Open access • Posted Content • DOI:10.1101/2020.02.17.953133

A predictive processing model of episodic memory and time perception

Zafeirios Fountas, Anastasia Sylaidi, Kyriacos Nikiforou, Anil K. Seth ...+2 more authors

Institutions: University College London, Imperial College London, University of Sussex

Published on: 17 Feb 2020 - bioRxiv (Cold Spring Harbor Laboratory)

Topics: Time perception, Perception, Recall, Episodic memory and Cognitive load

A predictive processing model of episodic memory and time perception

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

[1]Department of Computing, Imperial College London, London, UK
[2]Department of Informatics and Sackler Centre for Consciousness Science, University of Sussex, Sussex, UK
[3]Emotech Labs, London, UK
[4]Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, UK
[*]fountas@outlook.com

ABSTRACT

Human perception and experience of time is strongly affected by environmental context. When paying close attention to time, time experience seems to expand; when distracted from time, experience of time seems to contract. Contrasts in experiences like these are common enough to be exemplified in sayings like “time flies when you’re having fun”. Similarly, experience of time depends on the content of perceptual experience – more rapidly changing or complex perceptual scenes seem longer in duration than less dynamic ones. The complexity of interactions among stimulation, attention, and memory that characterise time experience is likely the reason that a single overarching theory of time perception has been difficult to achieve. In the present study we propose a framework that reconciles these interactions within a single model, built using the principles of the predictive processing approach to perception. We designed a neural hierarchical Bayesian system, functionally similar to human perceptual processing, making use of hierarchical predictive coding, short-term plasticity, spatio-temporal attention, and episodic memory formation and recall. A large-scale experiment with $\sim 13,000$ human participants investigated the effects of memory, cognitive load, and stimulus content on duration reports of natural scenes up to ~ 1 minute long. Model-based estimates matched human reports, replicating key qualitative biases including differences by cognitive load, scene type, and judgement (prospective or retrospective). Our approach provides an end-to-end model of duration perception from natural stimulus processing to estimation and from current experience to recalling the past, providing a new understanding of this central aspect of 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 and behaviour. Despite being a core topic of interest since the very beginning of psychology, the cognitive and neural processes that underlie human time perception remain largely unknown. Human perception of time is affected by many factors; experience on the scale of seconds is most prominently influenced by the content, complexity, and rate of change of experience[1–4] and whether attention can and is being directed toward the task of monitoring time[5–7]. Specifically, increasing cognitive load (by requiring attention to additional tasks beyond tracking time) has been reported to decrease the apparent duration of an interval when a person is simultaneously and actively attending to time (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[7] and Figure 1). Differences in time perception based on this interaction between cognitive load and prospective vs. retrospective duration judgements has led to suggestions that different mechanisms underlie the different scenarios: when actively attending to time (prospective time) the process is largely "attention" driven; conversely, when reflecting on a period of time after it has occurred (retrospective time) the process is largely driven by "memory"[5,7–9]. Reflecting these different approaches, models of time perception and episodic memory have developed largely independently since diverging several decades ago (approach in[10] versus[11], for example), precluding a unified account of time perception. To address this problem, here we present a single model of human time perception and episodic memory that integrates the influences on human time estimation of stimulus, attention, and whether the observer estimates duration prospectively or retrospectively. We validate model performance by showing that model-based estimates of naturalistic videos between 1 and 64 seconds match duration reports obtained from $\sim 13,000$ human participants regarding the same videos, for each of the key manipulations: knowingly attending to time (prospective judgements) or not (retrospective judgements), attending to only a single or multiple concurrent tasks (cognitive load), and for different types of natural stimulation (busy city scenes, less busy scenes on a university campus or leafy surrounds, or in a quiet office of cafe).

The present model is based on the proposal that the primary function of human perception is the classification of objects and events in the world and that estimating time can be accomplished simply by tracking the behaviour of neural networks that implement perceptual classification. It was recently demonstrated[4] that tracking salient events in the activity of an artificial image classification network while it processed videos of natural scenes (from 1-64 seconds in duration) can provide a basis for estimates of duration. In that study, model-based estimates replicated key features of human duration estimates of the exact same videos, including biases related to scene type (busy city scenes estimated as longer that quiet office scenes, for example; see also[12, 13]). The 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[14–18]. However, this previous model[4] had only very simple "memory", in the form of content-less tokens marking the occurrence of salient events in perception. Because it lacked the ability to form content-specific memories of an episode, it was incapable of estimating time in retrospect. 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, making use of both episodic and semantic memory concepts. Basing our approach on a predictive-processing account that allows for interactions between sensory input, attention, and previous experience, we provide a unified model of episodic memory and time perception.

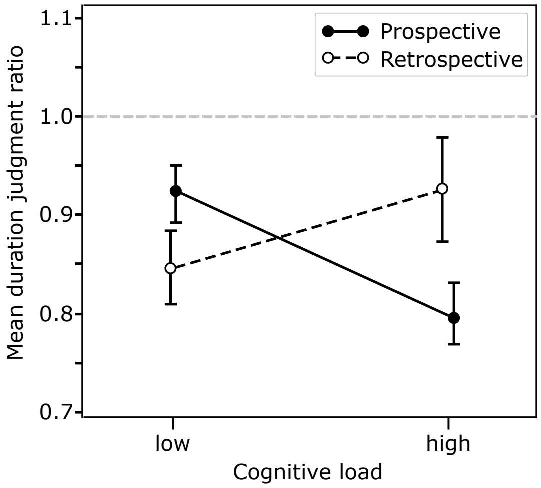


Figure 1. Relationship between prospective and retrospective duration judgments and their interaction with cognitive load (adapted from[7]). Duration judgement ratio is reported duration divided by physical duration. Error bars indicate standard error of the mean. Increasing cognitive load decreases reported duration in a prospective timing task, but increases reported duration in a retrospective timing task.

A predictive processing model of human episodic memory and time perception

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[14, 17, 19, 20]. The resulting 'top-down' generative model is used to predict new states and simulate the outcome of prospective actions[21]. 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[22], or as the result of the interplay between competing oscillatory bands[23, 24], as well as providing theoretical explanations for a wide range of cognitive and perceptual effects such as hallucinations[25], autism[26] and schizophrenia[27]. 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[28,29], 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 $P(x_t^n|x_t^{n-1})$.

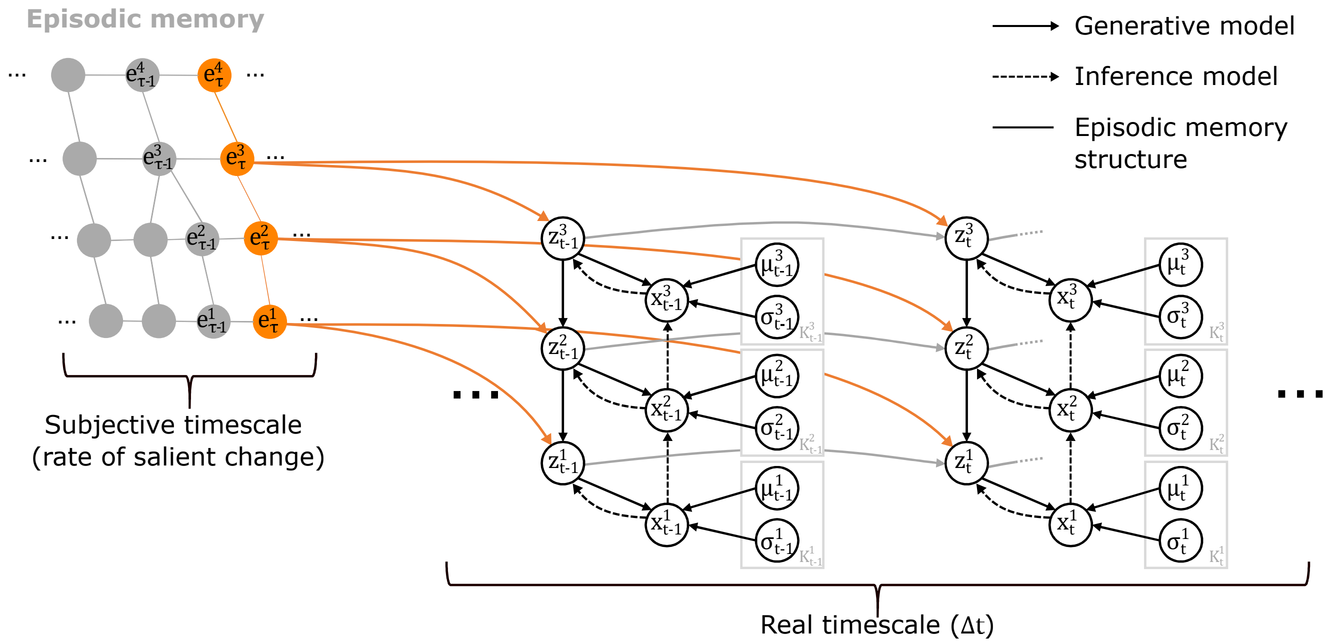


Figure 2. Probabilistic graphical representation of the predictive processing model described in the text. The random variables x_t^n, z_t^n correspond to different hierarchical representations of the input, continuous and categorical respectively, the variables $\mu_{t,i}^n, \sigma_{t,i}^n$ are the 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.

