Supplementary Materials

Accumulation of salient events in sensory cortex activity predicts subjective time.

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

Supplementary Table 1

Table S1.

Confirmatory GLM results

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

Supplementary Table 2

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

Supplementary Table 3

Table S3. Criterion parameters for each hierarchical layer

Layer	a	ϑ_{max} (SD above the mean)	ϑ_{min} (SD below the mean)
1	0.5	0.5	1
2	1	1	0.5
3	1.5	1.5	0

Supplementary Table 4

Table S4. Criterion parameters for 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.

Supplementary Figure 1

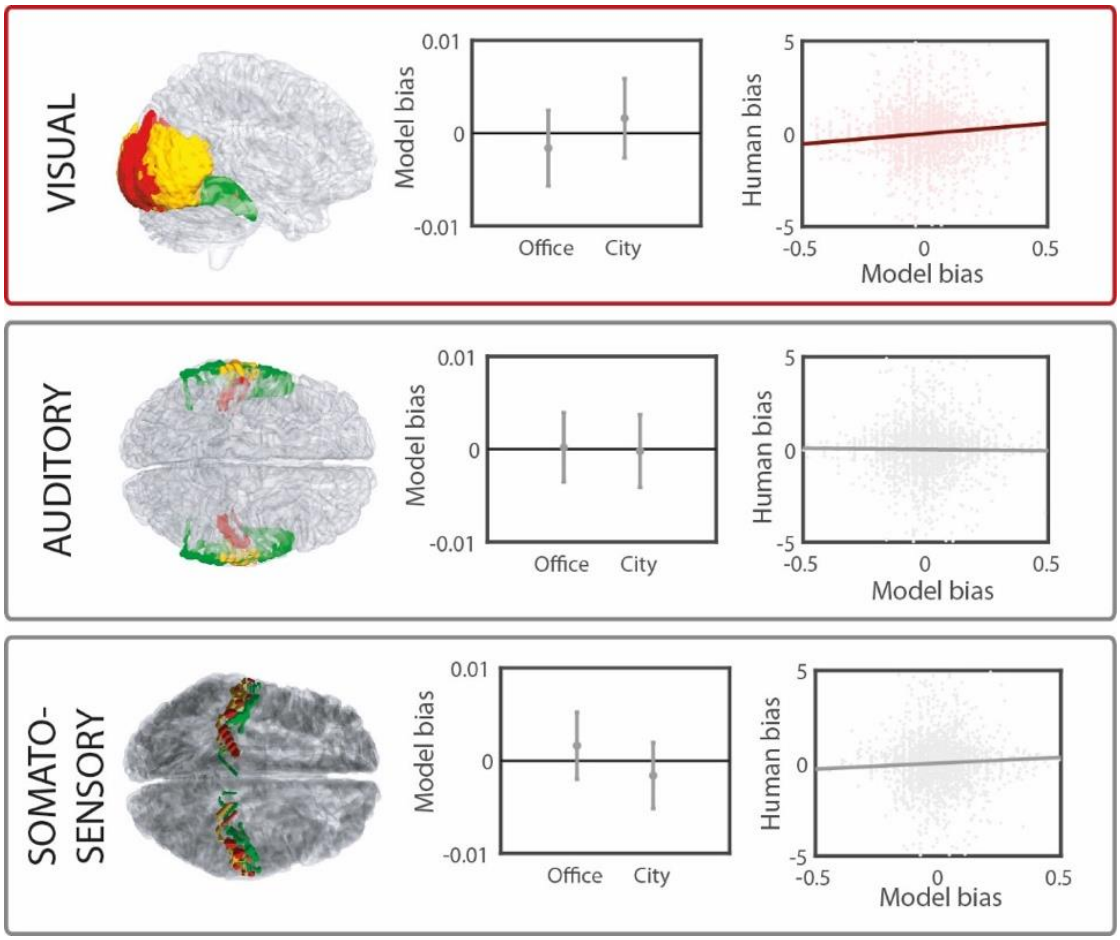

Figure S1. Normalized bias predicted by models trained on salient events (Euclidean distance) *in visual, auditory and somatosensory hierarchies.* **(Left)** Red, yellow and green clusters represent our hierarchical layers 1-3 respectively. **(Middle)** Differences in the models' normalized as a function 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.

