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]

[1] Sackler Centre for Consciousness Science, University of Sussex, UK [2] Department of Informatics and Engineering, 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

Abstract

Many contemporary models of time perception are based on the notion that our brain houses an internal “clock”, specialized for tracking duration. Here we show that specialized mechanisms are unnecessary, and that human-like duration judgements can be reconstructed from neural responses during sensory processing. Healthy human participants watched naturalistic, silent videos and rated their duration while fMRI was acquired. We constructed a computational model that predicts video durations from salient events in participants’ visual cortex activation. This model reproduced biases in participants’ subjective reports, whereas control models trained on auditory or somatosensory activity did not. Our data reveal that subjective time is inferred from information arising during the perception of our dynamic sensory environment, providing a computational basis for an end-to-end account of time perception.

Correspondence to: Maxine T. Sherman, m.sherman@sussex.ac.uk; Warrick Roseboom, wjroseboom@gmail.com

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*" and "*a watched pot never boils*", illustrating that complex contextual factors, such as exciting surroundings, strongly influence experiences of duration. How these factors are incorporated within the neural mechanisms underlying human time perception remains poorly understood. Most research on human time perception concentrates on evidencing a pacemaker-driven "internal clock" (*2*–*5*). "Internal clock" approaches posit the existence of an internal timekeeper, specialized for the perception of *objective* time. Here, regular physiological or neural processes produce rhythmic ticks like the hands of a clock which are counted by an accumulator, such that more ticks corresponds to more time (*6*). Within this approach, contextual biases leading to time "running fast" or "running slow" are simply driven by putative changes in pacemaker rate (*7*, *8*).

An alternative to "internal clock" approaches is that context-dependent distortions in time perception arise from the computations underlying perception of our dynamic sensory environment (*9*). Under this proposal, biases in duration emerge because (subjective) time is a function of accumulated salient environmental events, detected by perceptual classification networks. Therefore, temporal experience is not driven by regular pacemaker "ticks" unrelated to sensation, but rather by the non-temporal content of sensory experience. This proposal is reflected in ideas going back centuries (*10*–*12*).

Here, we tested whether human-like subjective duration judgements can be constructed from neural activity associated with perceptual classification of a dynamic sensory environment, examining this hypothesis in the dynamics of both an artificial classification network and human neuroimaging. Using functional magnetic resonance imaging (fMRI) and a fully pre-registered preprocessing and model-based analysis pipeline (osf.io/2zqfu), 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, 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).