The structure of the generative model (solid arrows in Figure 2) 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 $N\left(\mu_{t,i}^n, \Sigma_{t,i}^n\right)$ update over time with recursive Bayesian estimation. 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,

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

The selected component is indicated by the (discrete) random variable $z_t^n = \{z_{t,i}^n | i = 1, 2, ..., K\}$ 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 $\arg\max_i N\left(x_t^n|\mu_{t,i}^n, \Sigma_{t,i}^n\right)$. 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[30] 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

$$\mu_{t,i}^n = \begin{cases} \mu_{t-1,i}^n + K \cdot \xi_t^n, & \text{if } i \text{ selected.} \\ \mu_{t-1,k}^n + k_{\text{back}} \cdot \left(\tilde{\mu}_{t-1,i}^n - \mu_{t-1,i}^n\right), & \text{otherwise.} \end{cases}, \qquad \Sigma_{t,i}^n = \begin{cases} \left(1.0 - K\right) \cdot \left(\Sigma_{n,t,i}^n + Q\right), & \text{if } i \text{ selected.} \\ \Sigma_{t-1,k}^n + k_{\text{back}} \cdot \left(\tilde{\Sigma}_{t-1,i}^n - \Sigma_{t-1,i}^n\right), & \text{otherwise.} \end{cases} \tag{2}$$

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

$$\tilde{\mu}_{t,i}^n = \tilde{\mu}_{t-1,i}^n + k_{\text{base}} \cdot \left(\mu_{t-1,i}^n - \tilde{\mu}_{t-1,i}^n\right) \quad \text{and} \quad \tilde{\Sigma}_{t,i}^n = \tilde{\Sigma}_{t-1,i}^n + k_{\text{base}} \cdot \left(\Sigma_{t-1,i}^n - \tilde{\Sigma}_{t-1,i}^n\right). \tag{3}$$

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

$$\hat{x}_t^n \sim P\left(x_t^n, z_{t,i}^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n\right) = P\left(x_t^n | z_t^n\right) P\left(z_t^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n\right) = N\left(x_t^n | \mu_t^n, \Sigma_t^n, z_t^n\right) P\left(z_t^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n\right) \tag{4}$$

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

$$-\log P\left(x_t^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n\right) = -\log \sum_{i=1}^{K} P\left(x_t^n | z_t^n = z_i\right) P\left(z_t^n = z_i | z_{t-1}^n, z_t^{n+1}, e_\tau^n\right) \tag{5}$$

Splitting and merging categories Humans are able to rearrange taxonomic relations in their memory, which often involves learning new concepts or combining two existing ones into a single. 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[31,32]. Here, two components i and j at time t are merged iff

$$D_{JS}\left(N(\tilde{\mu}_{t,i}^n, \tilde{\Sigma}_{t,i}^n), N(\tilde{\mu}_{t,j}^n, \tilde{\Sigma}_{t,j}^n)\right) < T_{\text{merge}} \tag{6}$$

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

$$D_{KL}\left(N(\tilde{\mu}_{t,i}^n, \tilde{\Sigma}_{t,i}^n), N(\mu_{t,i}^n, \Sigma_{t,i}^n)\right) > T_{\text{split}} \tag{7}$$

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 $N(\tilde{\mu}_{t,i}^n, \tilde{\Sigma}_{t,i}^n)$ can be considered the *true* distribution of the representation of the component i and, thus, the equation (7) measures the amount of information lost when $N(\mu_{t,i}^n, \Sigma_{t,i}^n)$ is used instead of this true baseline. In contrast, equation (6) 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[33], 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[34] (see Supplementary Tables 2 and 3 for more details on how our model relates to mammalian 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 almost 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 first-person 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 deserve more to be encoded involve attention[35] and prediction error[36–38]. 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_t^n well enough. That is, iff

$$-\log P_{\text{evidence}}\left(x_t^n\right) > -\log T_t^n \qquad \iff \qquad P_{\text{evidence}}\left(x_t^n\right) < T_t^n \tag{8}$$

where $P_{\text{evidence}}\left(x_t^n\right) = P\left(x_t^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n\right)$ and $T_t^n \in (0,1)$ 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_t^n is exceeded, the tuple (z_t^n, x_t^n) is registered as a node $e_{\tau+1}^n$ (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 $P\left(x_t^n | z_{t-1}^n, z_t^{n+1}, e_\tau^n\right)$ is assigned as the weight of this node, while $e_{\tau+1}^n$ also has a direct connection to its preceding node e_τ^n 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.

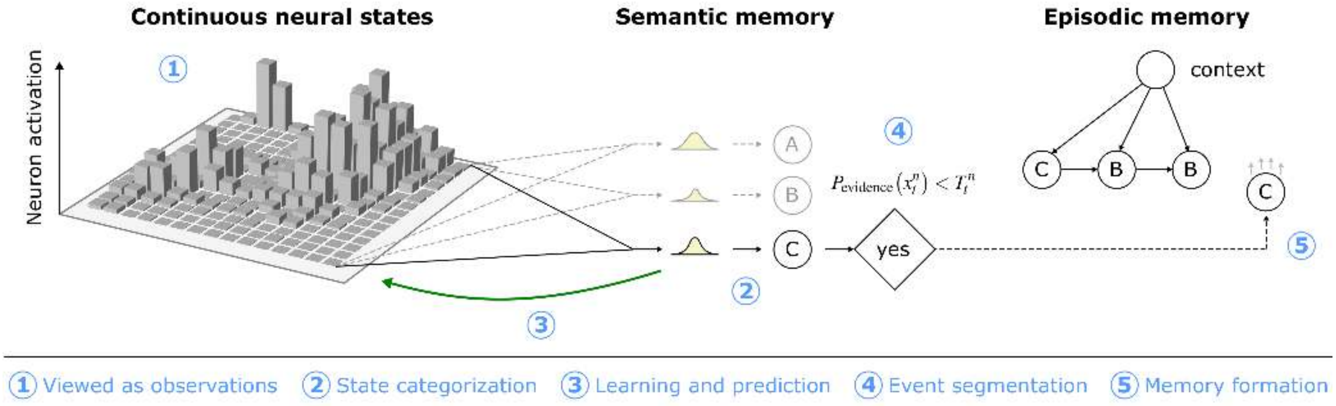


Figure 3. Diagram of the processes involved in episodic memory formation. At each time step, the neural state x_t^n is clustered using online Gaussian mixtures and prior information. An episodic event is generated when surprise of observing x_t^n exceeds the threshold of equation (8).

Relation to prospective timing The inequality (8) 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 rapid 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.

$$\text{prospective duration judgement} = f(\tau_{\text{end}}^1 - \tau_{\text{start}}^1, \tau_{\text{end}}^2 - \tau_{\text{start}}^2, \ldots, \tau_{\text{end}}^N - \tau_{\text{start}}^N) \tag{9}$$

where $\tau_{\text{start/end}}^n$ 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).

Episodic memory recall and retrospective timing

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[39]. For instance, when asked to report the time 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 1.

Algorithm 1: Tree construction during episodic memory recall

```
1 Create a tree $\mathrm{Tr}(e_{\mathbf{0}}^N)$ with root node $e_{\mathbf{0}}^N$ (the red node in Figure 4).
2 Append to $\mathrm{Tr}(e_{\mathbf{0}}^N)$ all nodes $e^n$ connected to $e_{\mathbf{0}}^N$, where $n < N$, with probability $\sim \mathrm{Bern}\big(1 - P_{\mathrm{evidence}}(e^n)\big)$.
3 For each $e_i^n \in \mathrm{Tr}(e_{\mathbf{0}}^N)$ 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 $P_{\mathrm{chil.}}(z_{k,t}^n)$.
4 If the estimated number is greater than the current number of children, append more nodes $e_j^{n-1}$ to $\mathrm{Tr}(e_{\mathbf{0}}^N)$, 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 $\exists\, e_i^n \in \mathrm{Tr}(e_{\mathbf{0}}^N)$ with $n > 1$ to which step 5 has not been applied yet, append $k$ blank nodes $e_j^{n-1}$ to $\mathrm{Tr}(e_{\mathbf{0}}^N)$, where $k$
    is either sampled from $P_{\mathrm{chil.}}(z_{k,t}^n)$ if $e_i^n$ is known, or from $P_{\mathrm{chil.}}(n)$ otherwise. (nodes with green dashed outline).
```

