A predictive processing model of episodic memory and time perception

Zafeirios Fountas[1,2,*], Anastasia Sylaidi[3], Kyriacos Nikiforou[4], Anil K. Seth[5,6], Murray Shanahan[4], Warrick Roseboom[5]

1 Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, London, UK
2 Emotech Labs, London, UK
3 Spike AI Research Labs, London, UK
4 Department of Computing, Imperial College London, London, UK
5 Department of Informatics and Sackler Centre for Consciousness Science, University of Sussex, Sussex, UK
6 Canadian Institute for Advanced Research (CIFAR) Program on Brain, Mind, and Consciousness, Toronto, Ontario, Canada

* fountas@outlook.com

Abstract

Human perception and experience of time is strongly influenced by ongoing stimulation, memory of past experiences, and required task context. When paying attention to time, time experience seems to expand; when distracted, it seems to contract. When considering time based on memory, the experience may be different than in the moment, exemplified by sayings like "time flies when you're having fun". Experience of time also depends on the content of perceptual experience – rapidly changing or complex perceptual scenes seem longer in duration than less dynamic ones. The complexity of interactions between attention, memory, and perceptual stimulation is a likely reason that an overarching theory of time perception has been difficult to achieve. Here, we introduce a model of perceptual processing and episodic memory that makes use of hierarchical predictive coding, short-term plasticity, spatio-temporal attention, and episodic memory formation and recall, and apply this model to the problem of human time perception. In an experiment with $\sim 13,000$ human participants we investigated the effects of memory, cognitive load, and stimulus content on duration reports of dynamic natural scenes up to ~ 1 minute long. Using our model to generate duration estimates, we compared human and model performance. Model-based estimates replicated key qualitative biases, including differences by cognitive load (attention), scene type (stimulation), and whether the judgement was made based on current or remembered experience (memory). Our work provides a comprehensive model of human time perception and a foundation for exploring the computational basis of episodic memory within a hierarchical predictive coding framework.

Author summary

Experience of the duration of present or past events is a central aspect of human experience, the underlying mechanisms of which are not yet fully understood. In this work, we combine insights from machine learning and neuroscience to propose a

combination of mathematical models that replicate human perceptual processing, long-term memory, attention, and duration perception. Our computational implementation of this framework can process information from video clips of ordinary life scenes, record and recall important events, and report the duration of these clips. To assess the validity of our proposal, we conducted an experiment with $\sim 13,000$ human participants. Each was shown a video between 1-64 seconds long and reported how long they believed it was. Reports of duration by our computational model qualitatively matched these human reports, made about the exact same videos. This was true regardless of the video content, whether time was actively judged or based on memory of the video, or whether the participants focused on a single task or were distracted - all factors known to influence human time perception. Our work provides the first model of human duration perception to incorporate these diverse and complex factors and provides a basis to probe the deep links between memory and time in human experience.

Introduction

The ability to estimate temporal properties of the world, such as how much time has elapsed since you started reading this paper, is key to complex cognition, planning, and behaviour. Human perception of time is affected by many factors; experience on the scale of seconds is prominently influenced by the content, complexity, and rate of change of experience [1–4]. The influence of stimulus complexity and rate of change can be seen in cases driven by very basic stimulus properties, such as the temporal frequency [5, 6], and can also be seen in very coarse, abstract differences in content. For example, human participants report videos of scenes filmed in a busy city as longer than those filmed walking around in the countryside, with both reported as longer than scenes of a quiet office [4, 7]. Time perception is also influenced by whether attention can and is being directed toward the task of monitoring time [8–10]. Specifically, increasing cognitive load (by requiring concurrent attention to additional tasks beyond tracking time) has been reported to decrease the apparent duration of an interval when a person is aware that attention to time is required for a given task (prospective time), while apparent duration is increased by cognitive load when reflecting on the duration of an interval after it has occurred (retrospective time; comprehensively reviewed in [10] and Figure 1). Differences in time perception based on this interaction between cognitive load and prospective versus retrospective duration judgements have led to suggestions that different mechanisms underlie the different scenarios: when actively attending to time (prospective time) the process is largely "attention" driven; when reflecting on a period of time after it has occurred (retrospective time) the process is largely driven by "memory" [8, 10–12]. Reflecting these different approaches, models of time perception and memory have developed largely independently since diverging several decades ago (approach in [13] versus [14], for example), precluding a unified account of time perception.

Here we present a model of human perceptual processing, memory, and time perception. Through this model, the influences of stimulus, attention, and content of memory on human time estimation can be integrated, via the two intertwined processes of memory formation and recall, providing a single overarching understanding of their relationship. We validate model performance in the domain of time perception by showing that model-based estimates of the duration of naturalistic videos between 1 and 64 seconds match duration reports obtained from $\sim 13,000$ human participants ($n = 6,977$ for the main results after removing all outliers) regarding the same videos for each of the key manipulations: knowingly tracking time (prospective judgements) or judging it from memory (retrospective judgements), attending to only a single or multiple concurrent tasks (cognitive load), and for different types of natural stimulation

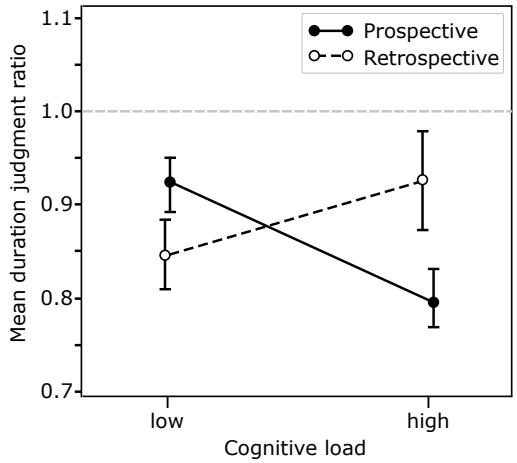


Fig 1. Relationship between prospective and retrospective duration judgments and their interaction with cognitive load (adapted from [10]). Duration judgement ratio is reported duration divided by physical duration. Error bars indicate standard error of the mean. Increasing cognitive load decreases reported duration in a prospective timing task, but increases reported duration in a retrospective timing task.

(busy city scenes, less busy scenes on a university campus or leafy surrounds, or in a quiet office of cafe).

Time perception via perceptual classification and long-term memory

The foundation of the present model rests upon the proposal that the primary function of human perception is the classification of objects and events in the world and that time can be estimated simply by tracking the behaviour of neural networks involved in perceptual classification. Using deep convolutional neural networks (CNN) as a model of biological perceptual classification has become common in recent years (see [15] for review). CNNs have a hierarchical structure both inspired by and broadly resembling the human/primate visual cortex [16–18], and have been shown to perform well in visual classification tasks [19, 20]. While reasonable discussion continues about the extent of similarity between biological classification networks and specific artificial network architectures [21, 22], the constellation of features exhibited by CNNs has been shown to have value for investigating the bases of human perception and cognition [15, 23].

In support of the proposal that time can be estimated on the basis of activity in perceptual classification networks, a recent study [4] demonstrated that by simply tracking salient events in the activity of a CNN while it processed videos of natural scenes (from 1-64 seconds in duration) it was possible to generate a basis for prospective estimates of duration. In that study, model-based duration estimates replicated key features of human estimates for the exact same videos, including biases related to scene type (busy city scenes estimated as longer than quiet office scenes, for example; see also [7, 24]). This model displayed conceptual similarities with the predictive processing account of perception wherein perception is proposed to operate as a function of both sensory predictions and current sensory stimulation, with perceptual content understood as the brain's "best guess" (Bayesian posterior) of the causes of current sensory input given the prior expectations or predictions [25–29].

However, this previous model [4] had only very simple "memory" that marked the occurrence of salient events in perception. Because this model lacks the ability to form

content-specific memories of an episode, it is impossible to use it to estimate time in *retrospect*, where the relevant content of memory (events that have occurred - taking a bus; seeing a cow) must be used to form an estimate. In order to resolve how human retrospective judgements of time might be accomplished, we extended on the core aspects of this previous model to provide the ability to record episodes of perceptual content as memory.

To understand the relation between memory and time in perception we first need to define a theoretical model of long-term memory. Explicit long-term memory is often divided into "semantic" and "episodic", reflecting the difference between "knowing" and "remembering" [30]. Whereas semantic memory maintains generalized (statistical) information about different concepts in the world, including taxonomic (i.e. similarity-based) and thematic (i.e. contiguity-based) relations [31], episodic memory stores individual events, including specific temporal contiguity relations between them [32]. These two memory systems are believed to be based on highly interdependent mechanisms, during both formation and recall [33]. In what follows, we first present a novel Bayesian model of episodic and semantic memory, defined under the predictive processing framework. We then focus on the proposed mechanisms for basic memory operations and their role in generating the inner sense of duration. Next, we compare the performance of this proposed model against human reports of duration in the order of seconds (up to 64 seconds). Finally, we discuss the relationship between our new model and prominent work on the neural foundations of episodic memory in event segmentation and temporal hierarchy in the brain.

Theoretical framework

Predictive processing neural architecture for perception

According to (Bayesian) predictive processing theories, the brain is constantly updating its internal representations of the state of the world with new sensory information x_t^0 (at time t) through the process of hierarchical probabilistic inference [25, 28, 34, 35]. The resulting 'top-down' generative model is used to predict new states and simulate the outcome of prospective actions [36]. Comparing predicted against actual sensory signals gives rise to prediction errors ξ_t^n which, in most instantiations of predictive processing, are assumed to flow in a 'bottom-up' direction from the sensory periphery ($n = 0$), towards higher hierarchical levels $n \in \{1, 2, ..\}$. Numerous connections with neurophysiology have been made over the years, including proposals viewing predictive processing as the fundamental function of canonical microcircuits in the cortex [37], or as the result of the interplay between competing oscillatory bands [38, 39], as well as providing theoretical explanations for a wide range of cognitive and perceptual effects such as hallucinations [40], autism [41] and schizophrenia [42]. In addition, this framework is uniquely qualified for studying the relation between perceptual change and human perception of time [4], as it provides a working explanation of the complex interplay between prior beliefs, sensory-based information and learning.

In this study, we define a predictive processing model that relies on feed-forward deep neural networks for the bottom-up flow of information (see *inference model* in Figure 2) and a novel stochastic process to account for the top-down flow of predictions (see *generative model* in the same figure). Unlike famous approaches that use neural networks to implement inference via amortization [43, 44], here neural activations of each hierarchical layer n represent single samples of the random variable x_t^n, rather than its distribution parameters. Hence, existing, pre-trained, feed-forward neural networks can be used to propagate bottom-up sensory signals or, in this case, to approximate the inference $P(x_t^n|x_t^{n-1})$.

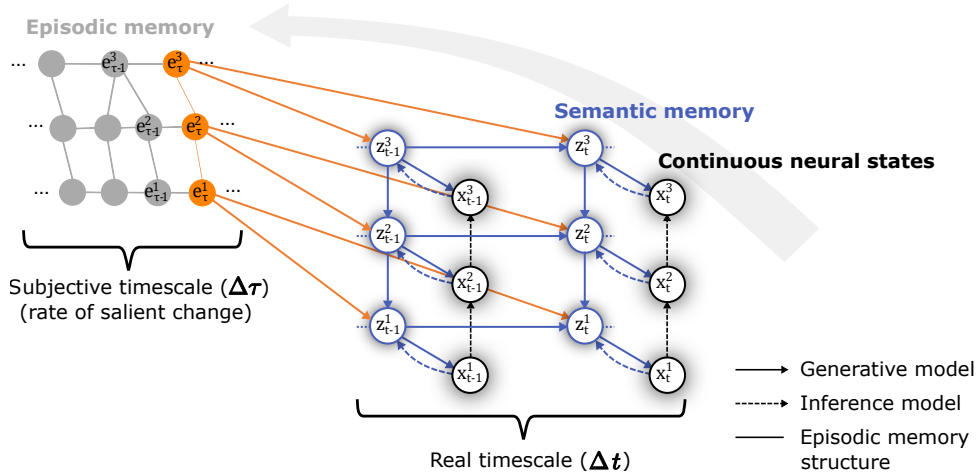


Fig 2. Simplified probabilistic graphical representation of the proposed predictive processing model described in the text, highlighting the distinction between episodic and semantic memory systems. The random variables x_t^n, z_t^n correspond to different hierarchical representations of the input, continuous and categorical respectively. All solid-line arrows represent conditional dependencies while, in particular, orange arrows highlight dependencies between variables with distributions that evolve over different time scales. For a full graphical representation of the model equations see S1 Fig.

The structure of the generative model (solid arrows in Figure 2 and S1 Fig) can be best described using the three major perceptual principles that it was designed to account for. First, humans are able to segregate raw sensory information into different categories of concepts based on similarity (often called taxonomic relations) and access these categories linguistically, answering for instance questions such as *what does an object X look like?* Likewise, the model presented here generates categorical hierarchical latent representations of the sensory signals it receives, in the form of probability distributions, that classify the current (continuous) neural states and create new predictions. Second, humans can associate sensory experiences with context and with other experiences over time (often called contiguity relations), and answer questions such as *do I expect to see object X under the current circumstances?* Our model also maintains state transition statistics, taking into account the hierarchy of the categorical representations and recent experiences. Finally, humans are able to maintain the general concept of an object X, without being continuously surprised after seeing a particular instance (e.g. a paper sketch) of this object. To account for this ability, each categorical representation in the model consists of a slowly-changing baseline distribution as well as a second, short-term bias.

To implement the first principle, the (continuous) random variable x_t^n follows a mixture of Gaussians, where the parameters of the individual Gaussian components $N\big(\mu_{t,i}^n, \Sigma_{t,i}^n\big)$ update over time with recursive Bayesian estimation[1] At each simulation time-step and layer n, a single component is selected, then used to sample a prediction $\bar{x}_t^n$, and finally updated according to the prediction error ξ_t^n. Hence,

$$x_t^n \sim \sum_{i=1}^{K} z_{t,i}^n N\big(\mu_{t,i}^n, \Sigma_{t,i}^n\big) \qquad \text{where} \quad z_{t,i}^n = \begin{cases} 1, & \text{if } i \text{ selected at time } t. \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The selected component is indicated by the (discrete) random variable

[1]All notation is defined again in the supplementary S4 Table.

$z_t^n = \{z_{t,i}^n | i = 1, 2, ..., K\}$ which allows the categorization of the model's (both sensory and inner) states. Given a current observation x_t^n, the value of z_t^n can be inferred from $\arg\max_i N\big(x_t^n | \mu_{t,i}^n, \Sigma_{t,i}^n\big)$. In addition, to cover the second principle, the tuple (x_t^n, z_t^n) is treated as a partially-observable hierarchical hidden Markov model over time t, with a discrete state-space and continuous observations. Apart from the previous and current higher order states, which are used to satisfy the Markovian property, the categorical variable z_t^n is also conditioned on the *last surprising state* that the system recorded. For now, we will assign the symbol e_τ^n to represent this state and later in the text we will see that e_τ^n corresponds to the last recorded event in episodic memory. Taking all together, z_t^n can be drawn from $P(z_t^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n)$. Finally, to address the third principle we propose an extension to the classical Kalman filter [45] that takes into account the differences between short- and long-term learning. Based on this proposal, each Gaussian component maintains two states, a short-term state $(\mu_{t,i}^n, \Sigma_{t,i}^n)$ and a baseline state $(\tilde{\mu}_{t,i}^n, \tilde{\Sigma}_{t,i}^n)$. Hence, the update process of each component becomes two-fold. The short-term states update according to

$$\mu_{t,i}^n = \begin{cases} \mu_{t-1,i}^n + K \cdot \xi_t^n, & \text{if } i \text{ selected.} \\ \mu_{t-1,k}^n + k_{\text{back}} \cdot \big(\tilde{\mu}_{t-1,i}^n - \mu_{t-1,i}^n\big), & \text{otherwise.} \end{cases} \tag{2}$$

$$\Sigma_{t,i}^n = \begin{cases} \big(1.0 - K\big) \cdot \big(\Sigma_{n,t,i}^n + Q\big), & \text{if } i \text{ selected.} \\ \Sigma_{t-1,k}^n + k_{\text{back}} \cdot \big(\tilde{\Sigma}_{t-1,i}^n - \Sigma_{t-1,i}^n\big), & \text{otherwise.} \end{cases} \tag{3}$$

where $K = \frac{\Sigma_{t,i}^n + Q}{\Sigma_{t,i}^n + R^2}$ is the Kalman gain, Q and R are free noise parameters and k_{back} is a parameter determining how quickly the short-term state converges back to its baseline, in the absence of new information. The update of the baseline states is given by

$$\tilde{\mu}_{t,i}^n = \tilde{\mu}_{t-1,i}^n + k_{\text{base}} \cdot \big(\mu_{t-1,i}^n - \tilde{\mu}_{t-1,i}^n\big) \quad \text{and} \quad \tilde{\Sigma}_{t,i}^n = \tilde{\Sigma}_{t-1,i}^n + k_{\text{base}} \cdot \big(\Sigma_{t-1,i}^n - \tilde{\Sigma}_{t-1,i}^n\big). \tag{4}$$

where k_{base} determines how quickly the baseline adapts to new information. By keeping this parameter low, the baseline updates at a slower pace and thus it is less prone to short-term biases.