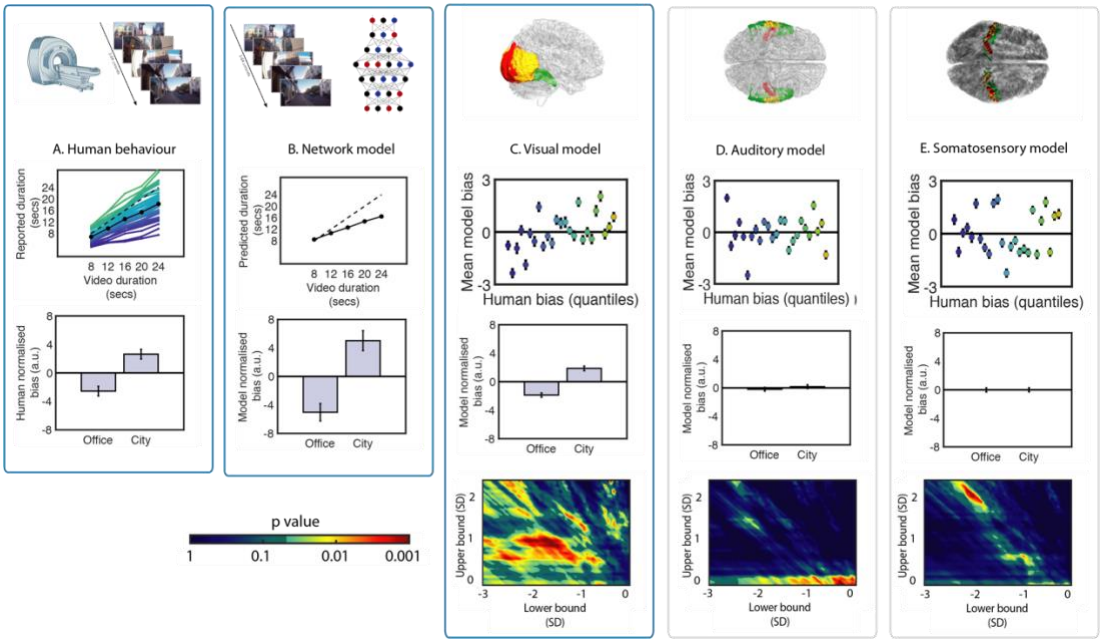

Figure 1. *Behavioral and modelling results.* **(A)** Top panel depicts participant-wise relationship between report and duration (colored lines) and the mean relationship (solid black line), relative to the line of unity (dashed line). Bottom panel shows relative under-/over-estimation of duration by human participants for office/city videos. **(B)** Top panel depicts relationship between model predicted and veridical durations when the model was trained on accumulated salient events in vide frames, detected by the classification neural network. Bottom panel shows 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 significance 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.

Participants could estimate time well, as indicated by a strong correlation between veridical and reported durations for each participant ($\bar{\rho} = 0.79 \pm 0.10$, Fig. 1A, top). As predicted, durations of city scenes were over-estimated and office scenes under-estimated, $t_{39} = 3.81$, $p < 0.001$, $M_{diff} = 5.18$ +/- 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, bottom).

Next we tested whether this effect of video type was reproduced by an artificial perceptual classification network. We fed each video clip participants viewed to a pre-trained image classification network (AlexNet (*13*)) and computed frame-to-frame Euclidean distances in network activity for each node. For each network layer,

distance was categorized as salient or not by an attention threshold with exponential decay (see supplementary methods). Salient events were accumulated at each layer and converted to estimates of duration in seconds via multiple linear regression by mapping them to veridical (not reported) durations (Fig 1B, top). As for human behavior, model-estimated and veridical durations were significantly correlated video-by-video, $\rho = 0.74$, $p < 0.001$, and model predictions exhibited a human-like effect of video type on estimation bias, $M_{diff} = 10.05 +/- 0.93$, $t(1018) = 5.39$, $p < 0.001$ (Fig 1B, bottom). These results demonstrate that simply tracking the dynamics of a perceptual classification network during exposure to natural scenes can produce human-like estimations (and distortions) of duration.