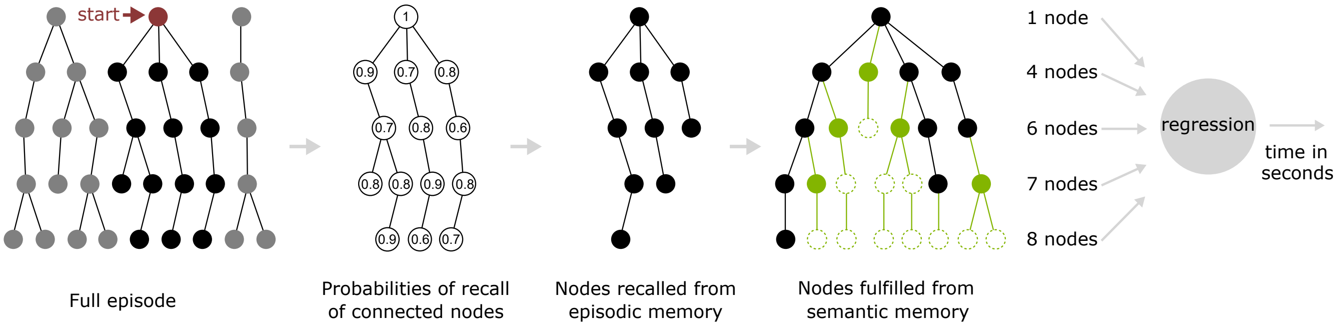


Figure 4. Algorithm of the steps required for episodic memory recall and retrospective duration judgements. Solid circles represent nodes with an assigned value to the categorical variable z_t^n either from the episodic (black) or the semantic memory (green). Circles with the dashed outline represent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

The tree $\mathrm{Tr}(e_{\mathbf{0}}^N)$ 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 $\tilde{\tau}_{\mathrm{tree}}^n$ be the number of nodes in each layer n of the tree. Whereas the nodes added in steps 1-2 of algorithm 1 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 nodes added in steps 3-5 contribute prior (semantic) information and thus fill any memory gaps.

Duration judgements of past events can be obtained in a similar fashion to the prospective case via the process of episodic recall. Given a past event represented with a node $e_{\mathbf{0}}^N$, 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. Hence,

$$\text{retrospective duration judgement} = f(\tilde{\tau}_{\mathrm{tree}}^1, \tilde{\tau}_{\mathrm{tree}}^2, \ldots, \tilde{\tau}_{\mathrm{tree}}^N) \tag{10}$$

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, 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, 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 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 equation (8) tends to decay over time to the value 1, by obeying the equation

$$T_t^n = 1 - T_{t-1}^n + c^{-1} e^{-\frac{D^n}{c}} - N\big(0, \sigma\big) \tag{11}$$

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 $\mathrm{Bern}\big(1-P_{\text{evidence}}(e^n)\big)$ 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 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, 12]; in these previous studies, each participant viewed a large number of videos) and reported the duration in seconds (see Figure 9). 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). Participants were excluded from further analysis if they reported that they guessed they would need to estimate the duration when doing the retrospective condition, or if they reported that they had been explicitly counting in any condition. The results for participants who reported they were counting (11% of participants) are presented in Supplementary Material (Supplementary Figure 10). For further details on the experimental procedure see Methods.

Figure 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 broadly following 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. 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[40, 41].

Figure 5B shows the mean duration judgement ratio (subjective versus physical duration) for all combinations of task (prospective/retrospective) and cognitive load (low/high). Reassuringly, the previously reported task-by-cognitive load interaction[7] is apparent in our data (compare Figure 1 and Figure 5B), though in our data it is of a smaller magnitude 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, 12, 42] (see also Figure 5C). Furthermore, the task-by-load interaction in our data becomes less straightforward when broken down by the different scene types that were used as stimuli (City, campus and outside, and office/cafe) (Figure 5C). 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 broadly replicates that seen in the previously reported results for different experiments using these same stimuli[4, 12]. Of greater interest, the nature of the task-by-load interaction also qualitatively changes with scene type. While the interaction for the City scene data 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 general shift towards prospective being longer than retrospective reports, regardless of cognitive load, reversing the relationship seen for reports regarding the Campus and outside scenes.

Computational model

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.

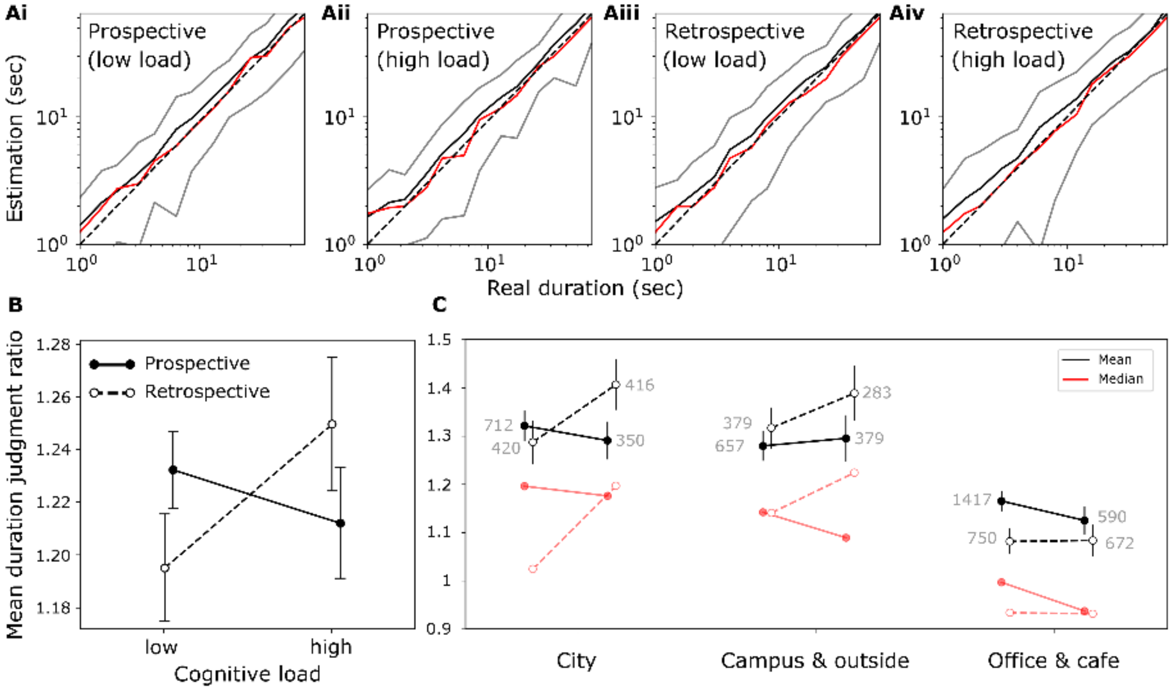


Figure 5. Human duration estimates according to task (pro/retrospective), cognitive load (low/high), and presented video scene (City, Campus and outside, or Office and cafe). **Ai-iv** Human duration estimation for task-cognitive load combination. The black curves represent the mean, the red is the median, 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).

The method to produce duration estimates was similar for trials from prospective and retrospective tasks. For trials completed by human participants as prospective, estimates were based on the model response to the initial input of the presented trial. 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 parameters c (for the effect of attention) and ε (for effort), described above in section Attention, effort and cognitive load. 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.

Duration judgements of the past and present replicate human behaviour

Initially, we trained a single linear regression model 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. Figure 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 load 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 (9-10) (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.

Interestingly, the overall behaviour of the model (shown in Figure 6B) replicates the interaction of task and cognitive load demonstrated both in previous work[7] (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.

