# Accumulation of Salient Perceptual Events Predicts Subjective Time.

Maxine T. Sherman[1,2,3*], Zafeirios Fountas[4,5], Anil K. Seth[1,2,6] & Warrick Roseboom[1,2*]

[1]Sackler Centre for Consciousness Science, University of Sussex, UK.

[2]Department of Informatics, University of Sussex, UK.

[3]Brighton and Sussex Medical School, University of Sussex, UK.

[4]Emotech Labs, London, UK.

[5]Wellcome Centre for Human Neuroimaging, University College London, London, UK

[6]Canadian Institute for Advanced Research, Azrieli Program on Brain, Mind, and Consciousness

* Maxine Sherman & Warrick Roseboom for correspondence.

**Email:** m.sherman@sussex.ac.uk; wjroseboom@gmail.com

**Author Contributions**

WR conceived of the study. MTS and WR designed and pre-registered the experiments and analyses. MTS collected, analyzed, and constructed models of human behavioral and neuroimaging data. ZF constructed the artificial network model and analyzed the data. MTS and WR wrote the manuscript. AKS and ZF provided critical revisions on the manuscript.

**Abstract**

Human experience of time exhibits systematic, context-dependent deviations from objective clock time. For example, time is experienced differently at work than on holiday. The cognitive and neural bases of how time perception interacts with the content of experience remain unclear, and leading explanations of human time perception are not equipped to explain this interaction. We propose an alternative account of human time perception, based on the dynamics of sensory processing. Our approach naturally links content of experience with time perception through a common foundation in basic sensory processing. We provide evidence for this proposal in model-based analyses of the dynamics of perceptual processing in an artificial neural network and in the activity of human sensory cortex. Healthy human participants watched naturalistic, silent videos and estimated their duration while fMRI was acquired. The same videos were used as stimuli for the artificial network. We constructed a computational model that predicted video durations from salient events in the activity of the artificial network, or in participants' visual cortex. The artificial network model reproduced human-like duration estimates, including biases in estimation depending on the content of a given video. Most importantly, the model based on human visual cortex activity reproduced trial-by-trial biases in our human participants' subjective reports, whereas control models trained on auditory or somatosensory activity did not. Together, our results reveal that human subjective time is based on information arising during the processing of our dynamic sensory environment, providing a computational basis for an end-to-end account of time perception.

**Significance Statement**

Our perception of time depends on the contents of experience, reflected in expressions such as "a watched pot never boils". Prevailing accounts of human time perception cannot provide good explanations for this. We tested a new explanation: time perception arises from the processing of our dynamic sensory environment. Supporting this theory, human participants' duration reports of silent videos correlated with estimates we reconstructed from activity in the visual cortex of their brain while they watched those videos. This was not a brain-wide phenomenon; reconstructions based on other brain regions did not correlate with subjective reports of duration. Our results validate our new theory, linking content of experience and perception of time through a common foundation in basic sensory processing.

**Introduction**

Our experience of time is characterized by strong distortions from objective "clock time" (1). These distortions are familiar enough to be reflected in common expressions like "*time flies when you're having fun*" or "*a watched pot never boils*" and illustrate that perception of duration is influenced by the content of sensory experiences. Despite this intuitive relationship between content and time, the neural processes through which it is instantiated are not clear. Moreover, none of the predominant explanations of human time perception provide adequate accounts for this phenomenon. Here we provide evidence from both an artificial perceptual system and model-based analysis of human neuroimaging data for an alternative approach to human time perception. Our account explicitly incorporates the content of experience, based on the neural processes that underlie non-temporal perceptual classification.

The two most well-established explanations of human time perception are the internal clock model (2–4) and the population clock approach (5, 6). In the internal clock approach, it is proposed that some regular physiological or neural process produces rhythmic ticks like the hands of a clock. Biases in experience such that time is "running fast" or "running slow" are putatively driven by changes in the rate of the clock "ticks" (7, 8). However, predictions of the relationship between content and time under the internal clock model are typically determined ad-hoc on a case-by-case basis,

because the relationship between content and time can be complicated by other factors like attention to time (9, 10). The population-clock account proposes that duration is 'read out' by neurons trained to relate time to the length of a trajectory through a network state-space. This explanation makes no explicit attempt to account for content-based biases in subjective duration, though ad-hoc accounts that relate specific sensory content with trajectory length are again available. Neither approach is able to intrinsically account for biases in subjective duration by content because both are rooted in the idea that the brain attempts to track clock time. While tracking clock time is likely to be useful for short, repetitive motor commands such as drumming (11–14) this is not necessarily the case for longer intervals, and indeed as established, subjective experience of longer intervals is characterized more by deviation from rather than adherence to clock time.

Many animal and human electrophysiology and neuroimaging studies have searched for potential neural regions or putative mechanisms underlying time perception (15, 16). This literature finds potential mechanisms in a diverse set of regions including pre-/supplementary motor areas (11, 14, 17), parietal areas (18, 19), insula (20), midbrain dopaminergic neurons (21), and hippocampus and entorhinal cortex (22–25). The diversity of neural regions apparently associated with human time perception supports the intuition underlying the population clock approach - that the dynamics of a neural system can be interpreted as reflecting time to an external observer because it is dynamic, i.e. evolves with objective clock time. However, for understanding the relationship between content of experience and subjective time this position is flawed. In trying to understand subjective time we are not interested in just *any* system dynamics that might be interpreted as time, but the *specific* neural dynamics that link the experience that is being judged (i.e. sensory content) to its subjective duration.

Here we asked whether human-like subjective duration judgements can be constructed from neural activity associated with perceptual classification of a dynamic sensory environment. We tested whether specific neural dynamics underlying perception of an event can form the basis of duration judgements for that experience. We investigated this hypothesis first in the dynamics of an artificial neural classification network while it was input naturalistic video stimuli, and then in human neuroimaging data recorded while participants viewed the same videos.

**Results**

Using functional magnetic resonance imaging (fMRI) and a fully pre-registered preprocessing and model-based analysis pipeline (osf.io/ce9tp), we measured BOLD activation while 40 human participants watched silent videos of natural scenes (8-24 seconds each) and made duration judgements on a visual analogue scale. Half of the videos depicted busy city scenes with many salient events, and the other half, office scenes with very few. We reasoned that if subjective time is constructed from the accumulation of salient events in sensory cortex, then videos with more salient events (city scenes) should be judged as lasting longer relative to videos with few (office scenes).

Participants' bias towards under- or over-reporting of duration was quantified using a (pre-registered) normalized bias measure, which for each level of duration t and each duration report for that duration $x_t$ was defined as:

$$\mathrm{bias}_x = \frac{x - \bar{x}_t}{\bar{x}_t}$$

Positive/negative values mean that durations were over-/under-estimated relative to the participant's mean for a given veridical video duration. Therefore, normalized bias here reflects idiosyncratic response patterns that are independent of clock time.

**Behavioral reports are biased by scene type.**

Participants could estimate duration well, as indicated by a strong correlation between presented and reported durations for each subject, both when computed trial-by-trial ($\bar{\rho} = 0.79 \pm 0.10$) and when averaged within duration categories ($\bar{\rho} = 0.99$, Fig. 1A, left). As predicted, durations of city scenes were relatively over-estimated and office scenes under-estimated, $t_{39} = 3.81$, $p < 0.001$, $M_{diff} = 5.18 \pm 1.36$, $BF_{H(0,10.5)} = 322$, confirming that natural scenes containing a higher density of salient events do indeed feel longer (Fig. 1A, right).

