Accumulation of Salient Perceptual Events Predicts Human 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 (m.sherman@sussex.ac.uk) & Warrick Roseboom (wjroseboom@gmail.com)

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. However, leading explanations of time perception are not equipped to explain these deviations. Here we test the idea that these deviations arise because time estimates are constructed by accumulating the same quantity that guides perception: salient events. To test this, healthy human participants watched naturalistic, silent videos and estimated their duration while fMRI was acquired. Using computational modelling, we show that accumulated events in visual, auditory and somatosensory cortex all predict 'clock time', but duration biases reflecting human *experience* of time could only be predicted from the region involved in modality-specific sensory processing: visual cortex. 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.

Introduction

Our experience of time is characterized by strong distortions from objective "clock time" (Eagleman, 2008). 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, and none of the predominant explanations of human time perception provide adequate accounts for this phenomenon.

The two most well-established explanations of human time perception are the internal clock model (Gu, van Rijn, & Meck, 2015; Matell & Meck, 2004; Treisman, 1963) and the population clock approach (Hardy & Buonomano, 2018; Karmarkar & Buonomano, 2007). The internal clock approach proposes 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" (Terhune, Sullivan, & Simola, 2016; Wearden, 2008). 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. This is because the relationship between content and time can be complicated by other factors, such as attention to time (Block, Hancock, & Zakay, 2010; Fountas et al., 2020). 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 again, ad-hoc accounts that relate specific sensory content with trajectory length are available. The problem with both of these approaches is that they are rooted in the idea that the brain attempts to track clock time. As a result, neither approach is able to intrinsically account for content-driven distortions in subjective duration. Furthermore, while tracking clock time is likely to be useful for short, repetitive motor commands such as drumming (Buonomano & Laje, 2010; Gámez, Mendoza, Prado, Betancourt, & Merchant, 2019; Goudar & Buonomano, 2018; Merchant, Pérez, Zarco, & Gámez, 2013), tracking clock time is not necessarily useful for longer intervals. Indeed, as established, subjective experience of longer intervals is characterized more by deviation from rather than adherence to clock time.

Both animal and human electrophysiology and neuroimaging studies have searched for potential neural regions or putative mechanisms underlying time perception (Allman, Teki, Griffiths, & Meck, 2014; Wiener, Turkeltaub, & Coslett, 2010). This literature finds potential mechanisms in a diverse set of regions including pre-/supplementary motor areas (Gámez et al., 2019; Merchant et al., 2013; Protopapa et al., 2019), parietal areas (Bueti & Walsh, 2009; Hayashi et al., 2015), insula (Wittmann, Simmons, Aron, & Paulus, 2010), midbrain dopaminergic neurons (Soares, Atallah, & Paton, 2016), and hippocampus and entorhinal cortex (Ezzyat & Davachi, 2014; Montchal, Reagh, & Yassa, 2019; Tsao et al., 2018; Umbach et al., 2020). The diversity of neural regions seemingly associated with human time perception supports the intuition underlying the population clock approach - that time is reflected in the dynamics of a neural system because those dynamics evolve 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 provide evidence for an alternative approach to human time perception. In contrast to traditional approaches, our proposal is that experienced duration is constructed from the neural and computational processes underlying perception. Therefore, whatever the processes through which the brain accomplishes perceptual classification, our model of subjective duration estimates can be constructed from the outputs of those processes (Ma & Peters, 2020). By estimating time from the processes underlying perceptual classification, we explicitly incorporate the content of experience into our model of subjective time.

To test our proposal we used model-based analysis of human functional neuroimaging, supplemented with analysis on an artificial perceptual classification network. To anticipate results, we first show that constructing duration judgements from the dynamics of an artificial neural classification network while it was input naturalistic video stimuli reproduces human-like biases. We then recorded human neuroimaging data while participants viewed the same videos, and show that constructing duration judgements from the dynamics of visual cortex during perceptual classification also produced participants' biases. Together, these results show that participants' subjective duration judgements can be constructed from neural activity associated with the perceptual classification of their dynamic sensory environment.

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 containing more salient events (city scenes) should be judged as lasting longer relative to videos with fewer (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 veridical ("clock") duration t and each duration report in that veridical duration condition x_t was defined as:

$$\mathrm{bias}_{\mathrm{x}} = \frac{\mathrm{x} - \overline{\mathrm{x}_{\mathrm{t}}}}{\overline{\mathrm{x}_{\mathrm{t}}}}$$

where $\overline{\mathrm{x}_{\mathrm{t}}}$ is the mean duration report for videos of length *t* seconds. Positive/negative values mean that individual duration reports 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 ($\bar{\rho} = 0.96$, Fig. 1A, left). As predicted , durations of city scenes were relatively over-estimated and office scenes under-estimated, $M_{diff} = 5.18 \pm 1.36$, 95% CI [1.81, 8.65], $t_{39} = 3.09$, $p = 0.004$, d = 0.50, $BF_{H(0,10.5)} = 33.8$, confirming that natural scenes containing a higher density of salient events do indeed feel longer (Fig. 1A, right). Note that this result shows that the *amount* of experienced time was lower for boring videos, not that time *passed faster* for boring videos.

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. The purpose here was to determine whether a neural network trained only for perceptual classification could be used to 'generate' human-like estimates of time for the same stimulus set presented to our human participants. If we can replicate human-like estimates, it would demonstrate that our simple, algorithmic approach to generating duration estimates from perceptual classification network activity is indeed effective.

We fed the same video clips that participants had viewed to a pre-trained hierarchical image classification network, AlexNet (Krizhevsky, Sutskever, & Hinton, 2017). For each network node, we computed frame-to-frame Euclidean distances in network activity. Then, separately for each network layer, each distance – or change in activation - was categorized as salient or not. This was achieved using an attention threshold with exponential decay that simply determined whether the change in node activation (the Euclidean distance) sufficient to be deemed salient (see Methods). Following ideas in episodic memory (Zacks, Kurby, Eisenberg, & Haroutunian, 2011) we call these salient