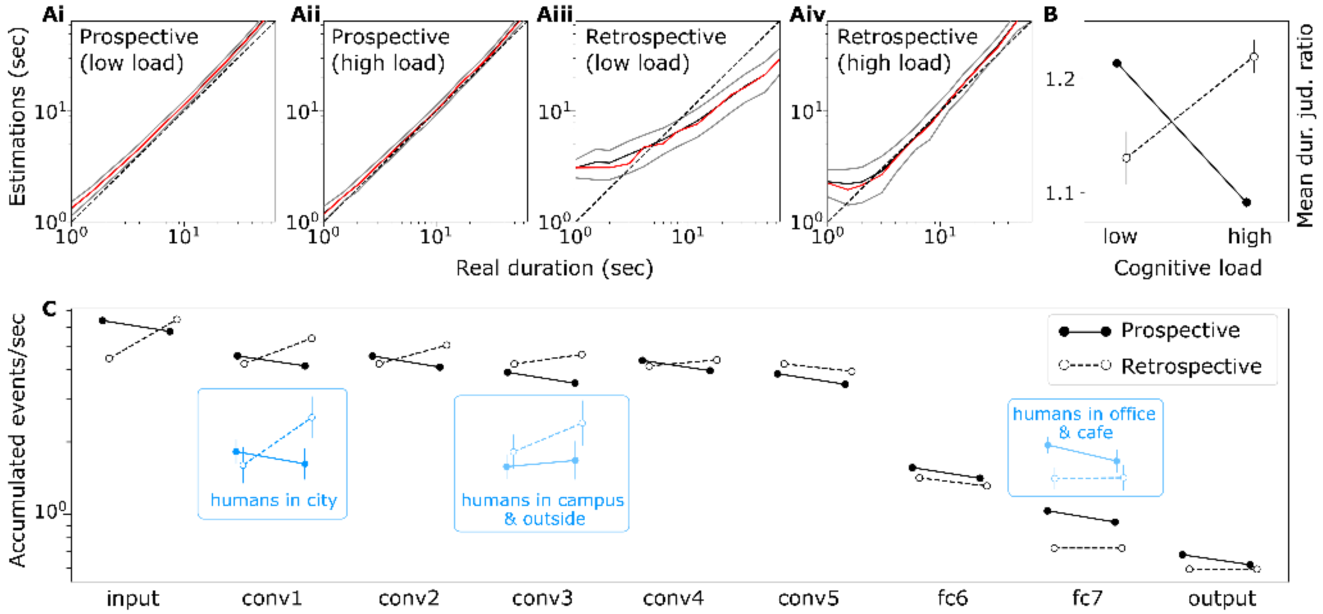


Figure 6. Model duration estimates in seconds or accumulated salient events by task (pro/retrospective) and cognitive load (high/low) for the same trials performed by human participants. **Ai-iv** Model duration estimates for task-cognitive load combination. The black curves represent the mean, the red 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.

The role of stimulus content in duration reports

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. Looking at the same breakdown by task and cognitive load as depicted in Figure 6B, but in accumulated salient changes (Figure 6C) rather than in seconds, and separated by hierarchical layer in the network, we see 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 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 significantly 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. 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, the representations of the first convolutional layer (*conv1*) seem to be used in highly dynamic videos of someone walking around a busy city, a higher convolutional layer (*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 above); 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 trained on all accumulation in all network layers. The best performing estimates were found for regressions trained to map between the accumulated salient events in a specific layer and physical duration using a 2nd 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 *appropriate* single layer accumulation (Figure 7C) compared to when it is trained on accumulation in all layers (Figure 7B).

Overall, these results show that it is possible to train 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, that can replicate the complex pattern of interactions between task, cognitive load, and scene-type that we found in human duration estimates for these same naturalistic video inputs. Assuming such a simple scenario, an interesting direction of investigation then becomes how the appropriate layer is selected. In Supplementary Material 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 (Supplementary Figure 12). 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.

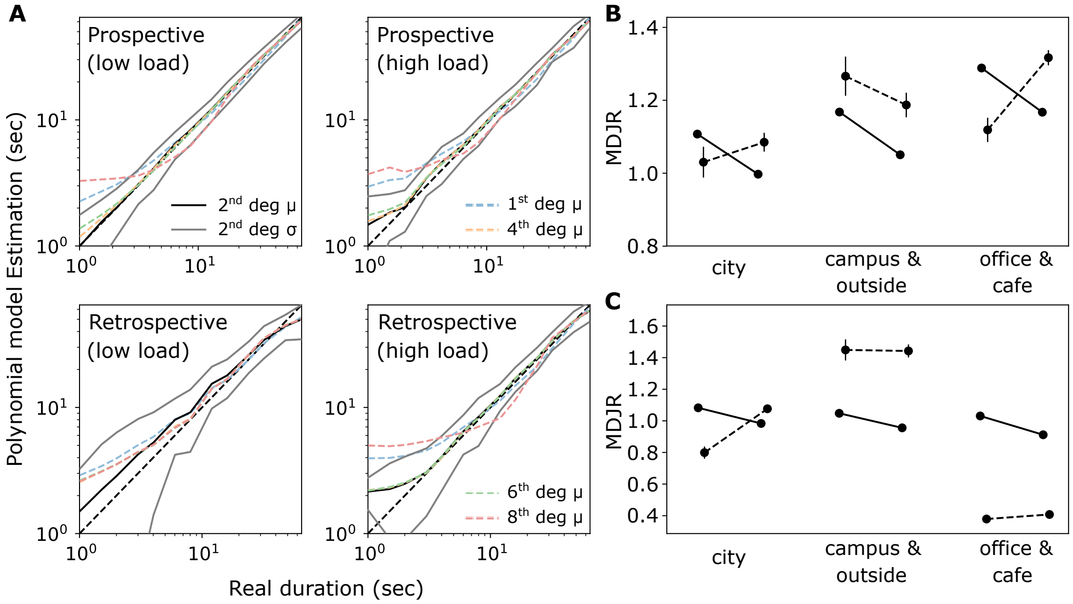


Figure 7. A: Model duration estimates for each task (pro/retrospective) and cognitive load (high/low) combination, using linear and polynomial regression trained on a single layer of salient event accumulations for different scene types, including layer *conv1*, *conv3* and *fc7* for trials that correspond 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 2nd degree polynomial and the grey line indicates ±1 standard deviation for estimates given by the 2nd degree polynomial regression. The blue, orange, green, red broken lines indicate the results of the same regression using 1st, 4th, 6th, 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 A (**C**).

The role of effort in memory recall

The next question we investigated is whether, according to our model, the amount of effort that humans spend 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 Algorithm 4). To assess the 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 $\varepsilon \in \{1, 2, ..., 70\}$. 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 8A 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.

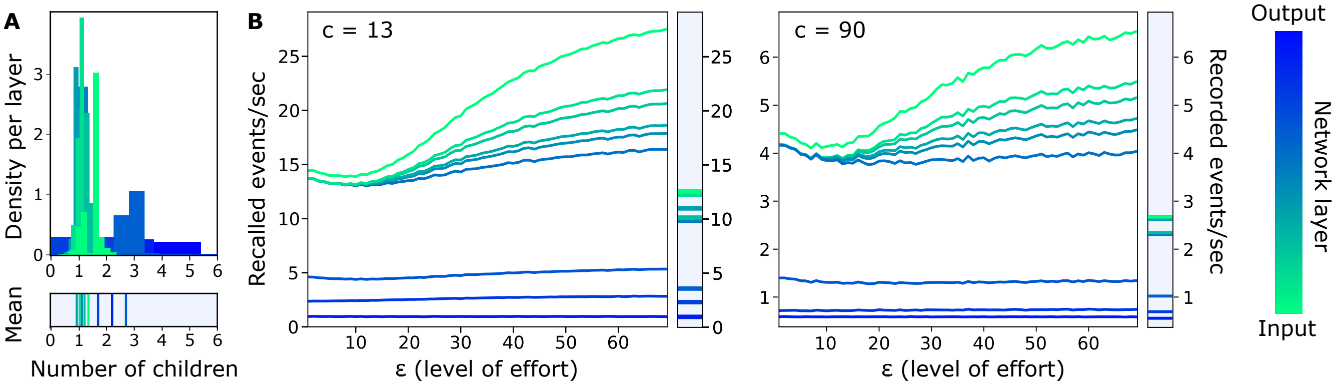


Figure 8. **A:** Distribution of children nodes recorded per layer over all components of this layer, used to calculate $P_{\text{chil.}}(n)$. **B:** Rate of recalling events in different layers of the model and for different levels of effort ε, defined as the number of recall attempts. The coloured 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 8B. 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. $c = \{13, 90\}$).

Overall, the model's behaviour displays the importance of effort in modelling the human biases found in the current study. It predicts that higher effort should generally lead to greater overestimation of the duration of recalled episodes and that visual episodes which are rich in content should be more overestimated. Further analysis in the supplementary section 1 predicts that the level of effort affects the relationship between cognitive load and the perception of duration, both in the past and present.

Discussion

In this work we present a predictive processing model of episodic memory and time perception. Using this model, we were able to reproduce the characteristic patterns of human duration judgement biases (estimated from ~ 7000 human participants, from an initial pool of $\sim 13,000$) in prospective and retrospective timing tasks, using as input only on the rate of salient events detected within a hierarchical perceptual processing network. With a single computational process that estimates duration both during ongoing perceptual input and based on recalled memory of those events, our model reconciles divergent approaches and experimental findings in the time perception literature under a single predictive processing framework to account for differences in human duration perception observed across tasks, cognitive load, attention, and stimulus content.

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. This approach follows from the intuitive basis that human perception of time is characterised more by its common deviations from veridicality, rather than acting like a clock that attempts to track elapsing physical time (e.g.[10,43]). 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 (9-10) 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 (see also Supplementary Figure 12).

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 [14]. Contemporary interpretations show that this mechanism optimizes energy efficiency in terms of information processing and storage[17,20], while it has been argued that it even constitutes a fundamental requirement of (biological) self-organized systems to achieve homeostatic control, or life[44–46]. 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 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 resourceful[47,48].