The next neural activation pattern of each layer n is predicted from the distribution

$$\hat{x}_t^n \sim P\big(x_t^n, z_{t,i}^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n\big) = P\big(x_t^n | z_t^n\big) P\big(z_t^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n\big) \tag{5}$$

$$= N\big(x_t^n | \mu_t^n, \Sigma_t^n, z_t^n\big) P\big(z_t^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n\big) \tag{6}$$

through ancestral sampling (sequential sampling of random variables each conditioned to the previous one). New feed-forward activations in layer n are then compared and yield the prediction error $\xi_t^n = \hat{x}_t^n - x_t^n$. Finally, the corresponding surprise can be calculated as the negative logarithm of model evidence for this layer and this time step

$$-\log P\big(x_t^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n\big) = -\log \sum_{i=1}^{K} P\big(x_t^n | z_t^n = z_i\big) P\big(z_t^n = z_i | z_{t-1}^n, z_t^{n+1}, e_\tau^n\big) \tag{7}$$

Splitting and merging categories Humans are able to rearrange taxonomic relations in their memory, which often involves learning new concepts or combining two existing concepts into a single one. To capture this effect, the predict-update mechanism described so far can be also seen as a high-dimensional online clustering method based on a mixture of Gaussians. A typical approach to obtain the number of clusters in these methods is via splitting and merging existing components over time [46, 47]. Here, two components i and j at time t are merged iff

$$D_{JS}\big(N(\tilde{\mu}_{t,i}^n, \tilde{\Sigma}_{t,i}^n), N(\tilde{\mu}_{t,j}^n, \tilde{\Sigma}_{t,j}^n)\big) < T_{\text{merge}} \tag{8}$$

where D_{JS} is the Jensen-Shannon divergence and $T_{\text{merge}} > 0$ is a threshold constant (hyper-parameter). In addition, the short-term state of a component i is split from its baseline representation if

$$D_{KL}\big(N(\tilde{\mu}^n_{t,i}, \tilde{\Sigma}^n_{t,i}), N(\mu^n_{t,i}, \Sigma^n_{t,i})\big) > T_{\text{split}} \tag{9}$$

where D_{KL} is the Kullback–Leibler divergence and $T_{\text{split}} > T_{\text{merge}}$, with T_{split} another threshold constant. In this case, D_{KL} was chosen over the symmetric D_{JS}, as the baseline $N(\tilde{\mu}^n_{t,i}, \tilde{\Sigma}^n_{t,i})$ can be considered the *true* distribution of the representation of the component i and, thus, Eq (9) measures the amount of information lost when $N(\mu^n_{t,i}, \Sigma^n_{t,i})$ is used instead of this true baseline. In contrast, Eq (8) measures the similarity between two hitherto independent representations.

Finally, the Bayesian model described here can also be viewed as the semantic memory of an agent [48], which maintains generalized (statistical) information about different concepts in the world, including taxonomic and contiguity relations. Rather than equating semantic memory to either the generative or inference models of the brain, we view this fundamental memory system as a knowledge storage that is involved in the processing of both top-down and bottom-up signals [49] (see S1 Table and S2 Table for more details on the relation between our model and the different memory systems). The learning process followed is recursive and allows the creation of new categorical components without the need to use the same observation twice for training. However, as past observations are discarded, it lacks the ability of off-line methods to reach solutions that are optimal for all previously seen data. This issue highlights the need for the second type of memory modeled here, that can maintain approximately intact information about surprising past observations (episodes) and use this information later to improve the predictive performance of the model.

Episodic memory formation based on perceptual surprise

Episodic memory refers to the brain's ability to store information related to specific past experiences in the form of events and mentally relive these events either voluntarily or due to intrinsic (*free recall*) or extrinsic (*cued recall*) stimulation. Evidence suggests that the criteria used to determine which parts of the current experience are more likely to be encoded involve attention [50] and prediction error [51–53]. In the context of our model, a neural event at time t and layer n is classified as salient (or surprising) when it is very unlikely according to the current state of the system. In other words, an event is classified as salient if the generative model is not able to predict x^n_t well enough. That is, iff

$$-\log P_{\text{evidence}}\big(x^n_t\big) > -\log T^n_t \iff P_{\text{evidence}}\big(x^n_t\big) < T^n_t \tag{10}$$

where $P_{\text{evidence}}\big(x^n_t\big) = P\big(x^n_t | z^n_{t-1}, z^{n+1}_t, e^n_\tau\big)$ and $T^n_t \in (0, 1)$ is a threshold describing the tolerated level of surprise. In a later section, we will see that this threshold can be dynamic and defines the level of attention the system pays to each hierarchical layer over time. When T^n_t is exceeded, the tuple (z^n_t, x^n_t) is registered as a node $e^n_{\tau+1}$ (τ being the current number of episodic nodes in layer n) to a weighted, directed and acyclic graph, shown on the left-hand side of Figure 2. The value of $P\big(x^n_t | z^n_{t-1}, z^{n+1}_t, e^n_\tau\big)$ is assigned as the weight of this node, while $e^n_{\tau+1}$ also has a direct connection to its preceding node e^n_τ and the current node in layer $n+1$. The complete graph of episodic nodes represents the full episodic memory of the system, and provides enough information to be partially or fully retrieved in the future. The overall process of memory formation is depicted in Figure 3.

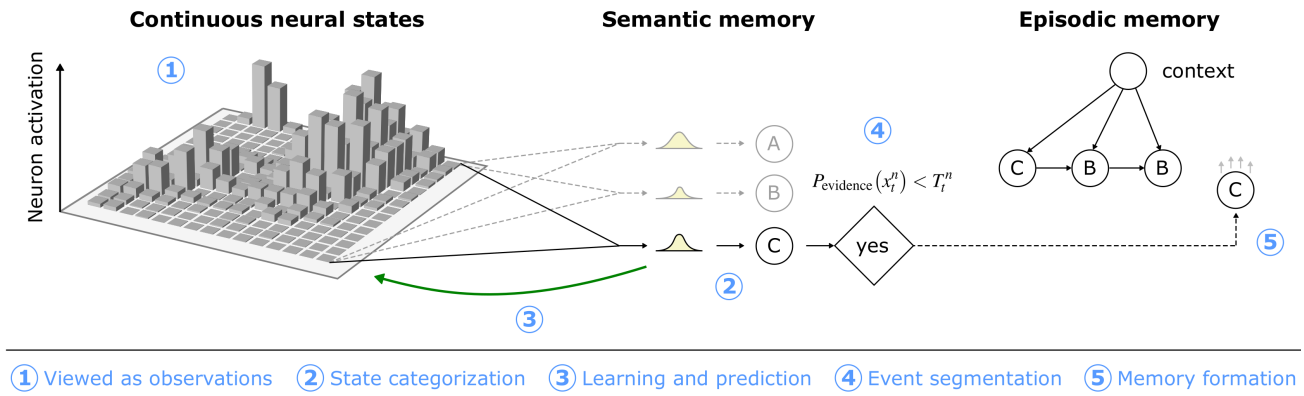


Fig 3. Diagram of the processes involved in episodic memory formation. At each time step, the neural state x_t^n is clustered using online Gaussian mixtures and prior information. An episodic event is generated when surprise of observing x_t^n exceeds the threshold of Eq (10).

Episodic memory recall

Information retrieval from specific episodic memories in humans can occur either spontaneously or triggered by a specific task and it is facilitated by the semantic memory [54]. For instance, when asked to report the duration of an event retrospectively, one needs to retrieve as much information related to this event as possible and fill any remaining gaps with statistical information, in order to accurately approximate the rate of perceptual change that originally occurred during this event. In this model, episodic recall is initially triggered from a single node $e_{\mathbf{0}}^N$ drawn from the episodic memory structure depicted on the left-hand side of Figure 2. The tree defined by all nodes connected to $e_{\mathbf{0}}^N$ in lower layers contains the complete amount of information that has been saved for the corresponding episode. In a single recall, the process followed is summarized in Algorithm 2.

Algorithm 1: Tree construction during episodic memory recall

1. Create a tree $\text{Tr}(e_{\mathbf{0}}^N)$ with root node $e_{\mathbf{0}}^N$ (*the red node in Figure 4*).
2. Append to $\text{Tr}(e_{\mathbf{0}}^N)$ all nodes e^n connected to $e_{\mathbf{0}}^N$, where $n < N$, with probability $\sim \text{Bern}\big(1 - P_{\text{evidence}}(e^n)\big)$.
3. For each $e_i^n \in \text{Tr}(e_{\mathbf{0}}^N)$ with component $z_{k,t}^n = e_i^n$, estimate the most likely number of children using the distribution of children recorded so far during this component as a prior $P_{\text{chil.}}(z_{k,t}^n)$.
4. If the estimated number is greater than the current number of children, append more nodes e_j^{n-1} to $\text{Tr}(e_{\mathbf{0}}^N)$, sampled from $P(e_j^{n-1}|e_{j-1}^{n-1}, e_i^n) \approx P(z_t^{n-1}|z_{t-1}^{n-1} == e_{j-1}^{n-1}, z_t^n == e_i^n, e_i^n)$ (*solid green nodes*).
5. While $\exists\ e_i^n \in \text{Tr}(e_{\mathbf{0}}^N)$ with $n > 1$ to which step 5 has not been applied yet, append k *blank* nodes e_j^{n-1} to $\text{Tr}(e_{\mathbf{0}}^N)$, where k is either sampled from $P_{\text{chil.}}(z_{k,t}^n)$ if e_i^n is known, or from $P_{\text{chil.}}(n)$ otherwise. (*nodes with green dashed outline*).

The tree $\text{Tr}(e_{\mathbf{0}}^N)$ represents the salient events that occurred during $e_{\mathbf{0}}^N$ and have been currently retrieved from memory. Counting the overall nodes per layer provides an approximation of the amount of salient change in each layer during this episode, and, therefore, also an approximation of the corresponding sense of the episode's duration. To maintain consistency in notation, let $\tilde{\tau}_{\text{tree}}^n$ be the number of nodes in each layer n of the tree. Whereas the nodes added in steps 1-2 of Algorithm 2 are the only ones taken from episodic memory, alone they do not provide enough information, as they can only be fewer or equal in number to the nodes originally registered. Instead, the remaining

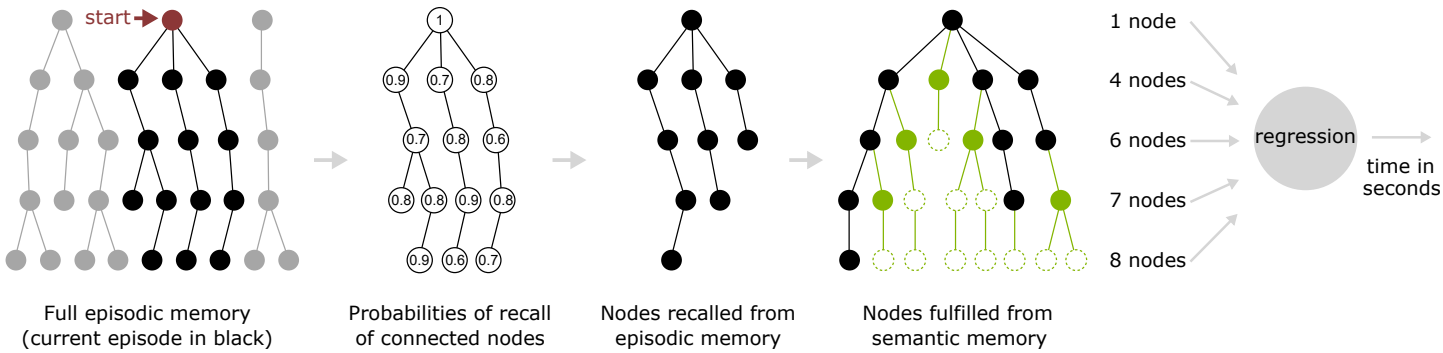


Fig 4. Algorithm of the steps required for episodic memory recall and retrospective duration judgements. Solid circles represent nodes with an assigned value to the categorical variable z_t^n either from the episodic (black) or the semantic memory (green). Circles with the dashed outline represent nodes whose component has not been determined during recall.

nodes added in steps 3-5 contribute prior (semantic) information and thus fill any memory gaps.

Estimating time

Prospective tasks The inequality (10) holds mainly at times when instances of surprise occur in the system, thus detecting the presence of new, salient events. Through this mechanism, new episodic memories are formed only when their contents provide non-redundant information that could not otherwise be predicted. In [4] it was claimed that the accumulation of events of perceptual change provided an intuitive and useful proxy for perceptual surprise and belief updating and therefore could be used as the basis for human subjective time perception. As these two cognitive processes (surprise/belief updating and time perception) share the same criteria for salient event detection across time and model hierarchy, we propose that the same saliency-detection mechanism can be shared in both cases. Hence, following the current notation, the duration of an event can be estimated prospectively as a function of the number of all recorded nodes during this event, i.e.

$$\text{prospective duration judgement} = f(\tau^1_{\text{end}} - \tau^1_{\text{start}}, \tau^2_{\text{end}} - \tau^2_{\text{start}}, \ldots, \tau^N_{\text{end}} - \tau^N_{\text{start}}) \quad (11)$$

where $\tau^n_{\text{start/end}}$ represents the number of episodic nodes in layer n that have been recorded at the (objective) beginning and the end of the target event respectively. After comparing different degree polynomials against human behaviour, we argue that the function $f()$ can be approximated by a simple linear combination of the number of nodes (see Figure 7 and related discussion below).

Retrospective tasks In order to obtain duration judgements of past events, it is necessary to recall (or at least examine) the content of the memory. Given a past event represented with a node $e^N_{\mathbf{0}}$, the episodic tree $\text{Tr}(e^N_{\mathbf{0}})$ that contains both episodic and semantic (filled) information (Figure 4) represents the system's current *best guess* of the salient events that occurred across the perceptual hierarchy during $e^N_{\mathbf{0}}$, thus defining the temporal boundaries of the corresponding episode. Therefore, following the same process as for prospective duration estimates, but based on the number or *recalled* salient events in each layer for an episode, it is possible to estimate retrospective duration as:

$$\text{retrospective duration judgement} = f(\tilde{\tau}^1_{\text{tree}}, \tilde{\tau}^2_{\text{tree}}, \ldots, \tilde{\tau}^N_{\text{tree}}) \quad (12)$$

In other words, the task here is to retrieve the maximum possible amount of information related to $e_{\mathbf{0}}^{N}$ and estimate the original number of salient features during this event, in order map the latter to human-comprehensible duration units. It is worth noting that the same function $f()$ can be used in both prospective and retrospective judgements as the task of mapping recorded salient events to durations remains the same in both cases.

Attention, recall effort and cognitive load

Two factors that affect human accuracy in forming and recalling episodic memories, and subsequently performing duration judgements, are the level of attention that is paid during episode formation, and the level of effort put into recall. Intuitively, recall effort maps onto the idea that a specific portion of an episode may not be immediately recallable and describes how persistent the model should be in trying to retrieve that specific piece of memory rather than moving on without it. Here, attention and recall effort are represented through a single free parameter each, which is fitted to our experimental results.

To capture the effects of attention over time, the value T_t^n of the surprise threshold $-\log T_t^n$ defined in Eq (10) tends to decay over time to the value 1, by obeying the equation

$$T_t^n = 1 - T_{t-1}^n + c^{-1} e^{-\frac{D^n}{c}} - N\big(0, \sigma\big) \tag{13}$$

where D^n indicates the number of simulation iterations since the last time the threshold in layer n was reset, c is a decaying time constant (free parameter) and σ indicates the level of Gaussian stochastic noise that corrupts this decay. If this threshold is exceeded, then $T_t^n \leftarrow 0$ and thus $-\log T_t^n \leftarrow +\infty$. In addition, the level of recall effort (represented by the letter ϵ) is defined as the number of attempts to retrieve information for $e_{\mathbf{0}}^{N}$, i.e. the number of times that the distribution $\mathrm{Bern}\big(1 - P_{\mathrm{evidence}}(e^n)\big)$ is sampled to determine whether each node of the complete recorded episodic experience will be recalled. Our model is built on the assumption that both attention and recall effort are modulated by cognitive load. Therefore, the influence of different levels of cognitive load on duration estimation can be fully represented by different combinations of c and ϵ.

Results

Human experiment

We used the online platform Amazon Mechanical Turk (MTurk) to recruit 12,827 participants. Each participant watched a single video (1 - 64 seconds in duration) of a natural scene such as walking around a city, walking in the countryside or a leafy campus, or in a quiet office or a cafe (the same video stimuli were also used in [4, 7]; in these previous studies, each participant viewed a large number of videos) and reported the duration in seconds (see Figure 8). Half the participants were not informed that they would need to estimate the video duration until after they had viewed it (retrospective task group) while half were told before the trial began (prospective group). Within each group, participants completed either a low or high cognitive load trial. In the prospective group, the low-load condition required that participants estimate video duration. For the retrospective low-load condition, participants were requested to use the cues in the video to estimate what time of day the video was taking place (note that this task instruction was given prior to the video being viewed). For both groups, the high-load condition added the requirement that participants should determine whether the person recording the video was by themselves or with another person (there was never another person traveling with the person recording the video, though strangers could certainly appear).