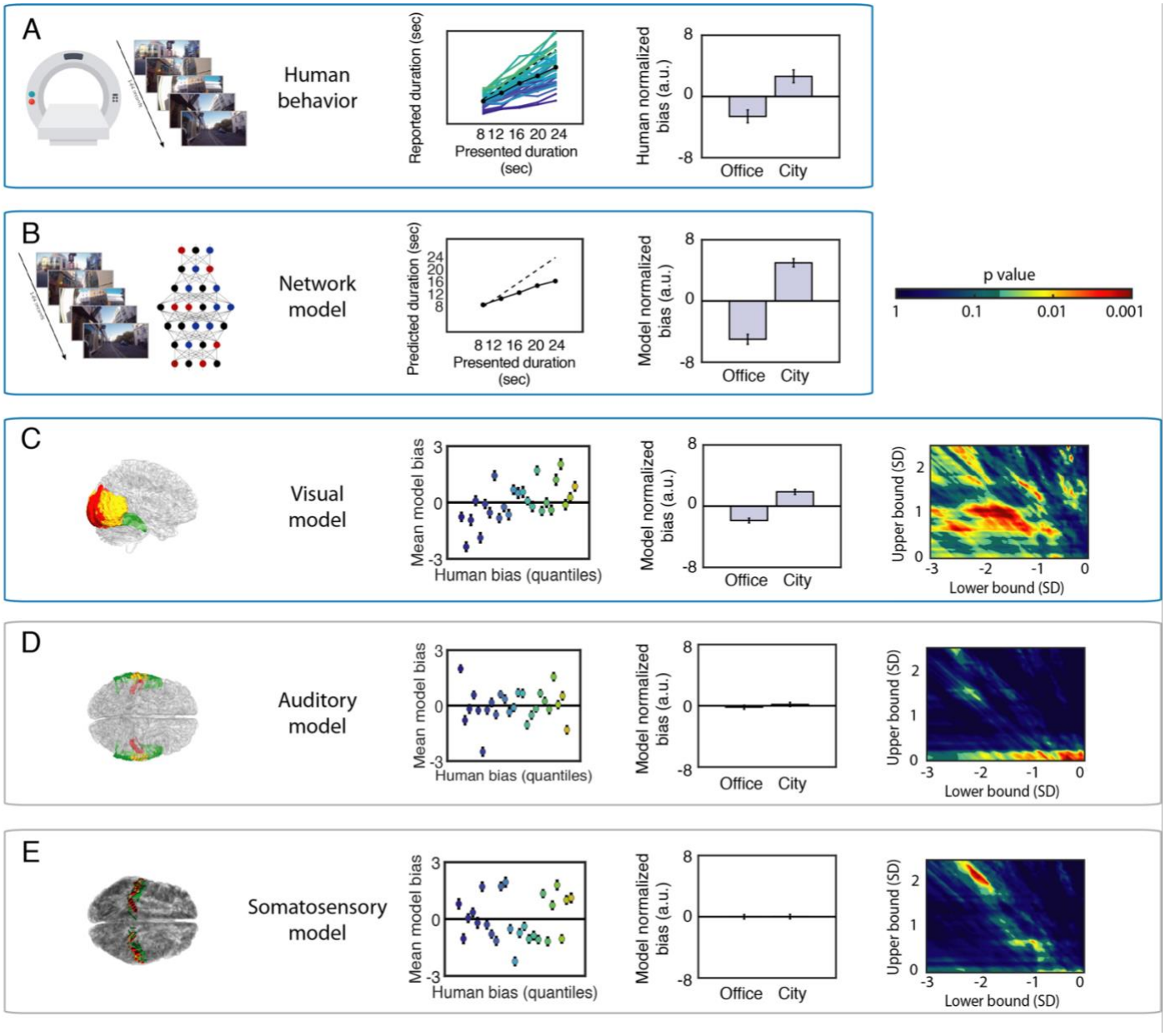

Figure 1. *Behavioral and modelling results.* **(A)** Behavior. Left: participant-wise relationship between report and duration (colored lines), mean relationship (solid black line), and the line of unity (dashed line). Right: relative under-/over-estimation of duration by human participants for office/city 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.

‘events’. Salient events (relatively large changes in activity) 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 veridical (not reported) durations (Fig 1B, left). This means that the model is attempting to reproduce “clock time” duration based on the input rather than the more trivial task of training the model to directly reproduce human estimates. Therefore, any human-like biases in estimates can be attributed to the behavior of the network in response to the input stimuli.

As for human behavior, model-produced duration reports were significantly correlated with the video duration $\rho(2329) = 0.73$, $p < 0.001$, indicating that we could effectively reproduce clock time from the dynamics of the classifier. But much more importantly, like our human participants, the model-produced differed as a function of video type: estimation bias was greater (i.e. reports relatively over-estimated) for busy city scenes than the office scenes, $M_{office} = -5.00 \pm 0.66$, $M_{city} = 4.99 \pm 0.55$, 95%CI = [8.31, 11.67], $t_{2329} = 11.65$, $p < 0.001$, $d = 0.48$ (Fig 1B, right). These results demonstrate that simply tracking the dynamics of a network trained for perceptual classification while it is exposed to natural scenes can produce human-like estimations (and distortions) of duration. Of course, we do not claim that AlexNet has subjective experiences of duration. Rather, we show that the node activity underlying perceptual classification processes contains the information relevant for constructing human-like *subjective* duration judgements.

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 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 1. This analysis provides a very coarse level insight into the differences in neural dynamics related to stimulation. However, it does not tell us anything about the suitability of our algorithmic approach to estimating subjective duration from human brain activity.

Reconstructing human-like duration reports from visual cortex BOLD.

For our key analysis we tested whether we could reproduce participants’ 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. If we can do this, then human subjective duration judgements may be constructed from brain activity associated with sensory processing (however that processing might be achieved).

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 S1). 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).

