior. A Principal Component Analysis (PCA) revealed a sub-space that explained most of the variance, containing a ramping and a bell-shaped component. When the neural state evolved slower through this space, rats were more likely to classify the interval as short (i.e., as if less time had passed), and when the neural state evolved quicker, rats responded 'long' more often (i.e., as if more time had passed).

We modelled the experiment by Gouvêa et al. (2015) with the spiking implementation of the LDN (d=3, N=600). We presented the network with a constant input for the duration of the sample interval. We assumed that the mean window size matched the maximum of the presented sample intervals, which was 2.4 seconds. This ensures that the network represents an overall accurate representation of elapsed time accross sample intervals. We generate behav-

ioral responses according to this readout at the end of the interval: if it was lower than 0.7 (which we set to match behavioral data), the model response was 'shorter', else the model responded 'longer'. Across trials, we assumed that the recurrent gain was distributed according to a normal distribution $\mathcal{N}(\mu = 11.5, \sigma = 0.15)$, qualitatively recreating the variability in 'neural speed' found in the empirical data. First, we found that the model was able to perform the task well, resembling the performance of the rats (Figure 14). Then, we visualize the normalized neural activity for the longest interval in a heatmap (Figure 15), where neurons are sorted according to their peak times. Individual neurons peaked during the start, middle and end of the interval, providing a fair match to the firing patterns in dorsal striatum found by Gouvêa et al. (2015). This time-cell activity by the LDN was first shown by Voelker and Eliasmith (2018) and is similar to time-cell activity found in hippocampus (Eichenbaum, 2014), entorhinal cortex (e.g., Heys & Dombeck, 2018), prefrontal cortex (e.g., Tiganj et al., 2017) and striatum (e.g., Mello et al., 2015).

Figure 13

Modelling interval categorization in Gouvêa et al. (2015)

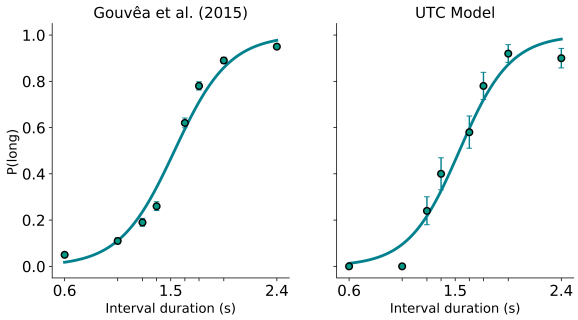


Note. On the *y*-axis, the first dimension of the LDN network, on the *x*-axis the time since interval onset. The decision boundary determines whether the interval is categorized as shorter (if lower than the decision boundary), or longer (if higher than the decision boundary).

Finally, following the analysis in Gouvêa et al. (2015), we performed a PCA on the simulated neural data for the 1.62s interval. In line with previous interval timing studies, the first component was a 'ramping' component and the second component was 'bell-shaped' (e.g., Emmons et al., 2017; Wang et al., 2018). Importantly, when the data was split on response, we observed that 'short' responses were associated with a slower trajectory through this PCA space, whereas 'long' responses were characterized by a faster trajectory. In other words, when the 'neural clock' moved faster, the UTC model estimated that more time had passed, similarly to the rats in Gouvêa et al. (2015). These results suggest that variability in 'neural speed' in our neural network is sufficient to qualitatively account for both behavioral and neural data in a time perception task.

Figure 14

Behavioral performance on the interval categorization task



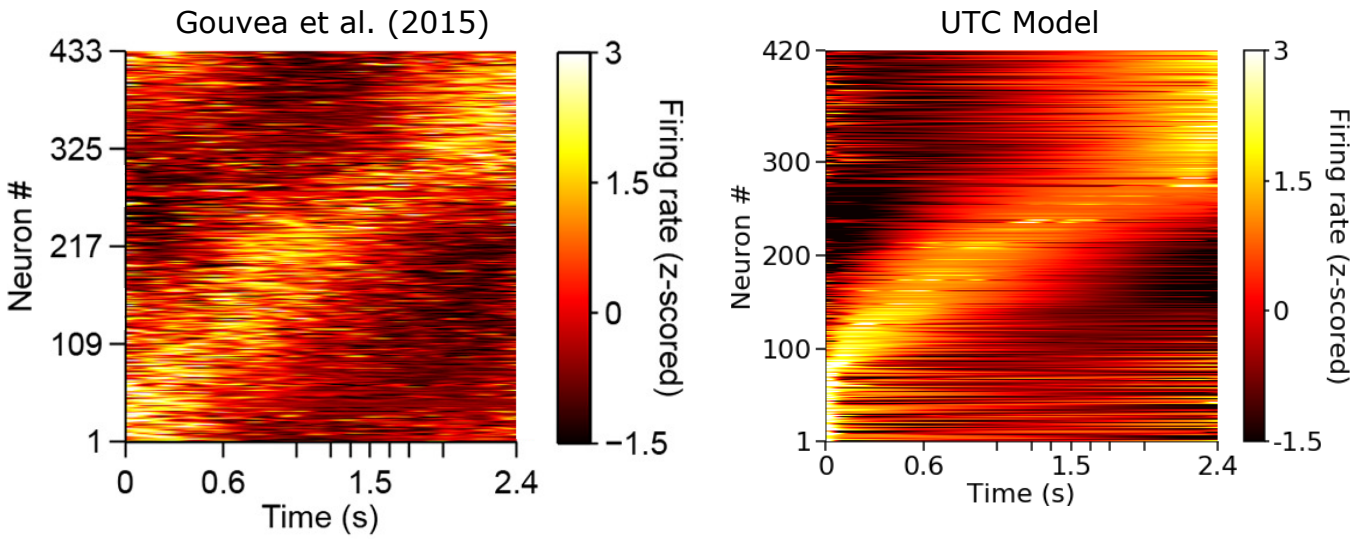
Note. On the *y*-axis, probability of long responses, on the *x*-axis the interval duration. Logistic regressions were fit to the data.

Forgetting of timing information accounts for the effect of interruptions

A prominent reason why the 'internal clock' metaphor is so appealing is that subjects seem to be able to start, stop, and reset their internal clock. This ability is most obvious from procedures where the timed signal is interrupted by gaps or distractors (for a review, see Buhusi & Meck, 2009a). In peak-interval procedures, subjects learn to respond when the timing signal (e.g., a light or a sound) has been on for a certain amount of time. When a gap or distractor is inserted into the timed signal, three patterns of behavior can be predicted from the perspective of an internal clock (Roberts & Church, 1978). First, subjects may not delay responding, as if 'running' their clock throughout the interrupting event. Second, subjects may delay their response by the duration of the interrupting event, as if they stopped or paused their internal clock and resumed timing after the event. Third, subjects may delay their response by the sum of the pre-event interval and the duration of the interrupting event, as if completely resetting accumulated time. However, these three patterns of behavior are not discrete possibilities but seem to exist on a run-stop-reset continuum (Buhusi & Meck, 2009a). When properties of the interrupting event are parametrically varied, such as the onset, duration or similarity to the timed signal, the delay in responding also varies continuously (Buhusi et al., 2006). This has prompted theoretical accounts to consider a more continuous mechanism that includes running, stopping and resetting behavior as special cases. A natural candidate for such a mechanism is memory decay (Cabeza de Vaca et al., 1994). Indeed, 'run' patterns can be observed when decay is much less than the rate

Figure 15

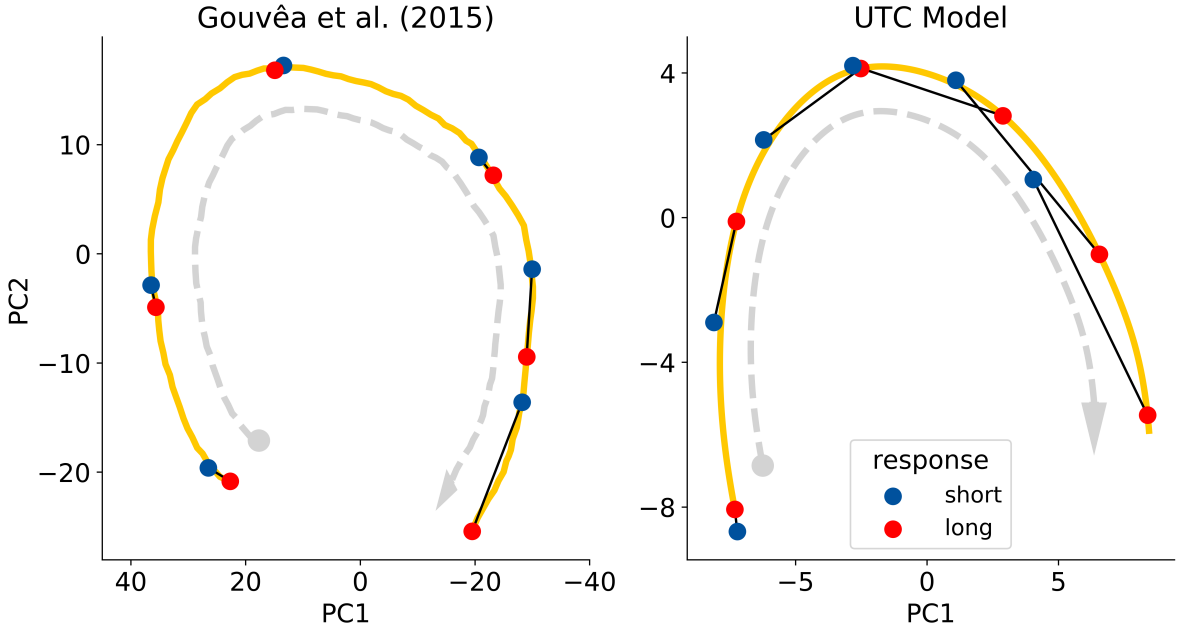
The UTC model captures neural dynamics of rodents performing a time perception task (Gouvêa et al., 2015)



Note. Individual neurons are plotted on the y-axis, time is represented on the x-axis, and color represents normalized firing rate. Individual neurons are sorted according to when their firing rate peaks. Neurons that fire early in the interval are plotted in the lower parts and neurons that fire later are plotted in the upper part.

Figure 16

Principal Component Analysis on simulated neural data for the 1.62 (middle) interval



Note. The first principal component (PC1; *x*-axis) shows a ramping profile over time, whereas the second principal component (PC2; *y*-axis) shows a bell-shaped profile. Neural data was split between 'short' and 'long' responses and their neural trajectories were projected onto the common PCs. The dots (red for 'short', blue for 'long') are evenly spaced time points between interval onset and offset. Connected dots represent the same intermediate time point. The trajectory moves faster for 'long' estimates (red) than for 'short' estimates (blue).

of ongoing accumulation, 'stop' patterns may be observed when decay and accumulation cancel out perfectly and 'reset' patterns may be observed when decay is much larger than accumulation (Buhusi & Meck, 2009a).

The UTC model has a natural connection to previous models that assume a decay of accumulated time during interrupting events. We model gap and distractor procedures with 2-dimensional LDNs (which provided the best fit to all modelled datasets). The speed parameter (θ^{-1}) in our network controls how quickly information is accumulated, but also how quickly it decays. In order to model how the stimulus content of the timed signal and the distractors control timing, we provide the network with a unit vector $\boldsymbol{t}$ that serves as the timed input. In order to read out the timing signal, we compute the similarity (i.e., dot product) between the vector $\boldsymbol{t}$ and the state of the network $\mathbf{X}$. This ensures similar behavior to integrating a one-dimensional step input. When a gap or distractor is introduced we assume that the similarity between the timed signal and the distractor can vary between approximately 0 and 1. For gaps, we assume that the gap has a similarity of 0 to the timed signal. That is, from the start to the end of the gap, the network receives no input (i.e., a vector that consists of zeroes) . When a distractor, which is also a D-dimensional vector, occurs, we provide it as an input to the network instead of the timing vector $\mathbf{t}$. Therefore, a distractor that is highly similar to the timed signal only has a small influence on accumulation and decay, since the driving input is highly similar to the timed signal (similarity close to 1). In contrast, a highly dissimilar distractor will have a large influence on accumulation and decay, since the driving input is highly dissimilar from the timed input (similarity close to 0).

For the fit to Buhusi (2012) we needed to make some assumptions about how sound intensity is presented as a vector. First, we assumed a power-law representation of stimulus intensity (Stevens, 1956) where the exponent was taken from an empirical study on sound intensity discrimination in rats (Pardo-Vazquez et al., 2019) and one scaling parameter (k) that we fitted for each experiment. This scaling parameter may capture differences in experimental setup that influence the magnitude of distractor effects [8]. We then converted stimulus space to a vector representation that can effectively deal with continuous quantities (Komer et al., 2019). For a more detailed description, see Appendix **??**.

Timing with Gaps

A natural consequence of forgetting mechanisms during the gap is that the timing of gaps has a large influence on timing behavior. These effects were systematically investigated by Cabeza de Vaca et al. (1994), who found that depending on the duration, onset, and offset of the gap, timing behavior varied between stopping and resetting. When we parametrically vary the duration, onset and offset of the gap in our model simulations, the delays are on a continuum between stopping and resetting, and provide a good quantitative fit to the empirical data (Figure 17, see Experiment 2 in Cabeza de Vaca et al. (1994)). When the onset or offset is fixed, but the duration is varied, we can see that as the duration of the gap increases, the accumulated time decays more, and the behavior tends to resemble a full reset. When only the location of gaps is varied, we can see that for later gaps, the accumulated time decays more strongly, since the accumulated time is larger at gap onset, resulting in a linear increase in peak shift.

Timing with Distractors

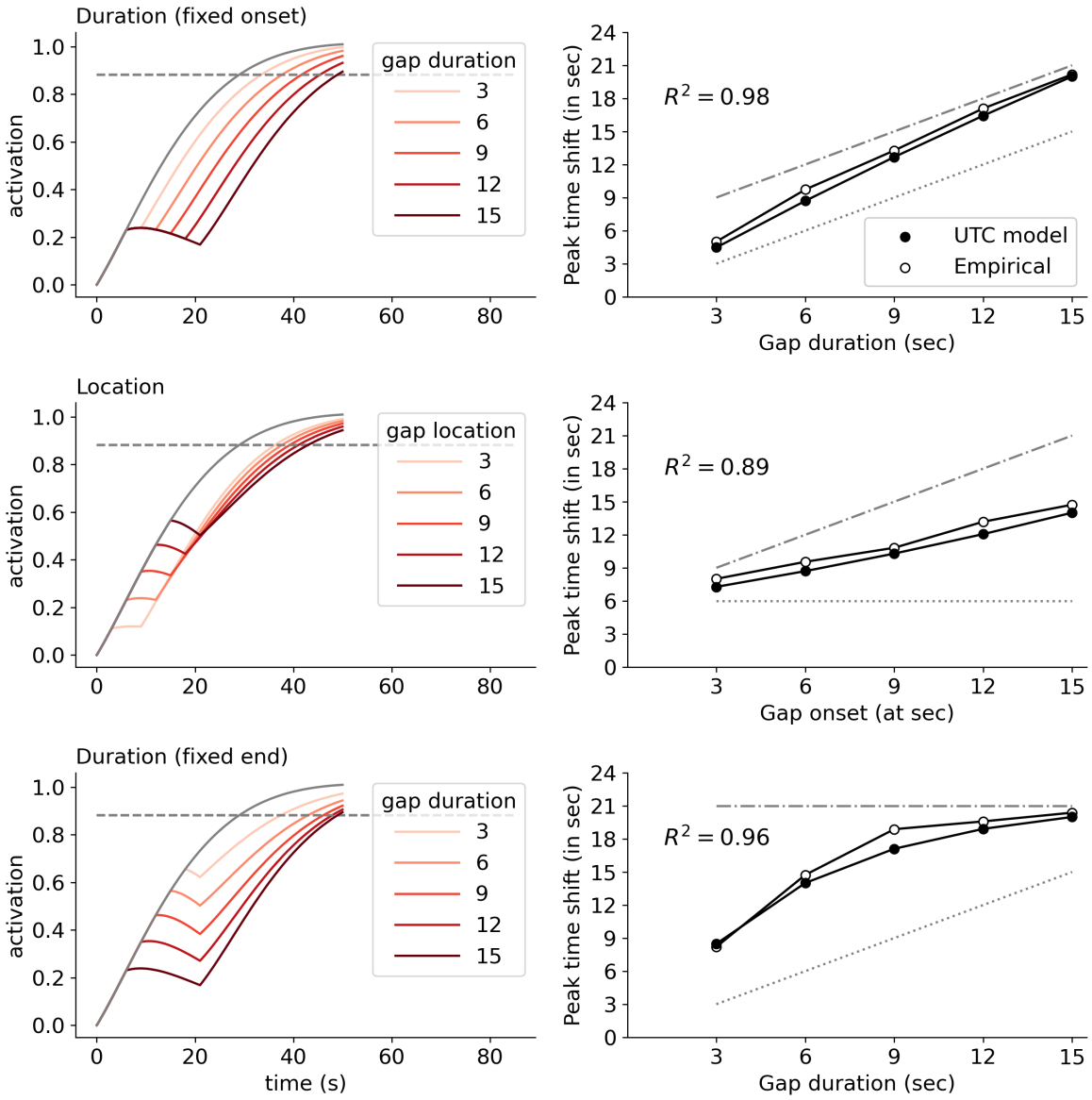
A similar explanation can be given for the effect of distractors on timing. The similarity of the distractor to the timed signal controls the magnitude of the delay in responding, such that distractors that are highly dissimilar to the timed signal produce large delays in responding (Buhusi, 2012). In Experiment 1 of Buhusi (2012), rats were trained to respond after a visual stimulus has been on for 30 seconds, while a 40 dB auditory stimulus was continuously presented in the background. This way, rats learned that the visual stimulus in addition to the auditory background noise was the signal to be timed. Then, distractors were introduced by increasing the loudness of the background noise after 20 seconds. Distractors that were similar to the 40 dB background noise (e.g., 55 dB) only had relatively small effects on behavior. However, as the distractor intensity increased (i.e., the similarity between the distractor and the signal decreased), the delay in responding also increased (Figure 18). Was this delay in responding only due to the absolute intensity of the distractor or did the similarity to the timed signal determine behavior? In Experiment 2, the 'similarity hypothesis' was tested more directly. Rats performed the same peak-interval task, but now the timed signal was a visual signal alongside a 70 dB white noise signal. In the inter-trial-intervals, a 40 dB white noise background was presented. Crucially, distractors could either be more or less loud than the 70 dB timed signal. Rats delayed their responses according to the similarity between the distractor and the 70 dB signal, regardless of whether distractors were more or less loud (Figure 18).