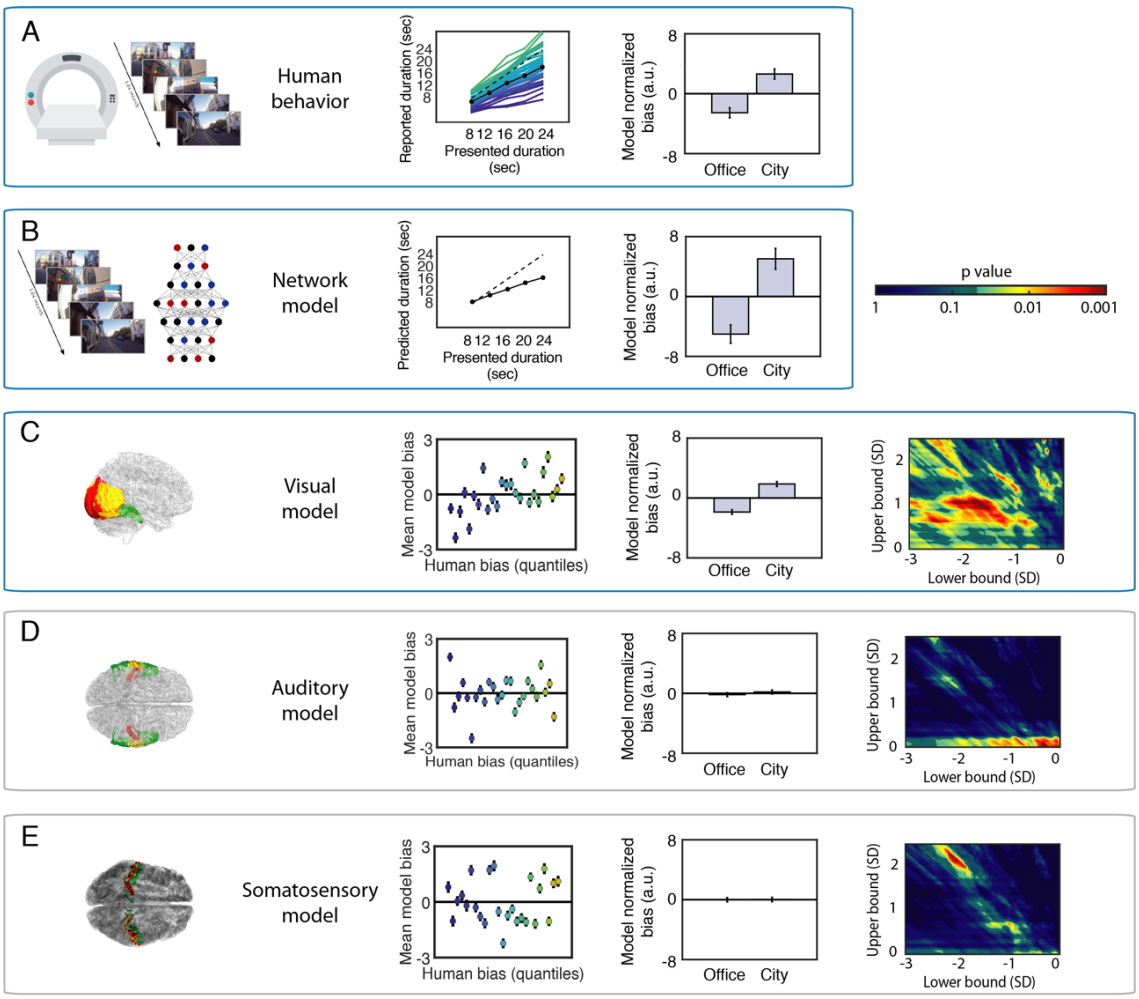


**Figure 1.** *Behavioral and modelling results.* **(A)** Behavioral results. Left: participant-wise relationship between report and duration (colored lines), mean relationship (solid black line), and the line of unity (dashed line). Right: relative under-/over-estimation of duration by human participants for office/city videos. **(B)** Artificial network model results. Left: relationship between model-predicted and veridical durations for this model, trained on accumulated salient events in video frames. Right: relative under-/over- estimation of duration for office/city scenes for this model. (**C-E, first column)** Mean normalized bias for the model trained on visual, auditory and somatosensory cortex activity respectively, as a function of 30 quantiles of human bias. Colors represent x-axis values. **(C-E, middle column)** Normalized bias as a function of video type. Only the visual cortex model exhibited a significant difference. **(C-E, right column)** Heat map depicting p-values for the association between human and model bias, as a function of minimum (x-axis) and maximum (y-axis) criterion values. Dark colors represent regions where the association was non-significant at $\alpha_{0.05}$ or negative. For all bar plots, error bars represent between- or within-participant SEM as appropriate.

**Reports generated by an artificial network model are biased by scene type**

Next, we tested whether this effect of scene type was reproduced by an artificial perceptual classification network. We fed the same video clips that participants had viewed to a pre-trained image classification network, AlexNet (26), and computed frame-to-frame Euclidean distances in network activity for each network node. Separately for each network layer, distance was categorized as salient or not by an attention threshold with exponential decay (see Methods). Salient events were accumulated at each layer and converted to estimates of duration in seconds via multiple linear regression by mapping the number of accumulated salient events to the presented (not reported) durations (Fig 1B, left). As for human behavior, model-produced durations were significantly correlated with the presented durations $\rho = 0.74$, $p < 0.001$, and model estimations exhibited the human effect of video type on estimation bias, $M_{diff} = 10.05 \pm 0.93$, $t_{1018} = 5.39$, $p < 0.001$ (Fig 1B, right). These results demonstrate that simply tracking the dynamics of a perceptual classification network while it is exposed to natural scenes can produce human-like estimations (and distortions) of duration.

**Neural correlates of scene type and estimation bias**

Using human fMRI data, we then examined the neural correlates of stimulation with these videos of natural scenes. Using a GLM, we tested effects on BOLD as a function of both how busy the scene was (office versus city) and participants' video-by-video estimation biases. Busy city

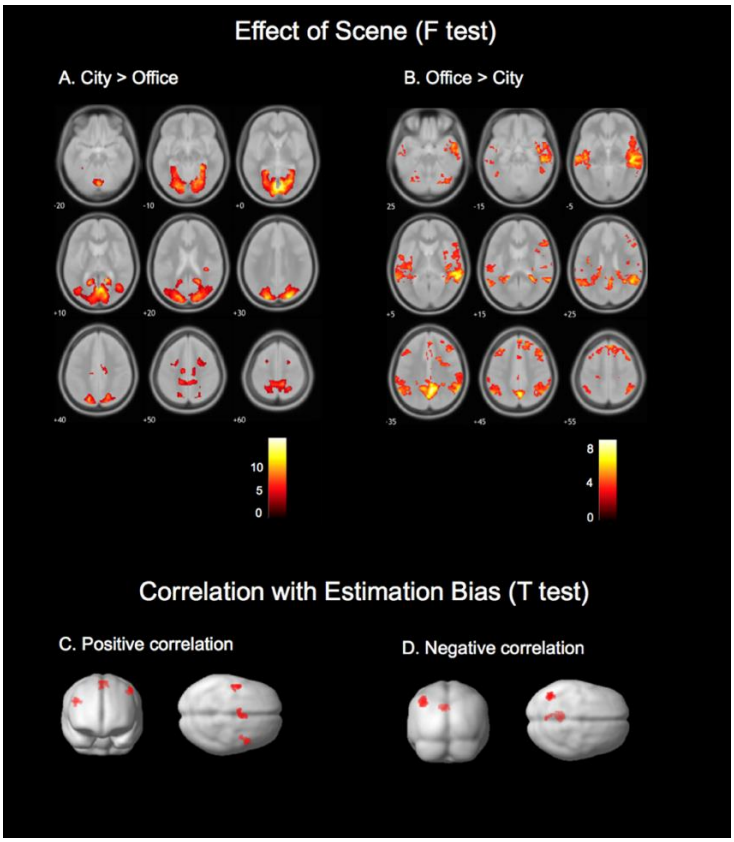


**Figure 2.** *Results from confirmatory GLM on BOLD (significant clusters only).* **A** Higher BOLD for city than office scenes: R lingual gyrus; bilateral midcingluate area; R insula; bilateral SFG. **B** Higher BOLD for office than city scenes: R precuneus; bilateral precentral gyrus; L MFG; bilateral cerebellum; L paracentral lobule; R SFG. **C** Positive correlation with normalized estimation bias: bilateral precentral gyrus; L SMA; R superior occipital gyrus. **D** Negative correlation with normalized estimation bias: L angular frontal gyrus; L MFG; L posterior cingulate.