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.[49–52]), though a formal description of event segmentation in episodic memory has only taken a prominent position more recently (e.g. event segmentation theory[53,54]). Extensive work in both humans[55–59] and animals[60,61] 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[58,59,61]. It has been suggested that activity reflecting event segmentation is based on prediction error[62–65], 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[5,66] - 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).

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 archetypal than processing at higher cortical areas (e.g. superior temporal sulcus and frontal eye field[56,67,68]). Beyond 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[65]). 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[69,70]. 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 more 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[55,56,71] while higher complexity events will be detectable in the dynamics of higher cortical regions[56]. 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[13]. While structurally equipped to produce many properties of episodic memory, our present model was developed explicitly to address human time perception. Further work is required to determine the depth to which our model (or modelling approach more generally) reproduces the many other known features of human episodic memory (see[65]).

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 1). 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[40,41,72]. 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[48]. 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 equation (8) 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[73], and connects the fundamental and timeless 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 (2-3) 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 accounts for only visual experiences. 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)[56], we anticipate that our approach is readily extendable to incorporate these additional features under a broader active inference approach[21].

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[74,75]. These two abilities are often referred to together as mental time travel[76]. Research in this field indicates a strong connection between recalling past experiences, semantic memory and imagining the outcomes in potential future scenarios[77,78], an undoubtedly crucial aspect of our decision making[21,79]. 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[80], dreaming[81,82], and hallucinations[83], 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[84].

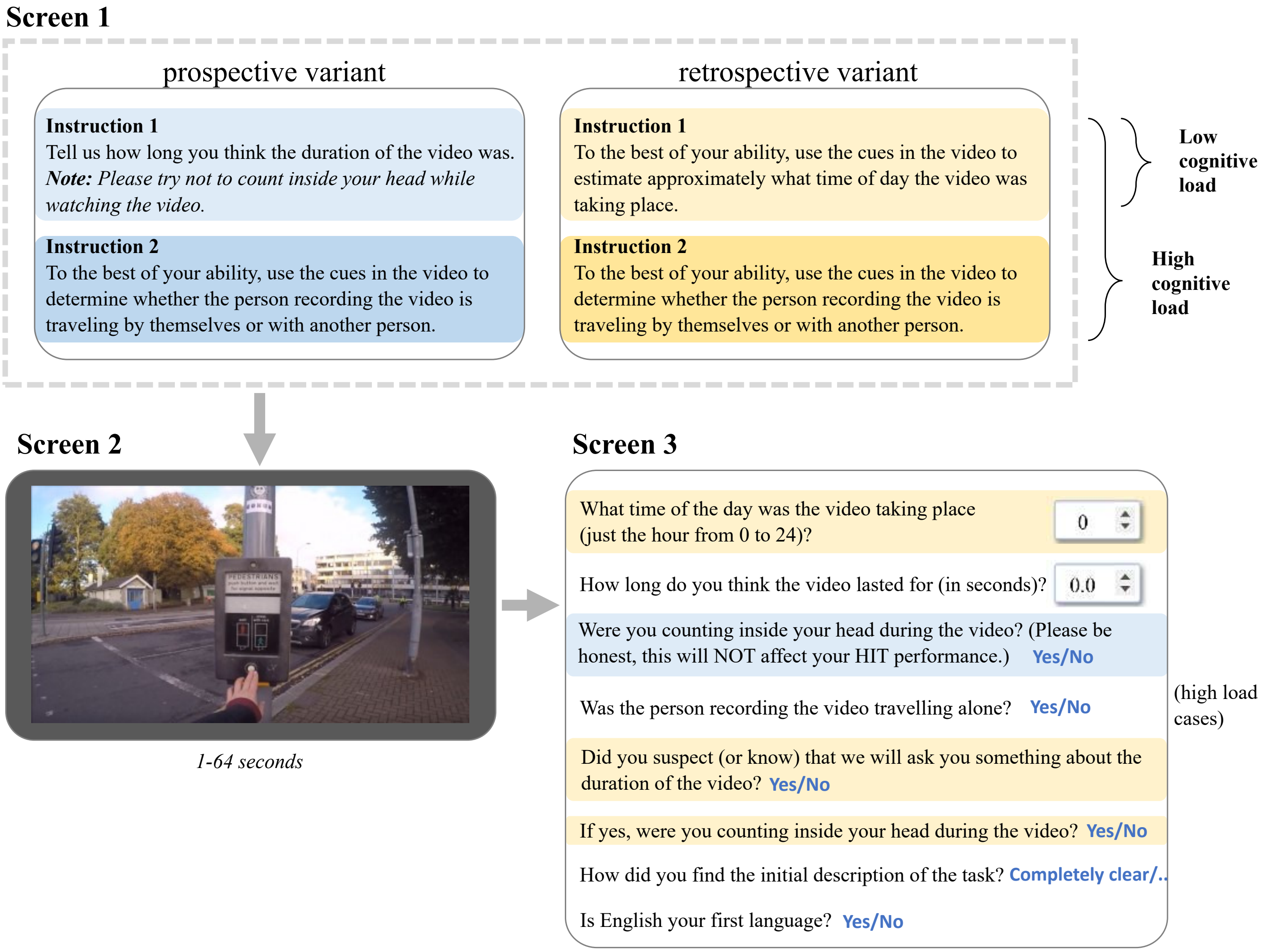
Methods

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 scale of their home computer monitor and then answer to a series of questions including a report of the perceived duration of this 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 aware 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[85]. 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[86].

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 9. 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.



Instruction 1
Tell us how long you think the duration of the video was. ***Note:*** *Please try not to count inside your head while watching the video.*

Instruction 2
To the best of your ability, use the cues in the video to determine whether the person recording the video is traveling by themselves or with another person.

Instruction 1
To the best of your ability, use the cues in the video to estimate approximately what time of day the video was taking place.

Instruction 2
To the best of your ability, use the cues in the video to determine whether the person recording the video is traveling by themselves or with another person.

What time of the day was the video taking place (just the hour from 0 to 24)? 0

How long do you think the video lasted for (in seconds)? 0.0

Were you counting inside your head during the video? (Please be honest, this will NOT affect your HIT performance.) Yes/No

Was the person recording the video travelling alone? Yes/No

Did you suspect (or know) that we will ask you something about the duration of the video? Yes/No

If yes, were you counting inside your head during the video? Yes/No

How did you find the initial description of the task? Completely clear/..

Is English your first language? Yes/No

Figure 9. The main steps of the trials including instructions the video and the final questionnaire, for the four variants of the task.

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 a reward 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 to the terms and conditions approved by the University of Sussex's research ethics committee (C-REC).

Out of the total number of participants, the 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

$$\frac{\text{playback time} - \text{real duration}}{\text{playback time}} \geq 1.2 \quad or \quad \text{playback time} - \text{real duration} \geq 0.8 \text{ seconds} \tag{12}$$

Furthermore, although participants of the prospective variant of the experiment were instructed to try not to count in seconds during the trial, when they were asked in the final questionnaire they reported they could not resist counting. Counting was also reported by a small group of participants of the retrospective variant, who suspected they will 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 $n = 6977$ non-counting participants (and trials). Outliers were also removed from the group of counting participants ($n = 809$) in order to produce the Supplementary Figure 10. 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 Supplementary Figure 11. The resulting dataset can be found in the repository [1].

Computational model implementation

As an implementation of the inference model $p(x_t^n|x_t^{n-1})$, as depicted in Figure 2, we used a well-studied convolutional network defined by Krizhevsky and colleagues[87]. Deep convolutional networks have hierarchical structure that resembles the human/primate visual cortex[69,70,88,89] have been shown to perform very well in visual classification tasks[87,90]. This network was pre-trained to classify natural images into 1000 different categories of objects and animals[91]. 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 equation 4 directly, the algorithm initially samples from $\mathbf{z} \sim P\left(z_t^n|z_{t-1}^n, z_t^{n+1}, e_\tau^n\right)$. 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 $\mathbf{x} \sim p(x_t^n|\mathbf{z}) = N\left(\mu_{t,\mathbf{z}}^n, \Sigma_{t,\mathbf{z}}^n\right)$. To approximate the model evidence for each Gaussian component, needed for equation (5), 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 $\sum_{t=0}^{3600}\sum_{n=1}^{8} -\log P\left(x_t^n|z_{t-1}^n, z_t^{n+1}, e_\tau^n\right)$ 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)[92]. The parameters for the attention and recall mechanisms were tuned following the principles of $c_{\text{low prosp}} < c_{\text{high prosp}} < c_{\text{retrospective}}$ and $\varepsilon_{\text{low load}} < \varepsilon_{\text{high load}}$, such that the overall relation between tasks and cognitive load levels, shown in Figure 5B, is approximated.

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}	$\varepsilon_{\text{low load}}$	$\varepsilon_{\text{high load}}$
46	52	90	90	0.1	0.1	1	70

Table 1. Model hyper-parameters

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 [2]. 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.