Buhusi (2012) fitted a resource-allocation model to the data. This model assumes that processing of the distractor and keeping track of time tap into the same limited pool of working memory resources. When more resources are spent for distractor processing, there are fewer resources left for timing. More specifically, Buhusi (2012) which assumed that the rate of was proportional to the similarity between the background noise and the distractor, obtaining a good fit in both experiments. The UTC model makes assumptions

[8] In the case of Buhusi (2012), the intensity of background noise in the inter-trial-interval varied, and may account for differences in the free scaling parameter k

Figure 17

Timing with gaps (Cabeza de Vaca et al., 1994): activity traces (left) and fit to empirical data (right)



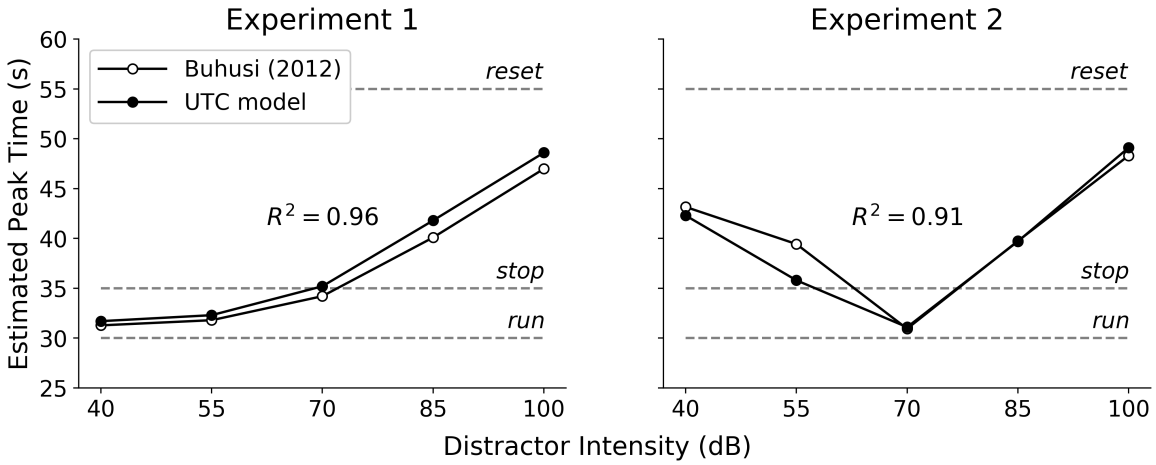
Note. Left: when the activity trace crosses the threshold (dashed line), we expect the peak in responding. The grey line represents baseline peak-interval trials without a gap. Right: when duration (top and bottom) and location (middle) are increased, the responses are increasingly delayed. The dotted line represents peak shifts expected for a 'stopping' pattern, dot-dashed line represents expected peak shifts for a 'reset' pattern.

similar to the resource-allocation account (see **??** for details) and obtains a reasonable quantitative fit to the data for both experiments (Figure 18).

The UTC model makes similar assumptions compared to the resource-allocation model by Buhusi (2012) about the relationship between distractor similarity and rate of decay. However, the UTC model proposes a different underlying mechanism. The resource-allocation model assumes that a limited resource is shared between working memory for time and working memory for other cognitive processes, implying that these working memory stores are, at least to some degree, functionally encapsulated (Buhusi & Meck, 2009a). In contrast, the UTC model assumes that both temporal and stimulus information is represented by the same neural population. Therefore, any effect of distractor similarity is not due to allocating resources to separate work-

Figure 18

Timing with distractors, fit to Buhusi (2012) experiment 1 (left; 40dB signal) and experiment 2 (right; 70 dB signal)



Note. As distractor similarity decreases peak times increase.

ing memory stores, but rather due to resource competition within a single working memory store.

One may argue that the effects of distractor similarity follow naturally from our account of vector representations within a single neural population, while a resource-allocation account would have to make additional assumptions about the role of distractor similarity. Unfortunately, behavioral data alone can not arbitrate between these theoretical possibilities. A more informative test may come from neural recordings. For instance, the UTC model strongly predicts that individual neurons (e.g., in the prefrontal cortex) are sensitive to both the timed signal and the distractor. Further, measures of 'neural similarity' in sensory areas should directly map onto the 'rate of neural decay' during distractor presentation. These hypotheses remain to be tested empirically.

The interdependence of integration and decay

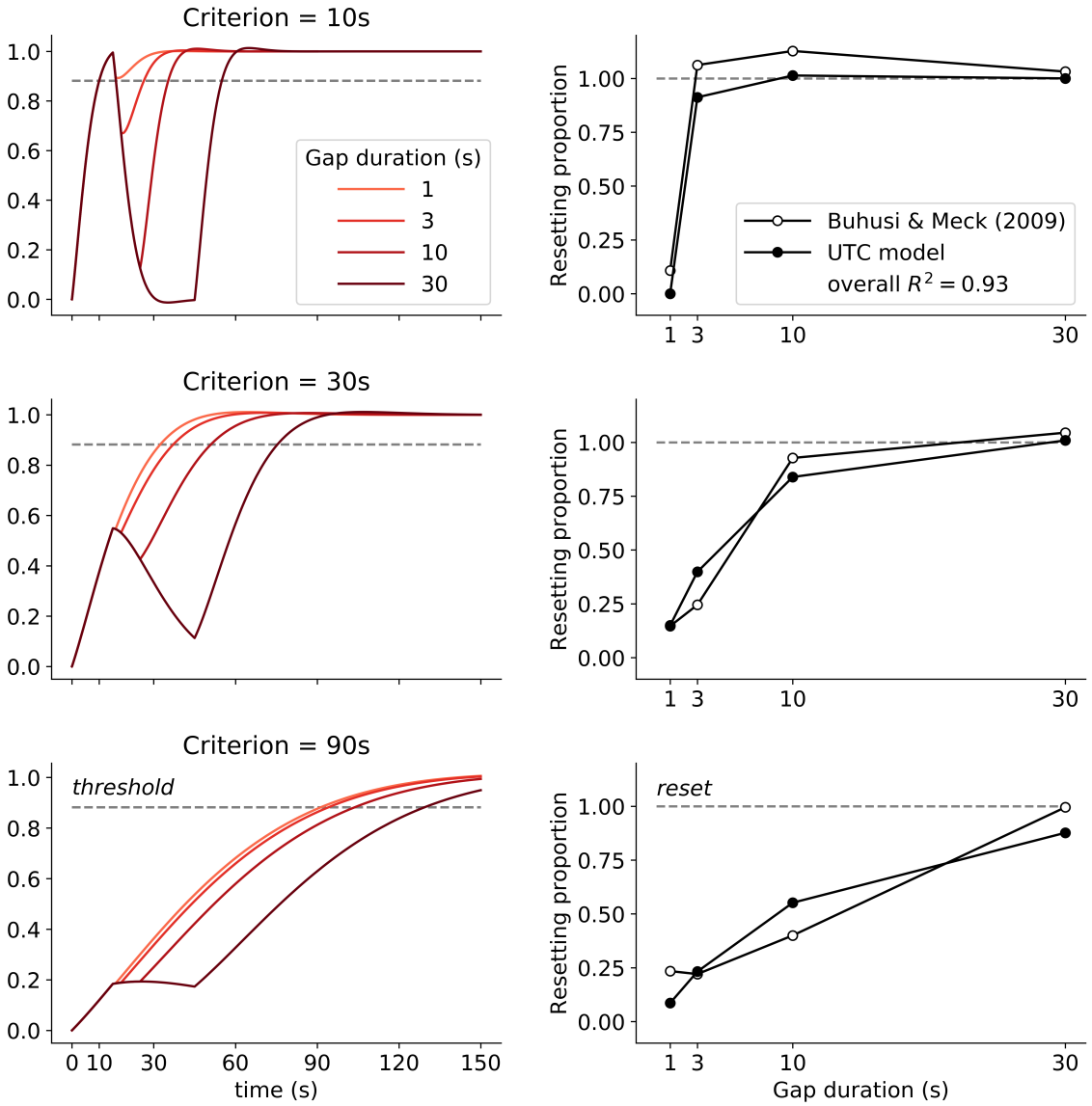
So far, the UTC model has been tested for scenarios that have also been accounted for by existing models. Indeed, Cabeza de Vaca et al. (1994) and Buhusi (2012) modeled their results successfully with memory decay mechanisms (Hopson, 1999, also see,). A strong prediction of the UTC model is that rate of integration and rate of decay are directly related through θ. When the rate of integration is high, the rate of decay is also high. Another crucial assumption of the model is that different intervals are timed by adapting θ to the appropriate timescale. Therefore, data from gap procedures where the target interval is varied could easily falsify these assumptions: Gaps should have a larger effect when shorter intervals are timed.

This question was directly addressed in an experiment by Buhusi and Meck (2009b). Rats were trained on a tri-peak procedure, where three different response levers were associated with 10, 30 and 90-second criteria. When the timed signal was presented, rats learned to respond after 10 seconds for the first, after 30 seconds for the second and after 90 seconds for the third lever. The authors observed that response times did not correlate between different levers, suggesting that different 'internal clocks' were running independently. Gaps of different durations (1, 3, 10 and 30 seconds) were introduced 15 seconds after stimulus onset. Crucially, the delay in responding depended on the length of the criterion. When the criterion was short (10 seconds), it appeared as if rats already reset their clock for relatively short gap durations. For longer criteria (30s and 90s), longer gaps were required for a full reset. The UTC model provides a good quantitative fit to the data, reproducing the finding that gaps have larger effects for shorter target times (Figure 19).

Interestingly, Buhusi and Meck (2009b) fitted the resource-allocation model to the data, using three internal clocks that ran at the same speed but had different response thresholds. To account for the findings, they made the additional assumption that each clock had separate resources that were reallocated during the gap, where the salience of the gap was proportional to the criterion time of each individual clock. Note that none of these assumptions is 'essential' to the resource-allocation model. Salience could be independent of criterion time and the clocks could run at different speeds without violating any of the core assumptions of the resource-allocation model. In contrast, the validity of the UTC model is highly constrained by the outcomes of this experiment. If it turned out that the effect of the gap was constant for different criteria (i.e., different speeds), at least one of our core model assumptions would be completely mistaken.

Figure 19

Model fit to Buhusi and Meck (2009b)



Note. Left panels depict activity traces for different gap durations and criteria. For the shortest criterion, the trace initially crosses the threshold, but as the duration of the gap is increased, the activity decays relatively quickly. When the criterion is longer, short gaps have less of an effect, since both integration and decay are slower. Right panels show the degree of resetting (as a percentage of a full reset) for different gap durations (x-axis) and criteria (hues of blue). Gaps have larger effects on shorter criteria.

Additionally, the UTC model suggests an alternative interpretation of some pharmacological and neurological effects on performance in gap procedures. For instance, the effect of dopamine agonists (e.g., methamphetamine) has been traditionally interpreted as independently increasing internal clock speed and impeding working memory (and therefore magnifying the effect of gaps). Conversely, dopamine antagonists (e.g., haloperidol) tend to decrease clock speed and attenuate the effect of gaps (Buhusi, 2003). Our UTC model, on the other hand, suggests that, all else being equal, any manipulation that increases the rate of integration will also increase the rate of decay. As a complementary example, lesions of the hippocampal system typically produce leftward shifts in responding in peak-interval procedures and larger resets in gap procedures (for a review, see Meck et al., 2013). Traditional accounts of these effects suggest that hippocampal lesions independently affect clock speed and working memory for temporal information (Meck

et al., 1984). Conversely, the UTC model predicts that horizontal shifts in timing functions - whether they are experimentally induced or reflect accurate timing - are systematically related to working memory for time.

The UTC also suggests how to model other phenomena in which 'working memory for time' plays a central role. For instance, systematic over- and underestimation have been found to depend on retention interval (Spetch & Wilkie, 1983; Wearden & Ferrara, 1993; Wearden et al., 2007; Wearden et al., 2002), order of sample and test stimulus (Bausenhart et al., 2016). The additional dimensions in the LDN network allow for maintenance of time intervals relatively accurately for at least the the interval between stimulus onset and the window size. By assuming that duration is both processed and stored through the same principles, the UTC model may be able to generate a constrained account of 'working memory for time'. More generally, it has been proposed that working memory for time taps into the same resources as other working memory functions, a phenomenon we turn to now.

Neural normalization explains effects of working memory load

Effects of interruptions already suggest that timing performance taps into a limited resource. Indeed, the UTC model assumes that both 'timing' and 'stimulus' information are represented by a common neural population, which allowed us to model the effects of distractors that were similar to the 'timing' input. More specifically, the UTC model assumes that both temporal and stimulus information are represented as vectors (see Figure 4 for the network architecture). Temporal and stimulus inputs are combined in a central input population ($\mathbf{x}_{\mathbf{input}}$) by adding the vectors together. Crucially, adding the vectors results in interference: when the temporal information is decoded from the network, the added 'noise' from stimulus information will ensure imperfect decoding. As a consequence, temporal information is integrated at a slower rate than would result from perfect decoding, explaining why prospective time estimates decrease in dual-task conditions, similar to the previously discussed gap and distractor paradigms.

An intuitive way to understand interference in the UTC model is to consider making a shopping list. In principle, we could have a shopping list with one entry per item. When we have many items, we just make a large list. However, if we only have limited resources (e.g., a small sheet of paper), a trade-off presents itself. We could write down some items, but once the next item does not fit on the list, we stop writing. Unfortunately, we would lose the rest of the items. Alternatively, we could write each item using smaller letters, but at the cost of the legibility of each item. In the case of our network, interference works like trying to write multiple items (vectors) on a limited piece of paper (neural population).

To illustrate how this interference works, consider a network that represents the sum of two orthogonal 2-dimensional vectors, $\mathbf{s}$ for *stimulus* and $\mathbf{t}$ for *time* (Figure 20). In the mathematical implementation of this network, no interference takes place: we can perfectly decode both $\mathbf{s}$ and $\mathbf{t}$ from the network that represents their sum. Note however, that $\mathbf{s}+\mathbf{t}$ is not unit length: the vector is not normalized [9]. Crucially, when we implement vector addition with a spiking neural network, a soft form of normalization takes place due to neural saturation. The tuning curves of our spiking neurons are such that firing rate is a decelerating function of input (Figure 20 and Figure B2). Higher inputs will produce progressively smaller increments of firing rate. With constant weights for decoding, a soft form of normalization thus takes place (Figure 20): Vector $\mathbf{s}+\mathbf{t}$ is almost reduced to unit length [10]. This, in turn, results in a loss of information: the dot product between $\mathbf{s}$ and $\mathbf{s}+\mathbf{t}$ is lower than one. In other words, when both stimulus- and temporal information are encoded in the same neural population, we lose some information about both. In effect, the summed inputs to our neural population share a common *representational* resource, where the normalization of the resulting vector puts a capacity limit on how well the original vectors can be decoded. This normalization mechanism underlying capacity limitations in working memory is similar to mechanisms used in other neural models of working memory (e.g., Bays, 2014; Bouchacourt & Buschman, 2019).

How is interference in our network related to the capacity limitations typically found in working memory? A consistent finding in the working memory literature is that memory variability is a power function of set size (e.g., Bays & Husain, 2008). To quantify how our network relates to this power function of set size, we simultaneously presented N orthogonal vectors to a spiking neural network, where N corresponds to the number of items in a typical working memory task. We then decoded one of the original vectors and used decoding accuracy (dot product) as a proxy for the memory precision measures reported in the literature. Decoding performance closely approximates a power function of set size, suggesting that the kind of interference in our network captures typical set size effects in working memory (Figure 21).

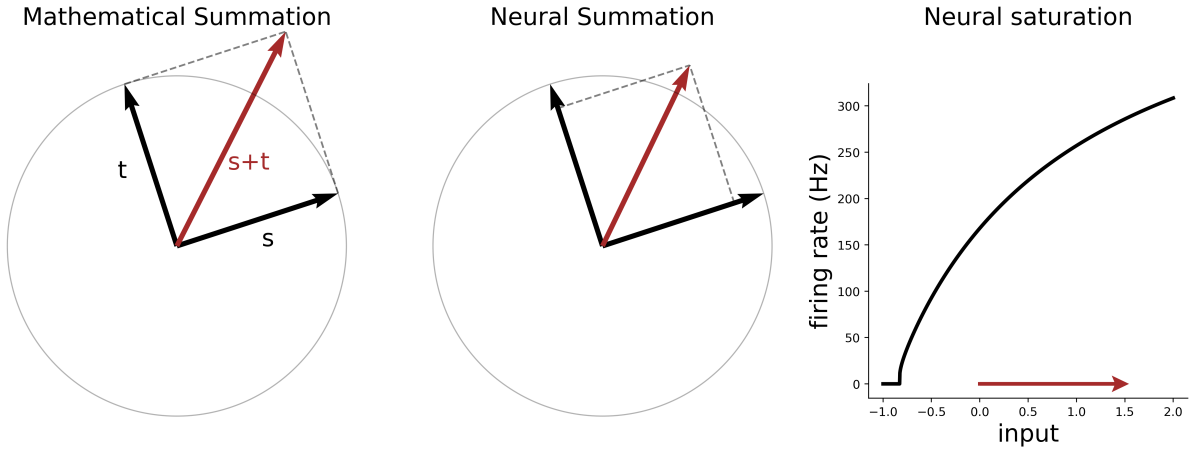
To demonstrate that set size effects in our network also capture interference between working memory and timing performance, we modelled an experiment by Polti et al. (2018). Participants performed an N-back working memory task, while also prospectively timing the duration of each

[9] Increasing the vector size each time a new vector is added is similar to increasing the size of a shopping list with each new item, for instance, by stacking sticky notes. The size (i.e., dimensionality) of the shopping list increases without distortion of its items.

[10] Having a normalized vector despite adding more vectors is similar to having a fixed-length shopping list

Figure 20

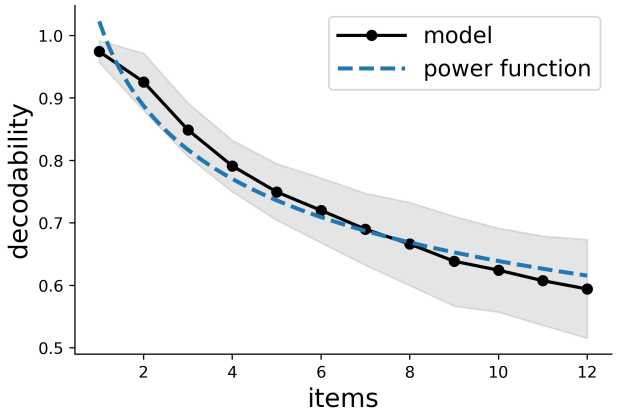
Normalization in spiking neural networks



Note. Left panel: vector addition without normalization. When unit vectors **s** and **t** are added, the resulting vector is not unit length. Both original vectors can be perfectly decoded from **s** + **t**. Middle panel: vector addition with normalization in a population of spiking neurons. The resulting vector is almost unit length due to soft normalization by spiking neurons. Right panel: neural saturation drives normalization. The tuning curve shows that firing rate is a decelerating function of input. Therefore, large values are compressed, leading to soft normalization.

Figure 21

Working memory capacity limitations



Note. When more items are simultaneously presented to the network, the decodability of the original vectors decreases approximately as a power function.