After excluding participants we believed to be outliers (see Participants section in Methods for details), we analysed the duration judgement ratio (reported duration divided by physically presented duration) in a 2x2 ANOVA with factors of task (prospective/retrospective judgement) and cognitive load (low/high). This analysis provided evidence that the task by cognitive load interaction ($F(1, 7782) = 5.34, p = .02$) previously reported in the literature (e.g. [10]) was present in our data. However, as some participants in this dataset reported that they were counting during estimation and we were interested in the role of external stimulation not manual counting in time perception, these participants were excluded from further analysis. The results for participants who reported they were counting (11% of participants) are presented in S2 Fig. For further details on the exclusion criteria see Methods.

Figures 5Ai-Aiv show the overall duration estimation results for each task (prospective/retrospective) and cognitive load (low/high). Our participants were capable of producing sensible duration estimates over the full 1-64 second range, with reports appearing broadly consistent with the scalar property/Weber's law in each case (across-participant variance in estimates was roughly proportional to duration). There was some evidence of overestimation for the short intervals, compared with previous results of prospective judgements made regarding the same stimuli wherein participants completed ~ 60 trials rather than a single trial (see Figure 3a in [4]). However, there was less evidence for underestimation of longer intervals, making the present data inconsistent with Vierordt's law/regression to the mean. The overestimation of shorter trials was likely due to the lower bound of report being limited to 0 seconds in duration, with no such corresponding upper limit for longer trials. Bounded limits in the report scale will affect duration judgements in this way [55]. Consequently, these results are consistent with the idea that Vierordt's law/regression to the mean effects in human reports about time are the result of sequential decision processes rather than anything inherent to time perception itself [56–58].

Figure 5B shows the mean duration judgement ratio for all combinations of task (prospective/retrospective) and cognitive load (low/high). The previously reported task by cognitive load interaction [10] is still apparent in our data (compare Figure 1 and Figure 5B), though in our data it is of a smaller magnitude (after excluding counting participants; 2x2 ANOVA ($F(1, 6973) = 3.44, p = .06$) and is shifted towards overestimation rather than underestimation. This shift towards overestimation was likely due to our experiment using more dynamic stimuli which are known to result in longer duration estimates [4, 6, 7] (see also Figure 5C). The reason we obtain a weak task by load interaction in our data (when excluding participants who were counting) becomes clear when considering the results broken down by the different scene types that were used as stimuli (City, campus and outside, and office/cafe) (Figure 5C) - the different scene types produce qualitatively different outcomes in duration estimation. First, regardless of task and load, the degree of duration overestimation was reduced particularly for the Office and cafe scenes (for City scenes mean duration judgement ratio is 1.327 ± 0.867, for Campus & outside scenes: 1.313 ± 0.832 and for Office & cafe scenes: 1.124 ± 0.735 were the overall means and standard deviations respectively). This pattern of results was reflected in a 2x2x3 ANOVA, with task, cognitive load, and scene type as factors, which revealed a significant main effect of scene ($F(2, 6965) = 53.07, p < 0.001$). This finding replicates previous, similar results indicating the difference in duration estimation for these naturalistic stimulus classes [4, 7, 24], but now also extends evidence for this effect to the retrospective task. Of greater interest is that the nature of the task by load interaction appears to qualitatively change with scene type. While the interaction for the City scene data (Figure 5C, *City*) was broadly similar to the overall pattern (Figure 5B), this was not the case for reports about the Office and cafe scenes. For these scenes, there was a

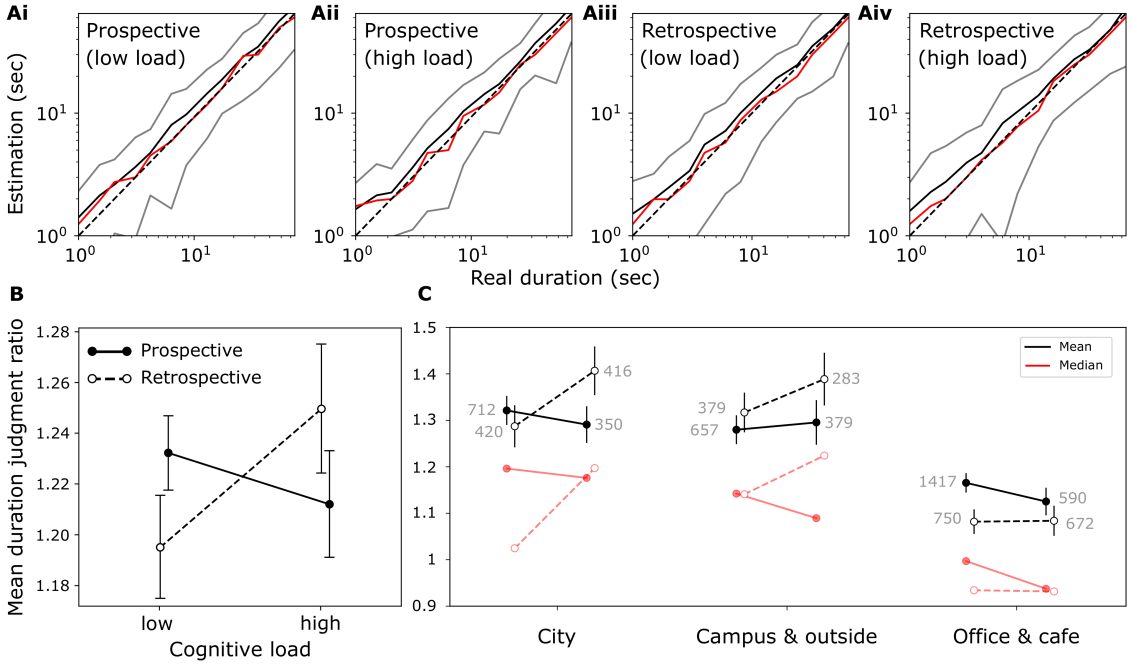


Fig 5. Human duration estimates according to task (pro/retrospective), cognitive load (low/high), and presented video scene (City, Campus and outside, or Office and cafe). **Ai-iv** Human duration estimation for task-cognitive load combination. The black curves represent the mean, the red is the median, and the gray is the standard deviation across all trials. **B** The mean duration judgement ratio (report versus physical duration) across all trials for each task-cognitive load combination. Broken lines/open markers indicate results from retrospective judgements, solid lines/filled markers indicate results from prospective judgements (compare with Figure 1). **C** as for **B**, but separated by scene type (City, Campus & outside, Office & cafe). The gray numbers denote the number of participants in each case, black (red) markers are means (medians).

general shift towards prospective estimates being longer than retrospective estimates, regardless of cognitive load. This pattern is the reverse of the relationship seen for reports regarding the Campus and outside scenes and was partially reflected in a significant task by scene interaction ($F(2, 6965) = 4.10, p = 0.017$) in the above 2x2x3 ANOVA. However, the three way (task by load by scene) interaction was not statistically significant ($F(2, 6965) = 0.36, p = 0.697$). Finally, the absence of a significant interaction of task and load ($F(1, 6965) = 3.59, p = 0.058$) when including scene as a factor reflects the similar absence in the overall analysis excluding scene as a factor, though is not interpretable beyond this superficial observation.

The computational model replicates human behaviour

We used the exact same trials as were completed by human participants and presented in the results above as input to our model. Initially, to acquire semantic statistics of all natural scenes used, the model was exposed to 34 different 200-frame-long videos, including all scene types. Then, the same resulting semantic memory was used as the initial point for all model-based trials. The method to produce duration estimates was similar for trials from prospective and retrospective tasks (see section Estimating time). For trials completed by human participants as prospective, estimates were based on the model response to the initial input of the presented trial, by using directly the number of recorded events per layer. Trials completed by human participants as retrospective were modelled based on model-recalled episodes of the presented trial. For both cases,

the difference in cognitive load was accounted for by fitting the free parameters c (for the effect of attention) and ϵ (for recall effort), described above in the section Theoretical framework, such that the relationship between model duration judgements in low and high cognitive loads follow the same slopes as the mean human duration estimates presented in Figure 5B. The parameters c and ϵ were not required to be the same for prospective and retrospective cases and, therefore, the model acted as four 'super-participants' each undergoing all trials of the four variants of the task. This process of fitting c and ϵ was the only part of parameter tuning in this study that was driven by human data and the resulting model behaviour that is reported in the following sections was recorded after this model fitting was performed (i.e. outside of these specific parameters, further observations of model performance are not simply a function of fine tuned fitting to human results). For further information regarding the method used to fit the free model parameters, see section Parameter tuning and S2 Appendix.

Duration judgements of the past and present

In the last part of model fitting, we trained a single linear regression model (using support vector regression) to map salient event accumulations to seconds on the prospective accumulation data and used it to produce time estimates for all four variants of the experiment. Figures 6Ai-iv depicts the model-produced estimates for each combination of task and cognitive load. Overall, the model was able to produce consistent estimates, with longer videos judged as longer in all cases. The variance of all reports was proportional to the reported duration, satisfying Weber's law, while, as expected, the variance of the retrospective estimates was greater. Although overestimated, prospective model reports were highly accurate without any signs of regression to the mean. The lack of this effect, which is evident in our previous modeling approaches [4], can be attributed to the fact that the regression model here was trained under a wider range of example durations including two sets of parameters representing high and low cognitive load. Retrospective reports were also overestimated but less accurate, indicating the need for both prospective and retrospective data to be used for training the regression function f defined in equations (11-12) (see also below and Figure 7 for results under different training regime). Compared to human results, the variability of the model estimates was substantially lower in all cases, especially in prospective estimates.

Finally, the overall behaviour of the model (shown in Figure 6B) replicates the interaction of task and cognitive load demonstrated both in previous work [10] (Figure 1) and our human results in Figure 5B, with duration estimates for prospective estimates decreasing with increased cognitive load while retrospective estimates increase with increasing cognitive load.

Stimulus content predicts which area of the perceptual hierarchy correlates more with human duration judgments

As in previous work [4], rather than looking at the model-based estimates made following the regression of accumulated salient events into seconds, we can also inspect the accumulations directly. In our conception, these accumulations directly underlie the 'sense of duration' and the translation into reports in seconds is used only to enable comparison with human data. In an exploratory analysis we looked at the same breakdown by task and cognitive load as depicted in Figure 6B, but in accumulated salient events (Figure 6C) rather than in seconds, and separated by hierarchical layer in the network, and we observed that the task/load interaction varies by layer. While the clear cross-over interaction seen in Figure 6B is present in each of the lower and middle

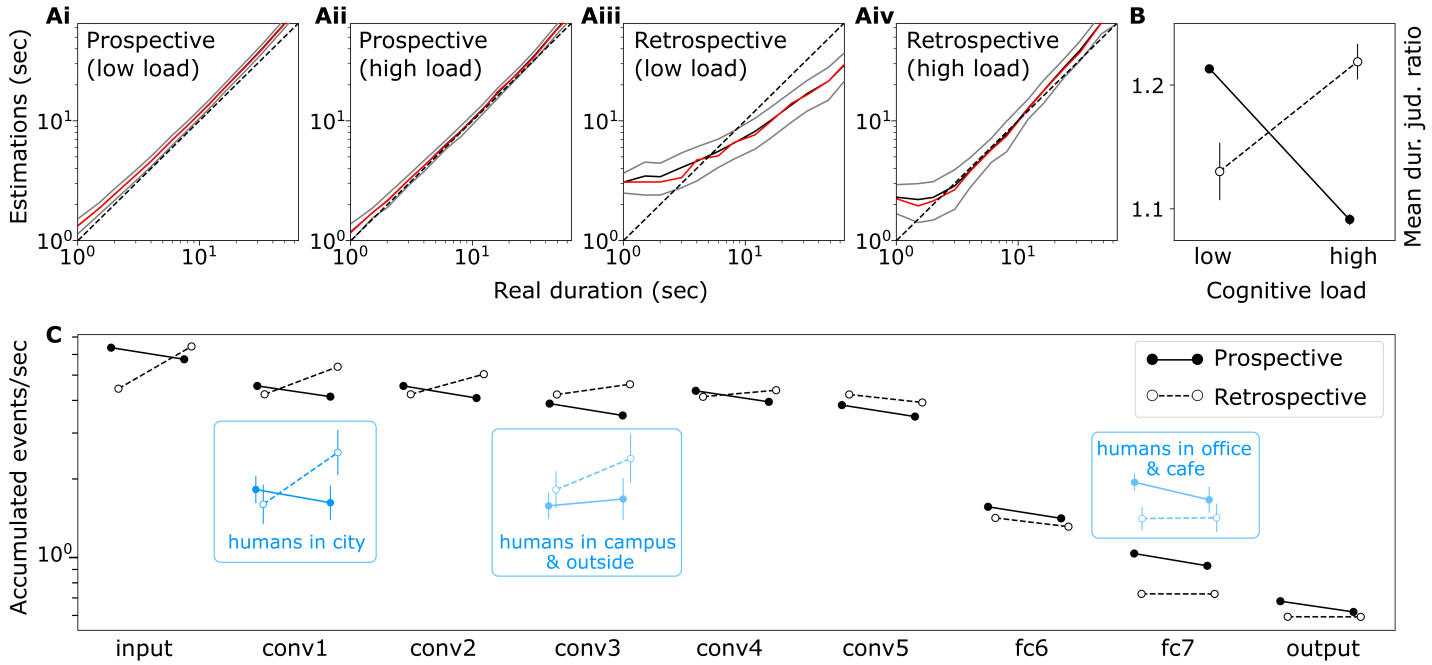


Fig 6. Model duration estimates in seconds or accumulated salient events by task (pro/retrospective) and cognitive load (high/low) for the same trials performed by human participants. **Ai-iv** Model duration estimates for task-cognitive load combination using a single linear regression trained only on prospective event accumulations. The black curves represent the mean, the red is the median, and the gray is the standard deviation across all trials. Estimates in seconds were obtained from a single linear model trained on objective video durations and the number of episodic nodes recorded for each trial. **B** The mean duration judgement ratio (estimate versus physical duration) over all trials for each task-cognitive load interaction, using the same linear model. Broken lines/open markers indicate results from retrospective judgements, solid lines/filled markers indicate results from prospective judgements. **C** The rate of accumulated salient events over time (per second) in the different network layers, across all scene types and separated by task and cognitive load. Note that the human data replicated from Figure 5C is in seconds and therefore is not meaningfully positioned on the y-axis - it is depicted for the purpose of comparing the task/load interaction in different scenes.

layers of the network (layers input through conv4), the interaction qualitatively changes at higher network layers (layer conv5 through output).

Most notable about these different interaction patterns is the remarkable similarities between the patterns observed in specific layers, and the patterns from human data in specific scenes. In the case of humans (Figure 5C) we see the clear cross-over task/load interaction for duration reports made about the dynamic city, while the interaction in reports made about other scenes is different. In Figure 6C we place those interaction patterns from the human results (in blue) next to the layers of the model that are qualitatively best matched. In S8 Fig we compare the cognitive load interaction between retrospective accumulated events/sec and retrospective human reports to further substantiate this match. These results suggest that duration estimation may be based on the activity in specific layers of the network for specific contexts; i.e., that humans may rely more on information captured in specific hierarchical levels of representation of an episode, given the context of this episode. For instance, stimulus representations similar to the first convolutional layer (*conv1*) of our network could be preferred for highly dynamic videos of someone walking around a busy city, representations in higher convolutional layers (such as *conv3*) in videos of medium information flow (walking around a university campus and outside), while more static videos seem to map to the last hidden layer of the network (*fc7*).

To further investigate this apparent relationship between network layer and scene type, we examined model duration estimation when accumulated salient events in only one layer of the network hierarchy were taken into account. Figure 7A shows that taking into account the single layer (identified as most similar in Figure 6C and S8 Fig); conv1 for City scenes, conv3 for Campus and outside, fc7 for Office and cafe scenes), estimation performance increases and becomes more similar to the human results than that shown in Figure 6A wherein the regression was fitted on all accumulation in all network layers. For a comparison between the two computational models in the same axis as the human data and the previous model in [4] see S7 Fig.

Model comparison using the Akaike information criterion (AIC) showed that the best performing model (both in terms of capturing human behaviour and estimating physical duration) trained to map between the accumulated salient events in a specific layer and physical duration using a 6th degree polynomial (solid black line in Figure 7A); however, note that even a linear model (broken blue line) performs well. Moreover, the interaction between task and cognitive load by scene is much more consistent with that seen in the human results (Figure 5C) when the regression (linear in this case) is trained only on the qualitatively matched single layer accumulation (Figure 7C) compared to when it is trained on accumulation in all layers (Figure 7B). All AIC scores for the different candidate models compared to both human reports and physical durations are reported in S3 Table.