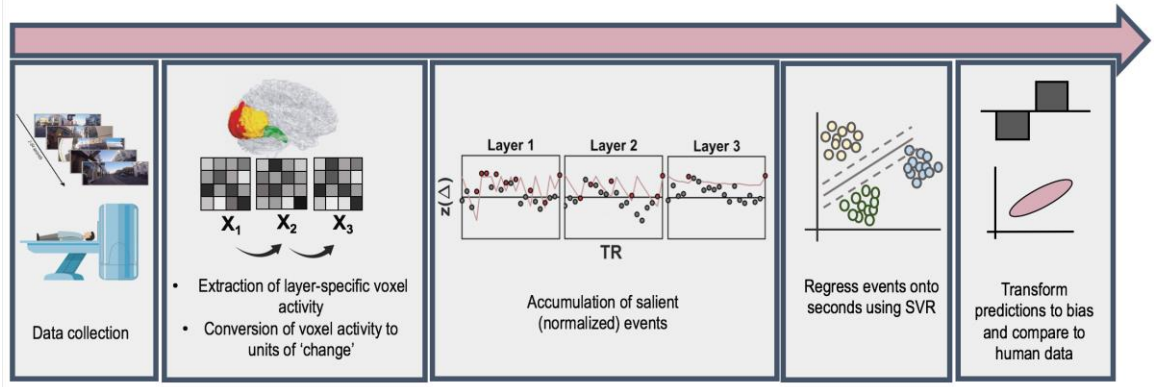


**Figure 3.** *Schematic of modelling analysis pipeline.* Following data collection, voxel-wise data was extracted and TR-by-TR changes (Euclidean distance or signed difference) computed. Total change in the ROI over TR was compared to an attention threshold that categorized events as salient or not. Accumulated salient events were regressed onto seconds and compared across condition and with human behavior.

scenes drove bilateral occipital lobe BOLD (as expected), as well as a set of frontal regions including right insula and bilateral superior frontal gyrus (Fig 2C). Furthermore, propensity to overestimate durations was correlated with higher right superior occipital gyrus activation (Fig 2D). Full GLM results are presented in Figure 2 and Table S1.

**Reconstructing human-like duration reports from visual cortex BOLD**

For our key analysis we tested whether we could reproduce human-like estimation biases from salient changes in BOLD activation, as we were able to do in a stimulus-driven fashion when using node activation changes in the artificial classification network. To do this, we defined a three-layer visual hierarchy *a priori* predicted to be involved in processing of the silent videos (see Fig.1 and supplementary methods). We selected regions such that lower layers reflect the detection and classification of low-level features (e.g. edge detection in primary visual cortex; V1), and higher layers, object-related processing (e.g. lateral occipital cortex; LOC). For control analyses, analogous hierarchies were built for auditory and somatosensory cortex (see Table S2). Because the stimuli we used were silent videos, our prediction was that only the model trained on the visual cortex hierarchy should reconstruct human duration reports from accumulated salient events (see pre-registration at osf.io/ce9tp).

We ran this key analysis in two ways: one was confirmatory (i.e. pre-registered) and one was exploratory. The pipeline is illustrated in Figure 3. In both analyses, for each participant voxel-wise patterns of BOLD were extracted from each TR (slice, or time point) in each hierarchical layer. Voxel-wise changes between each TR were calculated and then summed over all voxels in the layer, resulting in one value per TR. These 'change' values were standardized within-participant and compared to a criterion with exponential decay to classify the change value as a salient event or not, giving us the number of salient events detected by each layer for each video.

For the pre-registered analysis, change was quantified as Euclidean distance (as for the artificial network model), i.e.

$$\Delta_{TR} = \sum_{v} | X_{TR,v} - X_{TR-1,v} |$$

where $X_{TR,v}$ is activation in voxel *v* at slice *TR*. For the exploratory analysis, we tested an alternative algorithm for quantifying "change":

$$\Delta'_{TR} = \sum_{v} (X_{TR,v} - X_{TR-1,v})$$

i.e. signed difference. We did this because, at least in sensory cortices, BOLD may already reflect salient changes in perception (27), potentially in the form of "prediction errors". If this is the case, $\Delta_{TR}$ would reflect differences in change, rather than the change itself.

Finally, bias in model predictions were compared to human estimation bias. For our pre-registered analysis, we pooled human participants' data together to create a 'super-subject' by standardizing behavioral duration reports within-participant and re-computing estimation bias on the combined behavioral dataset. For the exploratory analysis, human estimation bias was computed separately for each of the 40 participants because pooling participants' data reduced the effect of video type on (human) estimation bias (see Fig. S1).

**Using Euclidean distance, estimation bias but not effects of scene type can be reconstructed from visual cortex BOLD**

The presented video duration could be predicted from salient events in all three confirmatory models to a similar degree (10-fold cross validation, $\bar{r}_{visual} = 0.93, \bar{r}_{auditory} = 0.95, \bar{r}_{somatosensory} = 0.94$), indicating that all models could reproduce "clock time".

Supporting our hypothesis, only the pre-registered model trained on visual salient events significantly reproduced the super-subject's estimation biases, $\beta_{2328} = 1.51$, $p = 0.015$; the models trained on salient events in auditory cortex, $\beta_{2328} = 0.87$, $p = 0.141$, and somatosensory cortex, $\beta_{2328} = 0.30$, $p = 0.339$, did not (Fig. S2). The visual model also predicted video-by-video biases better than other models (Fig. S3). This means that biases in subjective estimates of time can be predicted from neural activity associated with modality-specific perceptual processing.

While the visual model could reproduce human-like estimation biases video-by-video, neither the visual cortex model nor the two control models exhibited a significant difference in bias for city versus office scenes, though the visual cortex model alone exhibited the behavioral direction of difference numerically (visual: $M_{diff} = 0.19 \pm 13.96$, $t_{2329} = 0.33$, $p = 0.739$, auditory: $M_{diff} = -0.33 \pm 12.29$, $t_{2329} = -0.64$, $p = 0.522$, somatosensory: $M_{diff} = 0.16 \pm 13.09$, $t_{2329} = -0.30$, $p = 0.762$, see Fig. S3). Note that these *t*-tests were not pre-registered.

**Using Signed Difference, estimation bias and effects of scene type can be reconstructed from visual cortex BOLD**

Again, presented video duration could be predicted from salient events in all three exploratory models to a similar degree (10-fold cross validation, $\bar{r}_{visual} = 0.95, \bar{r}_{auditory} = 0.97, \bar{r}_{somatosensory} = 0.96$). However for this pipeline, in which salient events were determined from signed differences in voxel activity, linear mixed models revealed the visual model biases *did* strongly discriminate between office and city scenes, $M_{diff} = 3.75 \pm 0.23$, $\chi^2(1) = 85.06$, $p < 0.001$ (Fig. 1C, middle). They also remained correlated with participants' biases, $\beta = 0.02 \pm 0.008$, $\chi^2(1) = 5.62$, $p = 0.018$. This association is visualized in Fig. 1C (left) by plotting mean model bias as a function of 30 quantiles of human bias. Again, the visual model out-performed control models (Fig. S4).

While (exploratory) models trained on accumulated visual cortex salient events reproduced human behavior, again, biases from exploratory models trained on auditory and somatosensory salient events did not: they neither discriminated video type ($M_{diff} = 0.36 \pm 0.19$, $\chi^2(1) = 0.43$, $p = 0.514$, $M_{diff} = 0.02 \pm 0.21$, $\chi^2(1) = 0.46$, $p = 0.499$ respectively, see Fig 1D-E, middle), nor predicted trial-wise human normalized bias ($\beta = -0.003 \pm 0.006$, $\chi^2(1) = 0.20$, $p = 0.652$, $\beta = 0.002 \pm 0.007$, $\chi^2(1) = 0.11$, $p = 0.740$ respectively, Fig. 1D-E, left), underlining the specificity of visual cortex activity in predicting subjective time for silent videos.

**Discussion**

We have shown that subjective estimates of duration can be constructed from salient events in sensory cortex. In this study, for which stimuli were silent videos, this effect was only seen for models trained on salient events in visual cortex BOLD and not models based on somatosensory or auditory cortex BOLD. While we could reconstruct *veridical* clock time from all three sensory regions, only the information extracted by our visual model was related to subjective duration. These results reflect our initial proposal that divergence between objective and subjective time characterize natural experience of time, and show that the dynamics of sensory processing related to specific content - not just any neural dynamics - underlie subjective time. Our results were robust under a wide range of model parameter values (Fig. 1C-E, right), and, in combination with results from the perceptual classification network model and previous findings (10, 28), supports the idea that human time perception is based in the neural processes associated with processing the sensory context in which time is being judged.

Our approach provides an intuitive link between sensory content and subjective duration because the neural processes that are engaged in the processing of sensory content are the same processes that are used to build the estimate of time. While we have demonstrated this only for the visual modality, we expect that salient events from auditory cortex alone would predict subjective time in auditory-only contexts, and likewise for other modalities. This conclusion is in support of the idea that subjective time depends on distributed mechanisms (29), but that in each case they are naturally linked to sensory content by virtue of being determined by those content-related processes.