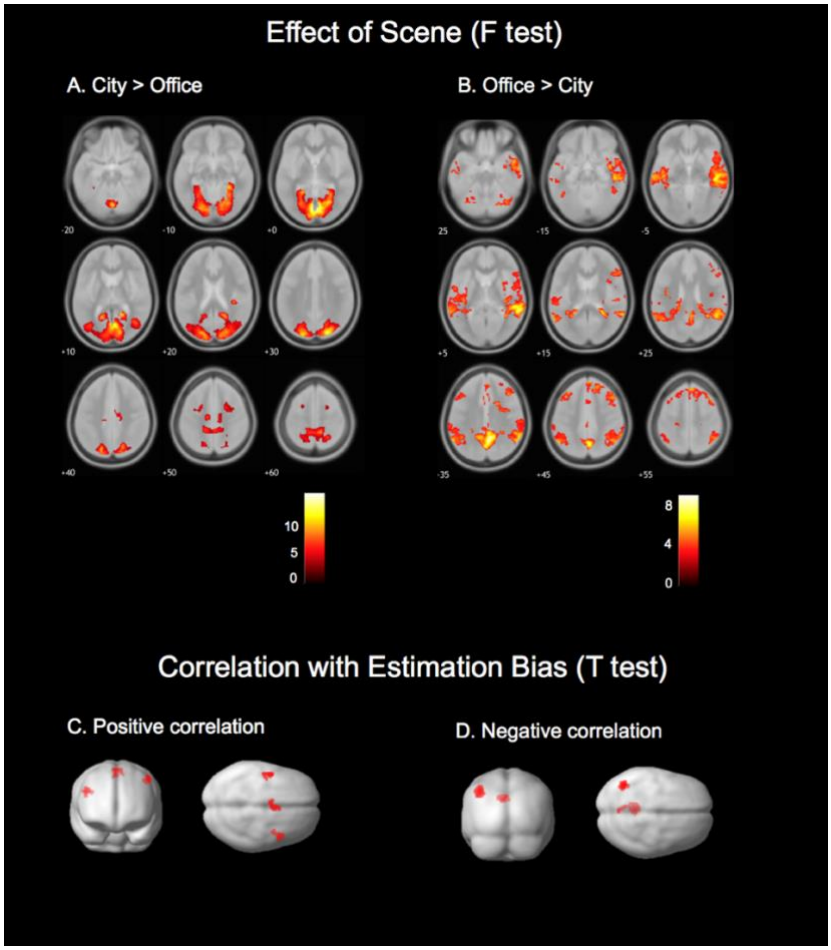

Figure 2. *Results from confirmatory GLM on BOLD (significant clusters only).* **A** Higher BOLD for city than office scenes: R lingual gyrus; bilateral midcingluate area; R insula; bilateral SFG. **B** Higher BOLD for office than city scenes: R precuneus; bilateral precentral gyrus; L MFG; bilateral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

Using human fMRI data, we then examined the neural correlates of stimulation; both how busy the scene was (office versus city) and estimation bias using a GLM on BOLD. As expected, busy city scenes drove bilateral occipital lobe BOLD as well as a set of frontal regions including right insula and bilateral superior frontal gyrus (Fig 2C). Estimation bias was positively correlated with visual activation in right superior occipital gyrus (Fig 2D), such that greater overestimation bias positively correlated with BOLD. Full GLM results are presented in Figure 2 and Table S1.

For our key analysis, we tested whether we could reproduce human-like duration biases from salient changes in BOLD activation, as we did for nodes in the artificial classification network (pipeline illustrated in Fig. 3). 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), such that lower layers reflect the detection and classification of low-level features (e.g. edge detection; primary visual cortex, V1), and higher layers, object-related processing (e.g. lateral occipital cortex; LOC). Analogous hierarchies were built for Auditory and Somatosensory cortex (see Table S2) for control analyses. Our prediction (see pre-registration at osf.io/2zqfu) was that only the model trained on the visual ROIs should predict human duration reports from accumulated salient events, because the stimuli were silent videos.

We ran this key analysis in two ways: one was confirmatory (i.e. pre-registered) and one was exploratory. In both cases, for each participant voxel-wise patterns of BOLD