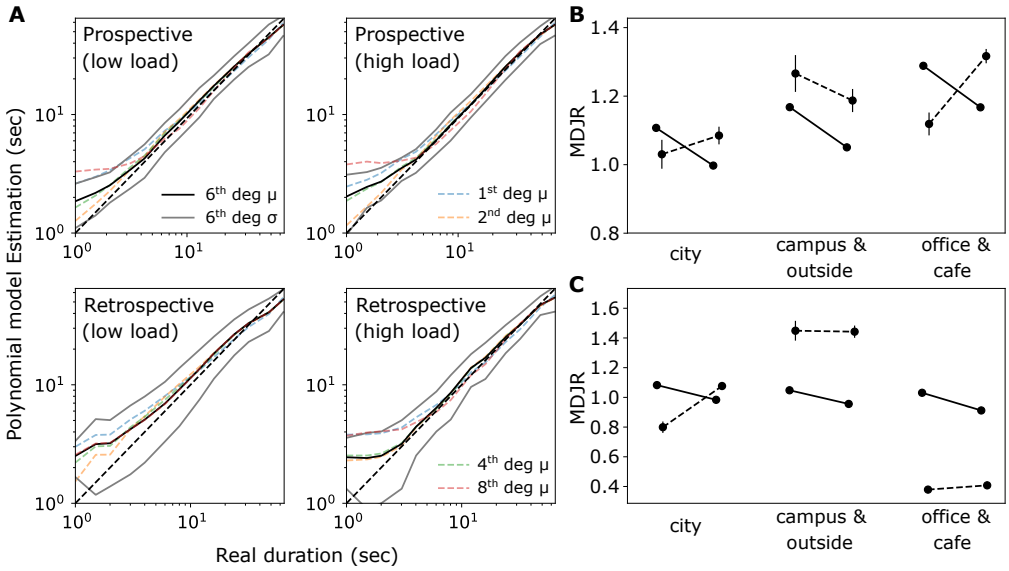


Fig 7. A: Model duration estimates for each task (pro/retrospective) and cognitive load (high/low) combination, using linear and polynomial regression trained on a single layer of salient event accumulations for different scene types, including layer *conv1*, *conv3* and *fc7* for trials that correspond to *city*, *campus & outside* and *office & cafe* respectively. In each figure panel, the solid black line shows the results for regression mapping between accumulated salient events and physical duration using a 6th degree polynomial and the grey line indicates ±1 standard deviation for estimates given by the 6th degree polynomial regression. The blue, orange, green, red broken lines indicate the results of the same regression using 1st, 2nd, 4th, and 8th degree polynomials. **B-C:** Relation between prospective and retrospective mean duration judgement ratio (MDJR) for different levels of cognitive load separated by scene type, via a single linear model using accumulated salient events in all layers (**B**) or just the layers described in Figure 6C (**C**).

Overall, these results show that it is possible to fit a simple linear model, using only a single layer per scene type as input (only two parameters x 3 scene types = 6 parameters in total) and real durations as the target (i.e. without using human data in fitting), that can replicate the complex pattern of interactions between task, cognitive load, and scene-type that we found in human duration estimates (Figure 5C) for these same naturalistic video inputs. Assuming such a simple scenario, an interesting direction of future research then becomes the questions of *how* (and *why*) a single appropriate layer is selected. In the Supporting Information we explore several possibilities for layer selection based on network behaviour in different layers in different contexts (i.e. scene types). These approaches include variability of accumulation in a layer, uniqueness of components for a given memory episode across layer, and accuracy of recall across layer (S5 Fig). All three approaches exhibit differences across network layer and therefore may provide potential foundations for a decision process that selects the specific network layer from which to estimate duration, given a specific context. Further research on this topic is required, while, due to the post-hoc nature of the results in this section, our conclusions should be further validated by another study with a different set of videos that replicate our findings.

The role of attention and recall effort

In [4], it was shown that attention c has an important effect when modeling prospective duration judgements. This relation is verified again here through the behaviour of the current computational model, as c was the only free parameter used to model differences in cognitive load. However, in retrospective judgements parameter tuning of c had minimal effect and was inadequate to produce model reports consistent with the human biases due to the cognitive load differences (see S6 Fig, especially B). Therefore, we kept the c fixed for both retrospective cases of cognitive load (values of all parameters are reported in Methods).

In contrast, our analysis indicates that the parameter ϵ, used to control the level of recall effort, has a substantial impact on the hierarchical structure of the memories being recalled and can be tuned to model human cognitive load biases (see S1 Appendix). It predicts that committing more effort to recalling an episode should generally lead to greater overestimation of the episode duration and that visual episodes which are rich in content should be more overestimated. Our current experiments however do not provide substantial evidence of a direct relation between high cognitive load and high recall effort and further work is required to study this relationship.

Discussion

The primary contribution of this work is a novel framework (and a computational implementation) for viewing human time perception via episodic memory in both prospective and retrospective cases. First, we designed a neural architecture that couples predictive processing, time perception, episodic and short-term memory. Additionally, we ran a large-scale human experiment where we recorded duration judgements from $\sim 13,000$ participants, under two manipulations including both prospective and retrospective timing. Using the proposed model, we were able to reproduce the characteristic patterns of human duration judgement biases (estimated from ~ 7000 of the human participants after excluding outliers) in all experimental manipulations, taking as input only the rate of salient events detected within a hierarchical perceptual processing network. Hence, we showed that using a single computational process to estimate duration both during ongoing perceptual input (prospective) and based on recalled memory of events (retrospective), our model reconciles divergent approaches

and findings in the time perception literature and accounts for observed differences in human duration perception across tasks, cognitive load, attention, and stimulus content.

A number of conceptual similarities connect the current proposal and the computational model introduced by Roseboom et al. [4]. Both studies are based on the fundamental proposal that time can be estimated in tracking activity in perceptual modalities [7, 24], they use the same type of natural stimulus, and the same mechanism for detecting salient events. Although in the previous model surprise was modelled by simple Euclidean distance between neuron activation in consecutive frames, here it is defined in a Bayesian fashion, as the negative log-model evidence. Crucially, the study in [4] did not account for the concepts of episodic, semantic or short-term memories and, therefore, was incapable of 1) reproducing results for tasks that rely on memory, i.e. retrospective tasks, and 2) viewing accumulation of salient changes for time estimation as fundamentally linked to event segmentation for memory formation. Finally, the current human experiment was designed for one trial per participant (as required to keep participants naive about the question of time for retrospective judgements), contained a manipulation of cognitive load, and checked for and excluded from analysis where necessary participants who reported that they were counting time. All these design differences provided for additional insights not possible based on the human data of previous studies, such as the lack of the characteristic regression to the mean in duration estimates (Vierordt's law).

Deep predictive processing with episodic memory

The literature contains various neural implementations of predictive processing, including deep generative models [59–61]. and models that aim to replace backpropagation with local learning rules [62, 63]. There are, however, a number of significant differences between existing computational models and the approach proposed here that are necessary for modelling retrospective timing tasks. Most importantly, the present model incorporates predictive coding with episodic memory (see Figure 2). Local prior predictions of the latent states z_t^n, which are part of the prediction-update loops in each hierarchical layer are jointly conditioned on episodic, short-term semantic and long-term semantic information, as shown in equations (5-6) and (4). Furthermore, the present model incorporates a semantic mechanism that categorizes continuous states into discrete indicative variables. These indicative variables can be viewed as "labels" of the continuous representations and recorded events. We argue that the processes of event segmentation and state categorization are fundamental for the formation of new episodic memories (see Figure 3 and next sections in the Discussion). Finally, the model presented here is Bayesian, in the sense that it is realized via a set of random variables, whose corresponding probability distributions are estimated via approximate Bayesian inference.

The subjective timescale of time perception and episodic memory

Our model shows that the dynamics of human perception could update over two distinct timescales; short-term and semantic memory systems follow the rate of changes in the real world (right-hand side of Figure 2), whereas the episodic memory system follows the *subjective* rate of captured salient events in a given hierarchical representation (left-hand side of Figure 2). We argue that human perception of duration (in all cases) resides entirely in the latter, hidden temporal structure and, therefore, has no direct connection to the physical passage of time [64]. This approach follows from the intuitive basis that human perception of time is characterised more by its common deviations from veridicality (see [24]), rather than acting like a clock that attempts to track elapsing physical time (e.g. [13, 65]). Without a direct connection to physical time,

human tasks that involve subjective duration reports must rely either on episodic memory formation (prospective time) or recall (retrospective time) processes. To communicate estimates of duration or relate subjective time to the physical world (like timing how long it may take to walk to your nearby bus stop), the brain need only employ a read-out network that learns to map the length of a memory to a standard unit, such as seconds. The emergent human-like performance in Figure 6B and Figure 7 using a linear model as the function $f(.)$ of equations (11-12) supports this view and highlights the potential simplicity of this read-out process, which could be performed by even single task-, content- and unit-specific neuronal cells. Such a read-out process would have a wide repertoire of available and partially overlapping resources across the perceptual hierarchy to potentially use. Therefore, it is reasonable to speculate that individual hierarchical layers might be employed, since this would require less computational resources. These layers could be different for different stimulus content, as in the case of our model shown in Figure 7, if we assume that different levels of perceptual representations might be more informative for different stimulus content. In busy city scenes for instance, where more high-layer events are recorded than on average, relying on a more consistent low perceptual layer could lead to a lower uncertainty in estimations.

It is worth stressing the difference between our proposal and proposals of mechanisms that involve the direct (or indirect) tracking of physical time. As with any other dimensions of experience, the human brain constantly receives redundant time-related information. For example, photoreceptor cells in the retina are driven by light, and low-level neural dynamics are governed by physical time as neurons are bound by electrophysiological constraints. However, most of the information received by the retina is lost in the pathways of the visual cortex; according to predictive processing, sensory information is propagated only or largely in the case that it cannot be predicted by the brain's model of the hidden causes of sensory input [25]. Contemporary interpretations show that this mechanism optimizes energy efficiency in terms of information processing and storage [35, 66, 67], while it has been argued that it even constitutes a fundamental requirement of (biological) self-organized systems to achieve homeostatic control, or life [68–70]. Here we view time as simply one more source of observations that the brain needs to reconcile with its internal model of the world. Indeed, our proposal takes time perception one step further from the physical world than other sensory modalities, as the cognitive architecture described in Figure 2 and S1 Fig does not require an explicit representation of time, or indeed any 'time sensors' (compare, e.g., vision). Instead, through the processes of event segmentation and episodic memory management, salient temporal information is encoded in a parsimonious fashion and remains available for most time-related tasks. Exceptions to this basis for temporal processing might include pattern timing tasks for short intervals, where neural network states, analogous to photoreceptor cells for vision, can be more useful [71, 72].

Finally, our model attempts to reconstruct previous stimulation based on a combination of what can be recalled from a given previous episode and the properties (in terms of stimulation) of similar episodes (referred to as semantic memory). However, neither the formation nor recall processes contain the contribution of any explicit knowledge of previous duration, such as the explicit knowledge that a previously experienced episode lasted 45 seconds. While humans have the ability to retain and recall such information, potentially in isolation of the content that may have originally underpinned the time estimation [73], it is beyond the scope of our present modelling work to incorporate such influences. While of great interest, such work would require a much deeper understanding of the interaction of experience, language, and memory than we currently possess, though this is likely not outside the scope of an extended predictive processing approach [74].

Time perception and episodic event segmentation

Perceived time in our model is determined by the frequency of occurrence of salient events over network layers for an epoch. These events can be interpreted as episode boundaries (at each layer) – the changes in content, with the relevant content type defined relative to network layer (more complex content at higher layers). In this way, our model provides a method for event segmentation in defining memory episode extent.

The relationship between time perception and memory segmentation goes back at least decades (e.g. [75–78]), though a formal description of event segmentation in episodic memory has only taken a prominent position more recently (e.g. event segmentation theory [79, 80]). Extensive work in both humans [81–85] and animals [86, 87] has identified (changes in) hippocampal activity as related to human reported (annotated) event boundaries and, more precisely, demonstrated the key role that (lateral) entorhinal cortex plays in the temporal arrangement of events for episodic memory [84, 85, 87]. It has been suggested that activity reflecting event segmentation is based on prediction error [88–91], placing the determination of event segmentation firmly within a predictive processing framework, in line with that demonstrated here.

Predictive processing has the distinct advantage over alternative modelling approaches that a definition of salient event (or event boundary) is possible without involving controversies about whether the processes that define time perception and episodic memory formation are related more to dynamics (or intensity) of external stimulation, or to internal states such as attention or task-specific context [8, 92] - instead, salience is always defined as a continual trade-off between stimulation and expectation. Consequently, while our model accommodates the powerful and well-established intuition that change in perceptual stimulation underlies subjective time [1, 2, 4], the context of stimulation and prior knowledge of similar scenarios are also fundamental for determining what constitutes *salience* (see also section Predictive processing neural architecture for perception). In addition, this formulation allows the current model to form salient events and, thus, produce prospective time estimates of an episode, even in cases that either noise or a static image is presented to the model across the entire episode duration. This is possible due to the decaying threshold of surprise in Eq (13) that captures the effects of attention in memory formation. In more complex implementations of the current model, this threshold is expected to play less significant role, as surprise across the perceptual hierarchy could be driven by more complex intrinsic processes, such as memory consolidation or mental time travel.

Finally, although the current model relies on the existence of event boundaries in continuous experience, it is worth noting that the thresholding mechanism described in Eq (10) and Figure 3 (step 4) could be replaced by a mechansim that regulates the originally recorded perceptual surprise $-\log P_{\text{evidence}}(x_\tau^n)$ based on the recency of the event $(t-\tau)$. Without an attention threshold, all instances of a given stimulus would be initially recorded in episodic memory, along with an initial memory weight. This memory weight could be further scaled to simulate the effect of attention. As the recorded instances become less recent, however, the memory weight of instances with low initial surprise will be reduced to zero creating boundaries between areas where the initially surprise was significantly higher, i.e. forming events. Hence, except for being significantly more computationally expensive, this approach has practically the same effect as temporal discretization with $T_t^n > 0$.

Hierarchy of temporal processing

Previous studies have highlighted the hierarchical nature of temporal processing in the brain, with the timescale of processing at lower cortical levels (e.g. primary sensory cortices) being shorter and more stereotyped than processing at higher cortical areas

(e.g. superior temporal sulcus and frontal eye field [82, 93, 94]). Beyond simply hierarchy alone, episodic memory is nested with more complex events comprised of smaller-scale and less complex events that provide internal structure to events as defined at the highest level (*partonomic* hierarchy [91]). The tree-based structure (see Figure 4) of our model is nested in its configuration and the perceptual classification network at the core of the network exhibits a hierarchical distribution of representational complexity (nodes in lower layers are selectively responsive to basic image features like edges, while nodes in higher layers are responsive to complex whole-object like features) – a feature shared with biological perceptual processing [16, 18]. The depth and distribution of representational complexity across our model hierarchy allows us to reproduce the differences in human duration report in different contexts – dynamic city scenes versus relatively static office scenes. Furthermore, we demonstrate that time perception may be accomplished by accumulating salient events in only some portion of the neural hierarchy, relevant to the ongoing context at the time (Figures 6C and 7B and C). This possibility is consistent with previous findings that the temporal dynamics of salient event transitions can be detected within the region of the brain relevant to the temporal scale of that event. For example, neural dynamics related to event boundaries in basic stimulus properties such as motion are detectable in lower level sensory areas like human MT [81, 82, 95] while higher complexity events will be detectable in the dynamics of higher cortical regions [82]. A context-related cortical specificity for salient events is also consistent with the finding that only the accumulation of salient events within the relevant sensory areas (ventral visual areas when watching silent videos) can reproduce human reports of subjective duration [24].

While structurally equipped to produce many properties of episodic memory, here we have only tested the time perception abilities of our model. Further work is required to determine the depth to which our model (or our modelling approach more generally) reproduces the many other known features of human episodic memory (see [91]). In particular, future work will need to address the precise degree to which important concepts in the memory literature such as primacy, and recency and contiguity [96–98] are produced in our memory model. Though it is worth noting that contiguity and forward asymmetry are preserved in principle by the model construction, with related events recorded and reconstructed in sequence within hierarchically nested episodes (see Figure 4). Further work is needed to fully relate these features of our model to findings from the literature that rely on word list tasks or similar rather than naturalistic visual sequences.

Interval timing via non-temporal Bayesian inference

Our process of episodic recall recursively estimates the number of children nodes for each recalled event (see step 3 of Algorithm 2). This process can be instantiated as a form of Bayesian inference over the duration of events, wherein new observations of the number of nodes for which an event can be parent updates the model's prior knowledge. It is worth noting that the mechanism of prospective timing proposed here is also consistent with this interpretation and could support inference, if prior semantic knowledge is combined with the current number of recorded (or recalled) events. Although the simulation results of the prospective task presented here are based only on current accumulations, as prior knowledge would have a negligible effect on single-trial experiments, this recursive inference process could potentially account for other known effects in the time perception literature, such as effects generally seen as regression to the mean of estimation [56–58, 99]. In addition, the Bayesian treatment of accumulated events allows for more sources of information to be integrated during estimations. This may be particularly important for very small intervals where more rhythmic mechanisms might be expected to provide meaningful information for duration related

tasks [72]. This proposal might be evidenced in various simple ways, such as finding differences in Vierordt's law/regression to the mean in a retrospective timing task, when episodic memory recall is manipulated. Finally, the accumulation of salient events defined by the thresholding mechanism in Eq (10) can be seen as an approximation of the amount by which the system's prior beliefs $P(z^n, \mu^n, \Sigma^n)$ update under a specific period of time. This view allows us to define an analytical form of perceptual change using tools from information geometry, such as the Fisher's information metric between belief states [100], and connects the fundamental notion of information to human perception of time, under the Bayesian framework.

Further study required

Like any mathematical model of a biological process, the present implementation is accompanied by a number of limitations. Since raw neural activation x_t^n, and not prediction errors ξ_t^n, is what propagates to higher layers in the network, the system can maintain the effect of a new sensory event for an unrealistic amount of time. Hence, the dynamical behaviour of the neural activity here should not be confused with biological neural activity, a fact that precludes predictions at that low level of description. This issue could potentially be counterbalanced by the existence of short-term plasticity in equations (2-4) that quickly reduces prediction errors, if ξ_t^n is thought to represent neural activation. However, more investigation is required to assess this relation.