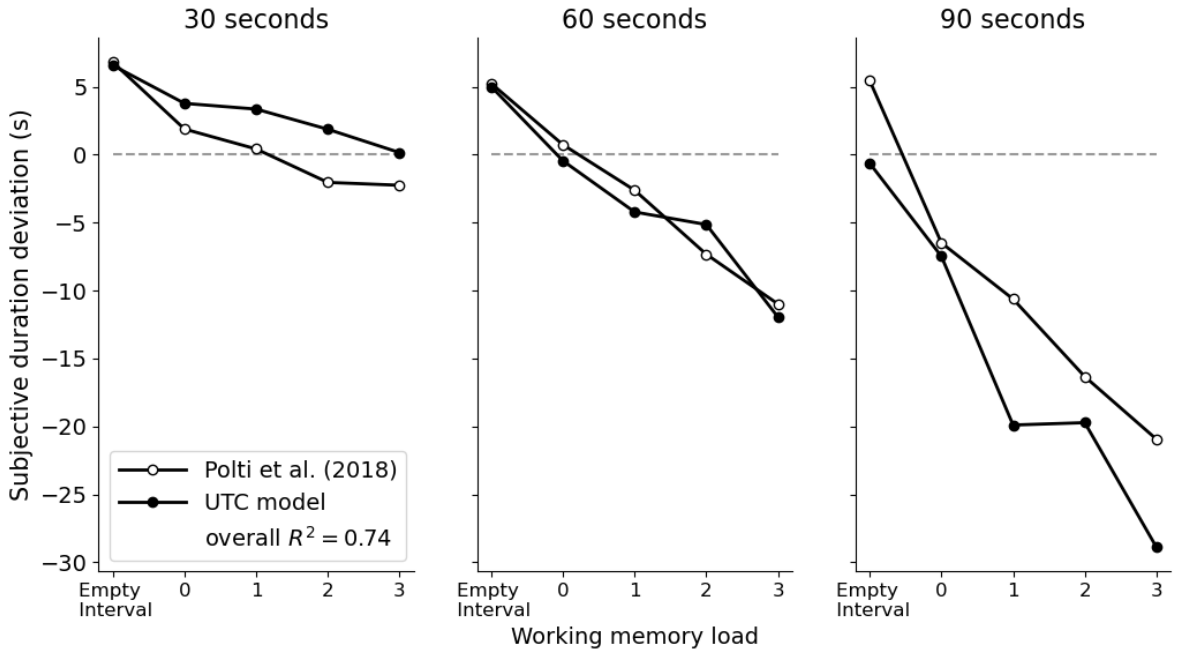
block (30, 60 or 90 seconds). During each block, letters were sequentially presented, and the start and end of the block were marked by the presentation of red dots. After each block, participants gave a verbal estimate of the interval between the red dots. In some blocks, participants timed an empty interval without any letters on the screen. In the rest of the blocks, working memory load was parametrically manipulated by requiring participants to either respond to a certain letter ('no load') or respond when the letter N positions back matched the letter currently on the screen. The authors found that intervals were increasingly underestimated as working memory load increased. Crucially, this effect scaled with the timed interval, suggesting that the rate of temporal accumulation was reduced by increasing working memory load.

We simulated the N-back experiment by assuming that the model (d=6) is presented with a constant temporal vector **t** and constant input that reflects, on average, items stored in working memory over the course of a trial. In empty interval conditions, we only presented **t**. In working memory conditions, we simultaneously presented one vector for the 'no load' condition, and $N + 1$ vectors for the 'N-back' conditions. As in the previous example, this produces interference between items and can possibly account for decreases in performance with increasing load. We only hand-tuned the window size ($\theta = 180s$), which was fixed across conditions, and we increased the number of neurons in the recurrent neural network to 1000 neurons, which improved the reliability of the time estimates. The UTC model captures both an increasing underestimation of time with higher loads and its dependence on the duration of the interval, providing a reasonable fit to the empirical data (Figure 22). These results suggest that the UTC model captures some important features of timing, working memory, and their interdependence in dual-tasking conditions.

Attentional gain explains effects of selective and divided attention on time estimation

As discussed previously, attention has multifaceted effects on time estimation. Selective attention to stimuli

Figure 22
Model fit to Polti et al. (2018)



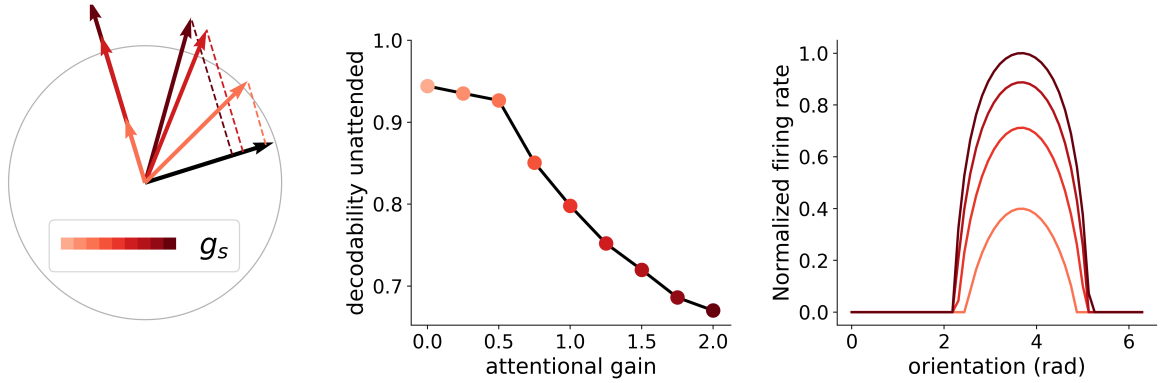
Note. Working memory load parametrically decreases subjective duration estimates, an effect that scales with the timed interval. The model (black dots) captures both features of the empirical data.

increases the perceived duration of those stimuli (Enns et al., 1999; Mattes & Ulrich, 1998; Yeshurun & Marom, 2008). Divided attention to time also increases time estimates (Casini & Macar, 1997; Franssen & Vandierendonck, 2002; Macar et al., 1994), but interferes with secondary task performance (for a review, see Brown, 2006). The UTC model views attention as multiplying (i.e., lengthening) stimulus vectors by some 'attentional' gain. In other words, when stimuli are attended, the network lengthens those vectors so that they can be decoded better. This mechanism accounts for the effect of selective attention. As an example, spatial selective attention improves stimulus processing of attended stimuli. The UTC model proposes that selectively attention works by multiplying stimulus vectors by an 'attentional gain' factor (g_s; see Figure 5). When stimuli need to be ignored, $g_s < 1$; when they need to be attended, $g_s > 1$. As a result, selectively attended stimuli have longer vectors, which allows for better readout of stimulus information. At the same time, this scaling of the length of the vector is preserved in the temporal window, resulting in longer time estimates. In the UTC, the size of this attention effect scales with the duration of the stimuli, as has been found in the literature (Mattes & Ulrich, 1998).

Crucially, the interference between 'timing' and 'stimulus' inputs (see previous section) explains the effects of divided attention. That is, when stimulus inputs are attended, they have longer vectors and gain a competitive advantage over the 'timing' inputs. Conversely, when stimulus inputs are 'ignored', they are shorter, tilting the advantage to the timing input. An intuitive way to understand how attentional gain biases the competition between stimulus and timing inputs is with our limited shopping list. If we want to put many items on a small, limited piece of paper, some items will become illegible. But if we should definitely not forget the milk, we could write 'milk' in a larger font. This comes at the cost of other items that have less space left. In our network, attending to stimulus inputs is like writing items (vectors) in a larger font (multiplying them), at the cost of other items that have less space left.

To illustrate how attentional competition works in the UTC model, again consider a network that represents the sum of two orthogonal 2-dimensional vectors. When one of the inputs is multiplied by an attentional gain factor (g_s), the decodability of the attended input increases and the decodability of the unattended input decreases (Figure 23). Therefore, our proposed mechanism captures the general finding that attention is competitive (Reynolds & Desimone, 1999). When the attended and unattended inputs are integrated in our network, it becomes clear that attended stimuli are judged as longer than unattended stimuli. This attentional mechanism does not just work on the level of vectors but is also in line with psychophysiological and neuro-

Figure 23
Attentional gain



Note. Left panel: Attentional gain (g_s) on the left vector is varied. The decodability of the original attended vector increases, while decodability of the unattended vector decreases (middle panel). Right panel: normalized tuning curve of example neuron. Attentional gain primarily influences the height of the tuning curve, not its width.

physiological work on attention (e.g., Hillyard et al., 1998; Treue, 2001) and previous modelling approaches in the Neural Engineering Framework (Bobier et al., 2014). For instance, our simple attentional gain mechanism qualitatively captures attentional modulations of tuning curves of individual neurons (e.g., McAdams & Maunsell, 1999): Attention primarily affects the height of the curve, not its width or position (Treue, 2001).

In Figure 5, we demonstrate a typical trial in a dual-tasking timing experiment. The UTC model assumes that 'attending to time' is nothing more than ignoring stimulus inputs (for a similar perspective, see Phillips, 2012). When stimulus inputs are ignored (i.e., lower attentional gain), the 'timing' input will suffer from less competition (see Figure 5). The net result is that timing inputs are more decodable, which increases time estimates, and stimulus inputs are less decodable, causing secondary task interference (see Figure 24. In sum, our attentional gain mechanism accounts for the bi-directional nature of paying divided 'attention to time': It both increases prospective time estimates, but also disrupts secondary task performance (Brown, 2006; Brown et al., 2013).

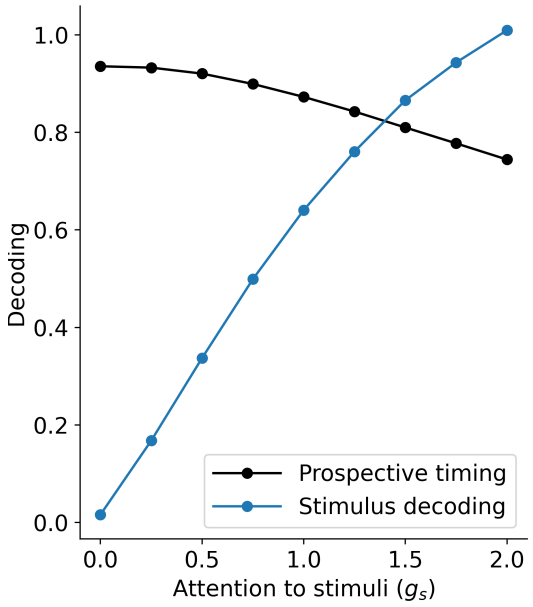
An important issue in the literature is how one can distinguish timing being disrupted by a lack of attention ('ignoring time') or a surplus of memory load ('overloading time'). The UTC model does not decisively answer this issue, but at least demonstrates why this issue is so difficult. Attentional gain and memory load work in concert to decrease the decodability of the temporal vector, shortening its subjective duration. As such, both at the level of neural representation and behavioral performance, attentional gain and load do similar things. However, the bi-directional interference predicted by the UTC can possibly dissociate 'ignoring' time and 'overloading' time. Ignoring time should decrease the length of the temporal vector, while increasing the length of the stimulus vectors. Increasing the load, however, should decrease the decodability of all vectors processed by the network, *including* the stimulus vectors. Experimentally, this means that ignoring time should hurt timing performance, while boosting performance on a concurrent task (Macar et al., 1994; Zakay, 1998). Increasing the load should be detrimental to performance on both the timing and concurrent task (Polti et al., 2018). It would be fruitful to see to what extent these effects can be teased apart in an experimental paradigm that combines both manipulations of load and attentional prioritization.

Integrating remembered content explains effects of contextual changes

As already discussed in the introduction, an interval with more changes appears to last longer. This is especially the case when the changes are actively processed (Mcclain, 1983; Predebon, 1996) or, in the case of retrospective timing, segmented (Poynter, 1983; Zakay et al., 1994). So far, the UTC model assumes that time estimates scale with how much change is encoded in a rolling temporal window, whether those changes consist of a constant 'timing' input or a varying 'stimulus' input. This same set of assumptions is also able to account for the finding that intervals with more changes are perceived as lasting longer. When more changes are encoded by the rolling temporal window, these changes (literally) add up to a longer time estimate of the interval that spans those changes. This is also true for retrospective time estimates, when no 'timing' input is present. When a retrospective time estimate is required at the end of an interval, the UTC model adds up the amount of 'stimulus' changes encoded inside the temporal window to generate an estimate (see Equation 2).

Figure 24

As more attention is paid to incoming stimuli (high g_s), prospective time estimates decrease while stimulus decodability increases



Note. When more attention is paid to time (low g_s), prospective estimates increase, which interferes with stimulus decodability.

What happens when we vary the number of stimuli presented to our network? We assessed this question by varying the number of stimuli without varying the total presentation time. We simulated the effect of the number of stimuli (N) by presenting the network with $1 - 10$ inputs in 1 second. The duration of the inputs was scaled such that the total duration of the inputs was always one second. Note that this way of presenting inputs resembles 'segmenting' the input. We then read out the network state at 1.2 seconds and infer a time estimate based on Equation 2. We also varied the number of LDN dimensions (d), which controls how precisely the inputs can be represented within the temporal window. Time estimates were a decelerating function of N (Figure 25). This pattern is consistent with the finding that retrospective time estimates increase with the number of perceived events (Block & Reed, 1978; Fountas et al., 2022; Lositsky et al., 2016; Mcclain, 1983; Predebon, 1996). Furthermore, the slope of this function crucially depends on the number of LDN dimensions, since more changes can be encoded by additional dimensions. What happens when incoming stimuli are more or less attended to? We can clearly see that for larger g_s, the signal strength increases, and therefore retrospective time estimates increase (Figure 5 and Figure 26). This behavior is broadly consistent with the finding that as stimulus inputs are more attended, retrospective time estimates increase (Block et al., 2010; Fountas et al., 2022).

Unified temporal coding accounts for differences between prospective and retrospective timing

In their seminal meta-analysis, Block et al. (2010) demonstrate that prospective time estimates decrease with increasing cognitive load, while retrospective estimates increase with increasing cognitive load. This interaction effect has been taken as evidence that prospective and retrospective timing are different kinds of processes. In line with this reasoning, previous models have assumed that cognitive load affects attention (prospective timing) and memory (retrospective timing) separately (Fountas et al., 2022; French et al., 2014). That is, when cognitive load increases, 'attention to time' is hindered, decreasing prospective estimates. Independently, increased cognitive load produces more remembered changes in memory, lengthening retrospective estimates.

The attentional mechanisms of the UTC model, however, can simultaneously explain how cognitive load modulates prospective, as well as retrospective estimates. The UTC model assumes that 'attending to time' is nothing more than ignoring 'stimulus' inputs. In cognitively demanding situations, more attention needs to be paid to incoming stimuli, resulting in stronger competition with the 'timing input'. The net result is that increasing cognitive load decreases prospective estimates (Figure 26). The effect of cognitive load on retrospective estimates is explained by the exact same mechanism. Cognitively demanding tasks require more attention to incoming stimuli, effectively boosting their representational precision (see Figure 5), increasing retrospective estimates that are based on the 'stimulus' inputs (Figure 26) [11]. Notably, the UTC model captures this interaction by only varying one parameter: attention to incoming stimuli, g_s [12]. Therefore, the UTC model makes constrained predictions regarding the relationship between cognitive load, attention and time estimates.

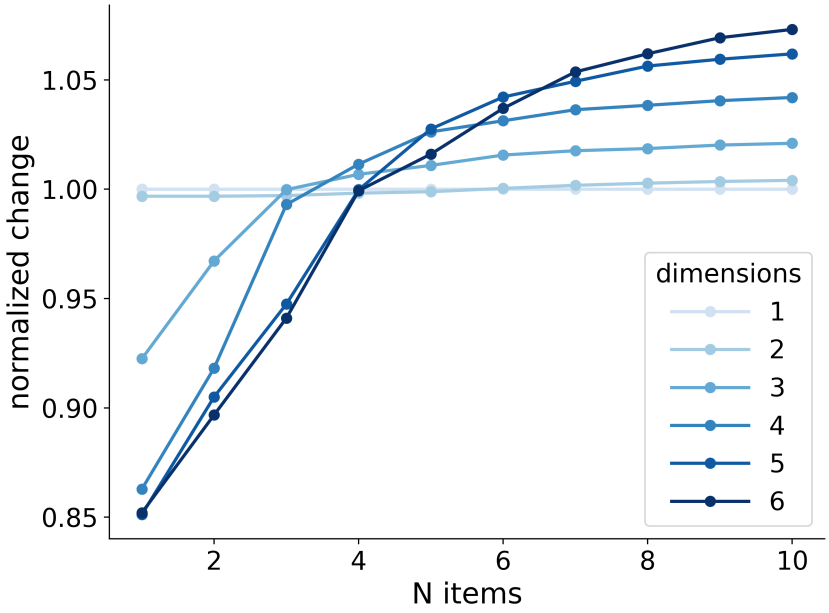
It should be noted that there are some possible boundary conditions for observing the interaction effect

[11] To match the scale of retrospective time estimates to the empirical data in Block et al. (2010), we only varied the intercept (i.e., time estimate if no changes are encoded) and slope (i.e., how much additional change increases estimates).

[12] Another explanation that is consistent with the UTC model is that as cognitive load increases, working memory load also increases. In that case, prospective estimates would decrease due to increasing working memory load. Retrospective estimates would increase, due to more items being encoded inside the temporal window. It is important to note, however, that whichever explanation turns out to be viable, it is only a *single* factor (attention or working memory load) that produces the 'cognitive load' effect in the UTC model.

Figure 25

Effect of the number of stimuli on change encoded in the network



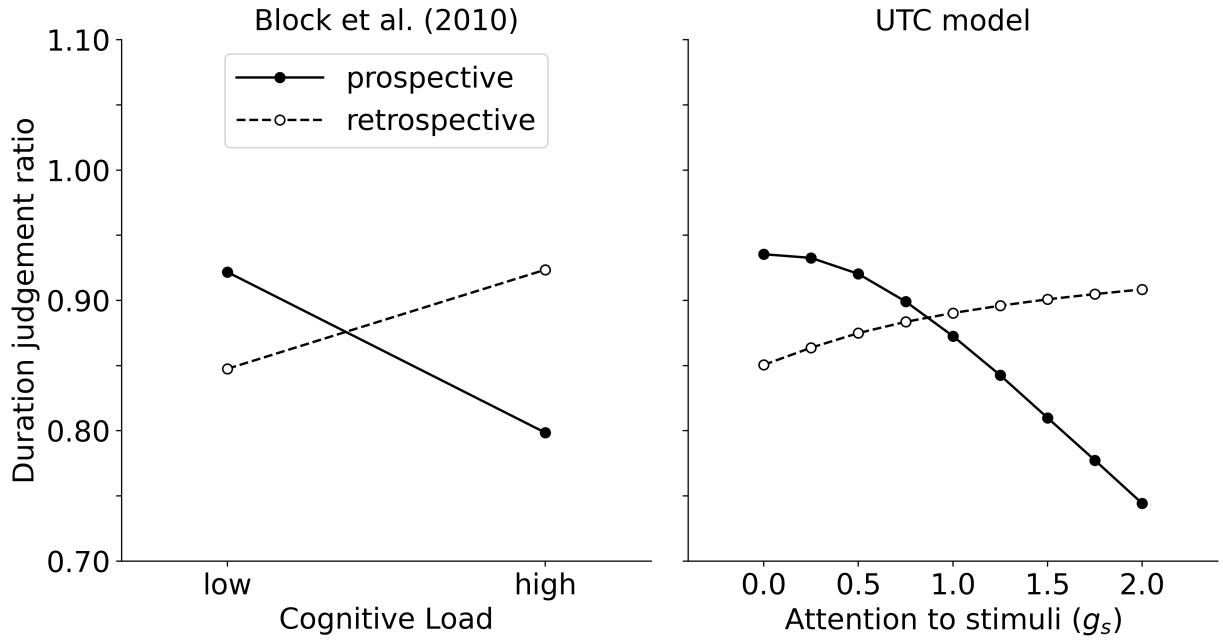
Note. When more stimuli are presented to the network in the same total time, the amount of change represented by the network increases. For low-dimensional networks, there is little effect of the number of stimuli, since multiple stimuli cannot be encoded by these networks. Change was normalized within each network across different *N*.

of cognitive load on time estimation in within-participant paradigms. For instance, Walker et al. (2022) found no evidence that cognitive load modulates prospective and retrospective retrospective time estimates of 8-minute and 58 minute intervals. This suggests a possible upper bound on the timing processes that are engaged for different time intervals. A recent study by Nicolaï et al. (2024) did not find a significant interaction between working memory load and duration judgment type (prospective vs. retrospective), using an online adaptation of the paradigm by Polti et al. (2018). In both studies, it is not clear how much evidence there was for the *absence* of an interaction effect, but future studies should attempt to replicate this interaction effect in controlled experimental settings.