Here we tested two possible metrics that could be used by the brain to link sensory content and time on a moment-to-moment basis: Euclidean distance (pre-registered) and signed difference (exploratory). In our data, subjective duration was best reconstructed using signed difference. Though both generated duration estimates that correlated with human behavior, only the latter differentiated video type. This superiority of signed difference is consistent with the view that BOLD already indexes detected environmental changes, in line with literature evidencing “surprise” or “prediction error” responses in sensory (27, 30, 31) and even frontal (32, 33) cortices. The idea that our model is based on an index of perceptual “surprise” – the difference between current sensory stimulation and expected stimulation based on an internal world model – is intriguing as it provides a natural link to the closely related topic of episodic memory (see also 10).

Prediction error has also been suggested to provide the basis for event boundaries in episodic memory (10, 34, 35) – transitions that segment some content (e.g. a cow) from some other content (e.g. a car) in continuous experience (36, 37). By emphasizing the importance of sensory content in time perception, our approach may provide the link between time perception and episodic memory that was lost by the content-free “clock” approaches. Further, by demonstrating a simple algorithm for how the stream of basic sensory processing is segmented into salient events, our approach may provide insight into how low-level sensory information is transformed into the temporally sequenced form of memory demonstrated by so-called “time cells” (24, 25, 38), linking the content of basic sensory processing with temporal properties of episodic memory within the powerful predictive coding approach (10, 39, 40).

In summary, we provide evidence for a simple algorithmic account of how duration perception can be achieved, in which the relevant information for time estimation can be obtained simply by tracking the dynamics of the relevant perceptual processing hierarchy. In this view, the processes underlying subjective time have their neural substrates in perceptual and memory systems, not systems specialized for time itself. Other recent studies of time perception (21, 23, 41, 42) have

attempted to correlate maps of neural activity with a specific timing-related behavior, responses, or physical elapsed durations. In contrast, we have taken a model-based approach to describe how sensory information arriving in primary sensory areas is transformed into subjective time. Our model provides a computational basis from which we can unravel how human subjective time is generated, encompassing every step from low level sensory processing to the detection of salient perceptual events, and further on to the construction and ordering of episodic memory. This end-to-end account of time perception represents a significant advance over homuncular accounts that depend on "clocks" in the brain.

## Materials and Methods

### Participants.

The study was approved by the Brighton and Sussex Medical School Research Governance and Ethics Committee (reference number ERA/MS547/17/1). Forty healthy, English speaking and right-handed participants were tested (18-43 years old, mean age = 22y 10mo, 26 females). All participants gave informed, written consent and were reimbursed £15 for their time.

### Procedure.

The experiment was conducted in one sixty minute session. Participants were placed in the scanner and viewed a computer visual display via a head-mounted eyetracker, placed over a 64-channel head coil. Eyetracker calibration lasted approximately five minutes and involved participants tracking a black, shrinking dot across nine locations: in the center, corners and sides of the visual display. Eyetracking data are not used in this manuscript due to technical failure.

Following calibration, we acquired six images reflecting distortions in the magnetic field (three in each of the posterior-to-anterior and anterior-to-posterior directions) and one T1-weighted structural scan.

Finally, functional echoplanar images (EPIs) were acquired while participants performed two to four blocks (time-permitting) of twenty trials, in which participants viewed silent videos of variable length and reported the duration of each video using a visual analogue scale extending from 0 to 40 seconds. A key grip was placed in each hand, and participants moved a slider left and right using a key press with the corresponding hand. Participants were not trained on the task prior to the experimental session.

### Experimental design and trial sequence.

Each experimental block consisted of 20 trials. On each trial a video of duration 8, 12, 16, 20 or 24 seconds was presented. For each participant, videos of the appropriate duration and scene category were constructed by randomly sampling continuous frames from the stimuli built for Roseboom et al. (2019). These videos depicted either an office scene or a city scene. Two videos for each duration and content condition were presented per block in randomized order.

### MRI acquisition and pre-processing (confirmatory).

Functional T2* sensitive multi-band echoplanar images (EPIs) were acquired on a Siemens PRISMA 3T scanner. Axial slices were tilted to minimize signal dropout from parietal, motor and occipital cortices. 2mm slices with 2mm gaps were acquired (TR = 800ms, multiband factor = 8, TE = 37ms, Flip angle = 52°). Full brain T1-weighted structural scans were acquired on the same scanner and were composed of 176 1mm thick sagittal slices (TR = 2730ms, TE = 3.57ms, FOV = 224mm x 256mm, Flip angle = 52°) using the MPRAGE protocol. Finally, we collected reverse-phase spin echo field maps, with three volumes for each of the posterior to anterior and anterior to posterior directions (TR = 8000ms, TE = 66ms, Flip Angle = 90°). Corrections for field distortions were applied by building

fieldmaps from the two phase-encoded image sets using FSL's TOPUP function. All other image pre-processing was conducted using SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/).

The first four functional volumes of each run were treated as dummy scans and discarded. Images were pre-processed using standard procedures: anatomical and functional images were reoriented to the anterior commissure; EPIs were aligned to each other, unwarped using the fieldmaps, and co-registered to the structural scan by minimizing normalized mutual information. Note that in accordance with HCP guidelines for multiband fMRI we did not perform slice-time correction (43). Following co-registration, EPIs were spatially normalized to MNI space using parameters obtained from the segmentation of T1 images into grey and white matter. Finally, spatially normalized images were smoothed with a Gaussian smoothing kernel of 4mm FWHM. Smoothed data were used for the GLM on BOLD; unsmoothed data were used for the brain-based modelling.

**Statistical analyses.**
All fMRI pre-processing, participant exclusion criteria, behavioral, imaging and computational analyses were comprehensively pre-registered while data collection was ongoing (https://osf.io/ce9tp/) but before it was completed. This analysis plan was determined based on pilot data from four participants, and was written blind to the data included in this manuscript. Analyses that deviate from the pre-registered analysis plan are marked as "exploratory". Pre-registered analyses are described as "confirmatory". Data are freely available to download at osf.io/2zqfu.

**fMRI statistical analysis (confirmatory).** At the participant level, BOLD responses obtained from the smoothed images were time-locked to video onset. BOLD responses were modelled by convolving the canonical haemodynamic response function with a boxcar function (representing video presentation) with width equal to video duration. Videos of office and city scenes were modelled using one dummy-coded regressor each. Each was parametrically modulated by normalized bias.

Data from each run was entered separately. No band-pass filter was applied. Instead, low-frequency drifts were regressed out by entering white matter drift (averaged over the brain) as a nuisance regressor (33, 44). Nuisance regressors representing the experimental run and six head motion parameters were also included in the first level models. Because of our fast TR, models were estimated using the 'FAST' method implemented in SPM.

Comparisons of interest were tested by running four one-sample t-tests against zero at the participant level for each variable of interest (video scenes, office scenes, and their normalized bias parametric modulator). Next, group-level F tests were run on those one-sample contrast images to test for effects of video type and the interaction between video type and normalized bias slope. A one-sample t-test against zero at the group level tested the slope of the normalized bias-BOLD relationship. All group-level contrasts were run with peak thresholds of $p < .001$ (uncorrected) and corrected for multiple comparisons at the cluster level using the FWE method. Clusters were labelled using WFU PickAtlas software (45, 46).

**Model-based fMRI (confirmatory).**

Our key prediction was that subjective duration estimates (for these silent videos) arise from the accumulation of salient (perceptual) events detected by the visual system, particularly within higher-level regions related to object processing. We tested this by defining a (pre-registered) three-layer hierarchy of regions to represent core features of the visual system:

Layer 1 was defined as bilateral V1, V2v and V3v, Layer 2 was defined as bilateral hV4, LO1 and LO2, and Layer 3 as bilateral VO1, VO2, PHC1 and PHC2 (clusters are depicted in Figure 1). For each layer, masks were constructed by combining voxels from each area, using the atlas presented in (47).