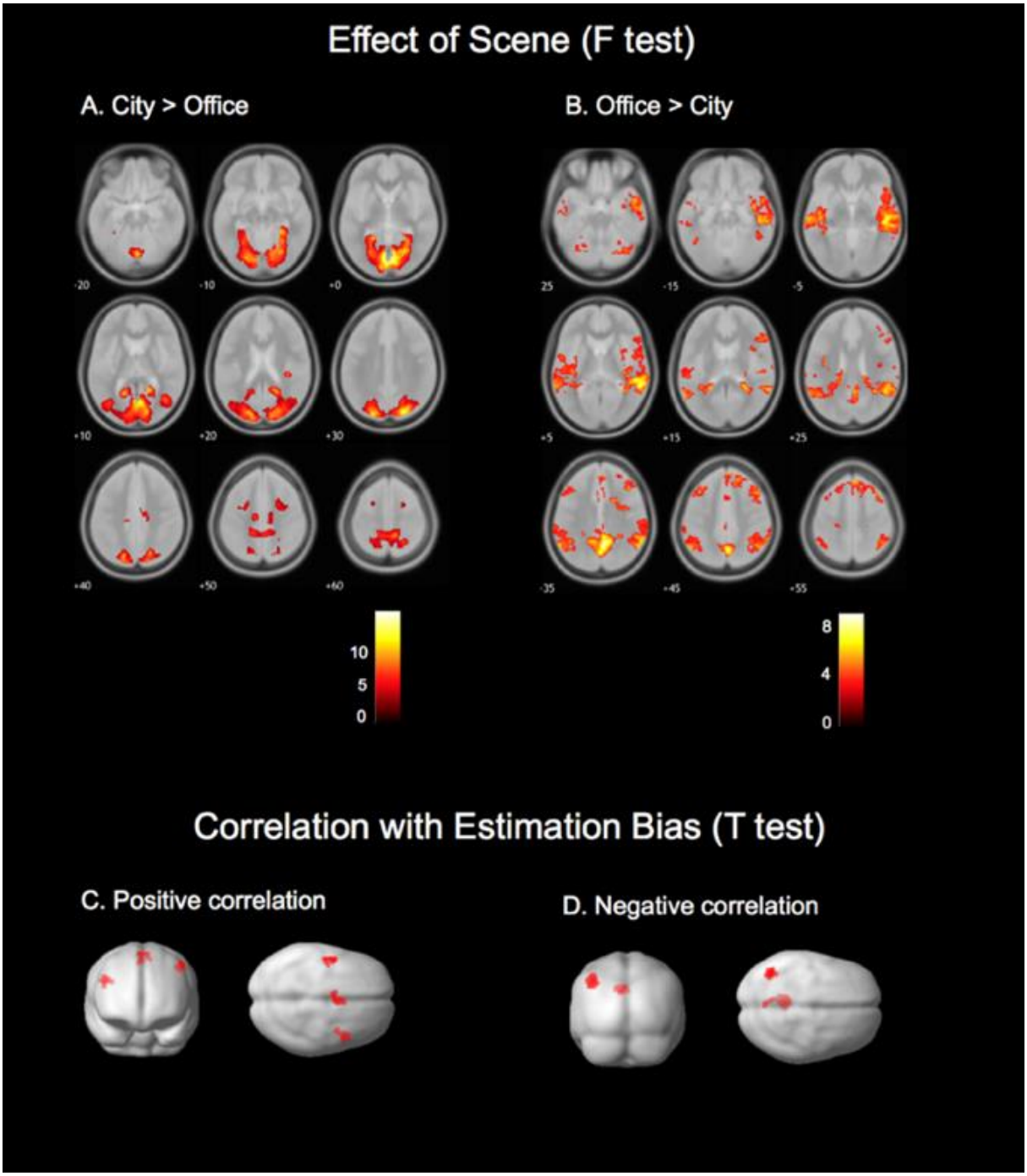

Figure 2. *Results from confirmatory GLM on BOLD (significant clusters only).* **A** Higher BOLD for city than office scenes: R lingual gyrus; bilateral midcingluate area; R insula; bilateral SFG. **B** Higher BOLD for office than city scenes: R precuneus; bilateral precentral gyrus; L MFG; bilateral 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.

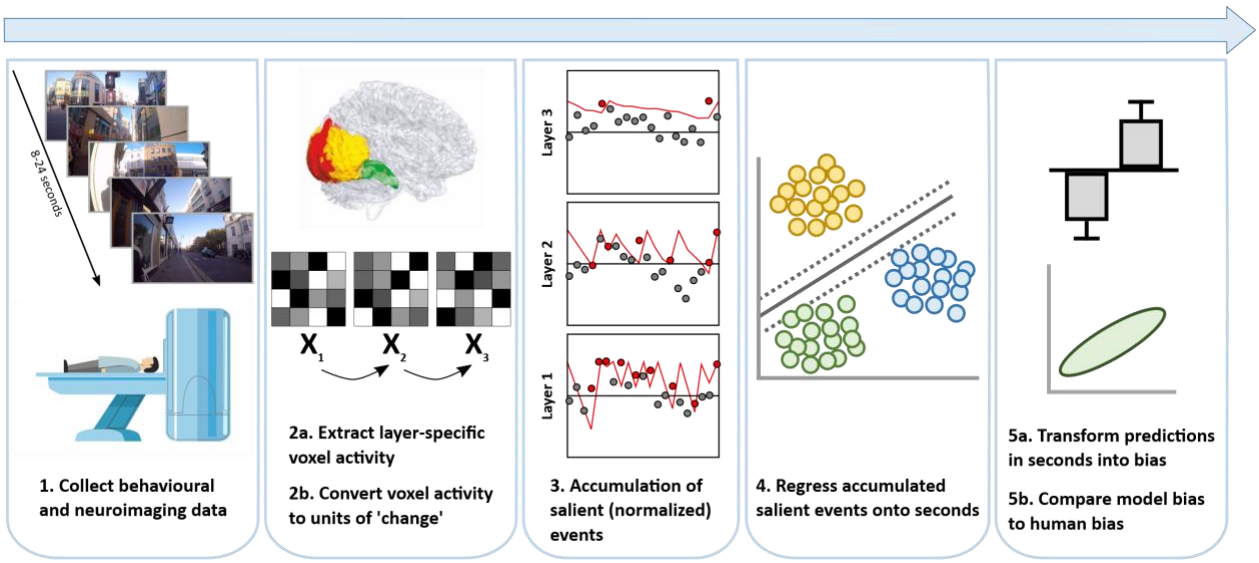

Figure 3. *Schematic of modelling analysis pipeline.* Following data collection, voxel-wise data was extracted and TR-by-TR changes (Euclidean distance or signed difference) computed. Total change in the ROI over TR was compared to an attention threshold that categorized events as salient or not. Accumulated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

We ran this key analysis in two ways: one was confirmatory (i.e. pre-registered) and one was exploratory (i.e. not pre-registered). The analysis 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 perceptual changes (Egner, Monti, & Summerfield, 2010), potentially in the form of "prediction errors". Therefore, while the model using Euclidean distance as the metric of change assumes that BOLD relates *directly* to neural activity (conceptually the same as "activation" of nodes in the artificial classification network), signed difference is more closely aligned with the idea that BOLD (in early sensory networks in this case) indicates (computational) *prediction error*.

We then used support vector regression to predict veridical video durations from accumulated salient events in layers 1, 2 and 3. This converted the accumulated salient

events in each layer to model duration "reports" in seconds so that they could be compared with human reports that were made in seconds.

Finally, bias in model predictions were compared to participants' estimation biases. 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). In both cases, model predictions were generated from the pooled accumulated changes. Participant data was standardized and pooled because the use of long stimulus presentation intervals (up to 24 seconds) meant that for each participant we could only obtain relatively few trials, insufficient to complete the analyses on a purely participant-by-participant basis.

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". The reproduction of clock time is trivial because, all else being equal, longer intervals will have more salient events. This is why our key analyses focus on reproducing the *subjective* biases present in the reports of human participants.

Our key hypothesis was that the visual cortex model alone would be able to reproduce participants' duration *biases*. Supporting this, only the pre-registered model trained on visual salient events significantly reproduced the super-subject's estimation biases trial-by-trial, $\beta_{2328} = 1.51$, $p = 0.015$; 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). Not only was the visual cortex regression coefficient a significant *predictor* of behavioral report, the visual cortex regression model was also a *better fit* to the trial-by-trial behavioral biases than the auditory or somatosensory cortex models (Fig. S3). These results mean that biases in subjective estimates of time can be predicted from neural activity associated with modality-specific perceptual processing. The processing is modality-specific because the video stimuli were silent, with no auditory or tactile stimulation.

While the visual model could reproduce participants' biases trial-by-trial, it did not reproduce the effect of video type (overestimation of duration for city scenes) despite a numerical trend in the predicted direction, $M_{diff} = 0.19 \pm 13.96$, 95%CI = [-0.94, 1.33], $t_{2329} = 0.33$, $p = 0.739$, d = 0.01. The control models could not reproduce the effect of video type either (auditory: $M_{diff} = -0.33 \pm 12.29$, , 95%CI = [-1.32, 0.67], $t_{2329} = -0.64$, $p = 0.522$, d = -0.03, somatosensory: $M_{diff} = 0.16 \pm 13.09$, , 95%CI = [-1.23, 0.90], $t_{2329} = -0.30$, $p = 0.762$, d = -0.01, 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.