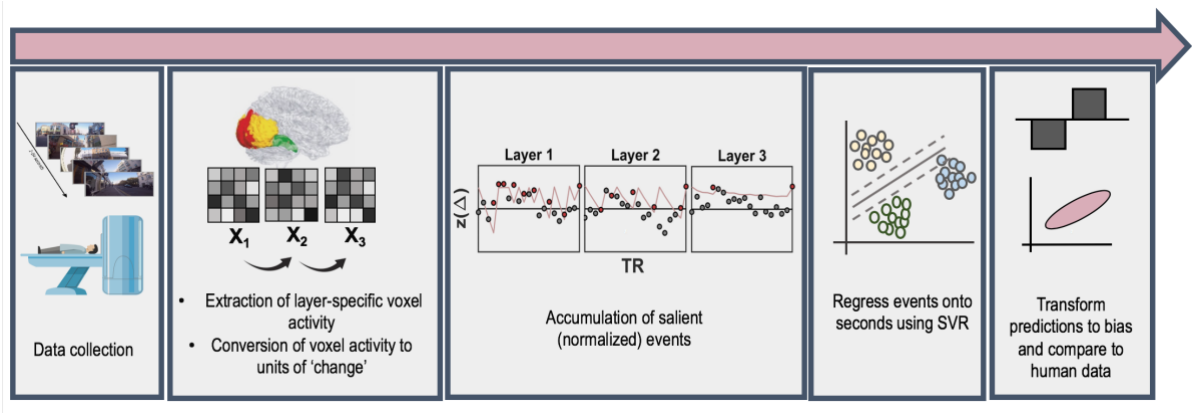

Figure 3. *Schematic of modelling analysis pipeline.* Following data collection, voxel-wise data was extracted and TR-by-TR changes (Euclidean distance or signed difference) computed. Total change in the ROI over TR was compared to an attention threshold that categorized events as salient or not. Accumulated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

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). For an additional exploratory analysis, change was TR_t - TR_{t-1} (i.e. signed difference) because, at least in sensory cortices, BOLD may already reflect salient changes in perception (*14*), potentially in the form of "prediction errors".

Separately for each of the three models, we mapped the number of salient events onto veridical video duration (not duration responses) using Epsilon-support vector regression (*15*). This delivered a model-predicted duration report for each trial, which we converted to normalized estimation bias for analysis.

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 estimates 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. S2).

Supporting our hypothesis, the pre-registered model trained on visual salient events significantly reproduced the super-subject's biases, $\beta_{2328} = 1.51$, $p = 0.015$, whereas 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. This means that biases in subjective estimates of time can only be predicted from neural activity associated with modality-specific perceptual processing. However, neither the visual cortex model nor the two control models exhibited a significant bias for overestimating busy city scenes, though the visual cortex model alone exhibited the correct direction of difference numerically (visual: $M_{diff} = 0.19$ +/- 13.96, $t_{2329} = 0.33$, $p = 0.739$, auditory:

M_{diff} = -0.33 +/- 12.29, t_{2329} = -0.64, p = 0.522, somatosensory: M_{diff} = 0.16 +/- 13.09, t_{2329} = -0.30, p = 0.762). Results are depicted in Fig. S1.

For our exploratory pipeline, for which salient events were determined from signed (not Euclidean) differences in voxel activity, linear mixed models revealed the visual model biases did strongly discriminate between office and city scenes, M_{diff} = 3.75 +/- 0.23, $\chi_2(1)$ = 85.06, p < 0.001 (Fig. 1C, bottom), and remained correlated with participants' biases, β = 0.02 +/- 0.008, $\chi_2(1)$ = 5.62, p = 0.018. This association is visualized in Fig. 1C (top) by plotting mean model bias as a function of 30 quantiles of human bias.

While (exploratory) models trained on accumulated visual cortex salient events reproduced human behavior, biases from exploratory models trained on auditory and somatosensory salient events neither discriminated video type (M_{diff} = 0.36 +/- 0.19, $\chi_2(1)$ = 0.43, p = 0.514, M_{diff} = 0.02 +/- 0.21, $\chi_2(1)$ = 0.46, p = 0.499 respectively) nor predicted trial-wise human normalized bias (β = -0.003 +/- 0.006, $\chi_2(1)$ = 0.20, p = 0.652, β = 0.002 +/- 0.007, $\chi_2(1)$ = 0.11, p = 0.740 respectively, Fig. 1D,E), underlining the specificity of visual cortex activity in predicting subjective time for silent videos.

To summarize, we could reconstruct our human participants' duration judgements from salient events detected by their visual cortex activity, but not from those detected by their auditory or somatosensory activity. Our results were robust under a wide range of model parameter values (Fig. 1C-E, bottom), and, in combination with the perceptual classification network model, support the notion that the basis of time perception is formed from the neural processes associated with processing the sensory context in which time is being judged (here, watching silent videos) rather than on the operation of a putative generic internal 'clock' (*16*, *17*).

Participants' duration judgements were best reconstructed from their BOLD activity when assuming that the relevant quantity for determining a salient event to be TR-by-TR difference, rather than Euclidean distance. This supports the view that BOLD already indexes detected environmental changes, in line with literature evidencing "surprise" or "prediction error" signals in sensory (*14*, *18*, *19*) and even frontal (*20*, *21*) cortices.