In addition, the network presented here is based only on visual processing. It does not currently have the ability to integrate information from different sensory modalities, to parse higher level visual information (such as the type of room of the current environment), to perform basic spatial navigation, nor to pay attention to only a specific part of a given stimulus. Nevertheless, given the nature of our predictive processing approach, and previous findings demonstrating the similar structure of neural activity related to event transitions in different sensory modalities (combinations of modalities) [82], we anticipate that our approach is readily extendable to incorporate these additional features under a broader active inference approach [36].

Finally, our human ability to reconstruct past experiences is considered closely intertwined to our ability to create completely fictitious scenarios, which either could have happened, or might happen in the future [101, 102]. These two abilities are often referred to together as mental time travel [103]. Research in this field indicates a strong connection between recalling past experiences, semantic memory and imagining the outcomes in potential future scenarios [104, 105], an undoubtedly crucial aspect of our decision making [36, 106]. The mechanism for episodic memory recall presented here can readily encompass the notion of imagination, if the selected root node of the tree structure $Tr(.)$ is not taken from existing episodic memory, but it is simply filled by a component z_i^n, as in the case of the *blank* nodes depicted in Figure 4. Under this small modification, the recall mechanism would trigger the generation of a tree where all remaining nodes are fulfilled based on prior (semantic) beliefs. Thus our model presents a platform with the necessary components to investigate a wide range of theories related to memory and time such as imagination [107], dreaming [108], and hallucinations [109], based only on such variations of how the episodic tree is constructed or recalled. Such a platform would likely prove highly valuable for both psychological and neuroscientific work on these topics, but also for the expanding fields of deep neural networks and reinforcement learning [110].

Methods

Human experiment

Experimental procedure The human experiment was conducted using the online platform Amazon Mechanical Turk (MTurk). The participants were asked to watch a single video using the full size of their home computer monitor and then answer a series of questions including a report of the perceived duration of the video. Each participant completed a single trial that lasted from 1 to 64 seconds. Any visual information regarding the real duration of the video was removed from all trials, apart from a description of the video using the expression 'very short', applied to all video durations. 50% of the participants were not instructed that the experiment involved duration judgements before the final questionnaire (see the description of the tasks bellow).

The platform The large number of participants required for this experiment (one participant per trial) makes it impracticable for a typical lab setting. Amazon's Mturk platform provides a satisfactory alternative approach [111]. It maintains a large and diverse worldwide pool of about 500,000 potential subjects, with a large portion of this pool shown to be a good representation of the general population of the United States [112].

The tasks The task comprised a single trial with four different variations, one of which was randomly assigned to each participant. These task variants were designed to allow for two different levels of cognitive load, as well as prospective and retrospective duration judgements. Trials comprised three successive screens. In screen 1, participants were told they will see a very short video and then, they will be asked a number of questions regarding the instructions given in Figure 8. In screen 2, the selected video was automatically played without any controls or further instructions and, once finished, the screen 3 spawned automatically with the corresponding questionnaire.

Participants Using the MTurk platform, we recorded data of $12,827$ unique individuals who performed one of the four variations of the experiment. Subjects were based in the United States, Canada, Australia, New Zealand, the EU and India and they received payment of US$0.25 per trial. Subjects were anonymous, over 18 years of age, they reported to have normal or corrected-to-normal eyesight and they provided their consent by agreeing in a web form to the terms and conditions approved by the University of Sussex's research ethics committee (C-REC; approval code ER/WJR22/9).

Out of the total number of participants, the primary results presented here excluded subjects who (1) were in the retrospective condition but reported that they realized during the trial that they would need to respond about duration, or (2) reported after the trial that they found the experiment instructions either 'completely unclear' or 'somewhat unclear'. To reduce the effect of the uncontrollable factors in the environment of the experiment, participants were also removed if (3) the video playback was different from the desired real duration for that trial, due to slow internet connectivity or any other factors. In particular, each trial was rejected if either

$$\frac{\text{playback time} - \text{real duration}}{\text{real duration}} > 1.2 \quad or \quad \text{playback time} - \text{real duration} \geq 0.8 \text{ sec} \tag{14}$$

Furthermore, although participants of the prospective variant of the experiment were instructed to try not to count in seconds during the trial, when asked in the final questionnaire some reported they could not resist counting. Counting was also reported

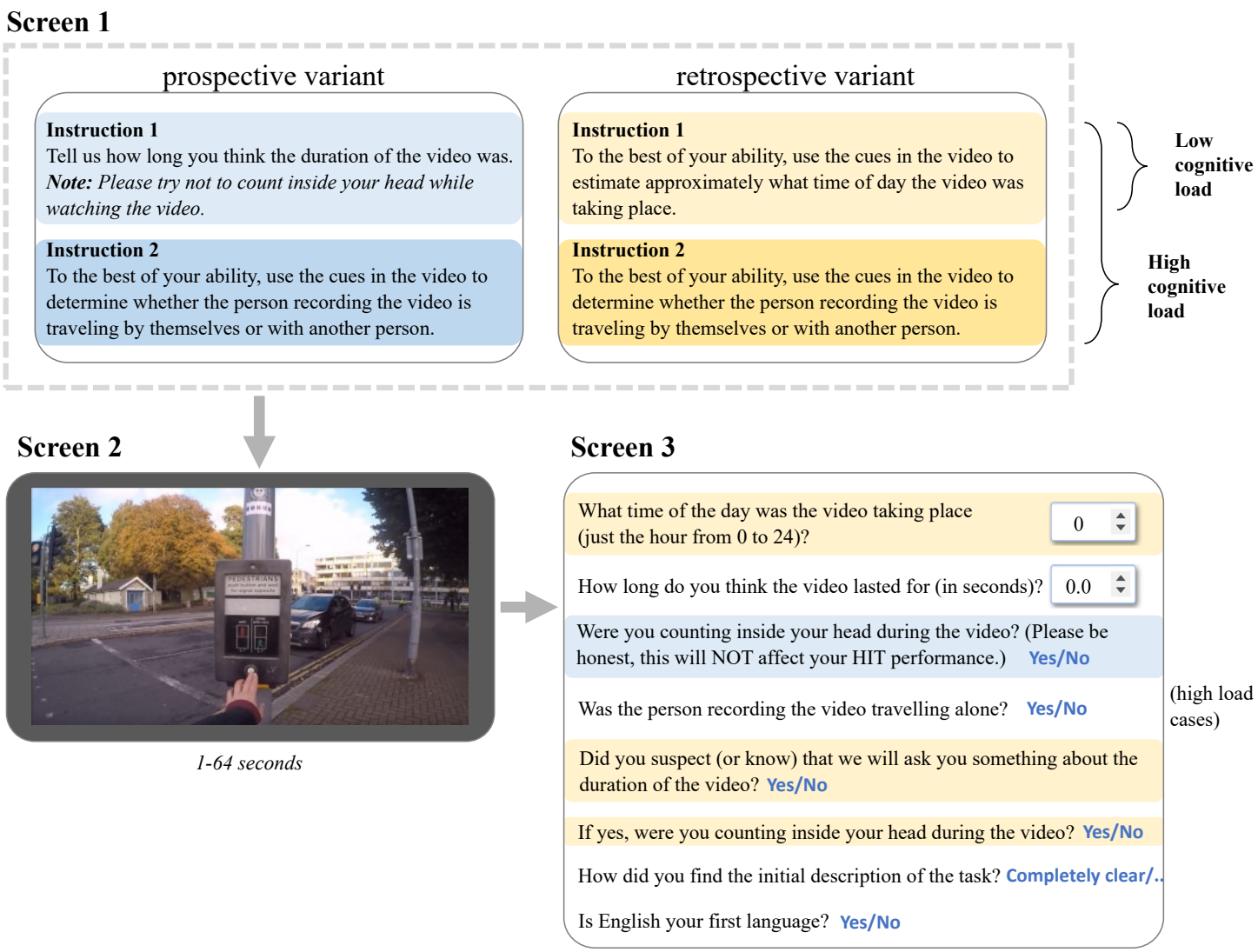


Fig 8. The main steps of the trials including instructions the video and the final questionnaire, for the four variants of the task.

by a small group of participants of the retrospective variant, who suspected they would be asked about the duration of the video. All these participants were also removed from the main sample.

After the exclusions (1-3), the sample comprised 8072 participants and, after removing 933 who reporting counting, the remaining sample comprised 7139 participants. Removing outliers for the reported durations (any duration such that report $\leq \frac{\text{real time}}{10}$ or report $\geq 10 \times$ real time) left the final sample used for the results throughout this manuscript, which comprised $n = 6,977$ non-counting participants (and trials) or $n = 7,786$ when counting participants are included. Outliers were also removed from the group of counting participants ($n = 809$) in order to produce S2 Fig. Outliers and counting participants were also removed from the four distinct groups illustrated in S3 Fig that compare the behaviour of participants that reported different level of understanding of the initial task instructions. Finally, we also recorded which participants were native English speakers, as well as the screen and web browser sizes, to ensure the trials were run in a wide field of view. An overview of the demographics of

the final sample can be seen in S4 Fig. The resulting dataset can be found in the repository [2].

Computational model implementation

Computational model implementation To implement the inference model $p(x_t^n|x_t^{n-1})$, as depicted in S1 Fig, we used a well-studied convolutional network defined by Krizhevsky and colleagues [19]. Deep convolutional networks have hierarchical structure inspired by and broadly resembling the human/primate visual cortex [16–18, 113], and have been shown to perform very well in visual classification tasks [19, 20]. This network was pre-trained to classify natural images into 1000 different categories of objects and animals [114]. It comprises 9 neuron layers of which 5 hidden layers are connected with convolutions (*conv1-5*) and 3 layers with all-to-all connections (*fc6-7* and *output*).

To predict the next activation pattern of each layer n, the method of ancestral sampling was used. Instead of sampling from the joint distribution described in Eq 5 directly, the algorithm initially samples from $\mathbf{z} \sim P\big(z_t^n|z_{t-1}^n, z_t^{n+1}, e_\tau^n\big)$. As the variable z_t^n is discrete, this distribution is approximated by a table that tracks the number of transitions to different values of z_t^n over time. Then, $\mathbf{z}$ is used to condition $\mathbf{x} \sim p(x_t^n|\mathbf{z}) = N\big(\mu_{t,\mathbf{z}}^n, \Sigma_{t,\mathbf{z}}^n\big)$. To approximate the model evidence for each Gaussian component, needed for Eq (7), the multivariate Gaussian probability density function is used instead. Finally, to approximate the initial covariance matrix Σ_i^n for each new component i, the overall covariance Σ_i^n for the layer n is tracked and used.

Parameter tuning The hyper-parameters related to the generative model (Table 1) were obtained via optimization with the double objective of (1) reducing the average surprise $\sum_{t=0}^{3600}\sum_{n=1}^{8} -\log P\big(x_t^n|z_{t-1}^n, z_t^{n+1}, e_\tau^n\big)$ and (2) maximizing clustering accuracy in the components of the output layer z^8. As a dataset, a 120-second video in an office environment was manually labeled, in order to capture the most prominent object per frame. The optimization method used was the multi-objective Covariance Matrix Adaptation Evolution Strategy (MO-CMA-ES) [115] a widely-used derivative-free algorithm, ideal for non-linear and non-convex parameter optimization. The parameters for the attention and recall mechanisms were also tuned using CMA-ES, following the principles of $c_{\text{low prosp}} < c_{\text{high prosp}} < c_{\text{retrospective}}$ and $\epsilon_{\text{low load}} < \epsilon_{\text{high load}}$, such that the overall relation between tasks and cognitive load levels, shown in Figure 5B, is approximated. These constraints represent the working assumptions that (1) less attention to the main task is expected in cases that humans are distracted with multiple tasks and that (2) human subjects are not expected to be more persistent in trying to retrieve a specific piece of memory, in cases that they were less distracted during the original experience. The complete model training process is described in pseudo-code in S2 Appendix.

Table 1. Model hyper-parameters

Generative model parameters							
k_{back}	k_{base}	Q	R	T_{split}	T_{merge}	T_{plastic}	σ
0.1	0.2	0.0002	2.0	1.5	0.1	0.5	0.02
Attention parameters				**Recall parameters**			
$c_{\text{low prosp}}$	$c_{\text{high prosp}}$	$c_{\text{low retro}}$	$c_{\text{high retro}}$	r	σ_{recall}	$\epsilon_{\text{low load}}$	$\epsilon_{\text{high load}}$
46	52	90	90	0.1	0.1	1	70

[2]Human dataset: `https://github.com/zfountas/prospective-retrospective-data`

During the training process described in S2 Appendix, the accumulated events for the four variants of the experiment, as well as regression weights for the model presented in Figure 6 were recorded. these accumulated events were later also used to fit the single-layer polynomial models presented in Figure 7. To train the single-layer models, the dataset of event accumulations was divided into 4 independent sub-sets, one per each variant of the experiment and the function f defined in equations (11-12) was fitted 4 times with all data of each 4 sub-sets, using again the real durations of the video and ordinary least-squares polynomial regression. As in the multi-layer regression model, the resulting 4 models were tested in human reports of the four experiment variants respectively.

Finally, the programming language used throughout the modeling implementation is Python 3.7 and the machine learning library TensorFlow 1.14. The source code can be found on the repository [3]. The analysis of the results was also performed in Python 3.7. The scripts used for the analysis of the human data can be found on the repository that contains the human dataset, while the scripts that perform analysis of the model behaviour, along with the resulting dataset of the computational model are included in the repository of the model's source code.

References

1. Selby-Bigge LA. A Treatise of Human Nature by David Hume, reprinted from the Original Edition in three volumes and edited, with an analytical index. Oxford: Clarendon Press; 1896.

2. Ornstein R. On the experience of time. Harmondsworth, UK: Penguin; 1969.

3. Poynter WD, Homa D. Duration judgment and the experience of change. Perception and Psychophysics. 1983;33(6):548–560.

4. Roseboom W, Fountas Z, Nikiforou K, Bhowmik D, Shanahan M, Seth AK. Activity in perceptual classification networks as a basis for human subjective time perception. Nature communications. 2019;10(1):267.

5. Kanai R, Paffen CLE, Hogendoorn H, Verstraten FAJ. Time dilation in dynamic visual display. Journal of Vision. 2006;6(12):8–8. doi:10.1167/6.12.8.

6. Linares D, Gorea A. Temporal frequency of events rather than speed dilates perceived duration of moving objects. Scientific Reports. 2015;5(1). doi:10.1038/srep08825.

7. Suárez-Pinilla M, Nikiforou K, Fountas Z, Seth AK, Roseboom W. Perceptual Content, Not Physiological Signals, Determines Perceived Duration When Viewing Dynamic, Natural Scenes. Collabra: Psychology. 2019;5(1):55. doi:10.1525/collabra.234.

8. Zakay D, Block RA. An Attentional·Gate Model of Prospective Time Estimation. In: I.P.A Symposium Liege; 1994. p. 167–178.

9. Block RA, Zakay D. Prospective and retrospective duration judgments: A meta-analytic review. Psychonomic Bulletin and Review. 1997;4(2):184–197.

10. Block RA, Hancock PA, Zakay D. How cognitive load affects duration judgments: A meta-analytic review. Acta psychologica. 2010;134(3):330–343.

[3] Source code of the model: `https://github.com/zfountas/prospective-retrospective-model`

11. Zakay D, Block RA. Temporal Cognition. Current Directions in Psychological Science. 1997;6(1):12–16.

12. Brown SW. Timing, Resources, and Interference: Attentional Modulation of Time Perception. In: Nobre AC, Coull JT, editors. Attention and Time. Oxford University Press; 2010. p. 107–121.

13. Matell MS, Meck WH. Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. Cognitive brain research. 2004;21(2):139–170.

14. Howard MW, MacDonald CJ, Tiganj Z, Shankar KH, Du Q, Hasselmo ME, et al. A Unified Mathematical Framework for Coding Time, Space, and Sequences in the Hippocampal Region. Journal of Neuroscience. 2014;34(13):4692–4707. doi:10.1523/JNEUROSCI.5808-12.2014.

15. Lindsay GW. Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. Journal of Cognitive Neuroscience. 0;0(0):1–15. doi:10.1162/jocn_a_01544.

16. Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology. 2014;10(11):e1003915.

17. Güçlü U, van Gerven MA. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience. 2015;35(27):10005–10014.

18. Kriegeskorte N. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annual review of vision science. 2015;1:417–446.

19. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.

20. Rawat W, Wang Z. Deep convolutional neural networks for image classification: A comprehensive review. Neural computation. 2017;29(9):2352–2449.

21. Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? BioRxiv. 2018; p. 407007.

22. Kietzmann TC, Spoerer CJ, Sörensen LKA, Cichy RM, Hauk O, Kriegeskorte N. Recurrence is required to capture the representational dynamics of the human visual system. Proceedings of the National Academy of Sciences. 2019;116(43):21854–21863. doi:10.1073/pnas.1905544116.

23. Ma WJ, Peters B. A neural network walks into a lab: towards using deep nets as models for human behavior; 2020.

24. Sherman MT, Fountas Z, Seth AK, Roseboom W. Accumulation of salient events in sensory cortex activity predicts subjective time. bioRxiv. 2020;doi:10.1101/2020.01.09.900423.

25. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience. 1999;2(1):79.

26. Friston K. A theory of cortical responses. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2005;360(1456):815–836. doi:10.1098/rstb.2005.1622.

27. Friston K, Kiebel S. Predictive coding under the free-energy principle. Philosophical Transactions of the Royal Society of London B: Biological Sciences. 2009;364(1521):1211–1221. doi:10.1098/rstb.2008.0300.

28. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences. 2013;36(3):181–204. doi:10.1017/S0140525X12000477.

29. Buckley CL, Kim CS, McGregor S, Seth AK. The free energy principle for action and perception: A mathematical review. Journal of Mathematical Psychology. 2017;81:55 – 79. doi:https://doi.org/10.1016/j.jmp.2017.09.004.

30. Tulving E, et al. Episodic and semantic memory. Organization of memory. 1972;1:381–403.

31. Mirman D, Landrigan JF, Britt AE. Taxonomic and thematic semantic systems. Psychological bulletin. 2017;143(5):499.

32. Healey MK, Long NM, Kahana MJ. Contiguity in episodic memory. Psychonomic bulletin & review. 2019;26(3):699–720.

33. Greenberg DL, Verfaellie M. Interdependence of episodic and semantic memory: evidence from neuropsychology. Journal of the International Neuropsychological Society: JINS. 2010;16(5):748.

34. Dayan P, Hinton GE, Neal RM, Zemel RS. The helmholtz machine. Neural computation. 1995;7(5):889–904.

35. Friston K. The free-energy principle: a unified brain theory? Nature reviews neuroscience. 2010;11(2):127.

36. Friston K, FitzGerald T, Rigoli F, Schwartenbeck P, Pezzulo G. Active inference: a process theory. Neural computation. 2017;29(1):1–49.

37. Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. Neuron. 2012;76(4):695–711.

38. Arnal LH, Wyart V, Giraud AL. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nature neuroscience. 2011;14(6):797.

39. Alamia A, VanRullen R. Alpha oscillations and traveling waves: Signatures of predictive coding? PLOS Biology. 2019;17(10):1–26. doi:10.1371/journal.pbio.3000487.

40. Corlett PR, Horga G, Fletcher PC, Alderson-Day B, Schmack K, Powers III AR. Hallucinations and strong priors. Trends in cognitive sciences. 2018;.

41. Palmer CJ, Lawson RP, Hohwy J. Bayesian approaches to autism: Towards volatility, action, and behavior. Psychological bulletin. 2017;143(5):521.

42. Sterzer P, Adams RA, Fletcher P, Frith C, Lawrie SM, Muckli L, et al. The predictive coding account of psychosis. Biological psychiatry. 2018;84(9):634–643.

43. Kingma DP, Welling M. Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR); 2014.

44. Rezende DJ, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:14014082. 2014;.

45. Kalman RE. A new approach to linear filtering and prediction problems. Journal of basic Engineering. 1960;82(1):35–45.

46. Ding C, He X. Cluster merging and splitting in hierarchical clustering algorithms. In: 2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE; 2002. p. 139–146.

47. Lughofer E, Sayed-Mouchaweh M. Autonomous data stream clustering implementing split-and-merge concepts–towards a plug-and-play approach. Information Sciences. 2015;304:54–79.

48. Steyvers M, Griffiths TL, Dennis S. Probabilistic inference in human semantic memory. Trends in cognitive sciences. 2006;10(7):327–334.

49. Binder JR, Desai RH. The neurobiology of semantic memory. Trends in cognitive sciences. 2011;15(11):527–536.

50. Uncapher MR, Wagner AD. Posterior parietal cortex and episodic encoding: insights from fMRI subsequent memory effects and dual-attention theory. Neurobiology of learning and memory. 2009;91(2):139–154.

51. Greve A, Cooper E, Kaula A, Anderson MC, Henson R. Does prediction error drive one-shot declarative learning? Journal of Memory and Language. 2017;94:149–165.

52. Jang A, Dillon D, Frank MJ, et al. Positive reward prediction errors strengthen incidental memory encoding. bioRxiv. 2018; p. 327445.

53. Rouhani N, Norman KA, Niv Y. Dissociable effects of surprising rewards on learning and memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2018;44(9):1430.

54. Weidemann CT, Kragel JE, Lega BC, Worrell GA, Sperling MR, Sharan AD, et al. Neural activity reveals interactions between episodic and semantic memory systems during retrieval. Journal of Experimental Psychology: General. 2019;148(1):1.

55. Damsma A, Schlichting N, van Rijn H, Roseboom W. Estimating Time: Comparing the Accuracy of Estimation Methods for Interval Timing; 2019. Available from: `psyarxiv.com/pg7bs`.

56. Petzschner FH, Glasauer S, Stephan KE. A Bayesian perspective on magnitude estimation. Trends in Cognitive Sciences. 2015;19. doi:10.1016/j.tics.2015.03.002.

57. Roseboom W. Serial dependence in timing perception. Journal of Experimental Psychology: Human Perception and Performance. 2019;45(1):100–110. doi:10.1037/xhp0000591.

58. Wiener M, Thompson JC, Coslett HB. Continuous Carryover of Temporal Context Dissociates Response Bias from Perceptual Influence for Duration. PLOS ONE. 2014;9(6):1–13. doi:10.1371/journal.pone.0100803.

59. Lotter W, Kreiman G, Cox D. Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:160508104. 2016;.

60. Whittington JC, Bogacz R. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. Neural computation. 2017;29(5):1229–1262.

61. Boutin V, Franciosini A, Chavane F, Ruffier F, Perrinet L. Sparse deep predictive coding captures contour integration capabilities of the early visual system. PLOS Computational Biology. 2021;17(1):1–31. doi:10.1371/journal.pcbi.1008629.

62. Canziani A, Culurciello E. Cortexnet: a generic network family for robust visual temporal representations. arXiv preprint arXiv:170602735. 2017;.

63. Song Y, Lukasiewicz T, Xu Z, Bogacz R. Can the brain do backpropagation?—exact implementation of backpropagation in predictive coding networks. Advances in neural information processing systems. 2020;33:22566.

64. Signorelli CM, Dündar-Coecke S, Wang V, Coecke B. Cognitive Structures of Space-Time. Frontiers in Psychology. 2020;11:2625.

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

66. Berkes P, Orbán G, Lengyel M, Fiser J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science. 2011;331(6013):83–87.

67. Chalk M, Marre O, Tkačik G. Toward a unified theory of efficient, predictive, and sparse coding. Proceedings of the National Academy of Sciences. 2018;115(1):186–191.

68. Conant RC, Ross Ashby W. Every good regulator of a system must be a model of that system. International journal of systems science. 1970;1(2):89–97.

69. Friston K. Life as we know it. Journal of the Royal Society Interface. 2013;10(86):20130475.

70. Seth AK, Tsakiris M. Being a beast machine: the somatic basis of selfhood. Trends in cognitive sciences. 2018;22(11):969–981.

71. Laje R, Buonomano DV. Robust timing and motor patterns by taming chaos in recurrent neural networks. Nature neuroscience. 2013;16(7):925.

72. Hardy NF, Buonomano DV. Neurocomputational models of interval and pattern timing. Current Opinion in Behavioral Sciences. 2016;8:250–257.

73. Burt CD. Reconstruction of the duration of autobiographical events. Memory & Cognition. 1992;20(2):124–132.

74. Quent JA, Henson RN, Greve A. A predictive account of how novelty influences declarative memory. Neurobiology of Learning and Memory. 2021;179:107382.

75. Block RA, Reed MA. Remembered duration: Evidence for a contextual-change hypothesis. Journal of Experimental Psychology: Human Learning & Memory. 1978;4(6):656–665. doi:10.1037/0278-7393.4.6.656.

76. Poynter WD. Duration judgment and the segmentation of experience. Memory & Cognition. 1983;11(1):77–82. doi:10.3758/bf03197664.

77. Poynter D. Chapter 8 Judging the Duration of Time Intervals: A Process of Remembering Segments of Experience. In: Advances in Psychology. Elsevier; 1989. p. 305–331. Available from: https://doi.org/10.1016/s0166-4115(08)61045-6.

78. Zakay D, Tsal Y, Moses M, Shahar I. The role of segmentation in prospective and retrospective time estimation processes. Memory & Cognition. 1994;22(3):344–351. doi:10.3758/bf03200861.

79. Zacks JM, Speer NK, Swallow KM, Braver TS, Reynolds JR. Event perception: A mind-brain perspective. Psychological Bulletin. 2007;133(2):273–293. doi:10.1037/0033-2909.133.2.273.

80. Radvansky GA, Zacks JM. Event boundaries in memory and cognition. Current Opinion in Behavioral Sciences. 2017;17:133–140. doi:10.1016/j.cobeha.2017.08.006.

81. Ezzyat Y, Davachi L. Similarity Breeds Proximity: Pattern Similarity within and across Contexts Is Related to Later Mnemonic Judgments of Temporal Proximity. Neuron. 2014;81(5):1179–1189. doi:10.1016/j.neuron.2014.01.042.

82. Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA. Discovering Event Structure in Continuous Narrative Perception and Memory. Neuron. 2017;95(3):709–721.e5. doi:10.1016/j.neuron.2017.06.041.

83. Ben-Yakov A, Henson RN. The Hippocampal Film Editor: Sensitivity and Specificity to Event Boundaries in Continuous Experience. The Journal of Neuroscience. 2018;38(47):10057–10068. doi:10.1523/jneurosci.0524-18.2018.

84. Montchal ME, Reagh ZM, Yassa MA. Precise temporal memories are supported by the lateral entorhinal cortex in humans. Nature Neuroscience. 2019;22(2):284–288. doi:10.1038/s41593-018-0303-1.

85. Bellmund JL, Deuker L, Doeller CF. Mapping sequence structure in the human lateral entorhinal cortex. eLife. 2019;8. doi:10.7554/elife.45333.

86. Bulkin DA, Sinclair DG, Law LM, Smith DM. Hippocampal State Transitions at the Boundaries between Trial Epochs. BioRxiv. 2018; p. 443077. doi:10.1101/443077.

87. Tsao A, Sugar J, Lu L, Wang C, Knierim JJ, Moser MB, et al. Integrating time from experience in the lateral entorhinal cortex. Nature. 2018;561(7721):57–62. doi:10.1038/s41586-018-0459-6.

88. Reynolds JR, Zacks JM, Braver TS. A Computational Model of Event Segmentation From Perceptual Prediction. Cognitive Science. 2007;31(4):613–643. doi:10.1080/15326900701399913.

89. Zacks JM, Kurby CA, Eisenberg ML, Haroutunian N. Prediction Error Associated with the Perceptual Segmentation of Naturalistic Events. Journal of Cognitive Neuroscience. 2011;23(12):4057–4066. doi:10.1162/jocn_a_00078.

90. Gershman SJ, Radulescu A, Norman KA, Niv Y. Statistical Computations Underlying the Dynamics of Memory Updating. PLoS Computational Biology. 2014;10(11):e1003939. doi:10.1371/journal.pcbi.1003939.

91. Zacks JM. Event Perception and Memory. PsyArXiv. 2019;doi:10.31234/osf.io/634dz.

92. Block RA. Prospective and Retrospective Duration Judgment: The Role of Information Processing and Memory. In: Time, Action and Cognition. Springer Netherlands; 1992. p. 141–152. Available from: https://doi.org/10.1007/978-94-017-3536-0_16.

93. Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N. A Hierarchy of Temporal Receptive Windows in Human Cortex. Journal of Neuroscience. 2008;28(10):2539–2550. doi:10.1523/jneurosci.5487-07.2008.

94. Hasson U, Chen J, Honey CJ. Hierarchical process memory: memory as an integral component of information processing. Trends in Cognitive Sciences. 2015;19(6):304–313. doi:10.1016/j.tics.2015.04.006.

95. Zacks JM, Braver TS, Sheridan MA, Donaldson DI, Snyder AZ, Ollinger JM, et al. Human brain activity time-locked to perceptual event boundaries. Nature Neuroscience. 2001;4(6):651–655. doi:10.1038/88486.

96. Howard MW, Kahana MJ. A Distributed Representation of Temporal Context. Journal of Mathematical Psychology. 2002;46(3):269–299. doi:10.1006/jmps.2001.1388.

97. Howard MW, Venkatadass V, Norman KA, Kahana MJ. Associative processes in immediate recency. Memory & Cognition. 2007;35(7):1700–1711. doi:10.3758/bf03193503.

98. Diamond NB, Levine B. Linking Detail to Temporal Structure in Naturalistic-Event Recall. Psychological Science. 2020;31(12):1557–1572. doi:10.1177/0956797620958651.

99. Rhodes D, Seth AK, Roseboom W. Multiple Duration Priors Within and Across the Senses. bioRxiv. 2018; p. 467027. doi:10.1101/467027.

100. Sengupta B, Tozzi A, Cooray GK, Douglas PK, Friston KJ. Towards a neuronal gauge theory. PLoS Biology. 2016;14(3):e1002400.

101. Hassabis D, Kumaran D, Maguire EA. Using imagination to understand the neural basis of episodic memory. Journal of neuroscience. 2007;27(52):14365–14374.

102. Zeidman P, Maguire EA. Anterior hippocampus: the anatomy of perception, imagination and episodic memory. Nature Reviews Neuroscience. 2016;17(3):173.

103. Tulving E. Memory and consciousness. Canadian Psychology/Psychologie canadienne. 1985;26(1):1.

104. Schacter DL, Addis DR, Hassabis D, Martin VC, Spreng RN, Szpunar KK. The future of memory: remembering, imagining, and the brain. Neuron. 2012;76(4):677–694.

105. Bulley A, Henry J, Suddendorf T. Prospection and the present moment: The role of episodic foresight in intertemporal choices between immediate and delayed rewards. Review of General Psychology. 2016;20(1):29–47.

106. Oettingen G, Sevincer AT, Gollwitzer PM. The Psychology of Thinking about the Future. The Guilford Press; 2018.

107. El Haj M, Moroni C, Samson S, Fasotti L, Allain P. Prospective and retrospective time perception are related to mental time travel: Evidence from Alzheimer's disease. Brain and Cognition. 2013;83(1):45–51.

108. Dement W, Kleitman N. Cyclic variations in EEG during sleep and their relation to eye movements, body motility, and dreaming. Electroencephalography and clinical neurophysiology. 1957;9(4):673–690.

109. Pienkos E, Giersch A, Hansen M, Humpston C, McCarthy-Jones S, Mishara A, et al. Hallucinations beyond voices: A conceptual review of the phenomenology of altered perception in psychosis. Schizophrenia bulletin. 2019;45(Supplement_1):S67–S77.

110. Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Blundell C, Hassabis D. Reinforcement learning, fast and slow. Trends in cognitive sciences. 2019;.

111. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on amazon mechanical turk. Judgment and Decision making. 2010;5(5):411–419.

112. Huff C, Tingley D. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. Research & Politics. 2015;2(3):2053168015604648.

113. Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, et al. Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS computational biology. 2019;15(4):e1006897.

114. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015;115(3):211–252.

115. Igel C, Hansen N, Roth S. Covariance matrix adaptation for multi-objective optimization. Evolutionary computation. 2007;15(1):1–28.

Supporting information

S1 Appendix. Analysis of the role of effort in memory recall.

S2 Appendix. Model training process.

S1 Fig. Full probabilistic graphical representation of the predictive processing model p. The random variables x_t^n, z_t^n correspond to different hierarchical representations of the input, continuous and categorical respectively, the variables $\mu_{t,i}^n, \sigma_{t,i}^n$ are the parameters of the Gaussian component i and the variables K_t^n represent the number of components at time t. All solid-line arrows represent conditional dependencies: black arrows denote dependencies between random variables within the same simulation time step, red arrows highlight dependencies between variables with distributions that evolve over different time scales and gray arrows denote dependencies across different time steps.

S2 Fig. Counting vs not-counting. Interaction in human duration reports tasks (pro/retrospective) and cognitive load for participants that reported they were counting (**A**), not counting (**B**), or both cases together (**C**), having excluded the rejected trials as described in methods.

S3 Fig. Level of understanding of the task. Interaction in human duration reports tasks (pro/retrospective) and cognitive load for participants depending on how clear they reported they found the initial instructions of task, having excluded the rejected trials as described in methods.

S4 Fig. Demographics of the human participants. This includes geographic location and first language.

S5 Fig. Statistical properties of the model related to episodic memory recall, per scene type and cognitive load level. The episodic memory model shows unique patterns for the 3 different scene types, although no single feature is clearly suitable by itself or superior to alternative solutions to directly select the network layers conv1, conv3 and fc7 as the source of information when estimating time in the corresponding scene types, as shown in Results. **A:** Standard deviation of recalled events per second. Regardless of cognitive load, lower hierarchical layers show far greater variability and, thus, are more affected by the prior of the overall number of events usually recorded. **B:** Number of unique components (different types of events) captured in single episodes. Here, there is a similar overall pattern with A, although more individual components are recorded in cases of high cognitive load. In addition, when comparing different scene types, the unique number of captured components is negatively correlated with the standard deviation of the recalled events. **C:** Accuracy of recalls measured as the number of episodic nodes (events) recalled with components assigned divided by the number of all events recalled in each episode. In cases of low cognitive load only very high-level events can be recalled accurately and estimation is based almost entirely on prior knowledge.