To determine events detected by the visual system over the course of each video, we extracted raw voxel activity for each TR in each layer from unsmoothed, normalized EPIs. Then, for each voxel v,

change was defined as the Euclidean distance between BOLD activation $x_v$ at volume TR and TR-1. The amount of change detected by the layer at any time point, denoted $\Delta_{TR}$, was then given by summing the Euclidean distances over all voxels such that:

$$\Delta_{TR} = \sum_{v} | X_{TR} - X_{TR-1} |$$

This process furnishes one value per layer for each TR of each trial for each participant. The next step was to categorize each value as a "salient" event or not and convert to an estimate of duration using an event detection, accumulation and regression model, as presented in Roseboom et al (48), for example, Figure 2. To do this, we first pooled participants' data by z-scoring the summed events $\Delta_{TR}$ within each participant and layer. Pooling was performed to increase statistical power of our subsequent regression analyses. Then, for each trial, TR-by-TR categorization of $\Delta_{TR}$ was achieved by comparing against a criterion with exponential decay, corrupted by Gaussian noise ε:

$$\vartheta_{TR} = ae^{-TR} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,0.05)$$

Only the parameter a took different values in each layer (see Table S3). The criterion decayed with each TR until either an event was classified as salient or until the video finished, after each of which the criterion reset to its starting point. Importantly, because the summed Euclidean distances $\Delta_{TR}$ were z-scored, the criterion has meaningful units corresponding to SDs above or below the mean. To account for potential head-motion artefacts, criterion updating ignored volumes where $\Delta_{TR}$ was greater than 2.5 (i.e. more than 2.5 SDs from the mean).

The final modelling step was to predict raw duration judgements (in seconds) from the BOLD-determined accumulation of salient events. This was achieved via Epsilon-support vector regression (SVR, implemented on python 3.0 using sklearn (49)) to regress accumulated events in each of the three layers onto the veridical video duration.

To evaluate whether the model could reproduce human-like reports of time from participants' BOLD activation, we converted the trial-by-trial model predictions to normalized bias. These were then compared to a human "super-subject": participants' duration judgements were z-scored within participants, then all participant data were pooled and converted to normalized bias. We created a super-subject to mirror the data pooling performed before training our SVR.

Trial-by-trial normalized bias values were compared across model and human using linear regression, fitting the model:

$$\text{behaviour}_t = \beta_0 + \beta_1 \text{model}_t$$

To test our a priori hypothesis that the model trained on visual cortex salient events positively correlates with subjective time, a (one-tailed) p-value for $\beta_1$ was calculated via bootstrapping, shuffling the behavioural data and refitting the regression line 10,000 times.

**Control models (confirmatory).**

The aforementioned steps were replicated on two alternative, control hierarchies. The purpose of these was to determine whether, if our hypothesis held for visual cortex, salient events accumulated by any sensory region is sufficient for predicting subjective time.

The first control hierarchy was auditory cortex, previously implicated in time perception but whose involvement in duration judgements should not be driven by visual stimuli, as in our study. Layers 1 and 2 were defined as Brodmann Area (BA) 41 and 42 respectively, both of which are located in primary auditory cortex. Layer 3 was posterior BA22 (superior temporal gyrus/Wernicke's Area).

The second control hierarchy was somatosensory cortex, which we reasoned should not be involved in duration judgements based on visual stimuli. Layer 1 was set as posterior and anterior BA 3, and

layers 2 and 3 were set as BA 1 and 2 respectively. These Brodmann areas correspond to the primary somatosensory cortex.

Masks for these two control analyses were constructed using WFU PickAtlas atlases (45, 46). As for our empirical analyses using visual cortex, for each of the two controls we estimated the relationship between the trial-by-trial normalized bias based on the model's predictions and based on z-scored participant data by fitting a linear regression line.

To test whether the visual cortex model out-performed the somatosensory and auditory cortex models we compared their log-likelihoods, obtained from the Matlab function *fitlm* (see Fig. S3). This evaluation of model performance was *not* pre-registered.

**Exploratory modelling.** We also ran an exploratory (i.e. not pre-registered) set of models. This was identical to the pre-registered analysis plan, apart from the following differences:

First, we transformed voxel-wise BOLD activation X to signed (i.e. raw) rather than unsigned changes:

$$\Delta'_{\mathrm{TR}} = \sum_{\mathrm{v}} (\mathrm{X}_{\mathrm{TR}} - \mathrm{X}_{\mathrm{TR}-1})$$

Using SVR as before, for each hierarchy we obtained model-predicted duration estimates in seconds. To avoid pooling participants' reports together, human judgements were not standardized. Instead, for each of our 40 participants we computed human and model normalized biases from the human reports and model predictions associated with their set of videos. In other words, normalized bias was computed 'within-participant'.

To test the association between video-by-video human and model bias while accounting within-participant variability we used a linear mixed model approach. Using R with the *lmer* and *car* packages, we fit the following random-intercept model:

$$\mathrm{bias}_{\mathrm{human}} \sim 1 + \mathrm{bias}_{\mathrm{model}} + (1|\mathrm{participant})$$

A chi-squared test (from the *car* function *Anova*) was used to determine the significance of the beta value for the fixed effect of $\mathrm{bias}_{\mathrm{human}}$.

To test the effect of video type (or scene) on model normalized bias, we fit the model:

$$\mathrm{bias}_{\mathrm{model}} \sim 1 + \mathrm{scene} + (1|\mathrm{participant})$$

Again, we used a chi-squared test to determine the significance of the beta for scene.

To test whether the model trained on visual cortex events out-performed the somatosensory and auditory models, we built the following control LMMs:

$$\mathrm{bias}_{\mathrm{human}} \sim 1 + (1|\mathrm{participant}) \text{ and } \mathrm{bias}_{\mathrm{model}} \sim 1 + (1|\mathrm{participant})$$

Using the *car* function *Anova*, we computed the difference in AIC between the main and control models for each hierarchy (see Fig. S4).

**Robustness analysis (exploratory).** To illustrate the robustness of our exploratory analysis to criterion parameters we reran the above analysis pipeline under varying values of $\vartheta_{min}$ and $\vartheta_{max}$. For layer 1 (where there should be most salient changes), $\vartheta_{min}$ took 50 linearly-spaced values between 3 SD and 0 SD below the mean. $\vartheta_{max}$ independently took 50 linearly-spaced values between 0 SD and 2.5 SD above the mean. We chose 2.5 SD here because this was the highest value z-scored BOLD could take before being discarded as a head motion artefact. For each $\vartheta_{min}$ and $\vartheta_{max}$ values for layer 1, the lower/upper bounds for layer 2 were $\vartheta_{min}$ + 0.5 and $\vartheta_{max}$ + 0.5 respectively. For layer 3, they were $\vartheta_{min}$ + 1 and $\vartheta_{max}$ + 1 respectively.

With these criteria, we obtained 250 datasets for each ROI. For each ROI and dataset we tested the association model predictions and human data by fitting the regression model:

$$bias_{human} = \beta_0 + \beta_1 \times bias_{model}$$

Heat maps depicted in Fig. 1 correspond to one-tailed p-values for $\beta_1$. This robustness analysis was *not* pre-registered.

**Artificial classification network-based modelling.** Frames from each video presented during the experiment were fed into the model presented in Roseboom et al (48). Instead of accumulating events based on changes in BOLD amplitude, salient events in the video frames themselves were detected by an artificial image classification network (Alexnet)(26). We used nine network layers (input, conv1, conv2, conv3, conv4, conv5, fc6, fc7, and output, where fc corresponds to a fully connected layer and conv to the combination of a convolutional and a max pooling layer). Node-wise Euclidean distances for each node were computed, then summed over all nodes in the layer giving us one value per video frame and layer. Each value was classified as a salient event or not using the same exponentially decaying criterion as before (see Table S4 for criterion values). Finally, accumulated salient events were mapped onto units of seconds using multiple linear regression.

**Acknowledgments**
Thank you to Charlotte Rae, Petar Raykov, Samira Bouyagoub, Chris Bird, and Mara Cercignani for their assistance with this project. Thanks also to Virginie van Wassenhove and Martin Wiener for comments on an earlier version of the manuscript. 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 Dr Mortimer and Theresa Sackler Foundation (MTS and AKS), which supports the Sackler Centre for Consciousness Science. AKS is also grateful to the Canadian Institute for Advanced Research (CIFAR) Azrieli Programme in Brain, Mind, and Consciousness.