In sum, what does the UTC model suggest about the distinction between prospective and retrospective timing; about whether it is a difference of degree or a difference of kind? We modelled prospective and retrospective timing by only assuming that in prospective contexts, an additional timing input is provided to the network, which only differs in representational content from the fluctuating, variable stimulus inputs. This can be viewed as a qualitative difference between prospective and retrospective timing, but the similarities are more salient. The model suggests that both forms of timing rely on the same representational and computational principles. Both temporal and stimulus information are represented as high-dimensional vectors and both types of information are encoded in a rolling temporal window, where the amount of change informs time estimates. In the case of retrospective timing, these changes reflect changes in stimulus content, whereas in the case of prospective timing, they reflect the integration of a constant input. And not *despite*, but precisely *because* of these similarities, does the UTC model explain why prospective and retrospective timing are differentially influenced by cognitive load. If temporal and stimulus information were represented separately, the UTC model would not be able to capture the competition between temporal and stimulus information. Further, supporting the notion that both types of timing rely on the same principles, the model suggests that cognitive load may only affect a single underlying process: attentional gain.

Figure 26

The effect of attention to stimuli (g_s) on prospective and retrospective timing reproduces the classic interaction effect of cognitive load (Block et al., 2010)



Note. On the y-axis, the duration judgement ratio ($\frac{t_{estimated}}{t_{target}}$), where one indicates perfectly accurate timing. For the Block et al. (2010) meta-analysis, cognitive load is plotted on the x-axis. For the UTC model, attention to stimuli (g_s) is plotted on the x-axis and assumed to scale with cognitive load. Prospective estimates decrease with more attention to stimuli due to working memory interference, whereas retrospective estimates increase with more attention to stimuli due to encoding of more change.

General Discussion

Here, we have pursued one possible answer to the question of how the brain represents and updates temporal information, proposing the Unified Temporal Coding (UTC) model. Instead of studying the behavior of neural networks that were trained extensively on timing tasks, we use a recurrent neural network, the Legendre Delay Network (LDN), whose connections are optimized from first principles to represent a flexible rolling window of input history. The LDN continually updates coefficients on temporal basis functions that together form the ever-changing representation of input history. The length of the rolling window, that is, how quickly inputs are encoded and forgotten, is controlled by the speed of this updating process. The UTC model puts forward clear and testable neural principles underlying temporal representation. Indeed, it can account for some fundamental behavioral and neural phenomena, such as (violations of) the scalar property, temporal scaling of neural responses and the effects of distracting events on timing.

The UTC model also scales naturally to more high-dimensional inputs and complex tasks. We make the crucial assumption that both temporal and stimulus information are represented in the same way, by the same neural population. We show that fundamental limits in simultaneously representing multiple temporal and stimulus inputs accounts for both limits in working memory capacity *and* time perception. Further, we implemented an attentional gain mechanism that not only accounts for attentional effects on prospective time estimation, but also retrospective time estimation, thereby providing a novel unification of these seemingly distinct forms of timing.

Comparison with previous models

In his famous *Principles of Psychology* (1890), William James closed his chapter on time perception with the question ‘To what cerebral process is the sense of time due?’. The ‘internal clock’ approach has long been the most formalized approach to answering this question, despite its focus on functional over more ‘cerebral’ explanations. In recent decades, a host of alternatives to the ‘internal clock’ have been proposed, such as oscillator and memory models, many of which are explicitly based on neural mechanisms. The most recent example of this increased emphasis on biological plausibility are recurrent neural network (RNN) mod-

els, providing a powerful lens through which to view timing and time perception. Interval timing is viewed as trajectories through complex neural spaces, quite unlike the monotonic and one-dimensional accumulation of 'ticks' assumed by 'internal clock' models. As a result of their biological plausibility and complexity, neural network models have provided strong accounts of neural phenomena and more complex timing phenomena like pattern timing (e.g., Hardy & Buonomano, 2016; Hardy et al., 2018). Despite these clear benefits over more traditional approaches, the representational and computational principles underlying the performance of neural network models are often obscure. The 'ticks' of a clock clearly represent the elapsed time since the onset - and the expected time until offset - of a relevant event. The same basic temporal information is often present in trained neural networks, however, the principles behind representing and continuously updating temporal information are ill-defined. Careful study of these artificial neural networks has proven productive (e.g., Bi & Zhou, 2020), however, their underlying principles are often inferred in hindsight rather than constructed from first principles.

How does the UTC model relate to the taxonomy of timing models put forward in the Introduction? First, the UTC model inherits prominent features of pacemaker-accumulator models, memory models and recurrent neural network models. In prospective timing contexts, a constant input is presented to the network, similarly to how a pacemaker provides a constant stream of 'ticks' to the accumulator. The first dimension, representing the mean of the input history, behaves like a (leaky) integrator, where the speed of integration is controlled by the recurrent gain, similarly to several PA-models (Simen et al., 2013). The UTC model also handles the scalar property, one-shot learning and adaptive neural 'speed' in similar ways. One major difference, however, is how time and stimulus information is represented.

With regard to stimulus representation, the UTC model also shows clear similarities with memory models, especially those that represent stimulus history on a continuous timeline, such as the TILT model (Howard et al., 2015; Shankar & Howard, 2012). The conjunctive representation of 'what' and 'when' is central to the ability of both models to account for behavioral and neural data, and as such they embody common principles. The main difference between UTC and TILT is that the timeline of UTC is bounded and flexible, whereas the timeline of TILT is (theoretically) infinite and fixed. One notable exception is Liu et al. (2019), who implemented TILT in a spiking neural network. The authors demonstrated that the timeline can be stretched or compressed by changing the gain of the tuning curves of individual neurons. As such, the principle of changing the dynamics of the network by scaling the recurrent gain is not new. However, the UTC does offer a novel perspective on what such a recurrent gain modulation implies for representing a rolling window of history: It exactly scales window length (θ). By extension, scaling the recurrent gain of the network has consequences for the filtering properties of the UTC model, but also the speed of encoding and forgetting in working memory.

The UTC model is also a Recurrent Neural Network and therefore has the same basic structure. However, it embodies clear representational and computational principles that are derived from optimally representing a rolling window of the past.

As we have shown, the principles embodied by the UTC model can account for complex neural signatures and their adaptive temporal scaling. Previous work has also demonstrated that this rolling window can model time-cell data well (Voelker & Eliasmith, 2018). However, it is also clear that some distinct features of neural data are not captured by the UTC model. For instance, stimulus-selective cells have been found in entorhinal cortex that exhibit a continuous spectrum of time-constants (Bright et al., 2020; Tsao et al., 2018). These cells have been predicted by the TILT model several years before they were found, and as such the TILT model likely has an edge over the UTC model in explaining the neural substrates of long-term retrospective timing. On the other hand, we believe that some complex features of the neural data in Wang et al. (2018), which are well captured by the UTC model, are not anticipated by the TILT model. Future modelling work should attempt at a formal comparison between how well these models capture neural data in different brain areas, given different task requirements and on different timescales.

Another prominent dimension along which models of timing are categorized is the 'dedicated versus intrinsic' axis (Ivry & Schlerf, 2008). Dedicated models propose that timing is implemented by a single, specialized, modality-independent neural mechanism, located in a specific network of brain areas, comprising the basal ganglia, thalamus, cerebellum and SMA (Merchant et al., 2013). In contrast, intrinsic models argue that *any* neural circuit with physiological- or population dynamics can tell time, suggesting multiple mechanisms underlie timing (Buonomano & Laje, 2010; Motanis et al., 2018). Hence, depending on the input modality or task, different neural structures will be involved in timekeeping. While some noteworthy attempts at integrating these views have been put forward (Merchant et al., 2013), these are mainly conceptual models that delineate which mechanisms, dedicated or intrinsic, obtain for certain timescales, modalities or task requirements. Hence, these proposals lack clear functional explanations as to why intrinsic or dedicated timing mechanisms may be employed in different situations. In contrast, the UTC model suggests a possible functional account of context-dependent temporal processing. Instead of assuming that certain neural mechanisms or brain areas are inherently 'intrinsic' or 'dedicated',

the UTC model proposes that 'intrinsic' networks may be controlled - by tightly regulating inputs, dynamics and outputs so as to tell time in a dedicated way. Neural networks that already generate complex temporal representations may be 'recruited' for interval timing by providing a constant input to the network, controlling the speed of the neural dynamics and tuning its readout to match task demands. This mechanism of 'recruiting' neural circuits for timekeeping is clear from the way UTC deals with prospective and retrospective timing: the network 'intrinsically'[13] encodes temporal information in retrospective timing tasks, however, these *same* dynamics are adapted to perform prospective timing tasks.

The UTC model also provides a different perspective on prospective and retrospective timing compared to existing models (Fountas et al., 2022; French et al., 2014). Both GAMIT and the Predictive Processing model assume that cognitive load affects two independent parameters: Attention for prospective timing and memory for retrospective timing. Conversely, the UTC model assumes that cognitive load only affects a single parameter: Attention to stimuli. Arguably, a separation of attention and memory would complicate explanations of other phenomena. For instance, neither GAMIT nor the Predictive Processing model explains why paying attention to timing interferes with secondary task performance. The UTC model suggests that stimulus and timing information compete within the same neural network, resulting in bi-directional interference.

The UTC provides a different explanation for the effect of cognitive load and explains some phenomena that may be beyond the scope of existing models. How could the UTC model still be tested against alternatives, like the Predictive Processing model? A major contrasting prediction relates to attention, stimulus encoding and timing. The Predictive Processing model assumes that as more attention is paid to timing, more stimuli are encoded, resulting in longer time estimates. Conversely, the UTC model assumes that as more attention is paid to timing, stimuli are *ignored*, resulting in longer time estimates. Clearly, the Predictive Processing model and the UTC model make qualitatively different predictions regarding stimulus processing, which would be worthwhile to test empirically.

Future Directions

The UTC model attempts to integrate phenomena across different forms of timing (prospective - retrospective) and levels of explanation (neurophysiological - cognitive). Here, we will briefly outline how the UTC model may be ideally situated to explain more complex forms of timing, implement alternative learning and adaptation rules, and extend to other phenomena in 'temporal cognition'.

Pattern Timing

When studying 'interval' timing, the temporal complexity of our surroundings and actions is easily overlooked. Humans and non-human animals are remarkably skilled at recognizing and producing complex temporal patterns (Hardy & Buonomano, 2016). For instance, in speech, a wealth of information is contained, not in the isolated timings of vowels, words or sentences, but rather their embedding in a complex, hierarchical temporal structure. This poses a fundamental issue for models of interval timing: Is complex timing somehow constructed from isolated intervals, or is interval timing derived from more complex temporal representations? The UTC is clearly consistent with the latter view: Interval timing is accomplished by integrating a constant signal and time estimates are based primarily on the first dimension (i.e., the mean of the temporal window), effectively ignoring more complex temporal patterns that are encoded by the network. Nevertheless, the LDN is optimized to time these more complex patterns and embodies clear principles of temporal representation. For instance, the dimensionality and window size of the LDN jointly control the upper bound on the frequency content that can be represented (Voelker & Eliasmith, 2018; Voelker, 2019). As more dimensions are added, higher frequencies in the input history can be approximated, since these higher dimensions themselves contain higher frequencies (see the temporal basis function in 2). Similarly, as window size is decreased, higher frequencies can be represented, since the entire temporal basis function is compressed in time. Whether and how these principles apply to flexible pattern timing remains an open question, but they are clearly consistent with the basic observation that humans can both perceive and produce complex temporal patterns at a range of timescales (Hardy & Buonomano, 2016). An exciting avenue of future research is to test whether the UTC model provides an intuitive account of an effect that combines the temporal complexity and flexibility of timing, the 'Weber-speed' effect: Intervals are produced more precisely when they are embedded in faster temporal motor patterns (Hardy et al., 2018; Slayton et al., 2020). Interestingly, Hardy et al. (2018) have demonstrated that the effect does not result from subdividing the interval. Instead, true warping of the neural dynamics, which is a central tenet of UTC, seemed to best account for the data.

Learning and Adaptation

The UTC model implements well-established learning rules (Simen et al., 2013) for adapting its window size

[13] Our network is optimized to represent a rolling window of input history, hence, it is less intrinsic than other theoretical proposals (e.g., Motanis et al., 2018). Nevertheless, in retrospective conditions, the network's mechanisms are clearly more 'intrinsic' than in prospective conditions.

in prospective timing tasks, matching both observed learning rates and adaptation of neural dynamics. Nevertheless, there are still some open questions about temporal learning that the UTC model has not addressed yet. For instance, how does the network learn to represent coefficients on a temporal basis function in the first place? And how does the network learn the appropriate window size based on reinforcements, or, alternatively, adapt window size based on the temporal statistics of its input? Here, we will discuss how the UTC may be able to address these issues.

The problem of how the recurrent dynamics of our network are learned in the first place goes to the heart of unsupervised learning. How can a neural network, without any teaching signal, learn to represent the structure of its input? In the case of RNN models of timing, several learning algorithms have been proposed that can learn to represent the temporal structure of its inputs (Laje & Buonomano, 2013; Liu & Buonomano, 2009). It would be interesting to see under which conditions these algorithms generate neural networks that are structurally and functionally similar to the one used by the UTC model (Voelker & Eliasmith, 2018).

As we have demonstrated, learning appropriate timescales for timing tasks appears more tractable. Similarly to the UTC, many of the learning rules employed in interval timing models are designed to speed up or slow down the behavioral or neural dynamics, as to produce shorter or longer intervals, respectively (Gavornik et al., 2009; Killeen & Fetterman, 1988; Luzardo et al., 2013; Mikhael & Gershman, 2019; Namboodiri & Shuler, 2016; Simen et al., 2011b; Wang et al., 2020). One-shot learning rules capture the rate of temporal learning well, however, it is not clear how they relate to the underlying neurophysiology. While the role of dopamine in interval timing is multifaceted (Fung et al., 2021), it has a clear role in temporal learning. How might the UTC model capture this central role of dopamine? The most straightforward solution is to include reinforcement learning (RL) for adapting the timescale of our network (Gershman et al., 2014; Mikhael & Gershman, 2019; Petter et al., 2018). For instance, a learning rule recently proposed by Mikhael and Gershman (2019) (and empirically supported by Jakob et al. (2022)) might have an intuitive mapping to components in our network.[14] This simple algorithm learns to predict when rewards will occur after a cue. When rewards are received earlier than expected, the 'pacemaker-rate' increases, ensuring that in the future, rewards are expected to occur earlier (and vice versa for rewards that occur later than expected). This learning rule operates on principles similar to the one-shot learning rules (although they may be slower). Additionally, it captures how pharmacological (Coull et al., 2011) and optogenetic (Howard et al., 2017; Soares et al., 2016; Toda et al., 2017) modulation of dopaminergic neurons affect interval timing, and why they sometimes seem to do so in opposite ways. As such, this learning rule would further allow the UTC model to generate constrained predictions about dopaminergic effects on timing (Mikhael & Gershman, 2019), especially as they relate to gap- and distractor procedures (Buhusi, 2003). It would also allow the UTC model to learn when rewards will occur from complex temporal sequences by mapping the coefficients on its temporal basis function to expected rewards, going beyond simple intervals.

In many cases, the relevant timescale is not clearly denoted by rewards or punishments, rendering reinforcement learning mechanisms ineffective. For instance, modulations in speaking rate are typically not accompanied by changes in reward rate. Nevertheless, humans show excellent performance on classifying time-warped speech, possibly through the temporal scaling of cortical responses (Lerner et al., 2014; for a different perspective, see Vagharchakian et al., 2012). How could our network adapt the length of its temporal window based on the rate of change in the input? Previous models have broadly proposed either neurophysiological (Gütig & Sompolinsky, 2009), or network (Goudar & Buonomano, 2018) mechanisms. It would be worthwhile to investigate whether these mechanisms can automatically adapt window size in the UTC model.

Temporal Cognition

'A central issue highlighted by the earlier discussed 'dedicated versus intrinsic' axis is the nature of temporal cognition. On the one hand, it is clear that subjective time is not completely abstracted from 'non-temporal' properties of an event. A host of stimulus features shape time perception, such as contrast, loudness, size, motion, numerosity, and several more (Matthews & Meck, 2016). Similarly, directing more attention to stimuli dilates their apparent duration. These effects can be unified by the processing principle, as proposed by Matthews and Meck (2016): 'subjective duration of a stimulus is related to the strength of its perceptual representation'. A mapping between 'perceptual strength' and the UTC can be readily made: perceptual strength scales with vector magnitude. When a stimulus has high perceptual strength, through low-level stimulus features or cognitive factors, the longer its vector, and therefore, the longer it appears to last.

The processing principle is already embodied by UTC to some extent. For instance, the UTC model assumes that attentional gain scales the vector magnitude of stimuli, increasing both their perceptual strength and their apparent

[14] This learning rule should have access to the reward-prediction error (RPE; for implementation in NEF, see Rasmussen et al., 2017), a subjective estimate of time (the decoded time estimates in our network), 'pacemaker rate' (θ^{-1}) and the temporal derivative of the estimated value (for an overview of temporal differentiation methods implemented in the NEF, see Tripp & Eliasmith, 2010).

duration. Meanwhile, the effects of attention on perceptual strength are also reflected on a neural level, where tuning curve height scales with perceptual strength. The UTC model also intuitively captures the effect of working memory load, which decreases vector magnitude, and therefore their perceptual strength and apparent duration. Reduced neural responses with increasing working memory load have been demonstrated in monkey electrophysiology (Buschman et al., 2011). It seems that the UTC model may be able to account for the effects of perceptual strength on subjective time through the intuitive mapping between perceptual strength and vector magnitude.

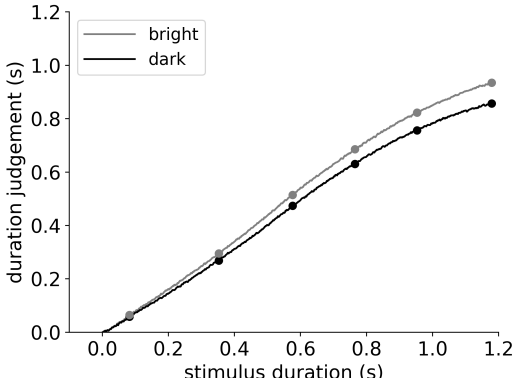
Future modelling work needs to demonstrate the feasibility of fully mapping the UTC to the processing principle, but here we provide a simple example. Matthews et al. (2011) found that stimulus contrast increases the judged stimulus duration. For instance, when a bright stimulus is presented against a dark background, it is perceived to last longer than a dim stimulus. Crucially, this effect scales with the duration of the stimulus. When we assume that the magnitude of the stimulus input scales with its contrast (dim=0.8, bright=0.9), the UTC model (d = 3) is able to capture overestimation of high-contrast inputs, an effect that scales with stimulus duration (see Figure 27).

An exciting avenue of future research would be to incorporate a realistic model of perceptual processing in the input stage of UTC, so if would be able to account for more complex stimulus-related effects. For instance, the UTC model has so far assumed some form of self-sustained activity underlies the perception of 'empty' intervals, practically equating the perception of 'filled' and 'empty' intervals. However, a consistent finding in the literature is that 'filled' intervals are perceived as lasting longer than 'empty' intervals (Wearden & Ogden, 2021). More realistic assumptions about the perceptual processes could elucidate these types of perceptual-temporal illusions. Such a model would, in turn make testable predictions about the neural dynamics underlying timing performance (see Toso et al., 2021).