Next, we analyzed the biases predicted from the exploratory model, in which salient events were determined from signed differences in voxel activity. 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, under this definition of salient event, 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' trial-by-trial 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. The association held under a wide range of reasonable attention threshold parameters (Fig. 1C, right), meaning that model performance in reproducing participant duration reports was robust to how salient events were categorized. 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 and right), 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 on a trial-by-trial basis from salient events in sensory cortex activity, where salient events are defined as relatively extreme changes in the neural responses to stimulation. 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 control models based on somatosensory or auditory cortex BOLD. While we could of course reconstruct *veridical* clock time from all three sensory regions (because those regions exhibited dynamic neural activity, regardless of the relation to ongoing visual stimulation), only the information extracted by the stimulus-relevant sensory model - our visual model - was related to subjective duration. These results reflect our initial proposal that the divergence between objective and subjective duration characterizes 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 (Fountas et al., 2020; Suárez-Pinilla et. al., 2019), support 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 (perceptual classification processes) are the same processes that are used to build the estimate of time. In other words, we show that the units upon which subjective duration estimates are constructed arise from the properties of sensory processing systems. 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. Of course, outside of the laboratory we judge time in multisensory contexts, and so we can still estimate the duration of some physical events when our eyes are closed or even if clinically deaf. Furthermore, because 'salient events' are events in the dynamics of the perceptual classification process rather than events in the physical world, visual imagery would also contribute to our experience of time, particularly in the absence of external stimulation. Our conclusion is in support of the idea that subjective time depends on distributed mechanisms (Paton & Buonomano, 2018), but that in each case they are naturally linked to sensory content by virtue of being determined by those content-related processes.

Our data do not speak to the question of *how* perceptual classification is achieved by the brain, and indeed, our results are largely invariant to this. Whether the artificial

classification network used is closely matched to biological vision in how it processes information is irrelevant; the algorithmic approach to estimating duration from network activity (artificial or human) produces outcomes consistent with the patterns seen in human subjective reports of time. The crucial point here is that, assuming we have a hierarchical, specialized system for perceptual classification - the common interpretation of primate ventral visual stream (Konen & Kastner, 2008; Ungerleider & Haxby, 1994; Van Essen & Maunsell, 1983) - whatever the computational processes underlying perception for the human brain are, the dynamics of perceptual systems implementing those processes can be used to construct a subjective experience of duration. This conclusion is best demonstrated by the fact that our model produced estimates consistent with biases in human reports regardless of whether applied to activation patterns of an artificial, feed-forward image classification network, or BOLD patterns in human ventral vision.

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). While the former assumes that BOLD activity indexes some raw quantity associated with sensory inputs, the latter assumes that BOLD already indexes change in sensory input, for example as perceptual prediction error. 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 (but not evidence for) the view that BOLD already indexes detected environmental changes, in line with literature evidencing “surprise” or “prediction error” responses in sensory (Egner et al., 2010; Langner et al., 2011; Todorovic, van Ede, Maris, & de Lange, 2011) and even frontal (Meyniel & Dehaene, 2017; Sherman, Seth, & Kanai, 2016) cortices.

The idea that our model may be based on an index of perceptual “surprise” is intriguing as it provides a natural link to the closely related topic of episodic memory (see also Fountas et al. 2020). In the episodic memory literature, prediction error, i.e. the difference between current sensory stimulation and expected stimulation, has been proposed as the basis for event boundaries (Fountas et al., 2020; Gershman, Radulescu, Norman, & Niv, 2014; Zacks et al., 2011) – transitions that segment some content (e.g. a cow) from some other content (e.g. a car) in continuous experience (Radvansky & Zacks, 2017; Zacks, Speer, Swallow, Braver, & Reynolds, 2007). 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” (MacDonald, Lepage, Eden, & Eichenbaum, 2011; Montchal et al., 2019; Umbach et al., 2020), linking the content of basic sensory processing with temporal properties of episodic memory within the powerful predictive coding approach (Fountas et al., 2020; Friston, 2010; Rao & Ballard, 1999).

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 (Harvey, Dumoulin, Fracasso, & Paul, 2019; Hayashi, van der Zwaag, Bueti, & Kanai, 2018; Soares et al., 2016; Tsao et al., 2018) 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. Sample size was determined according to funding availability.

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. For one participant and one block, only 11/20 trials were completed giving a total of 2331 trials across the entire dataset.

MRI acquisition and pre-processing (confirmatory).

Functional T2* sensitive multi-band echoplanar images (EPIs) were acquired on a Siemens PRISMA 3T scanner (2mm slices with 2mm gaps, TR = 800ms, multiband factor = 8, TE = 37ms, Flip angle = 52°). To minimize signal dropout from parietal, motor and occipital cortices, axial slices were tilted. Full brain T1-weighted structural scans were acquired on the same scanner using the MPRAGE protocol and consisting of 176 1mm thick sagittal slices (TR = 2730ms, TE = 3.57ms, FOV = 224mm x 256mm, Flip angle = 52°). 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. A standard image pre-processing pipeline was used: 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 (Glasser et al., 2013). After co-registration, EPIs were spatially normalized to MNI space using parameters obtained from the segmentation of T1 images into grey and white matter, then smoothed with a 4mm FWHM Gaussian smoothing kernel. Smoothed data were used for the GLM on BOLD only; 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 (Law et al., 2005; Sherman et al., 2016). 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 (Lancaster et al., 2000; Maldjian, Laurienti, Kraft, & Burdette, 2003).

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 (Wang, Mruczek, Arcaro, & Kastner, 2015).

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:

$$\Delta_{\mathrm{TR}} = \sum_{\mathrm{v}} | \mathrm{X}_{\mathrm{TR}} - \mathrm{X}_{\mathrm{TR}-1} |$$

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 (Roseboom et al., 2019), 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:

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

Only the parameter a took different values in each layer (see Table S2). 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 (SVR, implemented on python 3.0 using sklearn (Pedregosa et al., 2011)) 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 β_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 (Lancaster et al., 2000; Maldjian et al., 2003). 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:

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

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:

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

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.

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

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

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 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:

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

Heat maps depicted in Fig. 1 correspond to one-tailed p-values for β_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 (Roseboom et al., 2019). 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)(Krizhevsky et al., 2017). 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 S3 for criterion values). Finally, accumulated salient events were mapped onto units of seconds using multiple linear regression.