**References**

1. D. M. Eagleman, Human time perception and its illusions. *Curr. Opin. Neurobiol.* **18**, 131–6 (2008).
2. M. Treisman, Temporal discrimination and the indifference interval. Implications for a model of the “internal clock”. *Psychol. Monogr.* **77**, 1–31 (1963).
3. M. S. Matell, W. H. Meck, Cortico-striatal circuits and interval timing: Coincidence detection of oscillatory processes. *Cogn. Brain Res.* **21**, 139–170 (2004).
4. B. M. Gu, H. van Rijn, W. H. Meck, Oscillatory multiplexing of neural population codes for interval timing and working memory. *Neurosci. Biobehav. Rev.* **48**, 160–185 (2015).
5. N. F. Hardy, D. V Buonomano, Encoding Time in Feedforward Trajectories of a Recurrent Neural Network Model. *Neural Comput.* **30**, 378–396 (2018).
6. U. R. Karmarkar, D. V. Buonomano, Timing in the Absence of Clocks: Encoding Time in Neural Network States. *Neuron* (2007) https:/doi.org/10.1016/j.neuron.2007.01.006.
7. J. H. Wearden, Slowing down an Internal Clock: Implications for Accounts of Performance on four Timing Tasks. *Q. J. Exp. Psychol.* **61**, 263–274 (2008).
8. D. B. Terhune, J. G. Sullivan, J. M. Simola, Time dilates after spontaneous blinking. *Curr. Biol.* (2016) https:/doi.org/10.1016/j.cub.2016.04.010.
9. R. a Block, P. a Hancock, D. Zakay, How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychol. (Amst).* **134**, 330–43 (2010).
10. Z. Fountas, *et al.*, A predictive processing model of episodic memory and time perception. *bioRxiv*, 2020.02.17.953133 (2020).
11. H. Merchant, O. Pérez, W. Zarco, J. Gámez, Interval tuning in the primate medial premotor

cortex as a general timing mechanism. *J. Neurosci.* (2013) https:/doi.org/10.1523/JNEUROSCI.5513-12.2013.

12. D. V. Buonomano, R. Laje, Population clocks: Motor timing with neural dynamics. *Trends Cogn. Sci.* (2010) https:/doi.org/10.1016/j.tics.2010.09.002.
13. V. Goudar, D. V. Buonomano, Encoding sensory and motor patterns as time-invariant trajectories in recurrent neural networks. *Elife* (2018) https:/doi.org/10.7554/eLife.31134.
14. J. Gámez, G. Mendoza, L. Prado, A. Betancourt, H. Merchant, The amplitude in periodic neural state trajectories underlies the tempo of rhythmic tapping. *PLoS Biol.* (2019) https:/doi.org/10.1371/journal.pbio.3000054.
15. M. J. Allman, S. Teki, T. D. Griffiths, W. H. Meck, Properties of the Internal Clock: First- and Second-Order Principles of Subjective Time. *Annu. Rev. Psychol.* (2014) https:/doi.org/10.1146/annurev-psych-010213-115117.
16. M. Wiener, P. Turkeltaub, H. B. Coslett, The image of time: A voxel-wise meta-analysis. *Neuroimage* (2010) https:/doi.org/10.1016/j.neuroimage.2009.09.064.
17. F. Protopapa, *et al.*, Chronotopic maps in human supplementary motor area. *PLoS Biol.* (2019) https:/doi.org/10.1371/journal.pbio.3000026.
18. D. Bueti, V. Walsh, The parietal cortex and the representation of time, space, number and other magnitudes. *Philos. Trans. R. Soc. B Biol. Sci.* (2009) https:/doi.org/10.1098/rstb.2009.0028.
19. M. J. Hayashi, *et al.*, Time Adaptation Shows Duration Selectivity in the Human Parietal Cortex. *PLoS Biol.* (2015) https:/doi.org/10.1371/journal.pbio.1002262.
20. M. Wittmann, A. N. Simmons, J. L. Aron, M. P. Paulus, Accumulation of neural activity in the posterior insula encodes the passage of time. *Neuropsychologia* (2010) https:/doi.org/10.1016/j.neuropsychologia.2010.06.023.
21. S. Soares, B. V. Atallah, J. J. Paton, Midbrain dopamine neurons control judgment of time. *Science (80-. ).* **354**, 1273–1277 (2016).
22. Y. Ezzyat, L. Davachi, Similarity breeds proximity: Pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity. *Neuron* (2014) https:/doi.org/10.1016/j.neuron.2014.01.042.
23. A. Tsao, *et al.*, Integrating time from experience in the lateral entorhinal cortex. *Nature* (2018) https:/doi.org/10.1038/s41586-018-0459-6.
24. M. E. Montchal, Z. M. Reagh, M. A. Yassa, Precise temporal memories are supported by the lateral entorhinal cortex in humans. *Nat. Neurosci.* **22**, 284–288 (2019).
25. G. Umbach, *et al.*, Time cells in the human hippocampus and entorhinal cortex support episodic memory. *bioRxiv*, 2020.02.03.932749 (2020).
26. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Commun. ACM* (2017) https:/doi.org/10.1145/3065386.
27. T. Egner, J. M. Monti, C. Summerfield, Expectation and surprise determine neural population responses in the ventral visual stream. *J. Neurosci.* **30**, 16601–16608 (2010).
28. Marta Suárez-Pinilla, Kyriacos Nikiforou, Zafeirios Fountas, Anil Seth, Warrick Roseboom, Perceptual content, not physiological signals, determines perceived duration when viewing dynamic, natural scenes. *Collabra Psychol.* **5** (2019).
29. J. J. Paton, D. V. Buonomano, The Neural Basis of Timing: Distributed Mechanisms for Diverse Functions. *Neuron* (2018) https:/doi.org/10.1016/j.neuron.2018.03.045.
30. A. Todorovic, F. van Ede, E. Maris, F. P. de Lange, Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. *J. Neurosci.* **31**, 9118–23 (2011).

31. R. Langner, *et al.*, Modality-specific perceptual expectations selectively modulate baseline activity in auditory, somatosensory, and visual cortices. *Cereb. Cortex* **21**, 2850–2862 (2011).

32. F. Meyniel, S. Dehaene, Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proc. Natl. Acad. Sci. U. S. A.* (2017) https:/doi.org/10.1073/pnas.1615773114.

33. M. T. Sherman, A. K. Seth, R. Kanai, Predictions shape confidence in right inferior frontal gyrus. *J. Neurosci.* **36** (2016).

34. J. M. Zacks, C. A. Kurby, M. L. Eisenberg, N. Haroutunian, Prediction error associated with the perceptual segmentation of naturalistic events. *J. Cogn. Neurosci.* (2011) https:/doi.org/10.1162/jocn_a_00078.

35. S. J. Gershman, A. Radulescu, K. A. Norman, Y. Niv, Statistical Computations Underlying the Dynamics of Memory Updating. *PLoS Comput. Biol.* (2014) https:/doi.org/10.1371/journal.pcbi.1003939.

36. J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, J. R. Reynolds, Event perception: A mind-brain perspective. *Psychol. Bull.* (2007) https:/doi.org/10.1037/0033-2909.133.2.273.

37. G. A. Radvansky, J. M. Zacks, Event boundaries in memory and cognition. *Curr. Opin. Behav. Sci.* (2017) https:/doi.org/10.1016/j.cobeha.2017.08.006.

38. C. J. MacDonald, K. Q. Lepage, U. T. Eden, H. Eichenbaum, Hippocampal “time cells” bridge the gap in memory for discontiguous events. *Neuron* **71**, 737–749 (2011).

39. R. P. N. Rao, D. H. Ballard, Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).

40. K. J. Friston, The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–38 (2010).

41. M. J. Hayashi, W. van der Zwaag, D. Bueti, R. Kanai, Representations of time in human frontoparietal cortex. *Commun. Biol.* (2018) https:/doi.org/10.1038/s42003-018-0243-z.

42. B. M. Harvey, S. O. Dumoulin, A. Fracasso, J. M. Paul, A Network of Topographic Maps in Human Association Cortex Hierarchically Transforms Visual Timing-Selective Responses. *SSRN Electron. J.* (2019) https:/doi.org/10.2139/ssrn.3438365.

43. M. F. Glasser, *et al.*, The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* (2013) https:/doi.org/10.1016/j.neuroimage.2013.04.127.

44. J. R. Law, *et al.*, Functional magnetic resonance imaging activity during the gradual acquisition and expression of paired-associate memory. *J. Neurosci.* **25**, 5720–5729 (2005).

45. J. L. Lancaster, *et al.*, Automated Talairach Atlas labels for functional brain mapping. *Hum. Brain Mapp.* (2000) https:/doi.org/10.1002/1097-0193(200007)10:3<120::AID-HBM30>3.0.CO;2-8.

46. J. A. Maldjian, P. J. Laurienti, R. A. Kraft, J. H. Burdette, An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* (2003) https:/doi.org/10.1016/S1053-8119(03)00169-1.

47. L. Wang, R. E. B. Mruczek, M. J. Arcaro, S. Kastner, Probabilistic maps of visual topography in human cortex. *Cereb. Cortex* (2015) https:/doi.org/10.1093/cercor/bhu277.