Another salient feature of 'temporal cognition' is that cognitive processes do not only evolve through time. They are also sensitive to temporal contingencies in our environment and actively shape the temporal structure of our behavior. Hence, a central question is whether temporal representations are generated by a central mechanism from which cognitive processes inherit their temporal sensitivity and through which they exert influence on the timing of behavior, or whether temporal representations are a *built-in* feature of almost any cognitive process (Salet et al., 2022). Our 'recruitment' hypothesis proposes that, while temporal representations are likely a built-in feature of many cognitive processes, they are nevertheless systematically controlled, so that their inputs, outputs and dynamics track the temporal contingencies at hand. This hypothesis is relevant to many

Figure 27

Linking the UTC model to effects of stimulus intensity



Note. In Matthews et al. (2011), stimuli (bright or dark) were presented against a dark background, and participants had to verbally judge their duration. Higher contrast (bright) stimuli expanded subjective time, and its effect increased over stimulus duration. When we assume that vector magnitude scales with stimulus contrast, the UTC model is able to capture these effects.

cognitive processes, such as attention (Nobre & van Ede, 2017), episodic memory (Eichenbaum, 2014) and working memory (van Ede et al., 2017). Indeed, recent findings suggest that humans can speed up the rate at which they encode information in visual working memory when they expect little time to do so (de Jong et al., 2023). Here, we briefly discuss decision-making, a field in which several models have been developed that deal with adaptive timescales. We argue that even in the absence of explicit time estimation, the speed of evidence-accumulation and forgetting are adaptively controlled so as to track temporal contingencies in the environment, suggesting a promising avenue of future research for our UTC model.

When making decisions, we are often faced with noisy and uncertain situations, requiring the integration of multiple samples of evidence before committing to a choice. The most dominant decision-making models assume that evidence is accumulated without any forgetting over time, that is, perfect integration. However, most support for these 'perfect integration' models comes from static decision-making paradigms: The ideal starting point of evidence accumulation is known beforehand (i.e., at the start of a signal), and the underlying source of the signal remains constant throughout the trial. Indeed, perfect integration is an optimal strategy under such conditions (Bogacz et al., 2006). But when the environment is volatile, the underlying source of the signal may change frequently, and as a result, previous evidence may no longer be relevant. For instance, when deciding on the location of a potentially dangerous animal on the basis of sound, the animal may already have moved sig-

nificantly, rendering previous information obsolete. Glaze et al. (2015) show that an optimal observer forgets previous evidence more quickly as the environment changes more quickly (i.e., becomes more volatile). These principles are consistent with decision-making behavior in humans (Glaze et al., 2015; Ossmy et al., 2013) and rats (Piet et al., 2018). For instance, rats are able to optimally adapt their rate of evidence integration, tracking the volatility of the environment. When rats were moved from a highly volatile environment into a stable environment, their rate of forgetting increased; when subsequently placed back to a stable environment, their rate of forgetting decreased (Piet et al., 2018). Adaptive timescales also apply to extrapolating from the immediate past to the near future. Baumgarten et al. (2021) found that humans were able to accurately predict upcoming tones from sequences with naturalistic temporal patterns over a four-fold change in input rate. These adaptive predictions were supported by neural mechanisms, as measured by MEG, that integrated sensory evidence at flexible rates, ensuring that roughly a constant number of samples were integrated regardless of timescale. In sum, adaptive control of integration and forgetting is a central feature of decision-making and as such its mapping to a flexible window size in the UTC seems a promising avenue for future research.

Extending time in retrospect

The UTC model explains how retrospective duration judgments are made from integrating remembered stimulus content, and how the strength and number of remembered stimuli affect these judgments. However, the retrospective sense of time is arguably richer than that. For instance, humans can accurately estimate how recently something happened. The accuracy and speed of recency judgments have been explained in some detail by the TILT model (Tiganj et al., 2022), by assuming that stimuli are stored on a logarithmically compressed internal timeline. The UTC model also stores stimuli on an internal timeline, and as can be seen from Figure 3, the recency and relative order of stimuli can be readily decoded from the window. As such, the basic ingredients for recency judgements are present in the UTC model. However, we need to make additional assumptions on exactly how these temporal are read out. For instance, empirical evidence suggests backward (Tiganj et al., 2022) or forward scanning (Chan et al., 2009) through an internal timeline. Similarly, the UTC model represents the serial order of stimuli, and therefore has the basic capacity to tell which came first. Future modeling work should focus on how the UTC model may explain recency and order judgments, which are an important component of retrospective timing.

The UTC model offers a unified account of prospective and retrospective timing by assuming both result from the readout of a common memory network. Should we expect the UTC model to also account for phenomena in the wider memory literature? We believe this question remains to be addressed. However, at a minimum, the UTC tries to explain some rudimentary phenomena in working memory. For instance, why it has capacity limitations, how information is encoded and forgotten and how that is supported by neural dynamics. For instance, when humans expect to have little time for encoding information, they are able to increase their encoding speed (de Jong et al., 2023). The UTC model can explain these adaptive speedups intuitively by appealing to adaptations in recurrent gain, which scales the speed of encoding in the temporal window. Given its fit to the working memory phenomena considered in this manuscript, and its potential to explain the dynamics of working memory encoding and forgetting, we believe that the UTC model is at least promising as a neural model of working memory.

Whether the specific implementation of memory in the UTC model (i.e., the LDN network) would also extend to long-term memory remains to be seen. There is no in principle limit to the size of the temporal window, and multiple LDN networks can be stacked to substantially extend it (Voelker, 2019). Also, as we have shown, the window size can be learned very quickly, and could therefore also adapt to how long information needs to be stored. However, it seems plausible that durable long-term memories utilize a more durable format than ongoing neural activity patterns. Furthermore, long-term memory for when events happened likely depends on a multitude of non-temporal factors from which temporal information can be reconstructed. Clearly, such types of temporal judgements would benefit from integration with long-term associative and semantic memory, as has been proposed by the Predictive Processing model (Fountas et al., 2022).

However, even if the memory system of the UTC model would not extend to these longer timescales, it could still inform future theoretical approaches to retrospective timing in long-term memory. The most important principle embodied by the UTC is that, at a crucial stage in processing, prospective timing information is coded conjunctively with stimulus information; and as a result of limited representational resources, they compete. As a consequence, it makes clear predictions (and model prescriptions) for how attention, memory load, and concurrent prospective timing influence the strength of memories encoded in working memory. To the extent that strength in working memory determines long-term memory performance, UTC's principles of conjunctive, competitive coding could extend to duration information in long-term memory. This seems like a promising prospect, especially given the proposed mapping of the UTC to the processing principle (see previous section). In fact, long-term memory strength is modulated by several factors that influence working memory strength. For instance, attention to information in working memory affects later LTM performance (Jeanneret et al., 2023) and items encoded at higher set sizes

decrease LTM strength (Forsberg et al., 2021). As such, the conjunctive, competitive coding of the UTC would predict that attended items, and items that suffered less from competition of other items, would be remembered as having lasted longer. Conversely, to the extent that concurrent prospective timing impairs working memory strength, they should also impair long-term memory performance, and therefore their remembered duration. These predictions remain to be tested, but it demonstrates that the UTC can generate novel predictions based on some of its core principles.

Concluding Remarks

Here, we have proposed a neurocomputational model of timing, the Unified Temporal Coding (UTC) model, that aims to unify prospective and retrospective timing through theoretically well-grounded representational and computational principles. The UTC model explains conformity and violations of the scalar property, neural population dynamics underlying time perception and time production, timing behavior under normal and distracting conditions, common capacity limits in timing and working memory, and how timing depends on attentional modulations. Strikingly, by assuming that prospective and retrospective timing rely on the same principles and are implemented in the same neural circuit, our attentional gain mechanism can resolve the apparently paradoxical effect of cognitive load on prospective and retrospective timing. Further, the UTC model suggests that explicit interval timing does not depend on a dedicated mechanism, nor is it a simple by-product of intrinsic neural dynamics. Instead, adaptive interval timing behavior is accomplished by appropriately controlling the inputs, dynamics and outputs of neural circuits that are tuned to represent a flexible window of their input history. In sum, the UTC model embodies clear representational and computational principles, providing an initial attempt to unify time in passing, and time in retrospect.

References

Almeida, R., & Ledberg, A. (2010). A biologically plausible model of time-scale invariant interval timing. *Journal of Computational Neuroscience*, *28*(1), 155–175. https://doi.org/10.1007/s10827-009-0197-8

Bangert, A. S., Kurby, C. A., Hughes, A. S., & Carrasco, O. (2020). Crossing event boundaries changes prospective perceptions of temporal length and proximity. *Attention, Perception, & Psychophysics*, *82*(3), 1459–1472. https://doi.org/10.3758/s13414-019-01829-x

Bangert, A. S., Kurby, C. A., & Zacks, J. M. (2019). The influence of everyday events on prospective timing "in the moment". *Psychonomic Bulletin & Review*, *26*(2), 677–684. https://doi.org/10.3758/s13423-018-1526-6

Bangert, A. S., Reuter-Lorenz, P. A., & Seidler, R. D. (2011). Dissecting the clock: Understanding the mechanisms of timing across tasks and temporal intervals. *Acta Psychologica*, *136*(1), 20–34. https://doi.org/10.1016/j.actpsy.2010.09.006

Baumgarten, T. J., Maniscalco, B., Lee, J. L., Flounders, M. W., Abry, P., & He, B. J. (2021). Neural integration underlying naturalistic prediction flexibly adapts to varying sensory input rate. *Nature Communications*, *12*(1), 2643. https://doi.org/10.1038/s41467-021-22632-z

Bausenhart, K. M., Bratzke, D., & Ulrich, R. (2016). Formation and representation of temporal reference information. *Current Opinion in Behavioral Sciences*, *8*, 46–52. https://doi.org/10.1016/j.cobeha.2016.01.007

Bays, P. M. (2014). Noise in Neural Populations Accounts for Errors in Working Memory. *Journal of Neuroscience*, *34*(10), 3632–3645. https://doi.org/10.1523/JNEUROSCI.3204-13.2014

Bays, P. M., & Husain, M. (2008). Dynamic Shifts of Limited Working Memory Resources in Human Vision. *Science*, *321*(5890), 851–854. https://doi.org/10.1126/science.1158023

Beiran, M., Meirhaeghe, N., Sohn, H., Jazayeri, M., & Ostojic, S. (2023). Parametric control of flexible timing through low-dimensional neural manifolds. *Neuron*, *111*(5), 739–753.e8. https://doi.org/10.1016/j.neuron.2022.12.016

Bekolay, T., Laubach, M., & Eliasmith, C. (2014). A Spiking Neural Integrator Model of the Adaptive Control of Action by the Medial Prefrontal Cortex. *Journal of Neuroscience*, *34*(5), 1892–1902. https://doi.org/10.1523/JNEUROSCI.2421-13.2014

Bekolay, T., Bergstra, J., Hunsberger, E., DeWolf, T., Stewart, T. C., Rasmussen, D., Choo, X., Voelker, A. R., & Eliasmith, C. (2014). Nengo: A Python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics*, *7*. https://doi.org/10.3389/fninf.2013.00048

Bi, Z., & Zhou, C. (2020). Understanding the computation of time using neural network models. *Proceedings of the National Academy of Sciences*, *117*(19), 10530. https://doi.org/10.1073/pnas.1921609117

Bizo, L. A., Chu, J. Y., Sanabria, F., & Killeen, P. R. (2006). The failure of Weber's law in time perception and production. *Behavioural Processes*, *71*(2-3), 201–210. https://doi.org/10.1016/j.beproc.2005.11.006

Block, R. A. (1974). Memory and the experience of duration in retrospect. *Memory & Cognition*, *2*(1), 153–160. https://doi.org/10.3758/BF03197508

Block, R. A., George, E. J., & Reed, M. A. (1980). A watched pot sometimes boils: A study of duration experi-

ence. *Acta Psychologica*, *46*(2), 81–94. https://doi.org/10.1016/0001-6918(80)90001-3

Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica*, *134*(3), 330–343. https://doi.org/10.1016/j.actpsy.2010.03.006

Block, R. A., & Reed, M. A. (1978). Remembered duration: Evidence for a contextual-change hypothesis. *Journal of Experimental Psychology: Human Learning & Memory*, *4*(6), 656–665. https://doi.org/10.1037/0278-7393.4.6.656

Block, R. A., & Zakay, D. (1997). Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic Bulletin & Review*, *4*(2), 184–197. https://doi.org/10.3758/BF03209393

Bobier, B., Stewart, T. C., & Eliasmith, C. (2014). A Unifying Mechanistic Model of Selective Attention in Spiking Neurons. *PLoS Computational Biology*, *10*(6), e1003577. https://doi.org/10.1371/journal.pcbi.1003577

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765. https://doi.org/10.1037/0033-295X.113.4.700

Bouchacourt, F., & Buschman, T. J. (2019). A Flexible Model of Working Memory. *Neuron*, *103*(1), 147–160.e8. https://doi.org/10.1016/j.neuron.2019.04.020

Bright, I. M., Meister, M. L. R., Cruzado, N. A., Tiganj, Z., Buffalo, E. A., & Howard, M. W. (2020). A temporal record of the past with a spectrum of time constants in the monkey entorhinal cortex. *Proceedings of the National Academy of Sciences*, *117*(33), 20274–20283. https://doi.org/10.1073/pnas.1917197117

Brown, S. W. (1985). Time perception and attention: The effects of prospective versus retrospective paradigms and task demands on perceived duration. *Perception & Psychophysics*, *38*(2), 115–124. https://doi.org/10.3758/BF03198848

Brown, S. W. (2006). Timing and executive function: Bidirectional interference between concurrent temporal production and randomization tasks. *Memory & Cognition*, *34*(7), 1464–1471. https://doi.org/10.3758/BF03195911

Brown, S. W., & Boltz, M. G. (2002). Attentional processes in time perception: Effects of mental workload and event structure. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(3), 600–615. https://doi.org/10.1037//0096-1523.28.3.600

Brown, S. W., Collier, S. A., & Night, J. C. (2013). Timing and executive resources: Dual-task interference patterns between temporal production and shifting, updating, and inhibition tasks. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 947–963. https://doi.org/10.1037/a0030484

Brown, S. W., & Stubbs, D. A. (1992). Attention and Interference in Prospective and Retrospective Timing. *Perception*, *21*(4), 545–557. https://doi.org/10.1068/p210545

Bueti, D., & Buonomano, D. V. (2014). Temporal Perceptual Learning. *Timing & Time Perception*, *2*(3), 261–289. https://doi.org/10.1163/22134468-00002023

Buhusi, C. (2003). Dopaminergic Mechanisms of Interval Timing and Attention. In W. Meck (Ed.), *Functional and Neural Mechanisms of Interval Timing*. CRC Press. https://doi.org/10.1201/9780203009574.ch12

Buhusi, C. V. (2012). Time-sharing in rats: Effect of distracter intensity and discriminability. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*(1), 30–39. https://doi.org/10.1037/a0026336

Buhusi, C. V., & Matthews, A. R. (2014). Effect of distracter preexposure on the reset of an internal clock. *Behavioural Processes*, *101*, 72–80. https://doi.org/10.1016/j.beproc.2013.09.003

Buhusi, C. V., & Meck, W. H. (2009a). Relative time sharing: New findings and an extension of the resource allocation model of temporal processing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1525), 1875–1885. https://doi.org/10.1098/rstb.2009.0022

Buhusi, C. V., & Meck, W. H. (2009b). Relativity Theory and Time Perception: Single or Multiple Clocks? *PLoS ONE*, *4*(7), e6268. https://doi.org/10.1371/journal.pone.0006268

Buhusi, C. V., Paskalis, J.-P. G., & Cerutti, D. T. (2006). Time-sharing in pigeons: Independent effects of gap duration, position and discriminability from the timed signal. *Behavioural Processes*, *71*(2-3), 116–125. https://doi.org/10.1016/j.beproc.2005.10.006

Buonomano, D. V. (2000). Decoding Temporal Information: A Model Based on Short-Term Synaptic Plasticity. *The Journal of Neuroscience*, *20*(3), 1129–1141. https://doi.org/10.1523/JNEUROSCI.20-03-01129.2000

Buonomano, D. V., & Laje, R. (2010). Population clocks: Motor timing with neural dynamics. *Trends in Cognitive Sciences*, *14*(12), 520–527. https://doi.org/10.1016/j.tics.2010.09.002

Buonomano, D. V., & Mauk, M. D. (1994). Neural Network Model of the Cerebellum: Temporal Discrimination and the Timing of Motor Responses. *Neural Computation*, *6*(1), 38–55. https://doi.org/10.1162/neco.1994.6.1.38

Buschman, T. J., Siegel, M., Roy, J. E., & Miller, E. K. (2011). Neural substrates of cognitive capacity limitations. *Proceedings of the National Academy of Sciences*, *108*(27), 11252–11255. https://doi.org/10.1073/pnas.1104666108

Cabeza de Vaca, S., Brown, B. L., & Hemmes, N. S. (1994). Internal clock and memory processes in animal timing. *Journal of Experimental Psychology: Animal Behavior Processes*, *20*(2), 184–198. https://doi.org/10.1037/0097-7403.20.2.184

Cahoon, D., & Edmonds, E. M. (1980). The watched pot still won't boil: Expectancy as a variable in estimating the passage of time. *Bulletin of the Psychonomic Society*, *16*(2), 115–116. https://doi.org/10.3758/BF03334455

Casini, L., & Macar, F. (1997). Effects of attention manipulation on judgments of duration and of intensity in the visual modality. *Memory & Cognition*, *25*(6), 812–818. https://doi.org/10.3758/BF03211325

Chan, M., Ross, B., Earle, G., & Caplan, J. B. (2009). Precise instructions determine participants' memory search strategy in judgments of relative order in short lists. *Psychonomic Bulletin & Review*, *16*(5), 945–951. https://doi.org/10.3758/PBR.16.5.945

Coull, J. T., Vidal, F., Nazarian, B., & Macar, F. (2004). Functional Anatomy of the Attentional Modulation of Time Estimation. *Science*, *303*(5663), 1506–1508. https://doi.org/10.1126/science.1091573

Coull, J. T., Cheng, R.-K., & Meck, W. H. (2011). Neuroanatomical and Neurochemical Substrates of Timing. *Neuropsychopharmacology*, *36*(1), 3–25. https://doi.org/10.1038/npp.2010.113

Creelman, C. D. (1962). Human Discrimination of Auditory Duration. *The Journal of the Acoustical Society of America*, *34*(5), 582–593. https://doi.org/10.1121/1.1918172

Cueva, C. J., Saez, A., Marcos, E., Genovesio, A., Jazayeri, M., Romo, R., Salzman, C. D., Shadlen, M. N., & Fusi, S. (2020). Low-dimensional dynamics for working memory and time encoding. *Proceedings of the National Academy of Sciences*, 201915984. https://doi.org/10.1073/pnas.1915984117

Damsma, A., Schlichting, N., & van Rijn, H. (2021). Temporal Context Actively Shapes EEG Signatures of Time Perception. *The Journal of Neuroscience*, *41*(20), 4514–4523. https://doi.org/10.1523/JNEUROSCI.0628-20.2021

Damsma, A., Schlichting, N., van Rijn, H., & Roseboom, W. (2021). Estimating Time: Comparing the Accuracy of Estimation Methods for Interval Timing. *Collabra: Psychology*, *7*(1), 21422. https://doi.org/10.1525/collabra.21422

de Jong, J., Akyürek, E. G., & van Rijn, H. (2021). A common dynamic prior for time in duration discrimination. *Psychonomic Bulletin & Review*, *28*(4), 1183–1190. https://doi.org/10.3758/s13423-021-01887-z

de Jong, J., van Rijn, H., & Akyürek, E. G. (2023). Adaptive Encoding Speed in Working Memory. *Psychological Science*, 095679762311739. https://doi.org/10.1177/09567976231173902

Duggins, P., Stewart, T. C., Choo, X., & Eliasmith, C. (2017). The Effects of Guanfacine and Phenylephrine on a Spiking Neuron Model of Working Memory. *Topics in Cognitive Science*, *9*(1), 117–134. https://doi.org/10.1111/tops.12247

Egger, S. W., Le, N. M., & Jazayeri, M. (2020). A neural circuit model for human sensorimotor timing. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16999-8

Eichenbaum, H. (2014). Time cells in the hippocampus: A new dimension for mapping memories. *Nature Reviews Neuroscience*, *15*(11), 732–744. https://doi.org/10.1038/nrn3827

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A Large-Scale Model of the Functioning Brain. *Science*, *338*(6111), 1202–1205. https://doi.org/10.1126/science.1225266

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT Press.