S6 Fig. Effects of attention on event accumulation and recall per layer, for the prospective and retrospective tasks respectively. **A**: Slope between the number of accumulated events using low- and high-load hyper parameters. This figure illustrates the slope instead of the absolute number of events in the two task variants, as the absolute number of events is normalized when accumulations are mapped to seconds via linear regression. **B**: The error between the resulting slopes of the computational model and the slopes resulting from human time judgements. The green dots represent the points in the parameter space that were selected and used in Results. This figure shows that in event accumulation and recall with low effort, small differences between c_{low} and c_{high} are enough to replicate human behaviour. However, under a high amount of recall effort, attention needs to be either significantly lower, in which case human results can be replicated even when $c_{\text{low}} = c_{\text{high}}$, or c_{low} should become significantly lower.

S7 Fig. Model comparison. Human duration judgements versus the different versions of the computational model proposed here, as well as the model in [4] that uses "static" frames, for the four variants of the task. The coloured lines and shades areas represent the mean and standard error of the mean respectively.

S8 Fig. Recalled events per second versus human reports. Mean absolute error between the slope of recalled events per second for low and high cognitive load versus the slope of humans mean duration judgement ratio for low and high cognitive load. Human reports have been segmented per scene type while the graphs compare this error for all hierarchical layers of the neural network. The error bars show standard error of the mean.

S1 Table. Probability distributions and corresponding parameters associated with each of the two memory types. Semantic in green and episodic in orange at time t.

S2 Table. Parameters of the model that can be trained recursively in real-time, along with the training method used.

S3 Table. Comparison of the proposed single-layer models using the Akaike information criterion (AIC). All listed models were fitted to the real video durations of the same trials that were used for testing. Testing was performed with either real (objective) video durations or with human subjective duration judgements.

S4 Table. Glossary of terms and notation. This glossary includes all notation except the parameters found in S2 Table.

Acknowledgements

This work was supported by the European Union Future and Emerging Technologies grant (GA:641100) TIMESTORM—Mind and Time: Investigation of the Temporal Traits of Human-Machine Convergence and the Sackler Centre for Consciousness Science. The authors would like to thank Karl Friston for insightful comments on the theoretical model presented here and Chris Bird for comments on a previous version of the manuscript. This research has been approved by the Science and Technology Cross-Schools Research Ethics Committee (C-REC).

S1 Appendix

Analysis of the role of effort in memory recall

The final question we investigated is whether, according to our model, the amount of effort that humans commit to recall episodic memories affects their re-experienced sense of duration for these events. Recall effort here is defined as a mechanism that increases the chances of recorded salient events to be retrieved and used as branches of the tree structure shown in Figure 4 (black nodes). The more effort is devoted, the more particular components $z^n_{k,t}$ will be part of the recalled experience. Hence, more specific priors $P_{\text{chil.}}(z^n_{k,t})$ will be used in the recall process over the general prior $P_{\text{chil.}}(n)$, to estimate the number of children nodes per recalled node, and deeper in the tree structure $\text{Tr}(e^N_{\mathbf{0}})$ (see Figure 4). To assess how this mechanism affects the rate of recalling events, we ran a series of simulations over the complete set of human trials, where the system received the same video frames as humans and then it produced memory recalls for $\epsilon \in \{1, 2, ..., 70\}$. As with all simulation results presented here, $P_{\text{chil.}}(z^n_{k,t})$ was approximated as the average number of children nodes of the component $z^n_{k,t}$ in a single given trial, since only one trial was shown to each subject. In contrast, $P_{\text{chil.}}(n)$ was obtained from the distribution of children nodes $n-1$ for all components in layer n, after running simulations for 3300 frames including all scene types. Figure 9A illustrates the resulting mean values and distributions for all layers of the network. These were 1.2 ± 0.59 children nodes for the convolutional layers (*conv1-5*) and 2.9 ± 2.67 nodes for the fully connected layers (*fc6-7* and *output*), more than twice the values of the former type of layers.

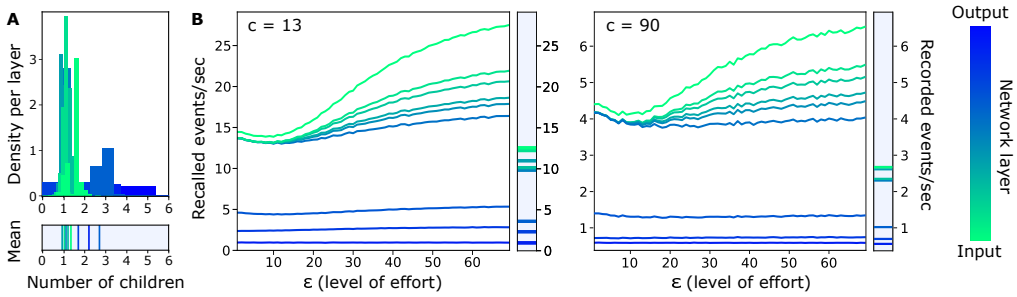


Fig 9. A: Distribution of children nodes recorded per layer over all components of this layer, used to calculate $P_{\text{chil.}}(n)$. **B:** Rate of recalling events in different layers of the model and for different levels of effort ϵ, defined as the number of recall attempts. The coloured lines represent the mean. Standard error was not included in this plot as it overlaps with the mean.

The overall effect of ϵ in the rate of recalling events can be seen in Figure 9B. As expected, the layers that are more sensitive to changes in ϵ are the lower convolutional ones, since recall is a recursive process that originates from the top (most contextual) layer and has a cumulative effect as it proceeds. For these low layers, high values of ϵ resulted in higher number of recalled events, and therefore longer time estimations. Interestingly, this relation was not monotonic and it was maintained for very different levels of attention (e.g. $c = \{13, 90\}$).

Further analysis Figure 9B illustrates that the relation between effort and accumulated events over time is consistent across different levels of attention and it is non-monotonic. An explanation for this complex behaviour can be summarised as follows. In high layers, where a small number of events occurs in a single episode leaving minimal prior knowledge for individual components $z^n_{k,t}$, the distribution $P_{\text{chil.}}(z^n_{k,t})$ is

closer to the real number of recorded events. That is, when trying to recall how many salient events happened during a memory, the richer this memory is in high-level contextual information, the more biased the recall will be by the actual experience. On the other hand, the estimated $P_{\text{chil.}}(z^n_{k,t})$ has large variance in low layers, since individual components are more likely to be used multiple times. Indeed, when $\epsilon = 1$, only the $\sim 5\%$ of the overall recalled events corresponded to nodes from the original episode with particular components z^n_k (black nodes in Figure 4), while when $\epsilon = 70$, the overall accuracy increased to more than $\sim 57\%$. In addition, the rate of recalled events showed an initial tendency to approach the real number of recorded events, indicated by the bars on the right-hand side of the sub-figures in Figure 9.B. This changed when $\epsilon > 7$, where the first components of the convolutional layers were recalled, leading to a massive increase in recalled events in these layers.

Relation to attention To further examine how recall effort interacts with attention, and affects the difference between the model's prospective and retrospective duration judgments, we ran simulations where low and high cognitive load trials were represented by different values of the attention decay time constant c, used in Eq (13). Low values of c cause the surprise threshold to decay faster, hence denoting higher attention. Additionally, ϵ was set to 1 in trials with low cognitive load while, in the opposite case, it was either also 1 or raised to 70. Clearly, when $c_{\text{low}} = c_{\text{high}}$ and if there is no effort difference, the rate of recalling events, and therefore duration judgements, is identical in both cases. When attention is considered higher in low cognitive load cases, i.e. $c_{\text{low}} < c_{\text{high}}$, the slope between duration judgements resembles the one found in human prospective judgements shown in Figure 5.B. The opposite slope (found in human retrospective judgements) can be achieved when higher effort is used in high cognitive load trials. This interplay between attention and recall effort was robust for a wide range of values of c. In fact, the effect of recall effort was significantly more visible in low layers of the network revealing the mechanism that leads to the different patterns of Figure 6.

For more information and a visualization of the model's behaviour, see S6 Fig.

S2 Appendix

Model training process

In the algorithm that follows, pll, phl, rll and rhl correspond to the four variants of the current experiment, i.e. prospective / retrospective and low / high cognitive load respectively, X_i is a matrix that contains accumulated events for $i \in \{\text{pll, phl, rll, rhl}\}$, $MDJR_i$ represents the mean duration judgement ratio in human trials for i, y a vector that contains the video durations of all trials and MSE stands for mean squared error.

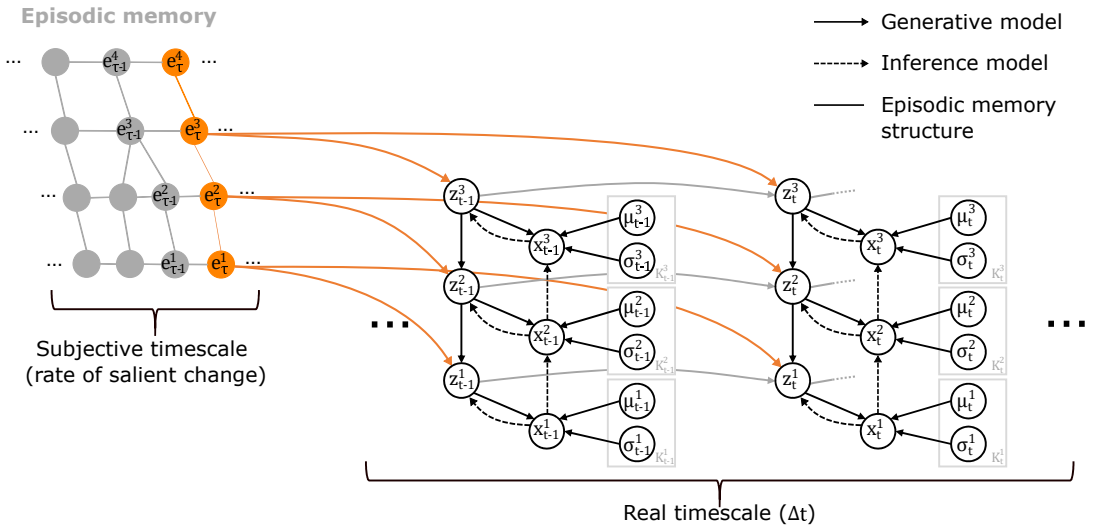
Algorithm 2: Complete model training

Train neural network weights and biases θ_n using a dataset of images and labels.
Train generative model parameters (Table 1) using MO-CMA-ES, along with labels of objects in the videos.
Train all parameters of semantic memory (green parameters in S1 Table), using only random samples from the recorded videos in an unsupervised fashion. For each type of parameters, the method described in S2 Table was used.
Fit the parameter set $(c_{\text{low prosp}}, c_{\text{high prosp}}, c_{\text{retro}}, \epsilon_{\text{low load}}, \epsilon_{\text{high load}})$ from Table 1, obeying preset value constraints described in Methods and using CMA-ES to minimize the cost function *similarity_to_human_slopes*().

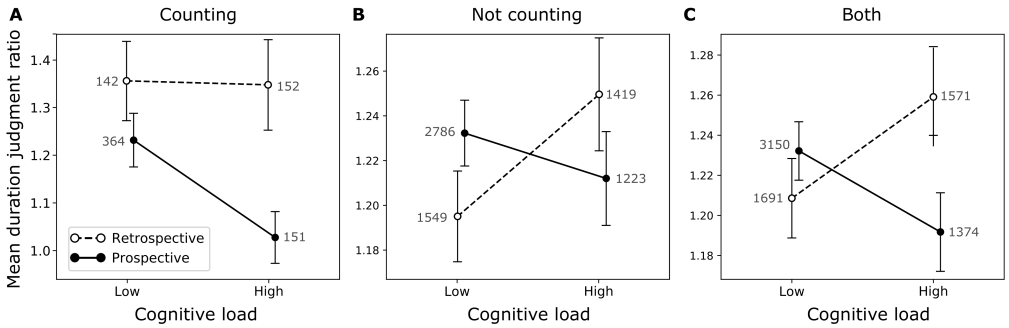
Algorithm 3: Cost function for free parameter optimization

function *similarity_to_human_slopes*(model, $c_{\text{pll}}, c_{\text{phl}}, c_{\text{retro}}, \epsilon_{\text{rll}}, \epsilon_{\text{rhl}}$):
    $X_{\text{pll}} \leftarrow$ model.memory_formation(videos of all trials $|c_{\text{pll}}$)
    $X_{\text{phl}} \leftarrow$ model.memory_formation(videos of all trials $|c_{\text{phl}}$)
    $X_{\text{prospective events}} = \text{concatenate}(X_{\text{pll}}, X_{\text{phl}})$
    $w = \text{linear_support_vector_regression}(X_{\text{prospective events}}, y)$
    $X_{\text{rll}} \leftarrow$ model.memory_recall(videos of all trials$|c_{\text{retro}}, e_{\text{rll}}$)
    $X_{\text{rhl}} \leftarrow$ model.memory_recall(videos of all trials$|c_{\text{retro}}, e_{\text{rhl}}$)
    $\text{slopes}^{\text{model}}_{\text{prosp}} = \left(\sum_n w_n * X_{n,\text{pll}}/y\right) / \left(\sum_n w_n * X_{n,\text{phl}}/y\right)$
    $\text{slopes}^{\text{model}}_{\text{retrosp}} = \left(\sum_n w_n * X_{n,\text{rll}}/y\right) / \left(\sum_n w_n * X_{n,\text{rhl}}/y\right)$
    $\text{slopes}^{\text{human}}_{\text{prosp}} = \text{MDJR}_{\text{pll}}/\text{MDJR}_{\text{phl}}$
    $\text{slopes}^{\text{human}}_{\text{retro}} = \text{MDJR}_{\text{rll}}/\text{MDJR}_{\text{rhl}}$
    return $\text{MSE}(\text{slopes}^{\text{model}}_{\text{prosp}}, \text{slopes}^{\text{human}}_{\text{prosp}}) + \text{MSE}(\text{slopes}^{\text{model}}_{\text{retro}}, \text{slopes}^{\text{human}}_{\text{retro}})$

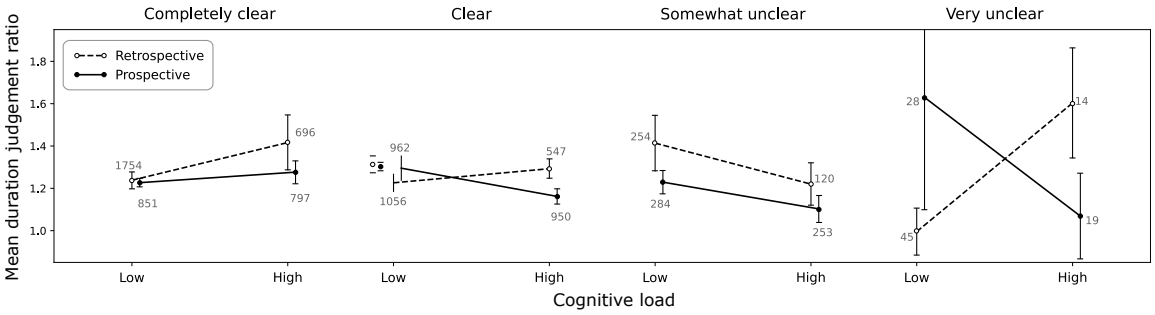
S1 Fig.



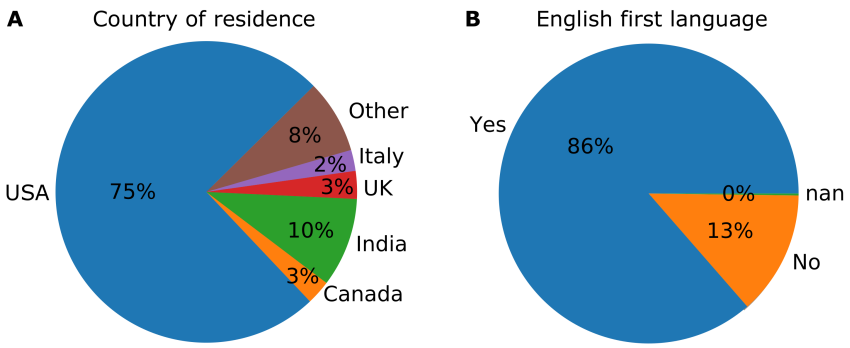
S2 Fig.