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

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

References

1. Selby-Bigge, L. A. *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* (Penguin, Harmondsworth, UK, 1969).

3. Poynter, W. D. & Homa, D. Duration judgment and the experience of change. *Percept. Psychophys.* **33**, 548–560 (1983).

4. Roseboom, W. *et al.* Activity in perceptual classification networks as a basis for human subjective time perception. *Nat. communications* **10**, 267 (2019).

5. Zakay, D. & Block, R. A. An attentional·gate model of prospective time estimation. In *I.P.A Symposium Liege*, 167–178 (1994).

6. Block, R. A. & Zakay, D. Prospective and retrospective duration judgments: A meta-analytic review. *Psychon. Bull. Rev.* **4**, 184–197 (1997).

7. Block, R. A., Hancock, P. A. & Zakay, D. How cognitive load affects duration judgments: A meta-analytic review. *Acta psychologica* **134**, 330–343 (2010).

8. Zakay, D. & Block, R. A. Temporal cognition. *Curr. Dir. Psychol. Sci.* **6**, 12–16 (1997).

9. Brown, S. W. Timing, resources, and interference: Attentional modulation of time perception. In Nobre, A. C. & Coull, J. T. (eds.) *Attention and Time*, 107–121 (Oxford University Press, 2010).

10. Matell, M. S. & Meck, W. H. Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Cogn. brain research* **21**, 139–170 (2004).

11. Howard, M. W. *et al.* A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *J. Neurosci.* **34**, 4692–4707, DOI: 10.1523/JNEUROSCI.5808-12.2014 (2014). https://www.jneurosci.org/content/34/13/4692.full.pdf.

12. Suárez-Pinilla, M., Nikiforou, K., Fountas, Z., Seth, A. K. & Roseboom, W. Perceptual content, not physiological signals, determines perceived duration when viewing dynamic, natural scenes. *Collabra: Psychol.* **5**, 55, DOI: 10.1525/collabra.234 (2019).

13. Sherman, M. T., Fountas, Z., Seth, A. K. & Roseboom, W. Accumulation of salient events in sensory cortex activity predicts subjective time. *bioRxiv* DOI: 10.1101/2020.01.09.900423 (2020). https://www.biorxiv.org/content/early/2020/01/09/2020.01.09.900423.full.pdf.

14. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. neuroscience* **2**, 79 (1999).

15. Friston, K. A theory of cortical responses. *Philos. Transactions Royal Soc. Lond. B: Biol. Sci.* **360**, 815–836, DOI: 10.1098/rstb.2005.1622 (2005). http://rstb.royalsocietypublishing.org/content/360/1456/815.full.pdf.

16. Friston, K. & Kiebel, S. Predictive coding under the free-energy principle. *Philos. Transactions Royal Soc. Lond. B: Biol. Sci.* **364**, 1211–1221, DOI: 10.1098/rstb.2008.0300 (2009). http://rstb.royalsocietypublishing.org/content/364/1521/1211.full.pdf.

17. Clark, A. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204, DOI: 10.1017/S0140525X12000477 (2013).

18. Buckley, C. L., Kim, C. S., McGregor, S. & Seth, A. K. The free energy principle for action and perception: A mathematical review. *J. Math. Psychol.* **81**, 55 – 79, DOI: https://doi.org/10.1016/j.jmp.2017.09.004 (2017).

19. Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. The helmholtz machine. *Neural computation* **7**, 889–904 (1995).

20. Friston, K. The free-energy principle: a unified brain theory? *Nat. reviews neuroscience* **11**, 127 (2010).

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

22. Bastos, A. M. *et al.* Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).

23. Arnal, L. H., Wyart, V. & Giraud, A.-L. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nat. neuroscience* **14**, 797 (2011).

24. Alamia, A. & VanRullen, R. Alpha oscillations and traveling waves: Signatures of predictive coding? *PLOS Biol.* **17**, 1–26, DOI: 10.1371/journal.pbio.3000487 (2019).

25. Corlett, P. R. *et al.* Hallucinations and strong priors. *Trends cognitive sciences* (2018).

26. Palmer, C. J., Lawson, R. P. & Hohwy, J. Bayesian approaches to autism: Towards volatility, action, and behavior. *Psychol. bulletin* **143**, 521 (2017).

27. Sterzer, P. *et al.* The predictive coding account of psychosis. *Biol. psychiatry* **84**, 634–643 (2018).

28. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)* (2014).

29. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082* (2014).

30. Kalman, R. E. A new approach to linear filtering and prediction problems. *J. basic Eng.* **82**, 35–45 (1960).

31. Ding, C. & He, X. Cluster merging and splitting in hierarchical clustering algorithms. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 139–146 (IEEE, 2002).

32. Lughofer, E. & Sayed-Mouchaweh, M. Autonomous data stream clustering implementing split-and-merge concepts–towards a plug-and-play approach. *Inf. Sci.* **304**, 54–79 (2015).

33. Steyvers, M., Griffiths, T. L. & Dennis, S. Probabilistic inference in human semantic memory. *Trends cognitive sciences* **10**, 327–334 (2006).

34. Binder, J. R. & Desai, R. H. The neurobiology of semantic memory. *Trends cognitive sciences* **15**, 527–536 (2011).

35. Uncapher, M. R. & Wagner, A. D. Posterior parietal cortex and episodic encoding: insights from fmri subsequent memory effects and dual-attention theory. *Neurobiol. learning memory* **91**, 139–154 (2009).

36. Greve, A., Cooper, E., Kaula, A., Anderson, M. C. & Henson, R. Does prediction error drive one-shot declarative learning? *J. Mem. Lang.* **94**, 149–165 (2017).

37. Jang, A., Dillon, D., Frank, M. J. *et al.* Positive reward prediction errors strengthen incidental memory encoding. *bioRxiv* 327445 (2018).

38. Rouhani, N., Norman, K. A. & Niv, Y. Dissociable effects of surprising rewards on learning and memory. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 1430 (2018).

39. Weidemann, C. T. *et al.* Neural activity reveals interactions between episodic and semantic memory systems during retrieval. *J. Exp. Psychol. Gen.* **148**, 1 (2019).

40. Petzschner, F. H., Glasauer, S. & Stephan, K. E. A bayesian perspective on magnitude estimation. *Trends Cogn. Sci.* **19**, DOI: 10.1016/j.tics.2015.03.002 (2015).

41. Roseboom, W. Serial dependence in timing perception. *J. Exp. Psychol. Hum. Percept. Perform.* **45**, 100–110, DOI: 10.1037/xhp0000591 (2019).

42. Linares, D. & Gorea, A. Temporal frequency of events rather than speed dilates perceived duration of moving objects. *Sci. Reports* **5**, DOI: 10.1038/srep08825 (2015).

43. Treisman, M. Temporal discrimination and the indifference interval: Implications for a model of the "internal clock". *Psychol. Monogr. Gen. Appl.* **77**, 1–31, DOI: 10.1037/h0093864 (1963).

44. Conant, R. C. & Ross Ashby, W. Every good regulator of a system must be a model of that system. *Int. journal systems science* **1**, 89–97 (1970).

45. Friston, K. Life as we know it. *J. Royal Soc. Interface* **10**, 20130475 (2013).

46. Seth, A. K. & Tsakiris, M. Being a beast machine: the somatic basis of selfhood. *Trends cognitive sciences* **22**, 969–981 (2018).

47. Laje, R. & Buonomano, D. V. Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nat. neuroscience* **16**, 925 (2013).

48. Hardy, N. F. & Buonomano, D. V. Neurocomputational models of interval and pattern timing. *Curr. Opin. Behav. Sci.* **8**, 250–257 (2016).

49. Block, R. A. & Reed, M. A. Remembered duration: Evidence for a contextual-change hypothesis. *J. Exp. Psychol. Hum. Learn. & Mem.* **4**, 656–665, DOI: 10.1037/0278-7393.4.6.656 (1978).

50. Poynter, W. D. Duration judgment and the segmentation of experience. *Mem. & Cogn.* **11**, 77–82, DOI: 10.3758/bf03197664 (1983).

51. Poynter, D. Chapter 8 judging the duration of time intervals: A process of remembering segments of experience. In *Advances in Psychology*, 305–331, DOI: 10.1016/s0166-4115(08)61045-6 (Elsevier, 1989).

52. Zakay, D., Tsal, Y., Moses, M. & Shahar, I. The role of segmentation in prospective and retrospective time estimation processes. *Mem. & Cogn.* **22**, 344–351, DOI: 10.3758/bf03200861 (1994).

53. Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S. & Reynolds, J. R. Event perception: A mind-brain perspective. *Psychol. Bull.* **133**, 273–293, DOI: 10.1037/0033-2909.133.2.273 (2007).