Emmons, E. B., De Corte, B. J., Kim, Y., Parker, K. L., Matell, M. S., & Narayanan, N. S. (2017). Rodent Medial Frontal Control of Temporal Processing in the Dorsomedial Striatum. *The Journal of Neuroscience*, *37*(36), 8718–8733. https://doi.org/10.1523/JNEUROSCI.1376-17.2017

Enns, J. T., Brehaut, J. C., & Shore, D. I. (1999). The Duration of a Brief Event in the Mind's Eye. *The Journal of General Psychology*, *126*(4), 355–372. https://doi.org/10.1080/00221309909595371

Faber, M., & Gennari, S. P. (2017). Effects of learned episodic event structure on prospective duration judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1203–1214. https://doi.org/10.1037/xlm0000378

Forsberg, A., Guitard, D., & Cowan, N. (2021). Working memory limits severely constrain long-term retention. *Psychonomic Bulletin & Review*, *28*(2), 537–547. https://doi.org/10.3758/s13423-020-01847-z

Fortin, C., & Schweickert, R. (2016). Timing, working memory and expectancy: A review of interference studies. *Current Opinion in Behavioral Sciences*, *8*, 67–72. https://doi.org/10.1016/j.cobeha.2016.01.016

Fountas, Z., Sylaidi, A., Nikiforou, K., Seth, A. K., Shanahan, M., & Roseboom, W. (2022). A Predictive Processing Model of Episodic Memory and Time Perception. *Neural Computation*, *34*(7), 1501–1544. https://doi.org/10.1162/neco_a_01514

Franssen, V., & Vandierendonck, A. (2002). Time estimation: Does the reference memory mediate the effect of knowledge of results? *Acta Psychologica*, *109*(3), 239–267. https://doi.org/10.1016/S0001-6918(01)00059-2

French, R. M., Addyman, C., Mareschal, D., & Thomas, E. (2014). GAMIT – A Fading-Gaussian Activation Model of Interval-Timing: Unifying Prospective and Retrospective Time Estimation. *Timing & Time Perception Reviews*, *1*(2), 17.

Fung, B. J., Sutlief, E., & Hussain Shuler, M. G. (2021). Dopamine and the interdependency of time perception and reward. *Neuroscience & Biobehavioral Reviews*, *125*, 380–391. https://doi.org/10.1016/j.neubiorev.2021.02.030

Gavornik, J. P., Shuler, M. G. H., Loewenstein, Y., Bear, M. F., & Shouval, H. Z. (2009). Learning reward timing in cortex through reward dependent expression of synaptic plasticity. *Proceedings of the National Academy of Sciences*, *106*(16), 6826–6831. https://doi.org/10.1073/pnas.0901835106

Gershman, S. J., Moustafa, A. A., & Ludvig, E. A. (2014). Time representation in reinforcement learning models of the basal ganglia. *Frontiers in Computational Neuroscience*, *7*. https://doi.org/10.3389/fncom.2013.00194

Getty, D. J. (1975). Discrimination of short temporal intervals: A comparison of two models. *Perception & Psychophysics*, *18*(1), 1–8. https://doi.org/10.3758/BF03199358

Gibbon, J. (1977). Scalar Expectancy Theory and Weber's Law in Animal Timing. *Psychological Review*, *84*(3), 279–325.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar Timing in Memory. *Annals of the New York Academy of Sciences*, *423*(Timing and Time Perception), 52–77. https://doi.org/10.1111/j.1749-6632.1984.tb23417.x

Gibbon, J., Malapani, C., Dale, C. L., & Gallistel, C. (1997). Toward a neurobiology of temporal cognition: Advances and challenges. *Current Opinion in Neurobiology*, *7*(2), 170–184. https://doi.org/10.1016/S0959-4388(97)80005-0

Glaze, C. M., Kable, J. W., & Gold, J. I. (2015). Normative evidence accumulation in unpredictable environments. *eLife*, *4*, e08825. https://doi.org/10.7554/eLife.08825

Gosmann, J., & Eliasmith, C. (2020). CUE: A unified spiking neuron model of short-term and long-term memory. *Psychological Review*. https://doi.org/10.1037/rev0000250

Goudar, V., & Buonomano, D. V. (2018). Encoding sensory and motor patterns as time-invariant trajectories in recurrent neural networks. *eLife*, *7*, e31134. https://doi.org/10.7554/eLife.31134

Gouvêa, T. S., Monteiro, T., Motiwala, A., Soares, S., Machens, C., & Paton, J. J. (2015). Striatal dynamics explain duration judgments. *eLife*, *4*. https://doi.org/10.7554/eLife.11386

Grondin, S. (2014). About the (Non)scalar Property for Time Perception. In H. Merchant & V. de Lafuente (Eds.), *Neurobiology of Interval Timing* (pp. 17–32). Springer New York. https://doi.org/10.1007/978-1-4939-1782-2_2

Grossberg, S., & Schmajuk, N. A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, *2*(2), 79–102. https://doi.org/10.1016/0893-6080(89)90026-9

Gütig, R., & Sompolinsky, H. (2009). Time-Warp–Invariant Neuronal Processing. *PLoS Biology*, *7*(7), e1000141. https://doi.org/10.1371/journal.pbio.1000141

Hardy, N. F., & Buonomano, D. V. (2018). Encoding Time in Feedforward Trajectories of a Recurrent Neural Network Model. *Neural Computation*, *30*(2), 378–396. https://doi.org/10.1162/neco_a_01041

Hardy, N. F., & Buonomano, D. V. (2016). Neurocomputational models of interval and pattern timing. *Current Opinion in Behavioral Sciences*, *8*, 250–257. https://doi.org/10.1016/j.cobeha.2016.01.012

Hardy, N. F., Goudar, V., Romero-Sosa, J. L., & Buonomano, D. V. (2018). A model of temporal scaling correctly predicts that motor timing improves with speed. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-07161-6

Henke, J., Bunk, D., Von Werder, D., Häusler, S., Flanagin, V. L., & Thurley, K. (2021). Distributed coding of duration in rodent prefrontal cortex during time reproduction. *eLife*, *10*, e71612. https://doi.org/10.7554/eLife.71612

Herbst, S. K., van der Meer, E., & Busch, N. A. (2012). Attentional Selection Dilates Perceived Duration. *Per-*

ception, *41*(8), 883–900. https://doi.org/10.1068/p7300

Heys, J. G., & Dombeck, D. A. (2018). Evidence for a subcircuit in medial entorhinal cortex representing elapsed time during immobility. *Nature Neuroscience*, *21*(11), 1574–1582. https://doi.org/10.1038/s41593-018-0252-8

Hicks, R. E., Miller, G. W., & Kinsbourne, M. (1976). Prospective and Retrospective Judgments of Time as a Function of Amount of Information Processed. *The American Journal of Psychology*, *89*(4), 719. https://doi.org/10.2307/1421469

Hillyard, S. A., Vogel, E. K., & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: Electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *353*(1373), 1257–1270. https://doi.org/10.1098/rstb.1998.0281

Hopson, J. W. (1999). Gap timing and the spectral timing model. *Behavioural Processes*, *45*(1-3), 23–31. https://doi.org/10.1016/S0376-6357(99)00007-8

Howard, C. D., Li, H., Geddes, C. E., & Jin, X. (2017). Dynamic Nigrostriatal Dopamine Biases Action Selection. *Neuron*, *93*(6), 1436–1450.e8. https://doi.org/10.1016/j.neuron.2017.02.029

Howard, M. W., Shankar, K. H., Aue, W. R., & Criss, A. H. (2015). A distributed representation of internal time. *Psychological Review*, *122*(1), 24–53. https://doi.org/10.1037/a0037840

Ivry, R. B., & Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends in Cognitive Sciences*, *12*(7), 273–280. https://doi.org/10.1016/j.tics.2008.04.002

Jakob, A. M. V., Mikhael, J. G., Hamilos, A. E., Assad, J. A., & Gershman, S. J. (2022). Dopamine mediates the bidirectional update of interval timing. *Behavioral Neuroscience*, *136*(5), 445–452. https://doi.org/10.1037/bne0000529

James, W. (1890). *The Principles of Psychology* (Vol. 1 and 2). Henry Holt; Company.

Jeanneret, S., Bartsch, L. M., & Vergauwe, E. (2023). To be or not to be relevant: Comparing short- and long-term consequences across working memory prioritization procedures. *Attention, Perception, & Psychophysics*, *85*(5), 1486–1498. https://doi.org/10.3758/s13414-023-02706-4

Killeen, P. R., & Fetterman, J. G. (1988). A behavioral theory of timing. *Psychological Review*, *95*(2), 274–295. https://doi.org/10.1037/0033-295X.95.2.274

Killeen, P. R., & Grondin, S. (2021). A trace theory of time perception. *Psychological Review*. https://doi.org/10.1037/rev0000308

Kladopoulos, C. N., Hemmes, N. S., & Brown, B. L. (2004). Prospective timing under dual-task paradigms: Attentional and contextual-change mechanisms. *Behavioural Processes*, *67*(2), 221–233. https://doi.org/10.1016/j.beproc.2003.12.004

Komer, B., Stewart, T. C., Voelker, A. R., & Eliasmith, C. (2019). A neural representation of continuous space using fractional binding. *41st Annual Meeting of the Cognitive Science Society*.

Komura, Y., Tamura, R., Uwano, T., Nishijo, H., Kaga, K., & Ono, T. (2001). Retrospective and prospective coding for predicted reward in the sensory thalamus. *Nature*, *412*(6846), 546–549. https://doi.org/10.1038/35087595

Kononowicz, T. W., & Penney, T. B. (2016). The contingent negative variation (CNV): Timing isn't everything. *Current Opinion in Behavioral Sciences*, *8*, 231–237. https://doi.org/10.1016/j.cobeha.2016.02.022

Kononowicz, T. W., van Rijn, H., & Meck, W. H. (2018). Timing and Time Perception: A Critical Review of Neural Timing Signatures Before, During, and After the To-Be-Timed Interval. In J. T. Wixted (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1–38). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119170174.epcn114

Laje, R., & Buonomano, D. V. (2013). Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nature Neuroscience*, *16*(7), 925–933. https://doi.org/10.1038/nn.3405

Lejeune, H., & Wearden, J. H. (2006). Scalar Properties in Animal Timing: Conformity and Violations. *Quarterly Journal of Experimental Psychology*, *59*(11), 1875–1908. https://doi.org/10.1080/17470210600784649

Lerner, Y., Honey, C. J., Katkov, M., & Hasson, U. (2014). Temporal scaling of neural responses to compressed and dilated natural speech. *Journal of Neurophysiology*, *111*(12), 2433–2444. https://doi.org/10.1152/jn.00497.2013

Lewis, P., & Miall, R. (2009). The precision of temporal judgement: Milliseconds, many minutes, and beyond. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1525), 1897–1905. https://doi.org/10.1098/rstb.2009.0020

Liu, J. K., & Buonomano, D. V. (2009). Embedding Multiple Trajectories in Simulated Recurrent Neural Networks in a Self-Organizing Manner. *Journal of Neuroscience*, *29*(42), 13172–13181. https://doi.org/10.1523/JNEUROSCI.2358-09.2009

Liu, Y., Tiganj, Z., Hasselmo, M. E., & Howard, M. W. (2019). A neural microcircuit model for a scalable scale-invariant representation of time. *Hippocam-*

pus, *29*(3), 260–274. https://doi.org/10.1002/hipo.22994

Liverence, B. M., & Scholl, B. J. (2012). Discrete events as units of perceived time. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(3), 549–554. https://doi.org/10.1037/a0027228

Lositsky, O., Chen, J., Toker, D., Honey, C. J., Shvartsman, M., Poppenk, J. L., Hasson, U., & Norman, K. A. (2016). Neural pattern change during encoding of a narrative predicts retrospective duration estimates. *eLife*, *5:e16070*. https://doi.org/10.7554/eLife.16070

Luzardo, A., Ludvig, E. A., & Rivest, F. (2013). An adaptive drift-diffusion model of interval timing dynamics. *Behavioural Processes*, *95*, 90–99. https://doi.org/10.1016/j.beproc.2013.02.003

Macar, F., Grondin, S., & Casini, L. (1994). Controlled attention sharing influences time estimation. *Memory & Cognition*, *22*(6), 673–686. https://doi.org/10.3758/BF03209252

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal "Time Cells" Bridge the Gap in Memory for Discontiguous Events. *Neuron*, *71*(4), 737–749. https://doi.org/10.1016/j.neuron.2011.07.012

Martinelli, N., & Droit-Volet, S. (2022). Judgment of duration and passage of time in prospective and retrospective conditions and its predictors for short and long durations. *Scientific Reports*, *12*(1), 22241. https://doi.org/10.1038/s41598-022-25913-9

Mattes, S., & Ulrich, R. (1998). Directed attention prolongs the perceived duration of a brief stimulus. *Perception & Psychophysics*, *60*(8), 1305–1317. https://doi.org/10.3758/BF03207993

Matthews, W. J., & Grondin, S. (2012). On the replication of Kristofferson's (1980) quantal timing for duration discrimination: Some learning but no quanta and not much of a Weber constant. *Attention, Perception, & Psychophysics*, *74*(5), 1056–1072. https://doi.org/10.3758/s13414-012-0282-3

Matthews, W. J., & Meck, W. H. (2016). Temporal cognition: Connecting subjective time to perception, attention, and memory. *Psychological Bulletin*, *142*(8), 865–907. https://doi.org/10.1037/bul0000045

Matthews, W. J., Stewart, N., & Wearden, J. H. (2011). Stimulus intensity and the perception of duration. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 303–313. https://doi.org/10.1037/a0019961

McAdams, C. J., & Maunsell, J. H. R. (1999). Effects of Attention on Orientation-Tuning Functions of Single Neurons in Macaque Cortical Area V4. *The Journal of Neuroscience*, *19*(1), 431–441. https://doi.org/10.1523/JNEUROSCI.19-01-00431.1999

Mcclain, L. (1983). Interval estimation: Effect of processing demands on prospective and retrospective reports. *Perception & Psychophysics*, *34*(2), 185–189. https://doi.org/10.3758/BF03211347

Meck, W. H., Church, R. M., & Matell, M. S. (2013). Hippocampus, time, and memory—A retrospective analysis. *Behavioral Neuroscience*, *127*(5), 642–654. https://doi.org/10.1037/a0034201

Meck, W. H., Church, R. M., & Olton, D. S. (1984). Hippocampus, time, and memory. *Behavioral Neuroscience*, *98*(1), 3–22. https://doi.org/10.1037/0735-7044.98.1.3

Mello, G. B., Soares, S., & Paton, J. J. (2015). A Scalable Population Code for Time in the Striatum. *Current Biology*, *25*(9), 1113–1122. https://doi.org/10.1016/j.cub.2015.02.036

Merchant, H., Harrington, D. L., & Meck, W. H. (2013). Neural Basis of the Perception and Estimation of Time. *Annual Review of Neuroscience*, *36*(1), 313–336. https://doi.org/10.1146/annurev-neuro-062012-170349

Mikhael, J. G., & Gershman, S. J. (2019). Adapting the flow of time with dopamine. *Journal of Neurophysiology*, *121*(5), 1748–1760. https://doi.org/10.1152/jn.00817.2018

Mita, A., Mushiake, H., Shima, K., Matsuzaka, Y., & Tanji, J. (2009). Interval time coding by neurons in the presupplementary and supplementary motor areas. *Nature Neuroscience*, *12*(4), 502–507. https://doi.org/10.1038/nn.2272

Motanis, H., Seay, M. J., & Buonomano, D. V. (2018). Short-Term Synaptic Plasticity as a Mechanism for Sensory Timing. *Trends in Neurosciences*, *41*(10), 701–711. https://doi.org/10.1016/j.tins.2018.08.001

Murray, J. M., & Escola, G. S. (2017). Learning multiple variable-speed sequences in striatum via cortical tutoring. *eLife*, *6*. https://doi.org/10.7554/eLife.26084

Namboodiri, V. M., & Shuler, M. G. (2016). The hunt for the perfect discounting function and a reckoning of time perception. *Current Opinion in Neurobiology*, *40*, 135–141. https://doi.org/10.1016/j.conb.2016.06.019

Namboodiri, V. M. K., Mihalas, S., & Hussain Shuler, M. G. (2016). Analytical Calculation of Errors in Time and Value Perception Due to a Subjective Time Accumulator: A Mechanistic Model and the Generation of Weber's Law. *Neural Computation*, *28*(1), 89–117. https://doi.org/10.1162/NECO_a_00792

Nicolaï, C., Chaumon, M., & Van Wassenhove, V. (2024). Cognitive effects on experienced duration and speed of time, prospectively, retrospectively, in and out of

lockdown. *Scientific Reports*, *14*(1), 2006. https://doi.org/10.1038/s41598-023-50752-7

Nobre, A. C., & van Ede, F. (2017). Anticipated moments: Temporal structure in attention. *Nature Reviews Neuroscience*, *19*(1), 34–48. https://doi.org/10.1038/nrn.2017.141

Ogden, R., Salominaite, E., Jones, L., Fisk, J., & Montgomery, C. (2011). The role of executive functions in human prospective interval timing. *Acta Psychologica*, *137*(3), 352–358. https://doi.org/10.1016/j.actpsy.2011.04.004

Ossmy, O., Moran, R., Pfeffer, T., Tsetsos, K., Usher, M., & Donner, T. H. (2013). The Timescale of Perceptual Evidence Integration Can Be Adapted to the Environment. *Current Biology*, *23*(11), 981–986. https://doi.org/10.1016/j.cub.2013.04.039

Pardo-Vazquez, J. L., Castiñeiras-de Saa, J. R., Valente, M., Damião, I., Costa, T., Vicente, M. I., Mendonça, A. G., Mainen, Z. F., & Renart, A. (2019). The mechanistic foundation of Weber's law. *Nature Neuroscience*, *22*(9), 1493–1502. https://doi.org/10.1038/s41593-019-0439-7

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsáki, G. (2008). Internally Generated Cell Assembly Sequences in the Rat Hippocampus. *Science*, *321*(5894), 1322–1327. https://doi.org/10.1126/science.1159775

Paton, J. J., & Buonomano, D. V. (2018). The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions. *Neuron*, *98*(4), 687–705. https://doi.org/10.1016/j.neuron.2018.03.045

Pérez, O., & Merchant, H. (2018). The Synaptic Properties of Cells Define the Hallmarks of Interval Timing in a Recurrent Neural Network. *The Journal of Neuroscience*, *38*(17), 4186–4199. https://doi.org/10.1523/JNEUROSCI.2651-17.2018