Using perceptual "surprise" – the difference between current sensory stimulation and expected stimulation based on an internal world model - as a common base for determining episodic memory event segmentation (*22*, *23*) and human time perception (*9*) unifies accounts of these aspects of human experience under a common process within the powerful predictive processing account of perception and cognition (*24–26*).

Furthermore, by considering salient events in our model as equivalent to event boundaries in episodic memory - transitions that segment some content (a cow) from some other content (a car) in continuous experience (*27*, *28*) – our model provides an account for how the internal structure of memory episodes (partonomic hierarchy; (*29*)) is generated across the hierarchy of temporal processing and representational complexity (*30*).

Other recent studies of time perception (*5*, *31–33*) 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 (pre-registered) model-based approach to describe how sensory information arriving in primary sensory areas can be transformed into units of subjective time. This 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 accumulation of salient 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.

Contributions

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

Acknowledgements

This work was supported by the European Union Future and Emerging Technologies grant (GA:641100) TIMESTORM – Mind and Time: Investigation of the Temporal Traits of Human-

Machine Convergence and the 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. Thanks to Charlotte Rae, Petar Raykov, Samira Bouyagoub, Chris Bird, and Mara Cercignani for their assistance with this project.

References

1. D. M. Eagleman, Human time perception and its illusions. *Curr. Opin. Neurobiol.* **18**, 131–6 (2008).

2. M. Treisman, N. Cook, P. L. N. Naish, J. K. MacCrone, The Internal Clock: Electroencephalographic Evidence for Oscillatory Processes Underlying Time Perception. *Q. J. Exp. Psychol. Sect. A*. **47**, 241–289 (1994).

3. M. S. Matell, W. H. Meck, Cortico-striatal circuits and interval timing: Coincidence detection of oscillatory processes. *Cogn. Brain Res.* **21** (2004), pp. 139–170.

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** (2015), pp. 160–185.

5. S. Soares, B. V. Atallah, J. J. Paton, Midbrain dopamine neurons control judgment of time. *Science (80-.).* **354**, 1273–1277 (2016).

6. M. Treisman, Temporal discrimination and the indifference interval. Implications for a model of the “internal clock”. *Psychol. Monogr.* **77**, 1–31 (1963).

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), , doi:10.1016/j.cub.2016.04.010.

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

10. R. Ornstein, *On the Experience of Time* (Penguin, Harmondsworth, UK, 1969).

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

12. L. A. Selby-Bigge, *A Treatise of Human Nature by David Hume, reprinted from the Original Edition in three volumes and edited, with an analytical index* (Clarendon Press, Oxford, 1896).

13. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Commun. ACM* (2017), doi:10.1145/3065386.

14. 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).

15. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* (2011).

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

17. 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).

18. 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).

19. R. Langner, T. Kellermann, F. Boers, W. Sturm, K. Willmes, S. B. Eickhoff, Modality-specific perceptual expectations selectively modulate baseline activity in auditory, somatosensory, and visual cortices. *Cereb. Cortex*. **21**, 2850–2862 (2011).

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

21. M. T. Sherman, A. K. Seth, R. Kanai, Predictions shape confidence in right inferior frontal gyrus. *J. Neurosci.* **36** (2016), doi:10.1523/JNEUROSCI.1092-16.2016.

22. 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), doi:10.1162/jocn_a_00078.

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

24. A. Clark, Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).

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

26. A. K. Seth, From Unconscious Inference to the Beholder's Share: Predictive Perception and Human Experience. *Eur. Rev.* (2019), doi:10.1017/S1062798719000061.

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

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

29. J. Zacks, Event perception and memory. *Annu. Rev. Psychol.* **71** (2019).

30. C. Baldassano, J. Chen, A. Zadbood, J. W. Pillow, U. Hasson, K. A. Norman, Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron* (2017), doi:10.1016/j.neuron.2017.06.041.

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

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

33. 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), doi:10.2139/ssrn.3438365.