Supplementary Figure 2

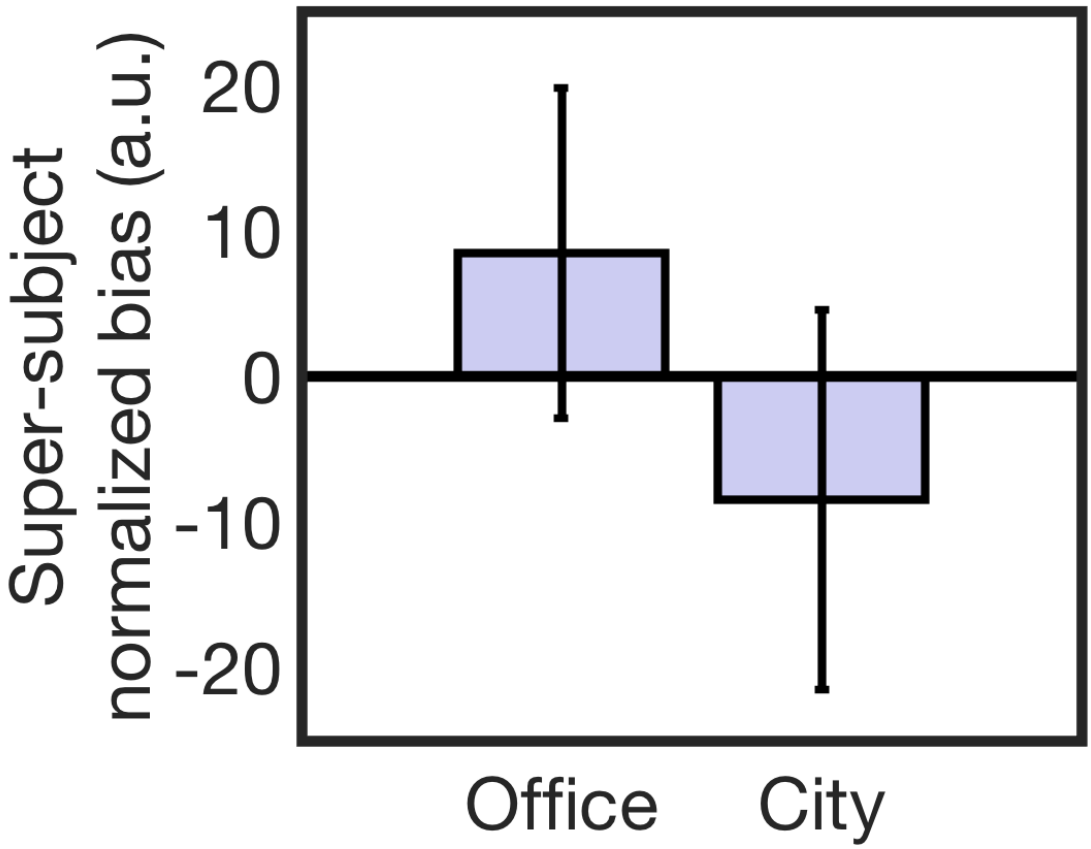

Figure S2. Normalized estimation bias computed on pooled ('super-subject') behavioral data, as a function of video scene (corresponding to the plots in row 3 of fig. 1).

Supplementary Methods

Participants

The study was approved by the Brighton and Sussex Medical School Research Governance and Ethics Committee. 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.

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/). 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” in the Results section. Pre-registered analyses are described as “confirmatory”. Data are freely available to download at osf.io/2zqfu.

Behavioral analyses. Participants' bias towards under- or over-reporting duration was quantified using normalized bias, which for each level of duration *t* and each duration report for that duration x_t is defined as:

$$bias_x = \frac{x - \bar{x_t}}{\bar{x_t}}$$

Positive/negative values mean that durations have been over-/under-estimated, relative to participants' mean duration report (for a given veridical video duration).

MRI acquisition and pre-processing. 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 (*34*). 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.

fMRI statistical analysis. 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 (*21*, *35*). 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 (*36*, *37*).

Model-based fMRI. 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 3). For each layer, masks were constructed by combining voxels from each area, using the atlas presented in (*38*).

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 Δ_{TR}, was then given by summing the Euclidean distances over all voxels such that:

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

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 (*9*), for example, Figure 2. To do this, we first pooled participants' data by z-scoring the summed events Δ_{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

Δ_{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 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 Δ_{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 Δ_{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 (SVM, implemented on python 3.0 using sklearn (*15*)) 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 SVM.

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

$$behaviour_t = \beta_0 + \beta_1 model_t$$

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 β_1 was calculated via bootstrapping, shuffling the behavioural data and refitting the regression line 10,000 times.

Control models

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 (*36*, *37*). 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.

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:

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

Using SVM 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 and the lmer and car packages, we fit the following random-intercept model:

$$bias_{human} \sim 1 + bias_{model} + (1|participant)$$

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

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

$$bias_{model} \sim 1 + scene + (1|participant)$$

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

Robustness analysis

To illustrate the robustness of our exploratory analysis to criterion parameters we reran the above analysis pipeline under varying values of ϑ_{min} and ϑ_{max}. For layer 1 (where there should be most salient changes), ϑ_{min} took 50 linearly-spaced values between 3 SD and 0 SD below the mean. ϑ_{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 ϑ_{min} and ϑ_{max} values for layer 1, the lower/upper bounds for layer 2 were ϑ_{min} + 0.5 and ϑ_{max} + 0.5 respectively. For layer 3, they were ϑ_{min} + 1 and ϑ_{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:

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

Heat maps depicted in Fig. 1 correspond to one-tailed p-values for β_1.

Artificial classification network-based modelling

Frames from each video presented during the experiment were fed into the model presented in Roseboom et al (*9*). 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)(*13*). 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.