54. Radvansky, G. A. & Zacks, J. M. Event boundaries in memory and cognition. *Curr. Opin. Behav. Sci.* **17**, 133–140, DOI: 10.1016/j.cobeha.2017.08.006 (2017).

55. Ezzyat, Y. & Davachi, L. Similarity breeds proximity: Pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity. *Neuron* **81**, 1179–1189, DOI: 10.1016/j.neuron.2014.01.042 (2014).

56. Baldassano, C. *et al.* Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721.e5, DOI: 10.1016/j.neuron.2017.06.041 (2017).

57. Ben-Yakov, A. & Henson, R. N. The hippocampal film editor: Sensitivity and specificity to event boundaries in continuous experience. *The J. Neurosci.* **38**, 10057–10068, DOI: 10.1523/jneurosci.0524-18.2018 (2018).

58. Montchal, M. E., Reagh, Z. M. & Yassa, M. A. Precise temporal memories are supported by the lateral entorhinal cortex in humans. *Nat. Neurosci.* **22**, 284–288, DOI: 10.1038/s41593-018-0303-1 (2019).

59. Bellmund, J. L., Deuker, L. & Doeller, C. F. Mapping sequence structure in the human lateral entorhinal cortex. *eLife* **8**, DOI: 10.7554/elife.45333 (2019).

60. Bulkin, D. A., Sinclair, D. G., Law, L. M. & Smith, D. M. Hippocampal state transitions at the boundaries between trial epochs. *BioRxiv* 443077, DOI: 10.1101/443077 (2018).

61. Tsao, A. *et al.* Integrating time from experience in the lateral entorhinal cortex. *Nature* **561**, 57–62, DOI: 10.1038/s41586-018-0459-6 (2018).

62. Reynolds, J. R., Zacks, J. M. & Braver, T. S. A computational model of event segmentation from perceptual prediction. *Cogn. Sci.* **31**, 613–643, DOI: 10.1080/15326900701399913 (2007).

63. Zacks, J. M., Kurby, C. A., Eisenberg, M. L. & Haroutunian, N. Prediction error associated with the perceptual segmentation of naturalistic events. *J. Cogn. Neurosci.* **23**, 4057–4066, DOI: 10.1162/jocn_a_00078 (2011).

64. Gershman, S. J., Radulescu, A., Norman, K. A. & Niv, Y. Statistical computations underlying the dynamics of memory updating. *PLoS Comput. Biol.* **10**, e1003939, DOI: 10.1371/journal.pcbi.1003939 (2014).

65. Zacks, J. M. Event perception and memory. *PsyArXiv* DOI: 10.31234/osf.io/634dz (2019).

66. Block, R. A. Prospective and retrospective duration judgment: The role of information processing and memory. In *Time, Action and Cognition*, 141–152, DOI: 10.1007/978-94-017-3536-0_16 (Springer Netherlands, 1992).

67. Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550, DOI: 10.1523/jneurosci.5487-07.2008 (2008).

68. Hasson, U., Chen, J. & Honey, C. J. Hierarchical process memory: memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–313, DOI: 10.1016/j.tics.2015.04.006 (2015).

69. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology* **10**, e1003915 (2014).

70. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. review vision science* **1**, 417–446 (2015).

71. Zacks, J. M. *et al.* Human brain activity time-locked to perceptual event boundaries. *Nat. Neurosci.* **4**, 651–655, DOI: 10.1038/88486 (2001).

72. Rhodes, D., Seth, A. K. & Roseboom, W. Multiple duration priors within and across the senses. *bioRxiv* 467027, DOI: 10.1101/467027 (2018).

73. Sengupta, B., Tozzi, A., Cooray, G. K., Douglas, P. K. & Friston, K. J. Towards a neuronal gauge theory. *PLoS Biol.* **14**, e1002400 (2016).

74. Hassabis, D., Kumaran, D. & Maguire, E. A. Using imagination to understand the neural basis of episodic memory. *J. neuroscience* **27**, 14365–14374 (2007).

75. Zeidman, P. & Maguire, E. A. Anterior hippocampus: the anatomy of perception, imagination and episodic memory. *Nat. Rev. Neurosci.* **17**, 173 (2016).

76. Tulving, E. Memory and consciousness. *Can. Psychol. canadienne* **26**, 1 (1985).

77. Schacter, D. L. *et al.* The future of memory: remembering, imagining, and the brain. *Neuron* **76**, 677–694 (2012).

78. Bulley, A., Henry, J. & Suddendorf, T. Prospection and the present moment: The role of episodic foresight in intertemporal choices between immediate and delayed rewards. *Rev. Gen. Psychol.* **20**, 29–47 (2016).

79. Oettingen, G., Sevincer, A. T. & Gollwitzer, P. M. *The Psychology of Thinking about the Future* (The Guilford Press, 2018).

80. 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 Cogn.* **83**, 45–51 (2013).

81. Dement, W. & Kleitman, N. Cyclic variations in eeg during sleep and their relation to eye movements, body motility, and dreaming. *Electroencephalogr. clinical neurophysiology* **9**, 673–690 (1957).

82. Damsma, A., van der Mijn, R. & van Rijn, H. Neural markers of memory consolidation do not predict temporal estimates of encoded items. *Neuropsychologia* **117**, 36–45 (2018).

83. Pienkos, E. *et al.* Hallucinations beyond voices: A conceptual review of the phenomenology of altered perception in psychosis. *Schizophr. bulletin* **45**, S67–S77 (2019).

84. Botvinick, M. *et al.* Reinforcement learning, fast and slow. *Trends cognitive sciences* (2019).

85. Paolacci, G., Chandler, J. & Ipeirotis, P. G. Running experiments on amazon mechanical turk. *Judgm. Decis. making* **5**, 411–419 (2010).

86. Huff, C. & Tingley, D. "who are these people?" evaluating the demographic characteristics and political preferences of mturk survey respondents. *Res. & Polit.* **2**, 2053168015604648 (2015).

87. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105 (2012).

88. Güçlü, U. & van Gerven, M. A. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).

89. Cadena, S. A. *et al.* Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology* **15**, e1006897 (2019).

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

91. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *Int. journal computer vision* **115**, 211–252 (2015).

92. Igel, C., Hansen, N. & Roth, S. Covariance matrix adaptation for multi-objective optimization. *Evol. computation* **15**, 1–28 (2007).

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). AKS is a Fellow of the Canadian Institute for Advanced Research (CIFAR) Azrieli Program on Brain, Mind, and Consciousness, to which he is grateful.

Supplementary Material

1 Further analysis for the effect of recall effort

Figure 8B of the main text 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 $\varepsilon = 1$, 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 $\varepsilon = 70$, 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 8.B. This changed when $\varepsilon > 7$, where the first components of the convolutional layers were recalled, leading to a massive increase in recalled events in these layers.

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 equation (11). 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 effort was robust for a wide range of values of c. In fact, the effect of 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 Supplementary Figure 13

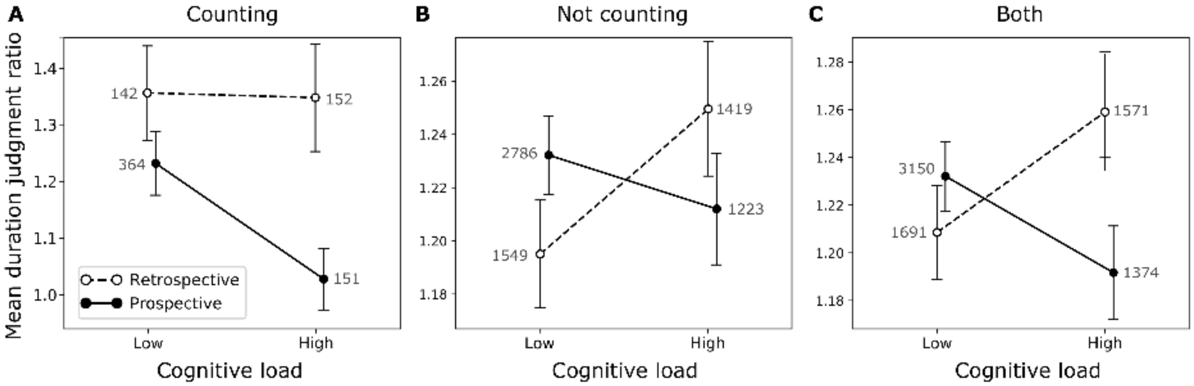


Figure 10. Interaction in human duration reports tasks (pro/retrospective) and cognitive load for participants that reported they were counting (**A**), not counting (**B**), or both cases together (**C**), having excluded the rejected trials as described in methods.