References

Allman, M. J., Teki, S., Griffiths, T. D., & Meck, W. H. (2014). Properties of the Internal Clock: First- and Second-Order Principles of Subjective Time. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-010213-115117

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

Bueti, D., & Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2009.0028

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

Eagleman, D. M. (2008). Human time perception and its illusions. *Current Opinion in Neurobiology*, *18*(2), 131–136. https://doi.org/10.1016/j.conb.2008.06.002

Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *The Journal of Neuroscience*, *30*(49), 16601–16608. https://doi.org/10.1523/JNEUROSCI.2770-10.2010

Ezzyat, Y., & Davachi, L. (2014). Similarity breeds proximity: Pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity. *Neuron*. https://doi.org/10.1016/j.neuron.2014.01.042

Fountas, Z., Sylaidi, A., Nikiforou, K., Seth, A. K., Shanahan, M., & Roseboom, W. (2020). A predictive processing model of episodic memory and time perception. *BioRxiv*, 2020.02.17.953133. https://doi.org/10.1101/2020.02.17.953133

Friston, K. J. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Gámez, J., Mendoza, G., Prado, L., Betancourt, A., & Merchant, H. (2019). The amplitude in periodic neural state trajectories underlies the tempo of rhythmic tapping. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.3000054

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

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., … Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2013.04.127

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

Gu, B. M., van Rijn, H., & Meck, W. H. (2015, January 1). Oscillatory multiplexing of neural population codes for interval timing and working memory. *Neuroscience and Biobehavioral Reviews*. Elsevier Ltd. https://doi.org/10.1016/j.neubiorev.2014.10.008

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

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

Hayashi, M. J., Ditye, T., Harada, T., Hashiguchi, M., Sadato, N., Carlson, S., … Kanai, R. (2015). Time Adaptation Shows Duration Selectivity in the Human Parietal Cortex. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.1002262

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

Karmarkar, U. R., & Buonomano, D. V. (2007). Timing in the Absence of Clocks: Encoding Time in Neural Network States. *Neuron*. https://doi.org/10.1016/j.neuron.2007.01.006

Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature Neuroscience*. https://doi.org/10.1038/nn2036

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*. https://doi.org/10.1145/3065386

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., … Fox, P. T. (2000). Automated Talairach Atlas labels for functional brain mapping. *Human Brain Mapping*. https://doi.org/10.1002/1097-0193(200007)10:3<120::AID-HBM30>3.0.CO;2-8

Langner, R., Kellermann, T., Boers, F., Sturm, W., Willmes, K., & Eickhoff, S. B. (2011). Modality-specific perceptual expectations selectively modulate baseline activity in auditory, somatosensory, and visual cortices. *Cerebral Cortex*, *21*(December), 2850–2862. https://doi.org/10.1093/cercor/bhr083

Law, J. R., Flanery, M. a, Wirth, S., Yanike, M., Smith, A. C., Frank, L. M., … Stark, C. E. L. (2005). Functional magnetic resonance imaging activity during the gradual acquisition and expression of paired-associate memory. *The Journal of Neuroscience*, *25*(24), 5720–5729. https://doi.org/10.1523/JNEUROSCI.4935-04.2005

Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: towards using deep nets as models for human behavior. *ArXiv*.

MacDonald, C. J., Lepage, K. Q., Eden, U. T., & Eichenbaum, H. (2011). Hippocampal “time cells” bridge the gap in memory for discontiguous events. *Neuron*, *71*(4), 737–749. https://doi.org/10.1016/j.neuron.2011.07.012

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

Suárez-Pinilla, M., Nikiforou, K., Fountas, Z., Seth, A.K. & Roseboom, W. (2019). Perceptual content, not physiological signals, determines perceived duration when viewing dynamic, natural scenes. *Collabra Psychology*, *5*(1).

Matell, M. S., & Meck, W. H. (2004, October). Cortico-striatal circuits and interval timing: Coincidence detection of oscillatory processes. *Cognitive Brain Research*. https://doi.org/10.1016/j.cogbrainres.2004.06.012

Merchant, H., Pérez, O., Zarco, W., & Gámez, J. (2013). Interval tuning in the primate medial premotor cortex as a general timing mechanism. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.5513-12.2013

Meyniel, F., & Dehaene, S. (2017). Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1615773114

Montchal, M. E., Reagh, Z. M., & Yassa, M. A. (2019). Precise temporal memories are supported by the lateral entorhinal cortex in humans. *Nature Neuroscience*, *22*(2), 284–288. https://doi.org/10.1038/s41593-018-0303-1

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Protopapa, F., Hayashi, M. J., Kulashekhar, S., Van Der Zwaag, W., Battistella, G., Murray, M. M., … Bueti, D. (2019). Chronotopic maps in human supplementary motor area. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.3000026

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*. https://doi.org/10.1016/j.cobeha.2017.08.006

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10195184

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

Sherman, M. T., Seth, A. K., & Kanai, R. (2016). Predictions shape confidence in right inferior frontal gyrus. *Journal of Neuroscience*, *36*(40). https://doi.org/10.1523/JNEUROSCI.1092-16.2016

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

Terhune, D. B., Sullivan, J. G., & Simola, J. M. (2016). Time dilates after spontaneous blinking. *Current Biology*. https://doi.org/10.1016/j.cub.2016.04.010

Todorovic, A., van Ede, F., Maris, E., & de Lange, F. P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(25), 9118–9123. https://doi.org/10.1523/JNEUROSCI.1425-11.2011

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

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

Umbach, G., Kantak, P., Jacobs, J., Kahana, M. J., Pfeiffer, B. E., Sperling, M., & Lega, B. (2020). Time cells in the human hippocampus and entorhinal cortex support episodic memory. *BioRxiv*, 2020.02.03.932749. https://doi.org/10.1101/2020.02.03.932749

Ungerleider, L. G., & Haxby, J. V. (1994). “What” and “where” in the human brain. *Current Opinion in Neurobiology*. https://doi.org/10.1016/0959-4388(94)90066-3

Van Essen, D. C., & Maunsell, J. H. R. (1983). Hierarchical organization and functional streams in the visual cortex. *Trends in Neurosciences*. https://doi.org/10.1016/0166-2236(83)90167-4

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

Wearden, J. H. (2008). Slowing down an Internal Clock: Implications for Accounts of Performance on four Timing Tasks. *Quarterly Journal of Experimental Psychology*, *61*(2), 263–274. https://doi.org/10.1080/17470210601154610

Wiener, M., Turkeltaub, P., & Coslett, H. B. (2010). The image of time: A voxel-wise meta-analysis. *NeuroImage*. https://doi.org/10.1016/j.neuroimage.2009.09.064

Wittmann, M., Simmons, A. N., Aron, J. L., & Paulus, M. P. (2010). Accumulation of neural activity in the posterior insula encodes the passage of time. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2010.06.023

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

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

Acknowledgments

General: 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.

Funding: 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.

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.

Competing interests: The authors declare no competing interests.

Data and materials availability: The pre-registration document, along with all data and analysis code are freely available to download at osf.io/2zqfu.