Petter, E. A., Gershman, S. J., & Meck, W. H. (2018). Integrating Models of Interval Timing and Reinforcement Learning. *Trends in Cognitive Sciences*, *22*(10), 911–922. https://doi.org/10.1016/j.tics.2018.08.004

Phillips, I. (2012). ATTENTION TO THE PASSAGE OF TIME: *Attention to the Passage of Time*. *Philosophical Perspectives*, *26*(1), 277–308. https://doi.org/10.1111/phpe.12007

Piet, A. T., El Hady, A., & Brody, C. D. (2018). Rats adopt the optimal timescale for evidence integration in a dynamic environment. *Nature Communications*, *9*(1), 4265. https://doi.org/10.1038/s41467-018-06561-y

Polti, I., Martin, B., & van Wassenhove, V. (2018). The effect of attention and working memory on the estimation of elapsed time. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-25119-y

Poynter, W. D. (1983). Duration judgment and the segmentation of experience. *Memory & Cognition*, *11*(1), 77–82. https://doi.org/10.3758/BF03197664

Poynter, W. D., & Homa, D. (1983). Duration judgment and the experience of change. *Perception & Psychophysics*, *33*(6), 548–560. https://doi.org/10.3758/BF03202936

Predebon, J. (1996). The effects of active and passive processing of interval events on prospective and retrospective time estimates. *Acta Psychologica*, *94*(1), 41–58. https://doi.org/10.1016/0001-6918(95)00044-5

Rasmussen, D., Voelker, A., & Eliasmith, C. (2017). A neural model of hierarchical reinforcement learning. *PLOS ONE*, *12*(7), e0180234. https://doi.org/10.1371/journal.pone.0180234

Remington, E. D., Narain, D., Hosseini, E. A., & Jazayeri, M. (2018). Flexible Sensorimotor Computations through Rapid Reconfiguration of Cortical Dynamics. *Neuron*, *98*(5), 1005–1019.e5. https://doi.org/10.1016/j.neuron.2018.05.020

Reynolds, J. H., & Desimone, R. (1999). The Role of Neural Mechanisms of Attention in Solving the Binding Problem. *Neuron*, *24*(1), 19–29. https://doi.org/10.1016/S0896-6273(00)80819-3

Rivest, F., & Bengio, Y. (2011). Adaptive Drift-Diffusion Process to Learn Time Intervals [arXiv: 1103.2382]. *arXiv:1103.2382 [q-bio]*. Retrieved February 17, 2022, from http://arxiv.org/abs/1103.2382

Roberts, S., & Church, R. M. (1978). Control of an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes*, *4*(4), 318–337. https://doi.org/10.1037/0097-7403.4.4.318

Roseboom, W., Fountas, Z., Nikiforou, K., Bhowmik, D., Shanahan, M., & Seth, A. K. (2019). Activity in perceptual classification networks as a basis for human subjective time perception. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-018-08194-7

Salet, J. M., de Jong, J., & van Rijn, H. (2022). Still stuck with the stopwatch. *Behavioral Neuroscience*, *136*(5), 453–466. https://doi.org/10.1037/bne0000527

Seifried, T., & Ulrich, R. (2011). Exogenous visual attention prolongs perceived duration. *Attention, Perception, & Psychophysics*, *73*(1), 68–85. https://doi.org/10.3758/s13414-010-0005-6

Shankar, K. H., & Howard, M. W. (2010). Timing using temporal context. *Brain Research*, *1365*, 3–17. https://doi.org/10.1016/j.brainres.2010.07.045

Shankar, K. H., & Howard, M. W. (2012). A Scale-Invariant Internal Representation of Time. *Neural Computation*, *24*(1), 134–193. https://doi.org/10.1162/NECO_a_00212

Shea-Brown, E., Rinzel, J., Rakitin, B. C., & Malapani, C. (2006). A firing rate model of Parkinsonian deficits in interval timing. *Brain Research*, *1070*(1), 189–201. https://doi.org/10.1016/j.brainres.2005.10.070

Sherman, M. T., Fountas, Z., Seth, A. K., & Roseboom, W. (2022). Trial-by-trial predictions of subjective time from human brain activity (M. B. Cai, Ed.). *PLOS Computational Biology*, *18*(7), e1010223. https://doi.org/10.1371/journal.pcbi.1010223

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences*, *17*(11), 556–564. https://doi.org/10.1016/j.tics.2013.09.009

Shimbo, A., Izawa, E.-I., & Fujisawa, S. (2021). Scalable representation of time in the hippocampus. *Science Advances*, *7*(6). https://doi.org/10.1126/sciadv.abd7013

Simen, P., Balci, F., deSouza, L., Cohen, J. D., & Holmes, P. (2011a). A Model of Interval Timing by Neural Integration. *Journal of Neuroscience*, *31*(25), 9238–9253. https://doi.org/10.1523/JNEUROSCI.3121-10.2011

Simen, P., Balci, F., deSouza, L., Cohen, J. D., & Holmes, P. (2011b). Interval Timing by Long-Range Temporal Integration. *Frontiers in Integrative Neuroscience*, *5*. https://doi.org/10.3389/fnint.2011.00028

Simen, P., Rivest, F., Ludvig, E. A., Balci, F., & Killeen, P. (2013). Timescale Invariance in the Pacemaker-Accumulator Family of Timing Models. *Timing & Time Perception*, *1*(2), 159–188. https://doi.org/10.1163/22134468-00002018

Simen, P., Vlasov, K., & Papadakis, S. (2016). Scale (in)variance in a unified diffusion model of decision making and timing. *Psychological Review*, *123*(2), 151–181. https://doi.org/10.1037/rev0000014

Singh, R., & Eliasmith, C. (2006). Higher-Dimensional Neurons Explain the Tuning and Dynamics of Working Memory Cells. *Journal of Neuroscience*, *26*(14), 3667–3678. https://doi.org/10.1523/JNEUROSCI.4864-05.2006

Slayton, M. A., Romero-Sosa, J. L., Shore, K., Buonomano, D. V., & Viskontas, I. V. (2020). Musical expertise generalizes to superior temporal scaling in a Morse code tapping task. *PLOS ONE*, *15*(1), e0221000. https://doi.org/10.1371/journal.pone.0221000

Soares, S., Atallah, B. V., & Paton, J. J. (2016). Midbrain dopamine neurons control judgment of time. *Science*, *354*(6317), 1273–1277. https://doi.org/10.1126/science.aah5234

Sohn, H., Narain, D., Meirhaeghe, N., & Jazayeri, M. (2019). Bayesian Computation through Cortical Latent Dynamics. *Neuron*, *103*(5), 934–947.e5. https://doi.org/10.1016/j.neuron.2019.06.012

Spetch, M. L., & Wilkie, D. M. (1983). Subjective shortening: A model of pigeons' memory for event duration. *Journal of Experimental Psychology: Animal Behavior Processes*, *9*(1), 14–30. https://doi.org/10.1037/0097-7403.9.1.14

Staddon, J. E., & Higa, J. J. (1999). Time and memory: Towards a pacemaker-free theory of interval timing. *Journal of the Experimental Analysis of Behavior*, *71*(2), 215–251. https://doi.org/10.1901/jeab.1999.71-215

Stevens, S. S. (1956). The Direct Estimation of Sensory Magnitudes: Loudness. *The American Journal of Psychology*, *69*(1), 1. https://doi.org/10.2307/1418112

Stewart, T. C., Bekolay, T., & Eliasmith, C. (2012). Learning to Select Actions with Spiking Neurons in the Basal Ganglia. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00002

Stewart, T. C., & Eliasmith, C. (2014). Large-Scale Synthesis of Functional Spiking Neural Circuits. *Proceedings of the IEEE*, *102*(5), 881–898. https://doi.org/10.1109/JPROC.2014.2306061

Stöckel, A., & Eliasmith, C. (2021). Passive Nonlinear Dendritic Interactions as a Computational Resource in Spiking Neural Networks. *Neural Computation*, *33*(1), 96–128. https://doi.org/10.1162/neco_a_01338

Stöckel, A., Stewart, T. C., & Eliasmith, C. (2021). Connecting Biological Detail With Neural Computation: Application to the Cerebellar Granule–Golgi Microcircuit. *Topics in Cognitive Science*, *13*(3), 515–533. https://doi.org/10.1111/tops.12536

Taatgen, N. A., van Rijn, H., & Anderson, J. (2007). An integrated theory of prospective time interval estimation: The role of cognition, attention, and learning. *Psychological Review*, *114*(3), 577–598. https://doi.org/10.1037/0033-295X.114.3.577

Tiganj, Z., Jung, M. W., Kim, J., & Howard, M. W. (2017). Sequential Firing Codes for Time in Rodent Medial Prefrontal Cortex. *Cerebral Cortex*, *27*(12), 5663–5671. https://doi.org/10.1093/cercor/bhw336

Tiganj, Z., Singh, I., Esfahani, Z. G., & Howard, M. W. (2022). Scanning a compressed ordered representation of the future. *Journal of Experimental Psychology: General*, *151*(12), 3082–3096. https://doi.org/10.1037/xge0001243

Toda, K., Lusk, N. A., Watson, G. D., Kim, N., Lu, D., Li, H. E., Meck, W. H., & Yin, H. H. (2017). Nigrotectal Stimulation Stops Interval Timing in Mice. *Cur-*

rent Biology, *27*(24), 3763–3770.e3. https://doi.org/10.1016/j.cub.2017.11.003

Toso, A., Fassihi, A., Paz, L., Pulecchi, F., & Diamond, M. E. (2021). A sensory integration account for time perception (S. J. Gershman, Ed.). *PLOS Computational Biology*, *17*(1), e1008668. https://doi.org/10.1371/journal.pcbi.1008668

Treisman, M. (1963). Temporal discrimination and the indifference interval: Implications for a model of the "internal clock". *Psychological Monographs: General and Applied*, *77*(13), 1–31. https://doi.org/10.1037/h0093864

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends in Neurosciences*, *24*(5), 295–300. https://doi.org/10.1016/S0166-2236(00)01814-2

Tripp, B. P., & Eliasmith, C. (2010). Population Models of Temporal Differentiation. *Neural Computation*, *22*(3), 621–659. https://doi.org/10.1162/neco.2009.02-09-970

Tsao, A., Sugar, J., Lu, L., Wang, C., Knierim, J. J., Moser, M.-B., & Moser, E. I. (2018). Integrating time from experience in the lateral entorhinal cortex. *Nature*, *561*(7721), 57–62. https://doi.org/10.1038/s41586-018-0459-6

Ulrich, R., Bausenhart, K., & Wearden, J. H. (2022). Weber's Law for Timing and Time Perception: Reconciling the Poisson Clock with Scalar Expectancy Theory (SET). *Timing & Time Perception*, 1–31. https://doi.org/10.1163/22134468-bja10055

Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C., & Dehaene, S. (2012). A Temporal Bottleneck in the Language Comprehension Network. *Journal of Neuroscience*, *32*(26), 9089–9102. https://doi.org/10.1523/JNEUROSCI.5685-11.2012

van der Mijn, R., & van Rijn, H. (2021). Attention Does Not Affect the Speed of Subjective Time, but Whether Temporal Information Guides Performance: A Large-Scale Study of Intrinsically Motivated Timers in a Real-Time Strategy Game. *Cognitive Science*, *45*(3). https://doi.org/10.1111/cogs.12939

van Ede, F., Niklaus, M., & Nobre, A. C. (2017). Temporal Expectations Guide Dynamic Prioritization in Visual Working Memory through Attenuated \textalpha Oscillations. *The Journal of Neuroscience*, *37*(2), 437–445. https://doi.org/10.1523/JNEUROSCI.2272-16.2016

van Rijn, H. (2014). It's time to take the psychology of biological time into account: Speed of driving affects a trip's subjective duration. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01028

van Rijn, H. (2016). Accounting for memory mechanisms in interval timing: A review. *Current Opinion in Behavioral Sciences*, *8*, 245–249. https://doi.org/10.1016/j.cobeha.2016.02.016

van Rijn, H., Kononowicz, T. W., Meck, W. H., Ng, K. K., & Penney, T. B. (2011). Contingent negative variation and its relation to time estimation: A theoretical evaluation. *Frontiers in Integrative Neuroscience*, *5*. https://doi.org/10.3389/fnint.2011.00091

van Wassenhove, V. (2009). Minding time in an amodal representational space. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1525), 1815–1830. https://doi.org/10.1098/rstb.2009.0023

Voelker, A., Kajić, I., & Eliasmith, C. (2019). Legendre Memory Units: Continuous-Time Representation in Recurrent Neural Networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/952285b9b7e7a1be5aa7849f32ffff05-Paper.pdf

Voelker, A. R., & Eliasmith, C. (2018). Improving Spiking Dynamical Networks: Accurate Delays, Higher-Order Synapses, and Time Cells. *Neural Computation*, *30*(3), 569–609. https://doi.org/10.1162/neco_a_01046

Voelker, A. R. (2019). *Dynamical systems in spiking neuromorphic hardware* (Doctoral dissertation) [https://uwspace.uwaterloo.ca/handle/10012/14625]. University of Waterloo. Waterloo, ON.

Waldum, E. R., & Sahakyan, L. (2013). A role for memory in prospective timing informs timing in prospective memory. *Journal of Experimental Psychology: General*, *142*(3), 809–826. https://doi.org/10.1037/a0030113

Walker, J. A., Aswad, M., & Lacroix, G. (2022). The impact of cognitive load on prospective and retrospective time estimates at long durations: An investigation using a visual and memory search paradigm. *Memory & Cognition*, *50*(4), 837–851. https://doi.org/10.3758/s13421-021-01241-7

Wang, J., Hosseini, E., Meirhaeghe, N., Akkad, A., & Jazayeri, M. (2020). Reinforcement regulates timing variability in thalamus. *eLife*, *9*, e55872. https://doi.org/10.7554/eLife.55872

Wang, J., Narain, D., Hosseini, E. A., & Jazayeri, M. (2018). Flexible timing by temporal scaling of cortical responses. *Nature Neuroscience*, *21*(1), 102–110. https://doi.org/10.1038/s41593-017-0028-6

Wearden, J. H., & Ferrara, A. (1993). Subjective shortening in humans' memory for stimulus duration [Publisher: Routledge]. *The Quarterly Journal of Ex-*

perimental Psychology Section B, *46*(2), 163–186. https://doi.org/10.1080/14640749308401084

Wearden, J. H., Goodson, G., & Foran, K. (2007). Subjective shortening with filled and unfilled auditory and visual intervals in humans? *Quarterly Journal of Experimental Psychology*, *60*(12), 1616–1628. https://doi.org/10.1080/17470210601121916

Wearden, J. H., & Lejeune, H. (2008). Scalar Properties in Human Timing: Conformity and Violations. *Quarterly Journal of Experimental Psychology*, *61*(4), 569–587. https://doi.org/10.1080/17470210701282576

Wearden, J. (2015). Passage of time judgements. *Consciousness and Cognition*, *38*, 165–171. https://doi.org/10.1016/j.concog.2015.06.005

Wearden, J., Parry, A., & Stamp, L. (2002). Is Subjective Shortening in Human Memory Unique to Time Representations? *The Quarterly Journal of Experimental Psychology Section B*, *55*(1b), 1–25. https://doi.org/10.1080/02724990143000108

Wearden, J. H., & Ogden, R. S. (2021). Filled-Duration Illusions. *Timing & Time Perception*, *10*(2), 97–121. https://doi.org/10.1163/22134468-bja10040

Yamazaki, T., & Tanaka, S. (2005). Neural Modeling of an Internal Clock. *Neural Computation*, *17*(5), 1032–1058. https://doi.org/10.1162/0899766053491850

Yeshurun, Y., & Marom, G. (2008). Transient spatial attention and the perceived duration of brief visual events. *Visual Cognition*, *16*(6), 826–848. https://doi.org/10.1080/13506280701588022

Zakay, D. (1998). Attention allocation policy influences prospective timing. *Psychonomic Bulletin & Review*, *5*(1), 114–118. https://doi.org/10.3758/BF03209465

Zakay, D., & Block, R. A. (1995). An attentional-gate model of prospective time estimation. In M. Richelle, V. de Keyser, G. d'Ydewalle, & A. Vandierendonck (Eds.), *Time and the dynamic control of behavior* (pp. 167–178). University of Liege Press.

Zakay, D., Tsal, Y., Moses, M., & Shahar, I. (1994). The role of segmentation in prospective and retrospective time estimation processes. *Memory & Cognition*, *22*(3), 344–351. https://doi.org/10.3758/BF03200861

Zhou, S., Masmanidis, S. C., & Buonomano, D. V. (2020). Neural Sequences as an Optimal Dynamical Regime for the Readout of Time. *Neuron*, *108*(4), 651–658.e5. https://doi.org/10.1016/j.neuron.2020.08.020

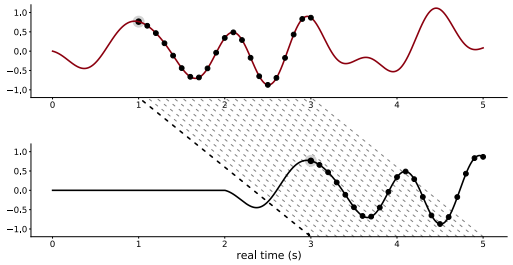
Appendix A

Legendre Delay Network

In the design of the Legendre Delay Network (LDN), we will argue from first principles: What would be the optimal algorithm for the problem of remembering the past[15]? First, let us specify what it means to remember a series of events as the ability to reproduce those events without distortion. The 'optimal' algorithm will be one that perfectly reproduces the past at an arbitrary time in the future without distortions. We also note that in the ideal case, this is true for continuous time: we cannot know in advance what moments in time are important or not, so we should apply our algorithm to every moment in time. However, storing all information at every continuous moment in time would require infinite resources, not just in a simulation of this algorithm, but in neural implementations. As this is impossible to attain, our algorithm needs to satisfy the constraint of *finite* resources.

Figure A1

Delaying a continuous-time input by two seconds



Note. This figure illustrates the challenge of delaying an input. The top panel shows a continuous-time signal that is delayed for two seconds, as shown in the bottom panel. While individual dots and lines represent how inputs are delayed, this is only done for clarity: The input is continuous in time, and therefore an infinite number of dots and lines would need to be displayed.

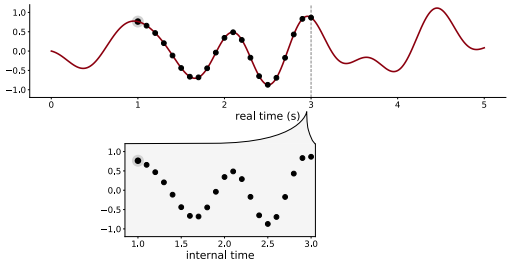
A quantification of this approach is shown in Figure A1 and A2. The continuous input events are $u(t)$ (see Figure A1). A perfect reproduction of those events at some later time would be $x(t) = u(t - \theta)$, where θ is the length of time between 'now' and that 'later time'. Essentially, we are storing the signal for the period θ. We can think of the system as 'delaying' the input by θ, providing an exact copy shifted by θ (see Figure A1). Notably, $x(t)$ would need to have infinite dimensions if it were to delay a single continuous-time input, regardless of how long we want to delay that input (see Figure A2).

The Legendre Delay Network (LDN) is a system that optimally solves this challenge given finite resources (Voelker et al., 2019). Specifically, it has provably the small-

[15] For a complete derivation (which is based on taking the Padé approximation of the Laplace transform of a pure delay ($u(t - \theta)$) and discussion of Legendre Memory, see Voelker (2019).