48. W. Roseboom, *et al.*, Activity in perceptual classification networks as a basis for human subjective time perception. *Nat. Commun.* (2019) https:/doi.org/10.1038/s41467-018-08194-7.

49. F. Pedregosa, *et al.*, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).

**Supplementary Figures**

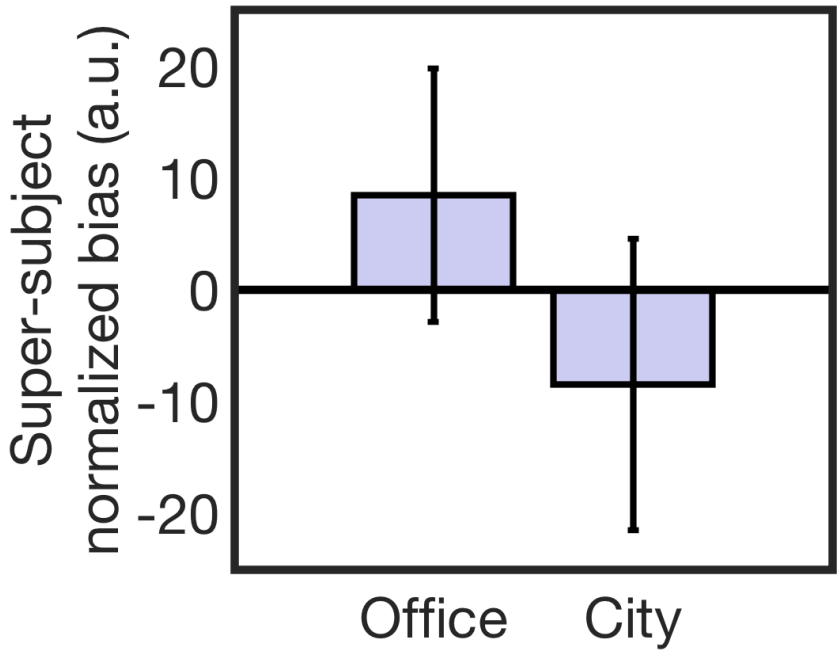


**Figure S1.** Normalized estimation bias computed on pooled ('super-subject') behavioral data, as a function of video scene.

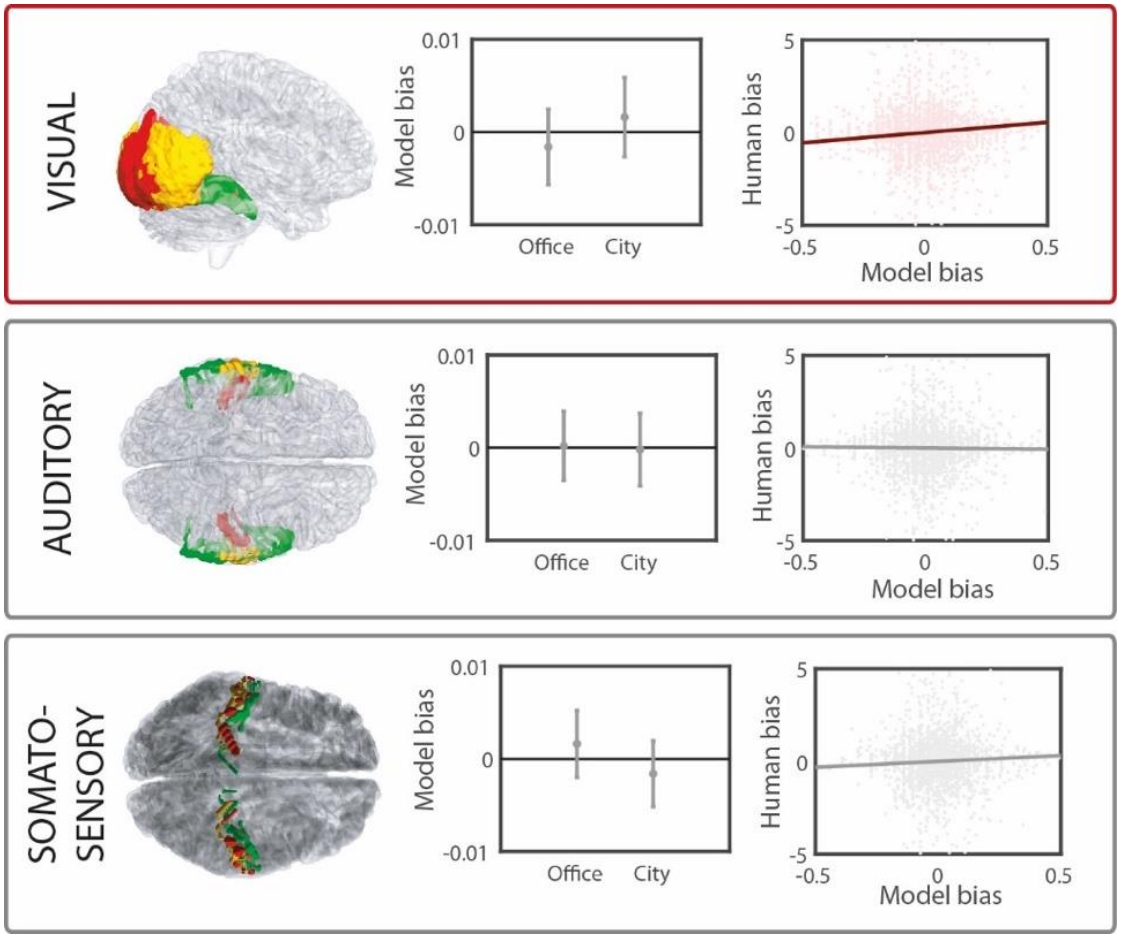


**Figure S2.** Normalized bias predicted by models trained on salient events (Euclidean distance) in visual, auditory and somatosensory hierarchies. **(Left)** Red, yellow and green clusters represent our hierarchical layers 1-3 respectively. **(Middle)** Differences in the models' normalized as a function of video type. Error bars represent +/- SEM. **(Right)** The association between the models' normalized bias and normalized bias from the pooled human data each video.

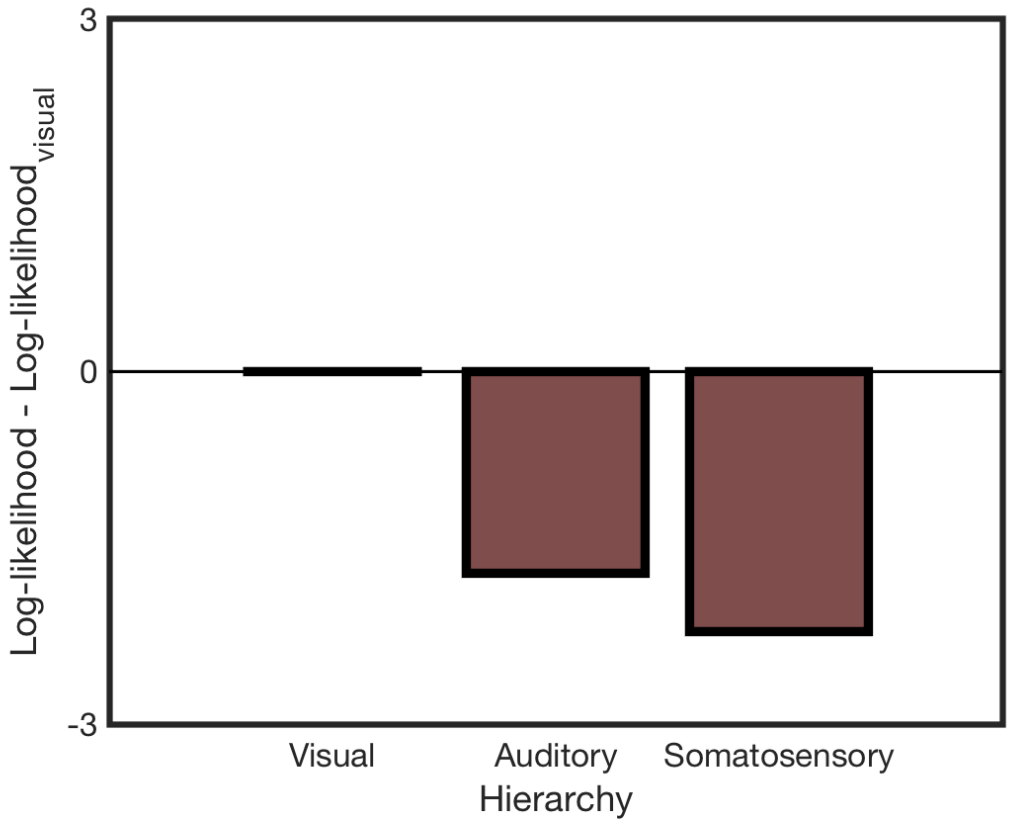


**Figure S3.** Model fits for the regression of predicted durations onto veridical durations (Euclidean Distance models). For each of the visual, auditory and somatosensory models, we regressed the model-predicted biases onto the human super-subject's biases. To compare the performance of the three regressions we compared their log-likelihoods (higher values indicate better model fits). The visual cortex regression outperforms the other two, as indicated by its higher log-likelihood. For visualization purposes, we plot the log-likelihood of the auditory cortex and somatosensory cortex regressions relative to the visual cortex regression log-likelihood.