Figure and Tables

Figures 1-3
Table 1

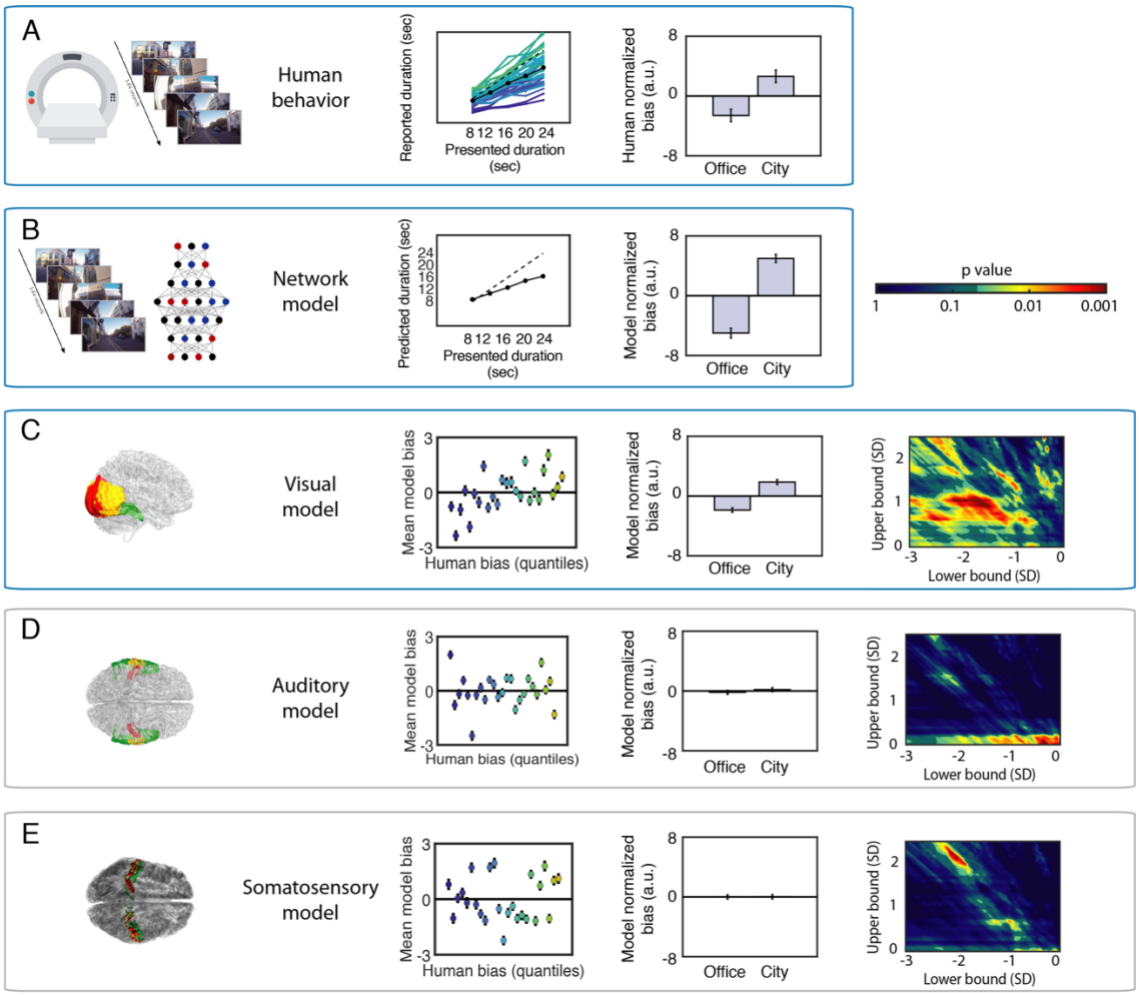

Figure 1. *Behavioral and modelling results.* **(A)** Behavioral results. Top: participant-wise relationship between report and duration (colored lines), mean relationship (solid black line), and the line of unity (dashed line). Bottom: relative under-/over-estimation of duration by human participants for office/city videos. **(B)** Artificial network model results. Top: relationship between model-predicted and veridical durations for this model, trained on accumulated salient events in video frames. Bottom: relative under-/over- estimation of duration for office/city scenes for this model. (**C-E, top)** 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)** Normalized bias as a function of video type. Only the visual cortex model exhibited a significant difference. **(C-E, bottom)** 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.

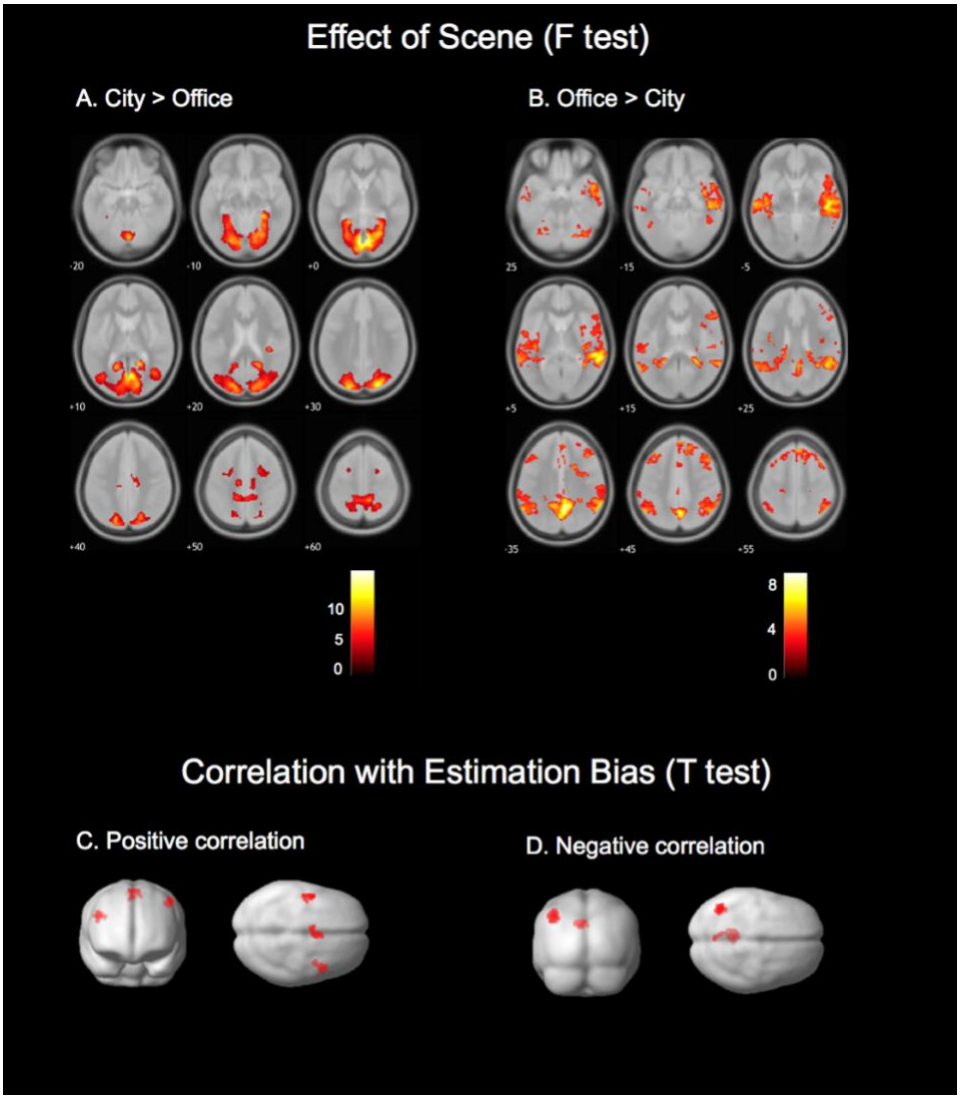

Figure 2. *Results from confirmatory GLM on BOLD (significant clusters only).* **A** Higher BOLD for city than office scenes: R lingual gyrus; bilateral midcingluate area; R insula; bilateral SFG. **B** Higher BOLD for office than city scenes: R precuneus; bilateral precentral gyrus; L MFG; bilateral 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.