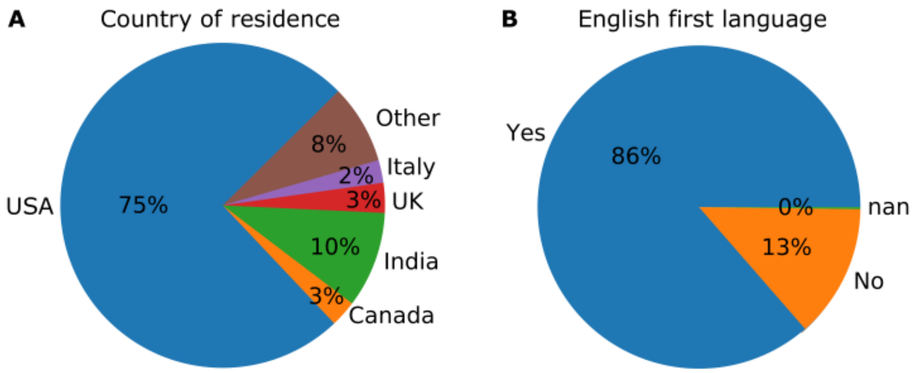


Figure 11. Demographics of the human participants including geographic location and first language.

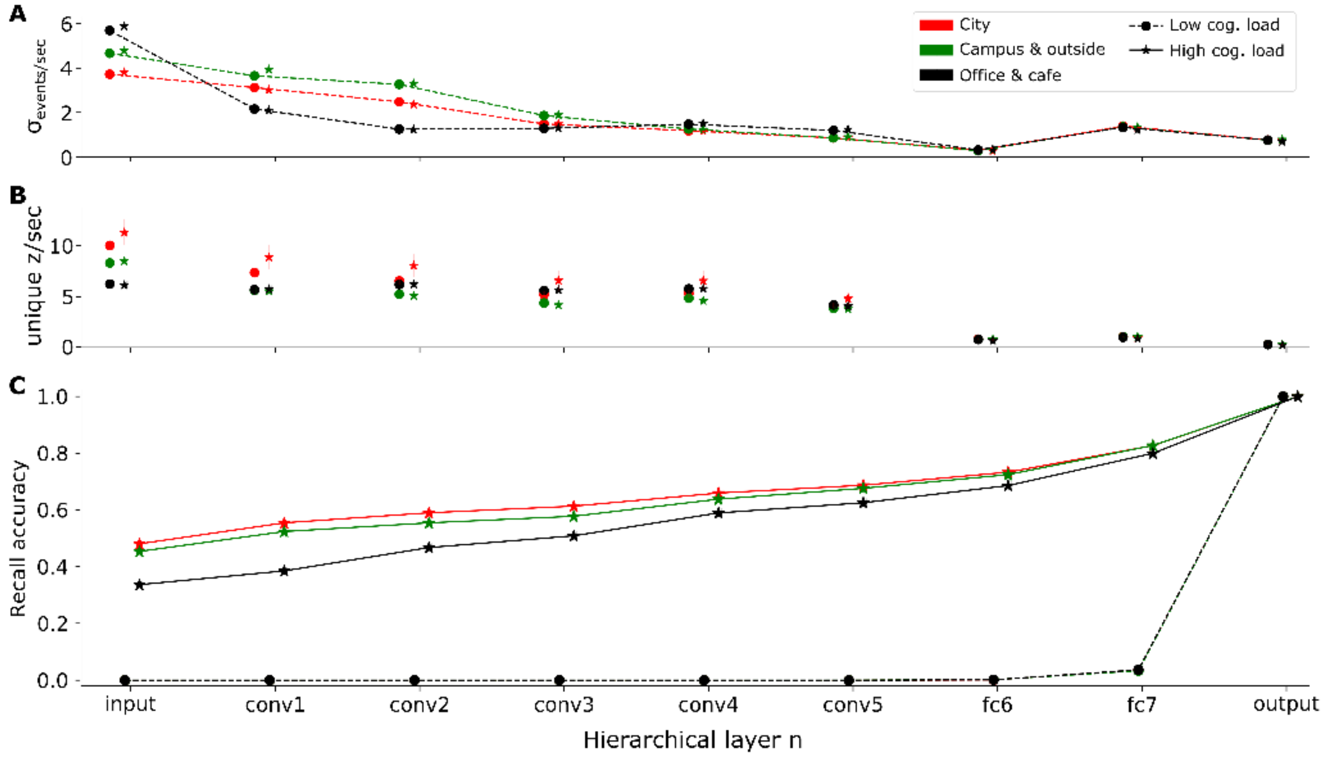


Figure 12. Statistical properties of the model related to episodic memory recall, per scene type and cognitive load level. The episodic memory model shows unique patterns for the 3 different scene types, although no single feature is clearly suitable by itself or superior to alternative solutions to 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.

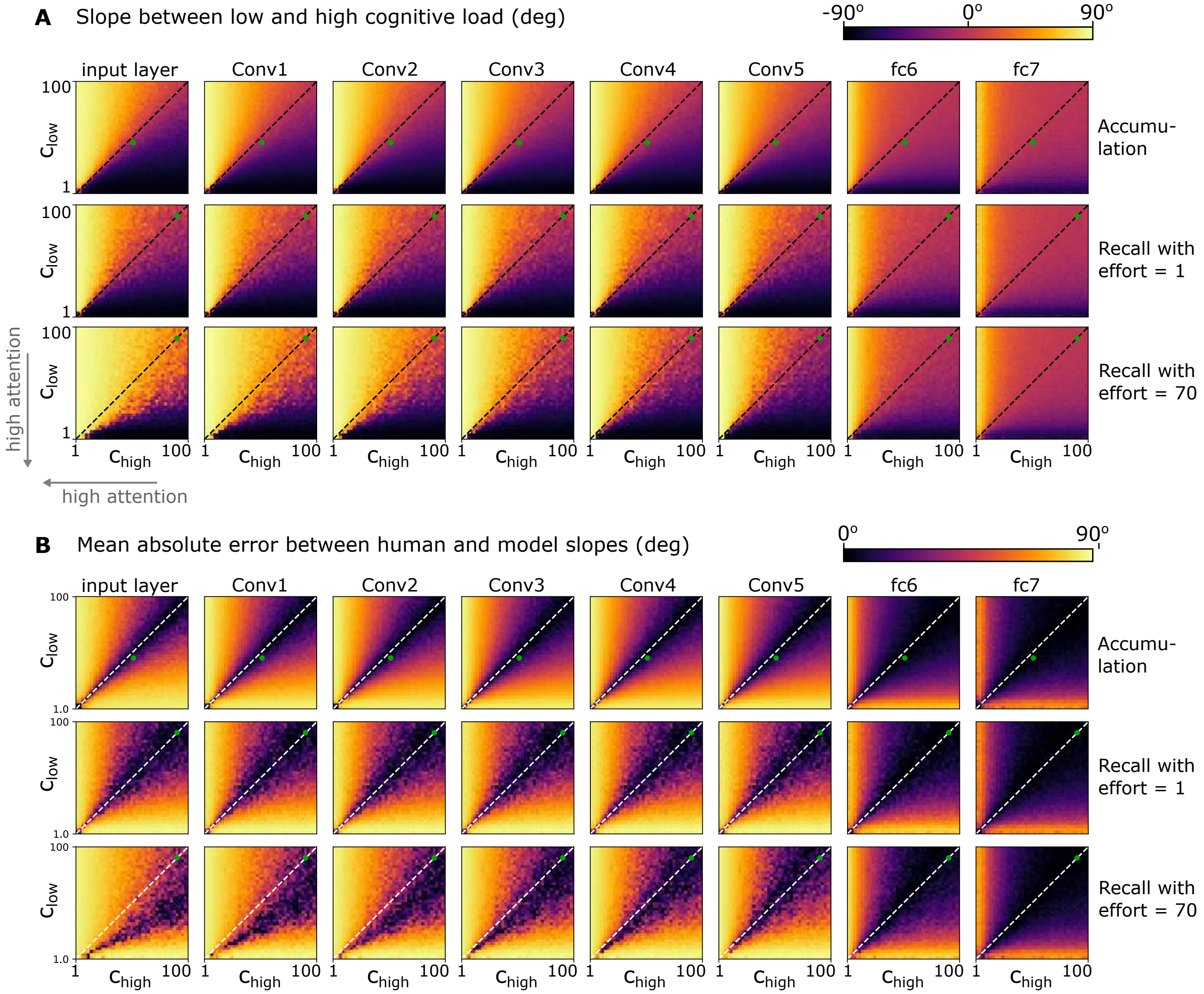


Figure 13. Effects of attention on salient event accumulation and recall per layer, for the prospective and retrospective tasks respectively. In accumulation and recall with low effort, small differences between c_{low} and c_{high} are enough to replicate human behaviour. However, under a high amount of 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. **A**: Slope between the number of events using low- and high-load hyper parameters. **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 5.

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

Table 2. Probability distributions and corresponding parameters associated with each of the two memory types (semantic in **green** and episodic in **orange**) at time t.

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
$W(e_\tau^n)$	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	

Table 3. Parameters of the model that can be trained recursively in real-time, along with the training method used.