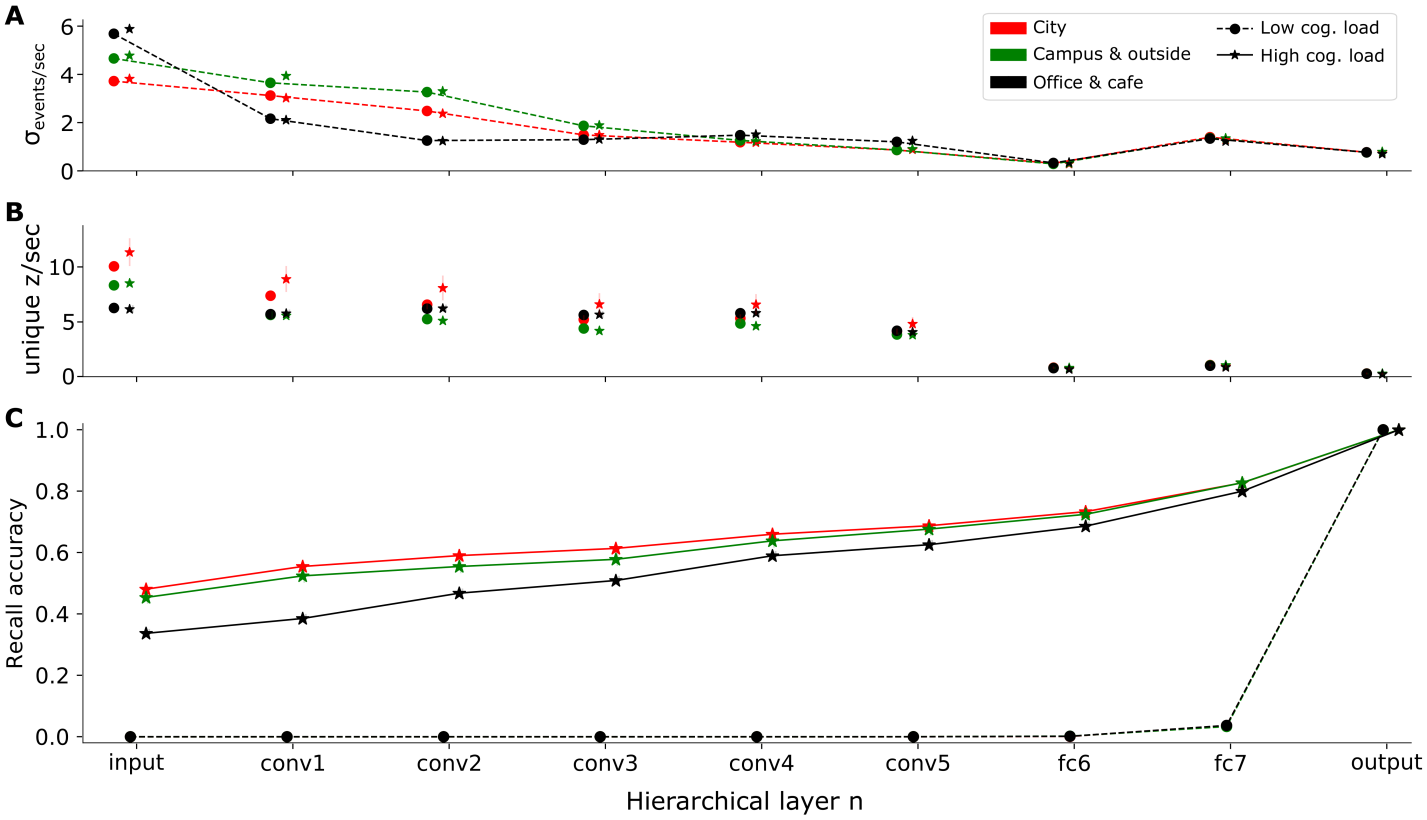
S3 Fig.



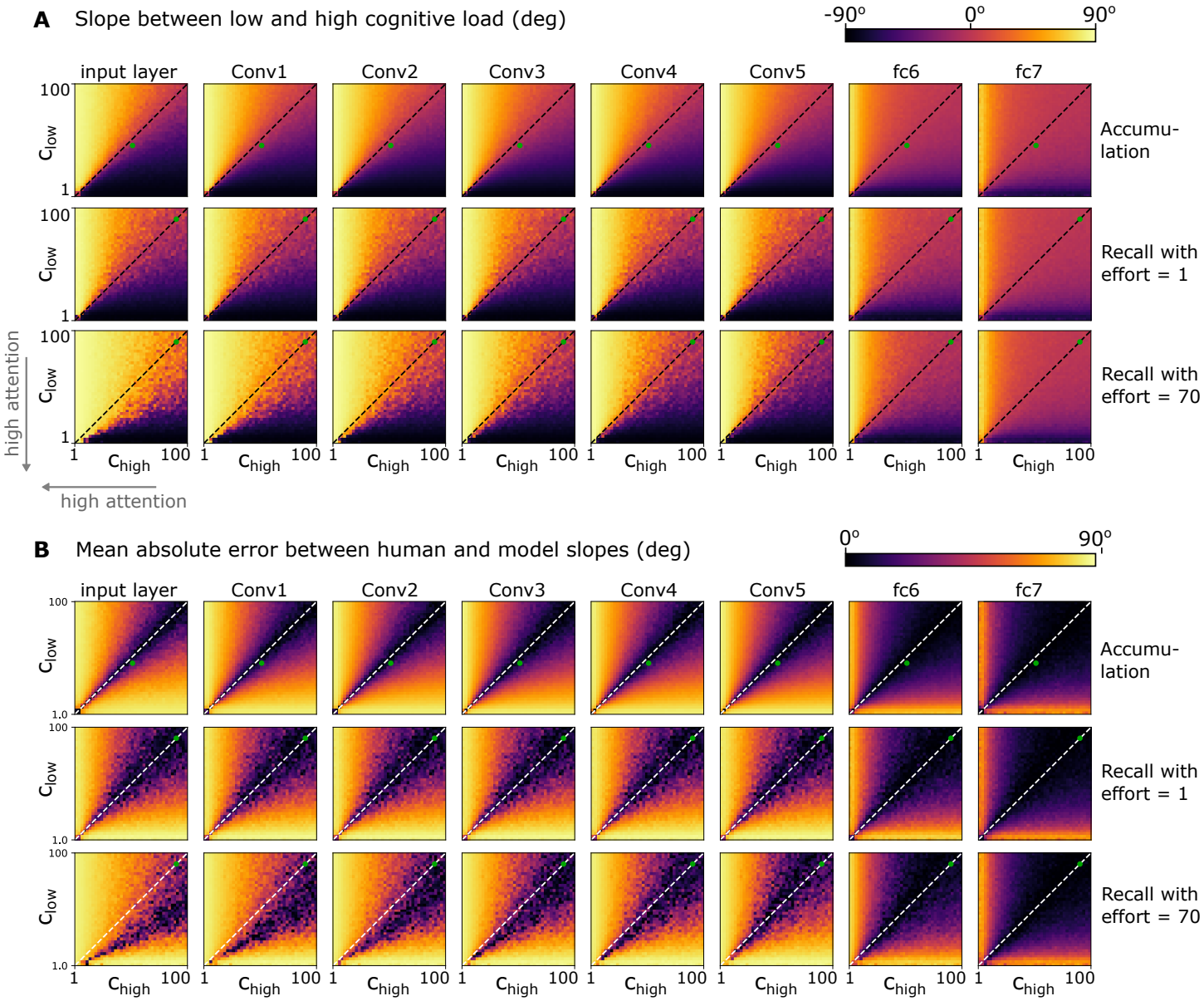
S4 Fig.


A
Country of residence
USA
75%
Other
8%
Italy
2%
UK
3%
India
10%
Canada
3%
B
English first language
Yes
86%
nan
0%
No
13%

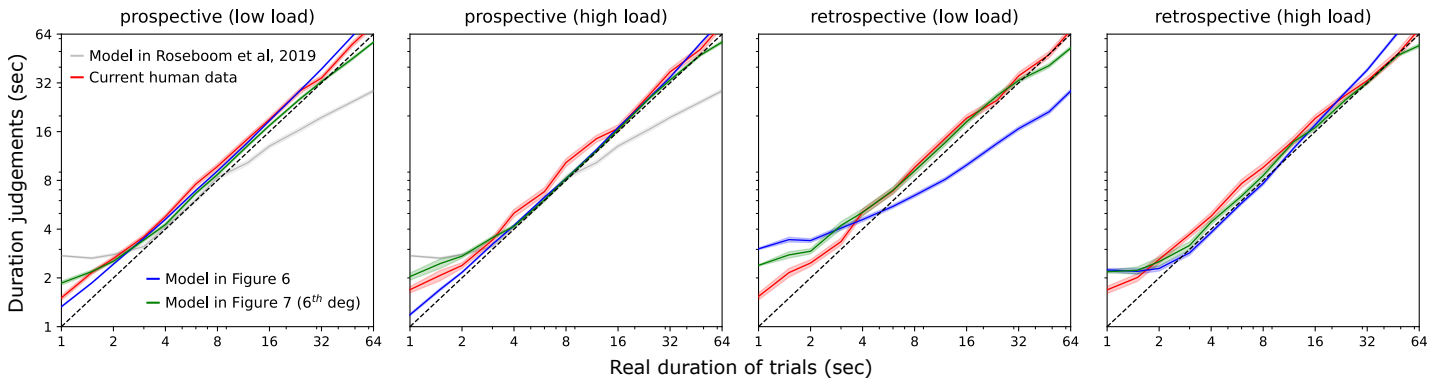
S5 Fig.


A
$\sigma_{events/sec}$
City
Campus & outside
Office & cafe
Low cog. load
High cog. load
B
unique z/sec
C
Recall accuracy
input
conv1
conv2
conv3
conv4
conv5
fc6
fc7
output
Hierarchical layer n

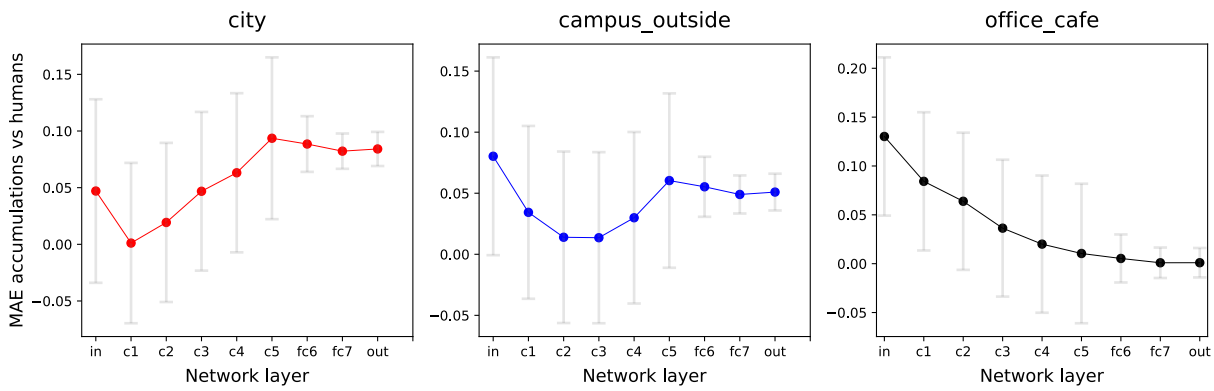
S6 Fig.


A
Slope between low and high cognitive load (deg)
-90°
0°
90°
input layer
Conv1
Conv2
Conv3
Conv4
Conv5
fc6
fc7
Accumu-
lation
Recall with
effort = 1
Recall with
effort = 70
C_low
C_high
1
100
high attention
high attention
B
Mean absolute error between human and model slopes (deg)
0°
90°
1.0

S7 Fig.


prospective (low load)
prospective (high load)
retrospective (low load)
retrospective (high load)
Model in Roseboom et al, 2019
Current human data
Model in Figure 6
Model in Figure 7 (6th deg)
Duration judgements (sec)
Real duration of trials (sec)

S8 Fig.


city
campus_outside
office_cafe
MAE accumulations vs humans
in
c1
c2
c3
c4
c5
fc6
fc7
out
Network layer

S1 Table

Model	Distributions	Associated trainable parameters
Generative	$P(z_t^n \| z_t^{n+1}, z_{t-1}^n, e_\tau^n) = Cat(\boldsymbol{A_t})$	$\boldsymbol{A_t}(z_t^{n+1}, z_{t-1}^n, e_\tau^n), \boldsymbol{K_t^n}, \boldsymbol{V} = \{e_0^0, e_1^0, ..., e_0^1, ...\}$
	$P(x_t^n \| z_t^n) = N(x_t^n \| \boldsymbol{\mu_{t,i}^n}, \boldsymbol{\Sigma_{t,i}^n}, z_t^n == i)$	$\boldsymbol{\mu_{t,i}^n}, \boldsymbol{\Sigma_{t,i}^n}$
Inference	$P(x_t^n \| x_t^{n-1}) = f_{\boldsymbol{\theta_n}}(x_{t-1}^{n-1})$	$\boldsymbol{\theta_n}$
	$P(z_t^n \| x_t^n) =$ $\text{argmax}_{i \in \boldsymbol{K_t^n}} \left[N(x_t^n \| \boldsymbol{\mu_{t,i}^n}, \boldsymbol{\Sigma_{t,i}^n})\right]$	$\boldsymbol{\mu_{t,i}^n}, \boldsymbol{\sigma_{t,i}^n}, \boldsymbol{K_t^n}$
Episodic recall	$P_{\text{evidence}}(e_\tau^n) = \boldsymbol{W}(e_\tau^n)$	$\boldsymbol{V} = \{e_0^0, e_1^0, ..., e_0^1, ...\}, \boldsymbol{W}(e_\tau^n)$
	$P_{\text{chil}}(z_{k,t}^n) = N(\boldsymbol{\mu_{\text{ch},k}}, \boldsymbol{\sigma_{\text{ch},k}})$	$\boldsymbol{\mu_{\text{ch},k}}, \boldsymbol{\sigma_{\text{ch},k}}$
	$P_{\text{chil}}(n) = N(\boldsymbol{\mu_{\text{ch}}^n}, \boldsymbol{\sigma_{\text{ch}}^n})$	$\boldsymbol{\mu_{\text{ch}}^n}, \boldsymbol{\sigma_{\text{ch}}^n}$

S2 Table

Symbol	Parameter description	Training method
$A_t()$	Transition probabilities table	Event counter
K_t^n	Number of components per layer	Splitting and merging Eq. 6-7
V	All episodic nodes	Episodic formation Eq. 8
$\mu_{t,i}^n, \Sigma_{t,i}^n$	Statistics of Gaussian component i	Kalman filter with short-term plasticity Eq. 2-3
θ_n	Neural network weights and biases	Back-propagation and gradient descent
$W(e_\tau^n)$	Weights of episodic nodes	Assignment of $P_{\text{evidence}}(x_t^n)$ value when event originally recorded.
$\mu_{\text{ch},k}, \sigma_{\text{ch},k}$	Number of children nodes for component k	Calculated recursively after every new episodic node.
$\mu_{\text{ch}}^n, \sigma_{\text{ch}}^n$	Number of children nodes for layer n	

S3 Table

Model	vs human reports	vs real durations	No. of parameters
Linear (single layer)	1115	-1613	2
Polynomial 2nd deg	1067	-4961	3
Polynomial 4th deg	635	-4997	5
Polynomial 6th deg	**417**	**-5361**	7
Polynomial 8th deg	1854	-35	9

S4 Table

Notation	Description
$n \in \{1, 2, ...\}$	Index of hierarchical layer.
t	Physical time (simulation time step).
τ	Subjective time defined as an index of episodic events.
x_t^0	Random variable (continuous) that represents sensory information, in this case frames of a video. This variable follows a mixture of Gaussians.
x_t^n	Random variable (continuous) that represents activity of the layer n of the feedforward neural network used, at time t. This variable follows a mixture of Gaussians.
ξ_t^n	Error between predicted and observed x_t^n.
z_t^n	Indicative (discrete) variable that represents which Gaussian component is used at time t.
e_τ^n	The τ-th event recorded in episodic memory.
$\mu_{t,k}^n, \Sigma_{t,k}^n$	The short-term parameters of the k-th Gaussian component.
$\tilde{\mu}_{t,k}^n, \tilde{\Sigma}_{t,k}^n$	The baseline (long-term) parameters of the k-th Gaussian component.
K	Kalman gain.
Q, R	Free noise parameters of Kalman filter.
k_{back}	Model parameter, represents how quickly short-term parameters of a Gaussian component converge back to their baseline.
k_{base}	Model parameter, represents how quickly the baseline parameters adapt to new information.
$D_{JS}(.)$	Jensen-Shannon divergence.
$D_{JS}(.)$	Kullback–Leibler divergence.
T_{merge}	Threshold constant that defines when two Gaussian components merge.
T_{split}	Threshold constant that defines when a Gaussian component splits into two (short-term and baseline parameters become independent).
T_t^n	Threshold variable describing the tolerated level of surprise.
$\text{Tr}(e)$	A tree defined as a subset of adjacent recorded episodic events with root node e.
$P_{\text{chil.}}(z_{k,t}^n)$	Prior belief of number of children nodes in layer $n-1$ when the parent node is known (and classified as $z_{k,t}^n$).
$P_{\text{chil.}}(n)$	Prior belief of number of children nodes per recalled node in layer n.
$P_{evidence}$	Equivalent to $P(x_t^n \| z_{t-1}^n, z_t^{n+1}, e_\tau^n)$.
$f(.)$	Function that maps event accumulations to human readable units of time (e.g. seconds).
τ_{start}^n	The number of episodic nodes in a layer n at the beginning of an episode that is defined by the tree $\text{Tr}(e^{n^2})$, where $n^2 > n$.
τ_{end}^n	Same as before, for nodes at the end of an episode.
$\tilde{\tau}_{\text{tree}}^n$	The number of nodes in each layer n of a tree $\text{Tr}(e)$.
D^n	Number of simulation iterations since the last time the threshold T_t^n was reset.
c	Decaying time constant for threshold T_t^n. Used to model the effect of attention.
ϵ	Recall effort.
$N(.)$	Gaussian distribution.
$Bern(.)$	Bernoulli distribution.