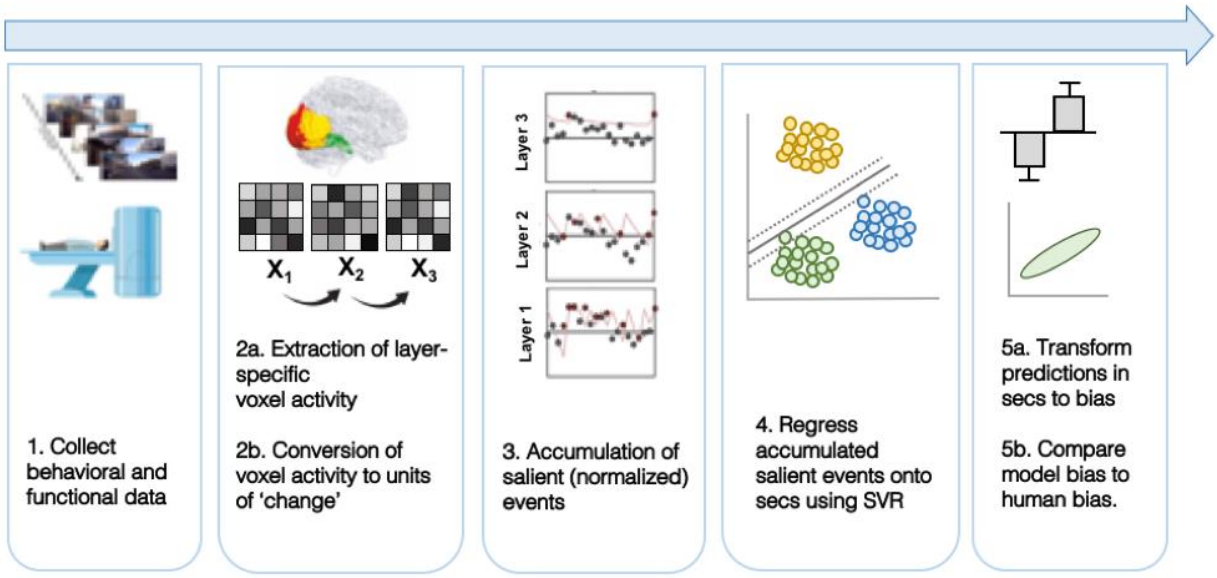

Figure 3. *Schematic of modelling analysis pipeline.* Following data collection, voxel-wise data was extracted and TR-by-TR changes (Euclidean distance or signed difference) computed. Total change in the ROI over TR was compared to an attention threshold that categorized events as salient or not. Accumulated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

Table 1. 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

Supplementary Materials

Figures S1 to S4
Tables S1 to S3

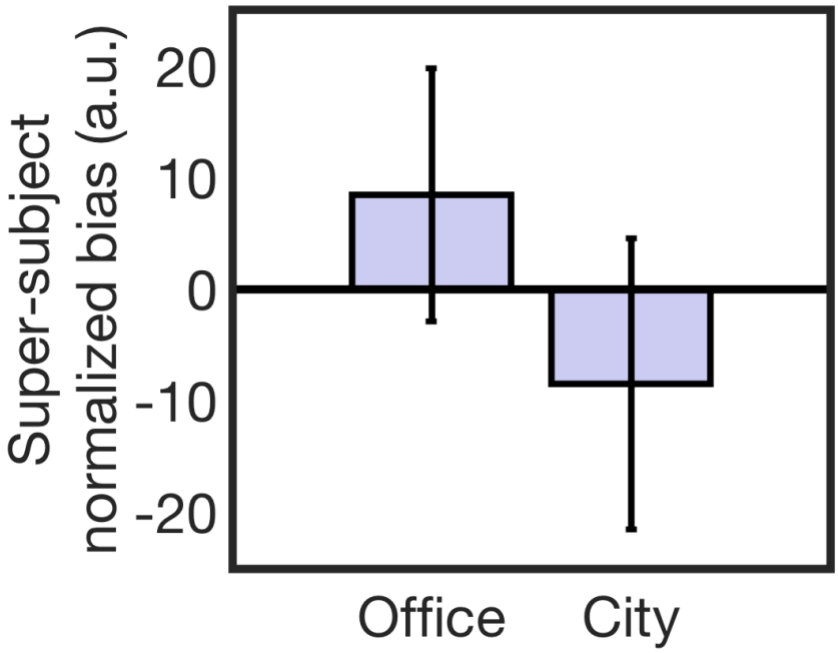

Figure S1. Normalized estimation bias computed on pooled ('super-subject') behavioral data, as a function of video scene.

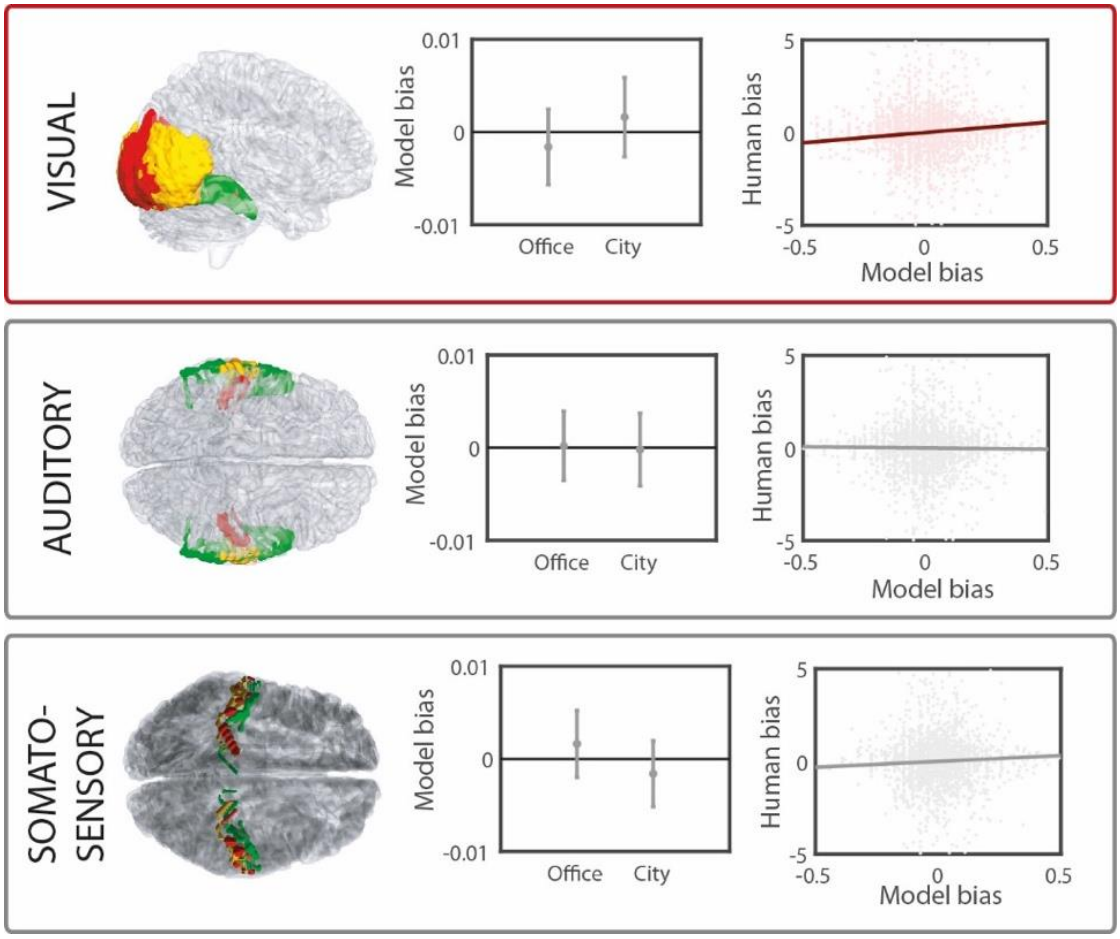

Figure S2. Normalized bias predicted by models trained on salient events (Euclidean distance) in visual, auditory and somatosensory hierarchies. **(Left)** Red, yellow and green clusters represent our hierarchical layers 1-3 respectively. **(Middle)** Differences in the models' normalized as a function 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.