Figure A2

Delaying an input for two seconds using infinite memory capacity



Note. One way to solve the delay challenge is to store each input for exactly two seconds. However, if this solution is applied for continuous-time signals, we would need to store an infinite number of inputs, regardless of the length of the delay.

est possible error given a specific number of resources compared to a perfect delay. These resources can be thought of as neurons in a neural network or, in the context of the LDN, as dimensions in a function space. Specifically, instead of implementing a perfect delay with an infinite storage capacity ($x(t) = u(t-\theta)$) we *approximate* this delay using a finite function space $\mathcal{P}$ that is defined over the interval $(t, t-\theta)$:

$$\sum_{i=0}^{d-1} \mathcal{P}_i\left(\frac{\theta'}{\theta}\right) x_i(t) \approx u(t-\theta) \tag{3}$$

where d is the highest dimension in the function space (i.e., the order of our approximation, or the number of resources available), θ is the length of the delay and θ' are the values between 0 and θ, and x_i is the vector containing coefficients on the function space.

The dimensions in function space that are optimal for approximating a delay are Shifted Legendre polynomials: $\mathcal{P}$:

$$\mathcal{P}_i(r) = (-1)^i \sum_{j=0}^{i} \binom{i}{j}\binom{i+j}{j}(-r)^j \tag{4}$$

These Legendre polynomials can be interpreted as temporal basis functions that represent a rolling window of the last θ seconds of input history, similarly to how sines and cosines can form a basis for signals in the frequency domain. We can then approximate input history by taking a linear combination (i.e., weighted sum) of the temporal basis functions. The coefficients on each polynomial (i.e., the weights, x), are generated so that their sum is an approximation of input history (see Figure 2). The formal dynamical system that generates x on-the-fly is described in the main text ().

Appendix B

Neural Engineering Framework

This algorithm can be implemented in the Neural Engineering Framework (NEF) (Eliasmith & Anderson, 2003), as implemented in the python library Nengo (Bekolay, Bergstra, et al., 2014). The NEF uses three general principles to implement computations in neural networks: representation, transformation, and dynamics (for detailed discussions of these principles, see Eliasmith & Anderson, 2003; Stewart & Eliasmith, 2014; Stöckel & Eliasmith, 2021; Voelker & Eliasmith, 2018). The NEF has been used to build models of working memory (Duggins et al., 2017; Gosmann & Eliasmith, 2020; Singh & Eliasmith, 2006), long-term memory (Gosmann & Eliasmith, 2020), attention (Bobier et al., 2014), action selection (Stewart et al., 2012) and reinforcement learning (Rasmussen et al., 2017), and a large-scale cognitive architecture, SPAUN (Eliasmith et al., 2012). Notably, the NEF has also been used to construct models that are able to track time (Bekolay, Laubach, et al., 2014; Singh & Eliasmith, 2006; Stöckel et al., 2021). For instance, Bekolay, Laubach, et al. (2014) constructed a 'double-integrator' network that was able to track elapsed time, which optimized task performance in a simple reaction-time experiment. Stöckel et al. (2021) have used the NEF to construct a biologically detailed model of the cerebellar circuits underlying eyeblink conditioning, using the LDN. Here, we briefly review the methods of the NEF focusing on those aspects that are relevant to the UTC model.

Principle 1: Representation

The NEF assumes that the representation of a variable can be described in terms of its encoding and decoding. The encoding process describes how an input $\mathbf{x}$ is captured by the system. In our case, the input is captured by spiking neural activity. For decoding, the system needs to explain how, given the neural activity a that is *encoding* $\mathbf{x}$, a downstream neuron can have access to the value of $\mathbf{x}$. Together, encoding and decoding define the representation of the variable.

Our algorithm has several variables that we would like to represent with the activity of spiking neurons. For instance, the coefficients $\mathbf{x}(t)$ are explicitly represented and updated by the algorithm. Principle 1 in the NEF, Representation, provides methods for capturing such representations. Specifically, if we want to represent a vector $\mathbf{x}$ with neurons, the activity of the neurons $\mathbf{a}$ should reflect changes in $\mathbf{x}$ over time. We can describe this relationship as follows:

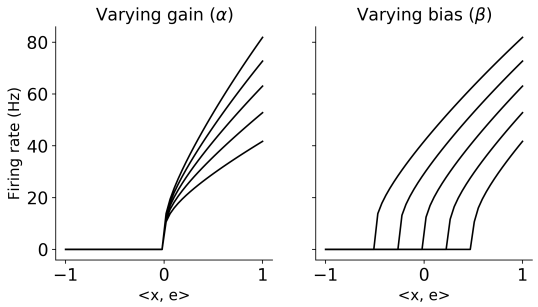
$$a_i(\mathbf{x}) = G_i[J_i(\mathbf{x})] \tag{5}$$

where $G_i[.]$ is a non-linear function that generates neural spikes and J_i is the input current to the neuron's soma. NEF allows for explaining how such abstractions map to biological processes, like somatic currents, neural spiking, tuning curves, and synaptic transmission.

To begin, we will discuss how neurons *encode* information into spike trains. Representing $\mathbf{x}$ requires that the neurons should be sensitive to changes in $\mathbf{x}$. This is implemented by assuming that each neuron i is associated with a preferred input, represented by a randomly chosen unit-length encoder $\mathbf{e}_i$. The more similar $\mathbf{x}$ is to $\mathbf{e_i}$, the higher the input current (J_i) received by neuron i, and thus the higher its firing rate. The unit-length encoders can be chosen to match recorded neuron tuning curves, but for this model, we use the default method of choosing these randomly.

Figure B1

Gain (α) and bias (β) determine the slope and intercept of neural tuning curves



Note. The input scalar $\mathbf{x}$ is plotted on the x-axis and firing rate (in Hz.) is plotted on the y-axis.

The current J_i is determined by a randomly chosen gain α_i, which determines the slope of the response function, and bias β_i, which determines the intercept of the response function (Figure B1). As for the encoders, these could be set to match known neuron tuning in future work. The full equation for the current driving the neural nonlinearity is thus:

$$J_i(\mathbf{x}) = \alpha_i \langle \mathbf{x}, \mathbf{e_i} \rangle + \beta_i \tag{6}$$

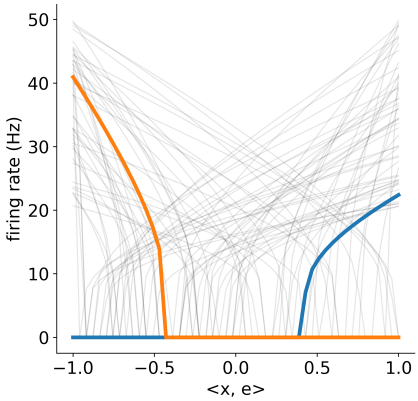
where $\langle . \rangle$ is the dot product (Figure B2).

As shown in Figure B3, these equations and random choices of neuron parameters serve to capture the known heterogeneity of neural systems. Distributing parameters in this manner makes it feasible to test possible neuron responses in specific applications. For instance, we will show later that assuming heterogeneous tuning curves captures the observed heterogeneity of firing patterns in timing experiments (see section Changes in window size explain temporal scaling in complex neural patterns).

In all simulations reported in this manuscript, we use the leaky integrate-and-fire (LIF) neuron model (but the NEF extends to more complex neuron models, see (Duggins et al., 2017)). In the LIF neuron, each time the membrane voltage V crosses some threshold V_{thresh}, the neuron generates a spike and resets to its resting state for the duration of the refractory period τ_{ref}. We can represent the spiking activity of a neuron as a series of delta functions, where each delta

Figure B2

Tuning curves for individual neurons in a population



Note. Heterogeneity in encoders (sensitivity to direction of $\mathbf{x}$, i.e., positive or negative values), intercepts (values when the neuron starts firing) and gains (slope of tuning curves) allow for efficient representation of $\mathbf{x}$. For instance, the orange neuron is sensitive to negative values, whereas the blue neuron starts firing more when positive values are represented. The orange neuron has a steeper slope (i.e., a higher gain) and also a different intercept (i.e., different bias).

function is located at a spike time (Figure B3), t_m, giving the train of spikes as:

$$a_i = \sum_m \delta(t - t_m) \tag{7}$$

Given the neural activity a that is *encoding* $\mathbf{x}$, downstream neurons should be able to infer the value of $\mathbf{x}$. The process of the downstream neuron extracting that information is called *decoding*. However, downstream neurons do not have direct access to spiking activity of upstream neurons. Notably, as spikes arrive at the end of the sending neuron's axon, they result in neurotransmitters being released across the synaptic cleft, which induces a post-synaptic current (PSC) on the dendrites of the receiving neuron. We model this PSC as a simple exponentially decaying current:

$$h(t) = e^{-t/\tau_{\text{PSC}}} \tag{8}$$

where τ_{PSC} is the time-constant of the PSC's decay. Biologically, τ_{PSC} depends on how long the ion channels (of the post-synaptic neuron) opened by the neurotransmitter remain in their open state. For instance, τ_{PSC} is around 5ms for AMPA, 10ms for GABA, and 100ms for NMDA receptors. Given some spike train that encodes $\mathbf{x}$, we can characterize the raw, unweighted current that could be injected into a cell as:

$$a_i = \sum_m h(t - t_m). \tag{9}$$

This equation can be thought of as applying the PSC model to each spike as it arrives and summing up the result (Figure B3). Notably, this filtered input does not map to observable currents, but is instead a useful theoretical construct that captures temporal decoding. In other words, it describes what kind of temporal variability is evident in the incoming spike train from one presynaptic cell.

Where the above describes the input of a single cell, $\mathbf{x}$ is obviously encoded by a *population* of cells. Decoding of a representation that may be distributed over a population of neurons requires 'spatial' decoding. Because different neurons may encode different parts of the $\mathbf{x}$ space, considering all of them is essential for fully decoding the information in the encoding population. Specifically, the NEF suggests how to solve for a set of optimal spatial decoders $\mathbf{d}$, using regularized least-squares optimization, while taking into account some level of noise. Combining the spatial and temporal decoding that optimally decodes out an estimate of our original vector $\mathbf{x}$ gives the estimate $\hat{\mathbf{x}}$, that is:

$$\hat{\mathbf{x}} = \sum_{i,m}^{N,M} h(t - t_{im})\mathbf{d}_i \tag{10}$$

where N is the number of neurons in the encoding population, M is the number of spikes, i indexes the neurons, m indexes the spikes, $h(t)$ is the PSC of the receiving neuron, $\hat{\mathbf{x}}$ is an estimate of our original input vector $\mathbf{x}$ and $\mathbf{d}_i$ is a decoder for neuron i to optimally represent $\mathbf{x}$ (Figure B3).

As shown in Figure B3, we have fully mapped the process of encoding and decoding an input variable $\mathbf{x}$ to neurobiological processes, thus specifying an implementation method for the representations in our algorithm.

Principle 2: Transformation

Principle 2 of the NEF, Transformation, describes how to implement linear and nonlinear computations with the represented variables. That is, it specifies how a neurobiological system can transform some representation of $\mathbf{x}$ to some function of $\mathbf{x}$, $f(\mathbf{x})$. Usefully, Principle 1, Representation, is a special case of Principle 2, Transformation: When we represented a vector $\mathbf{x}$, we defined the loss function of our regularized-least squared problem as the difference between $\mathbf{x}$ and our estimate $\hat{\mathbf{x}}$. However, we can think of transforming $\mathbf{x}$, as decoding out a certain function of $\mathbf{x}$, $f(\mathbf{x})$ from our neural activity. Hence, the loss function of our regularized least-squares problem becomes the difference between $f(\hat{\mathbf{x}})$ and $f(\mathbf{x})$. The resulting decoders $\mathbf{d}^f$ will compute the function $f(\mathbf{x})$ (Figure B3).

The above describes a general method for computing arbitrary functions of the variables represented using Principle 1, without introducing new neurobiological mappings. That is, the connection weights between neurons in subsequent populations combine the encoding and decoding/transformations of the NEF. Hence, connection weights can be analytically derived given the desired function $f(\mathbf{x})$, without assuming that encoders and decoders have some neurobiological analogue.

Principle 3: Dynamics

To this point, we have described how to characterize the transformations and representations in our algorithm. However, we have not yet described how to implement the core linear dynamical system (Equation 1). Principle 3 of the NEF exploits Principles 1 and 2 to implement arbitrary dynamical systems (of which our network, see equation 1, is an instance) in spiking neurons:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t) \tag{11}$$

Principle 3, Dynamics, tells us how to implement a dynamical system of this form in a recurrent spiking neural network. Under mild assumptions, the dynamics of the spiking recurrent neural network are dominated by the PSC ($h(t) = e^{-t/\tau_{\text{PSC}}}$). In other words, we can treat the PSC as the dynamical primitive of our dynamical system (Eliasmith & Anderson, 2003). To get the dynamics defined in Equation 1, we need to map the equation with the dynamical primitive of integration (Equation 1) to an equivalent equation with a dynamical primitive of the PSC. Or, more intuitively speaking, we need to take into account the fact that information decays over time because of the PSC, while it is perfectly remembered by integration. To illustrate, when a spike would arrive at some process that integrates perfectly, no information would be forgotten. But when that same spike arrives at a post-synaptic neuron, the information that this spike carries is lost over time because of the PSC (see Figure B3). Therefore, we have to figure out how strongly we should 'remind' the state-vector of its previous state to precisely counteract this PSC 'forgetting'. In order to do this, we should adjust the dynamics matrix $\mathbf{A}$ and input matrix $\mathbf{B}$. For linear systems, this adjustment is proposed by the NEF as follows (for a full derivation, see Eliasmith & Anderson, 2003)):

$$\mathbf{A}' = \tau\mathbf{A} + \mathbf{I} \tag{12}$$

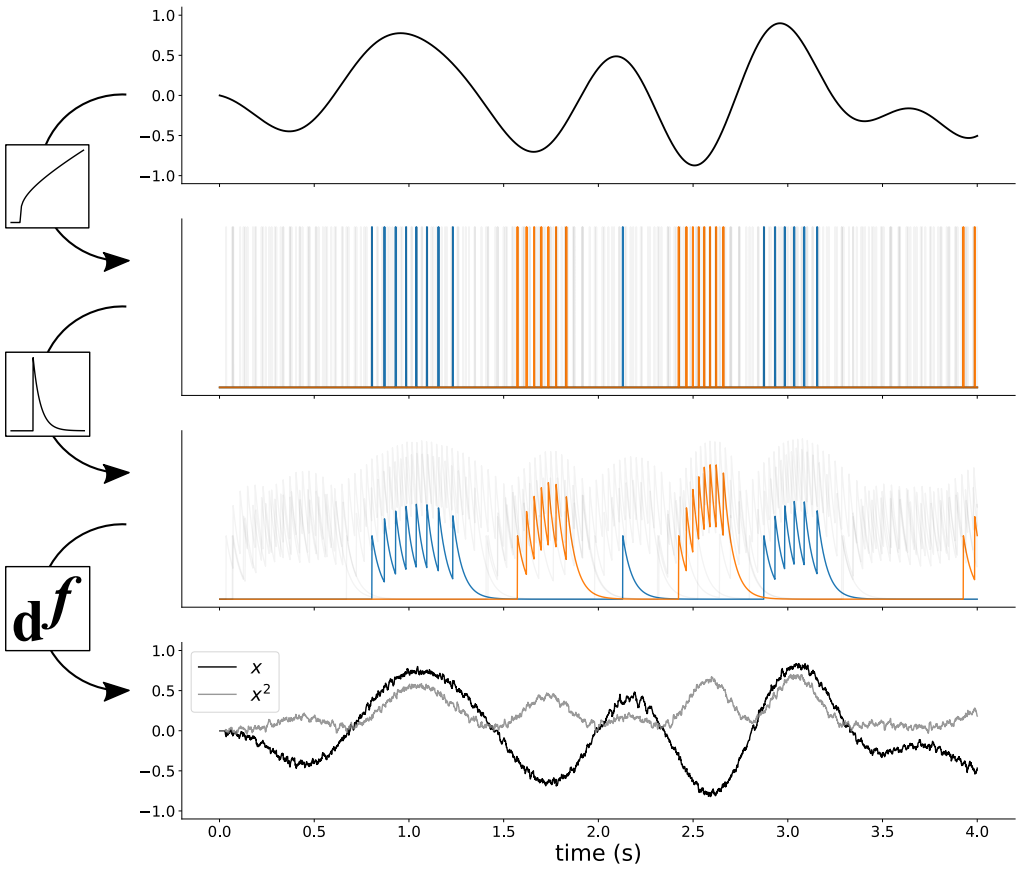
$$\mathbf{B}' = \tau\mathbf{B} \tag{13}$$

where $\mathbf{A}'$ is our neural recurrent transformation and $\mathbf{B}'$ is our neural input transformation.

For our specific network, we want to be able to control θ on-the-fly as well, in order to encode intervals that vary widely in timescale. If we solve for $\dot{\mathbf{x}}$ in Equation 1, we see that $\mathbf{A}$ and $\mathbf{B}$ should be multiplied by θ^{-1}. That is, the recurrent gain on $\mathbf{A}$ and $\mathbf{B}$ is inversely proportional to θ (Voelker, 2019). For instance, if we want θ to be 2 seconds, we should multiply the recurrent gain we already have (τ) with a multiplication factor ($\theta^{-1} = 0.5$). We introduce a neural population θ^{-1} that represents this multiplication factor on the recurrent gain. This allows us to adaptively control the recurrent gain and therefore θ (Figure 4).

Figure B3

Representation and Transformation with the Neural Engineering Framework (NEF)



Note. We feed an input signal (top row) into a population of spiking neurons. The tuning curves describe how this input drives the spiking frequency of individual neurons. For instance, the blue neuron is sensitive to positive inputs and the orange neuron to negative inputs (second row; also see previous figure). A downstream population of neurons receive post-synaptic potentials (PSC) for which we use a lowpass filter (third row). The original signal (black) or a transformation of the original signal (x^2; grey curve) can be read out by applying an optimal set of decoders $\mathbf{d}^{\mathbf{f}}$.

This completes our characterization of the implementation of all elements of our algorithm using the NEF. The resulting model is a recurrent neural network consisting only of standard LIF spiking neurons with connection weights between them that are determined by the **A**, **B**, **d**, and **e** matrices, and a simple exponential synapse model.

Appendix C
Modelling Buhusi (2012)

For the fit to Buhusi (2012), we need to model how objective sound intensity maps onto the vectors that our neural populations represent. First, we assumed a power-law mapping between sound intensity (I, in W/m^2) to subjective loudness (L; Stevens, 1956):

$$L = kI^m \tag{14}$$

where k is a free scaling parameter that was fit individually for each experiment (experiment 1: $k = 2.5$; experiment 2: $k = 5$), and we used the estimated exponent $m = 0.09$ from (Pardo-Vazquez et al., 2019), who modeled sound intensity discrimination in rats using a power-law

function. Subsequently, we converted the experimental values (40 - 100 dB) on this subjective loudness scale to Spatial Semantic Pointers (Komer et al., 2019). These vectors can represent continuous values by exponentiating a vector with a real value (in our case, subjective loudness):

$$SSP = \mathbf{x}^L \tag{15}$$

where SSP is a spatial semantic pointer and $\mathbf{x}$ is a unitary vector (which doesn't change length when circular convolution is applied). Vectors are exponentiated by first taking their Fourier transform $\mathcal{F}\{\cdot\}$, then doing an element-wise exponentiation on those complex numbers and then doing the inverse Fourier transform:

$$\mathbf{x}^L = \mathcal{F}^{-1}\{\mathcal{F}\{\mathbf{x}\}^L\} \tag{16}$$

These vectors, in turn, exhibit is a smooth function of subjective similarity with respect to physical similarity (Komer et al., 2019).