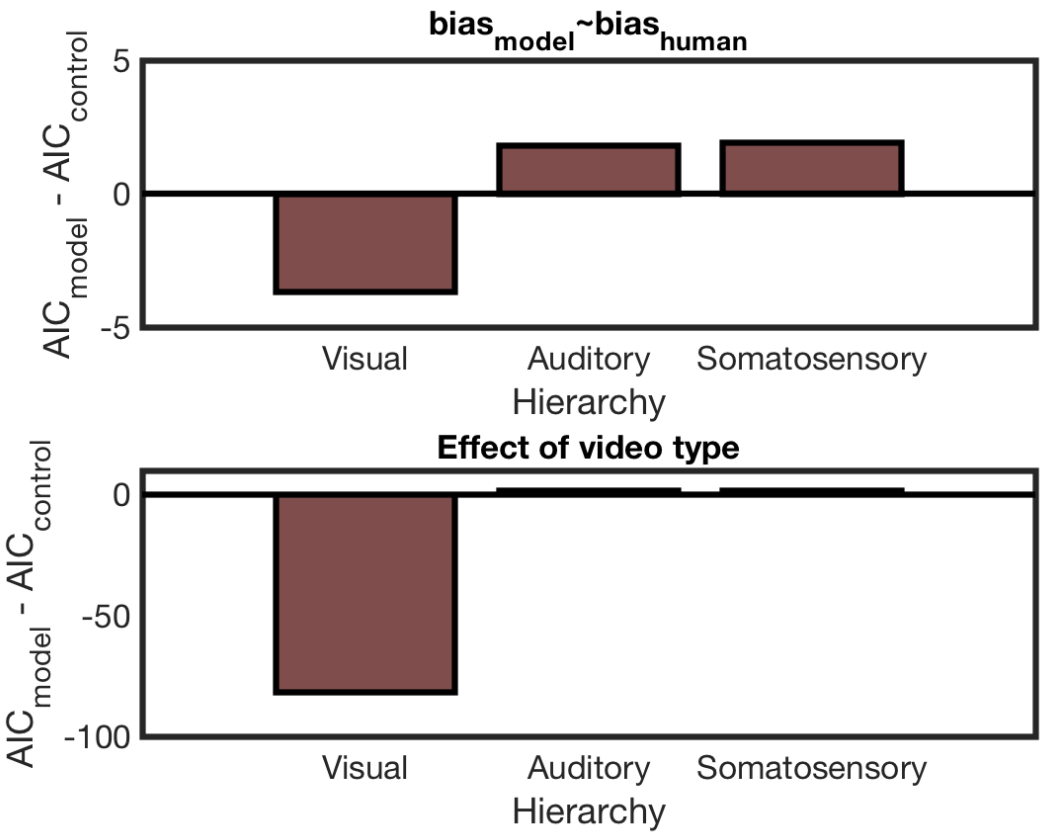


**Figure S4.** Model fits for the linear mixed models (Signed Difference analyses). **Top.** For each of the visual, auditory and somatosensory models, we constructed an LMM with human bias as the outcome and the model-predicted biases as a fixed effect. These LMMs tested the video-by-video correlations between predicted and human bias. These LMMs were compared to control models that did not have the model-predicted bias as a fixed effect, using AIC. The visual cortex LMM outperformed the auditory and somatosensory cortex LMMs, as indicated by the greater decrease in AIC. **Bottom.** To test whether the visual, auditory or somatosensory models generated predicted durations that discriminated video type, we ran LMMs with the fixed effect Video Type. These were compared to control LMMs that did not have this fixed effect, using AIC. The visual cortex LMM out-performed the auditory and somatosensory cortex LMMs as indicated by the greater decrease in AIC.

**Supplementary Tables**

**Table S1.** Significant clusters revealed by confirmatory GLM on BOLD

| Region | Size ($cm^3$) | *T or F* | $P_{FWE}$ | Peak MNI | | |
|---|---|---|---|---|---|---|
| | | | | x | y | z |
| **City > Office (two-tailed)** | | | | | | |
| R Lingual Gyrus | 121.37 | 14.69 | < 0.001 | 6 | -70 | 4 |
| L Midcingulate Area | 1.20 | 8.86 | 0.002 | -10 | -20 | 46 |
| R Insula | 0.6 | 7.91 | 0.049 | 36 | -28 | 22 |
| R Midcingulate Area | 1.66 | 6.73 | <0.001 | 12 | -12 | 44 |
| R Superior Frontal Gyrus | 2.11 | 6.44 | < 0.001 | 24 | 0 | 54 |
| L Superior Frontal Gyrus | 1.26 | 5.94 | 0.001 | -22 | 0 | 54 |
| **Office > City (two-tailed)** | | | | | | |
| R Precuneus | 101.80 | 9.48 | < 0.001 | 6 | -56 | 36 |
| R Precentral Gyrus | 1.86 | 6.45 | < 0.001 | 24 | -26 | 66 |
| L Middle Frontal Gyrus | 4.02 | 6.42 | < 0.001 | -32 | 30 | 48 |
| R Cerebellum 1 | 2.39 | 6.12 | < 0.001 | 46 | -62 | -26 |
| L Precentral Gyrus | 1.18 | 5.82 | 0.002 | -22 | -28 | 62 |
| L Cerebellum 6 | 1.06 | 5.51 | 0.003 | -22 | -70 | -22 |
| L Paracentral Lobule | 0.82 | 4.92 | 0.013 | -6 | -12 | 68 |
| L Superior Frontal Sulcus | 0.93 | 4.44 | 0.007 | -14 | 30 | 54 |
| **Positive correlation with normalized bias** | | | | | | |
| R Precentral gyrus | 0.90 | 4.88 | 0.002 | 38 | 2 | 30 |
| L Precentral gyrus | 0.71 | 4.86 | 0.006 | -48 | 0 | 54 |
| L Supplementary motor area | 0.82 | 4.53 | 0.003 | 0 | 0 | 64 |
| R Superior Occipital Gyrus | 0.48 | 4.03 | 0.041 | 24 | -64 | 46 |
| **Negative correlation with normalized bias** | | | | | | |
| L Angular Gyrus | 1.46 | 5.54 | < 0.001 | -40 | -64 | 44 |
| L Middle Frontal Gyrus | 1.66 | 5.00 | < 0.001 | -2 | -44 | 30 |
| L Posterior Cingulate | 0.62 | 4.96 | 0.013 | -28 | 24 | 56 |

**Table S2.** Definition of hierarchies for each sensory cortex model

| | Visual | Auditory | Somatosensory |
|---|---|---|---|
| Layer 1 | V1, V2v, V3v | BA41 | BA3 |
| Layer 2 | hV4, LO1, LO2 | BA42 | BA1 |
| Layer 3 | VO1, VO2, PHC1, PHC2 | BA22 | BA2 |

**Table S3.** Criterion parameters for each hierarchical layer of the sensory cortex models

| **Layer** | $a$ | $\vartheta_{max}$ (SD above the mean) | $\vartheta_{min}$ (SD below the mean) |
|---|---|---|---|
| **1** | 0.5 | 0.5 | 1 |
| **2** | 1 | 1 | 0.5 |
| **3** | 1.5 | 1.5 | 0 |

**Table S4.** Criterion parameters for the artificial network model

| Layer | tmax | tmin |
|---|---|---|
| conv1 | 39973 | 0 |
| conv2 | 11601 | 0 |
| conv3 | 5515 | 0 |
| conv4 | 3244 | 0 |
| conv5 | 1117 | 0 |
| fc6 | 124 | 0 |
| fc7 | 31 | 0 |
| output | 0.33 | 0 |

For all layers, $\alpha = 0.001/T_{max}$, $\tau = 6.6T_{max}$ while training and $\tau = 10T_{max}$ while testing.