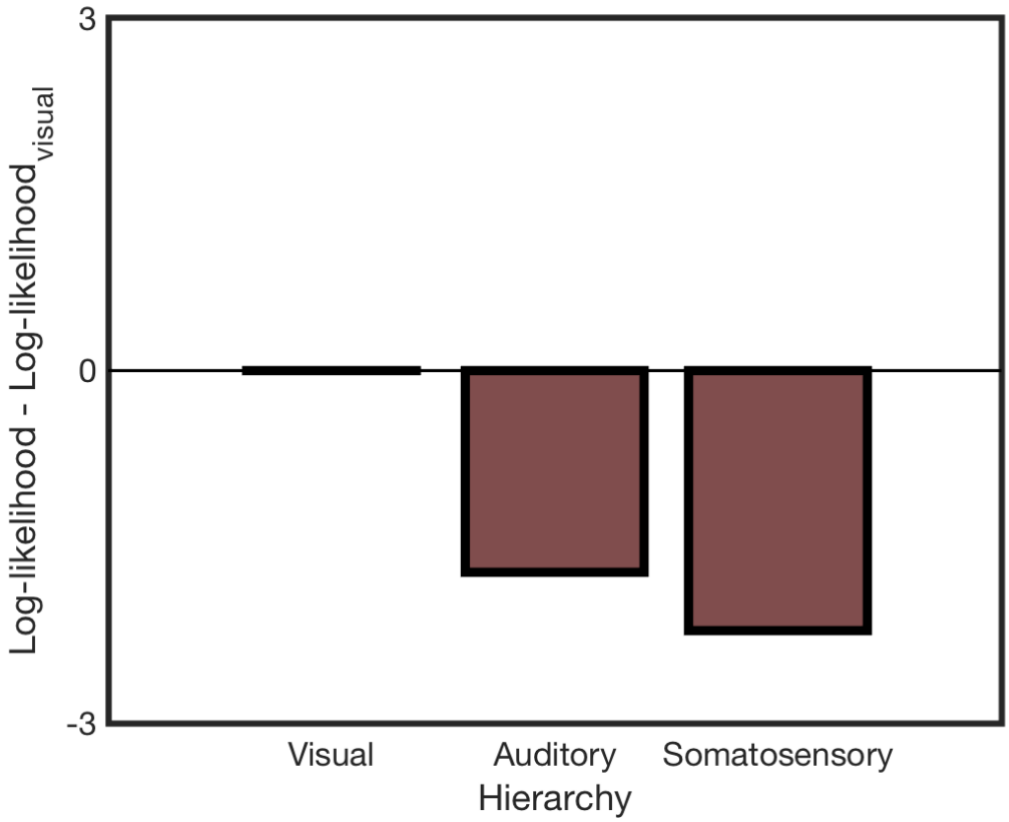

Figure S3. Model fits for the regression of predicted durations onto veridical durations (Euclidean Distance models). For each of the visual, auditory and somatosensory models, we regressed the model-predicted biases onto the human super-subject's biases. To compare the performance of the three regressions 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.

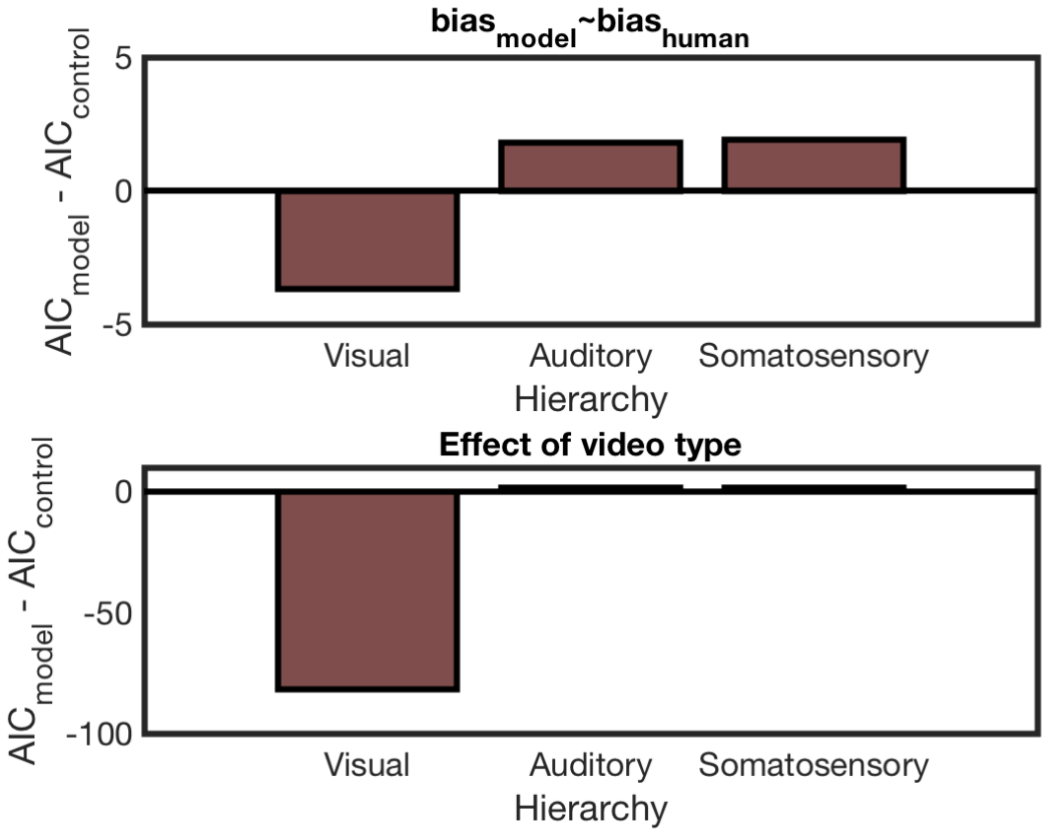

Figure S4. Model fits for the linear mixed models (Signed Difference analyses). **Top.** For each of the visual, auditory and somatosensory models, we constructed an LMM with human bias as the outcome and the model-predicted biases as a fixed effect. These LMMs tested the video-by-video correlations 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.

Table S1. 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 S2. Criterion parameters for each hierarchical layer of the sensory cortex models

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

Table S3. 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.