34. M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, M. Jenkinson, The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* (2013), doi:10.1016/j.neuroimage.2013.04.127.

35. J. R. Law, M. a Flanery, S. Wirth, M. Yanike, A. C. Smith, L. M. Frank, W. a Suzuki, E. N. Brown, C. E. L. Stark, Functional magnetic resonance imaging activity during the gradual acquisition and expression of paired-associate memory. *J. Neurosci.* **25**, 5720–5729 (2005).

36. J. L. Lancaster, M. G. Woldorff, L. M. Parsons, M. Liotti, C. S. Freitas, L. Rainey, P. V. Kochunov, D. Nickerson, S. A. Mikiten, P. T. Fox, Automated Talairach Atlas labels for functional brain mapping. *Hum. Brain Mapp.* (2000), doi:10.1002/1097-0193(200007)10:3<120::AID-HBM30>3.0.CO;2-8.

37. 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), doi:10.1016/S1053-8119(03)00169-1.

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

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

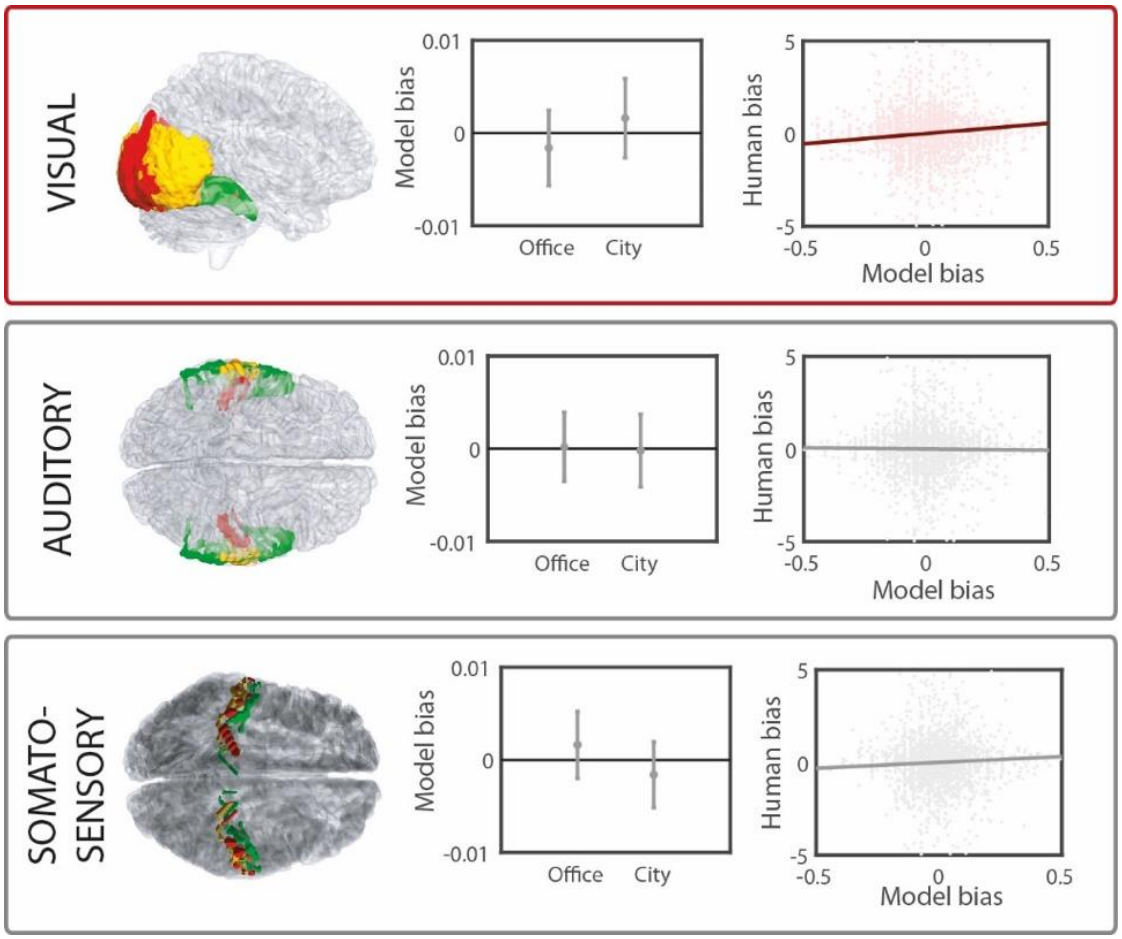

Figure S1. Normalized bias predicted by models trained on salient events (Euclidean distance) *in visual, auditory and somatosensory hierarchies.* **(Left)** Red, yellow and green clusters represent our hierarchical layers 1-3 respectively. **(Middle)** Differences in the models' normalized as a function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

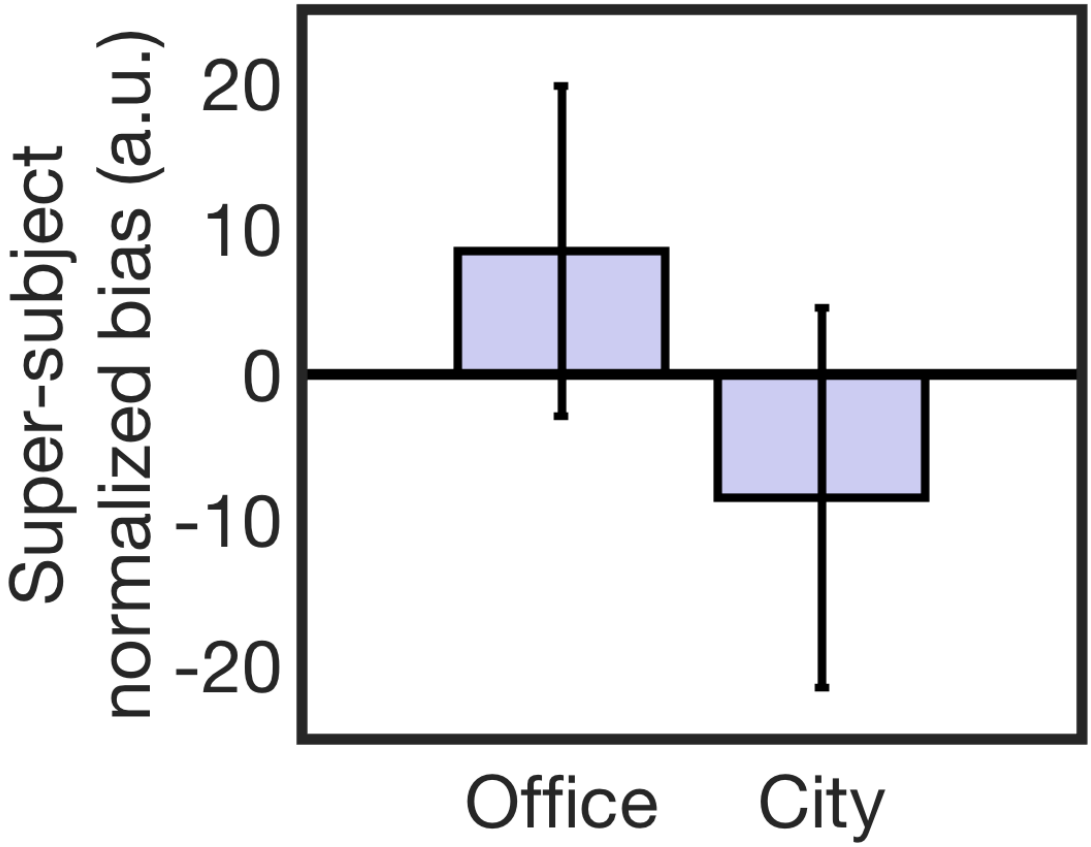

Figure S2. Normalized estimation bias computed on pooled ('super-subject') behavioral data, as a function of video scene (corresponding to the plots in row 3 of fig. 1).

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:

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

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.