To integrate or not to integrate: Temporal dynamics of Bayesian Causal Inference

Máté Aller[1], Uta Noppeney[1]

[1]Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, Birmingham, United Kingdom

Corresponding author:
Máté Aller

Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom

Email: MXA342@bham.ac.uk

Abstract

To form a percept of the environment, the brain needs to solve the binding problem – inferring whether signals come from a common cause and be integrated, or come from independent causes and be segregated. Behaviourally, humans solve this problem near-optimally as predicted by Bayesian Causal Inference; but, the neural mechanisms remain unclear. Combining Bayesian modelling, electroencephalography (EEG), and multivariate decoding in an audiovisual spatial localization task, we show that the brain accomplishes Bayesian Causal Inference by dynamically encoding multiple spatial estimates. Initially, auditory and visual signal locations are estimated independently; next, an estimate is formed that combines information from vision and audition. Yet, it is only from 200 ms onwards that the brain integrates audiovisual signals weighted by their bottom-up sensory reliabilities and top-down task-relevance into spatial priority maps that guide behavioural responses. Critically, as predicted by Bayesian Causal Inference, these spatial priority maps take into account the brain's uncertainty about the world's causal structure and flexibly arbitrate between sensory integration and segregation. The dynamic evolution of perceptual estimates thus reflects the hierarchical nature of Bayesian Causal Inference, a statistical computation, crucial for effective interactions with the environment.

Introduction

In our natural environment our senses are exposed to a barrage of sensory signals: the sight of a rapidly approaching motor truck, its looming motor noise, the smell of traffic fumes. How the brain effortlessly merges these signals into a seamless percept of the environment remains unclear. Critically, the brain faces two fundamental computational challenges: First, we need to solve the 'binding' or 'causal inference' problem - deciding whether signals come from a common cause and thus should be integrated, or instead be treated independently [1,2]. Second, when there is a common cause, the brain should integrate them taking into account their uncertainties [3,4]).

Hierarchical Bayesian Causal Inference provides a rational strategy to arbitrate between sensory integration and segregation in perception [2]. Bayesian Causal Inference explicitly models the potential causal structures that could have generated the sensory signals i.e., whether signals come from common or independent sources. In line with Helmholtz' notion of 'unconscious inference', the brain is then thought to invert this generative model during perception [5]. In case of a common signal source, signals are integrated weighted in proportion to their relative sensory reliabilities (i.e. forced fusion [3,4,6–8]). In case of independent sources, they are processed independently (i.e. full segregation [9,10]). Critically, on a particular instance the brain does not know the world's causal structure that gave rise to the sensory signals. To account for this causal uncertainty a final estimate (e.g. object's location) is obtained by combining the estimates under the two causal structures (i.e. common vs. independent sources) weighted by each causal structure's posterior probability (for other decisional strategies see [11]).

A large body of psychophysics research has demonstrated that human observers combine sensory signals near-optimally as predicted by Bayesian Causal Inference [2,11–14]. Most prominently, when locating events in the environment observers gracefully transition between sensory integration and segregation as a function of audiovisual spatial disparity [10]. For small spatial disparities, they integrate signals weighted by their reliabilities leading to crossmodal spatial biases [15], for larger spatial disparities audiovisual interactions are attenuated. A recent fMRI study showed how Bayesian Causal Inference is accomplished within the cortical hierarchy [12,14]: While early auditory and visual areas represent the signals on the basis that they are generated by independent sources (i.e. full segregation), the posterior parietal sulcus integrates sensory signals into one unified percept (i.e. forced fusion). Only at the top of the cortical hierarchy, in anterior parietal cortex, the uncertainty about the world's causal structure is taken into account and signals are integrated into a spatial estimate in line with Bayesian Causal Inference.

The hierarchical organization of Bayesian Causal Inference raises the intriguing question of how these computations evolve dynamically over time in the human brain. How does the brain merge spatial information that is initially coded in different reference frames and representational formats? While the brain is likely to update all spatial estimates by passing messages forwards and backwards across the cortical hierarchy [16–18], the unisensory estimates may precede the computation of the Bayesian Causal Inference estimate.

To characterize the neural dynamics of Bayesian Causal Inference we presented human observers with auditory, visual and audiovisual signals that varied in their spatial disparity in an auditory and visual spatial localization task, whilst recording their neural activity with EEG. First, we employed cross-sensory decoding and temporal generalization matrices of the unisensory auditory and visual signal trials to characterize the emergence and temporal stability of spatial representations across the senses. Second, combining psychophysics, EEG, Bayesian inference modelling, we temporally resolved the evolution of unisensory segregation, forced fusion and Bayesian causal inference estimates in multisensory perception.

Results

To determine the computational principles that govern multisensory processing we presented 13 participants with synchronous audiovisual spatial signals that varied in their spatial disparity and visual reliability (Fig. 1A, B). On each trial, participants reported their perceived location of either the auditory or the visual signal. In addition, we included unisensory auditory and visual signal trials under auditory or visual report, respectively.

Combining psychophysics, EEG and computational modelling, we addressed two questions: First, we investigated when and how human observers form spatial representations from unisensory visual or auditory inputs, which generalize across the two sensory modalities. Second, we studied the computational principles and neural dynamics that mediate the integration of audiovisual signals into spatial representations that take into account the observer's uncertainty about the world's causal structure consistent with Bayesian Causal Inference.

Shared and distinct neural representations of space across vision and audition

Participants were able to locate unisensory auditory and visual signals reliably as indicated by a significant Pearson correlation between participant's location responses and the true signal source location for both unisensory auditory (across subjects mean $\pm$ SEM: 0.88 $\pm$ 0.05) and visual (across subjects mean $\pm$ SEM: 0.95 $\pm$ 0.05) conditions. Next, we investigated how the brain dynamically encodes the location of unisensory auditory or visual signals using multivariate pattern decoding (Fig 2). Specifically, we trained a support vector regression (SVR) model on the EEG topography at each time point and used this learnt mapping to decode the spatial location from EEG topographies at this and all other time points. The decoding accuracy was expressed as the Pearson correlation coefficient between the true and the decoded stimulus locations and entered into a so-called temporal generalization matrix that illustrates the stability of decoding patterns across time [19]. Spatial locations were successfully decoded from EEG topographies from 55 ms onwards for visual stimuli and from 80 ms onwards for auditory stimuli as indicated by the correlation coefficients along the diagonal of each matrix (Fig 3, auditory: left, bottom; visual: right, top matrix). Moreover, the temporal generalization matrices suggest that the visual spatial representations were initially transient (i.e. 55 – 150 ms) reflecting the initial EEG components and later more sustained (i.e. from about 250 ms). By contrast, auditory spatial representations as expressed in EEG topographies were sustained from about 100 ms onwards. Critically, in addition to temporal generalization within each sensory modality, we also investigated the extent to which the SVR decoding model generalized across sensory modalities throughout post-stimulus time. While earlier neural representations were more specific to each particular sensory modality, the SVR model was able to generalize from audition to vision and vice versa from 150 to 400 ms. This cross-sensory generalization across visual and auditory-evoked EEG topographies suggests that at those later stages the brain forms spatial representations that are at least partly shared across sensory modalities.

Computational principles of audiovisual integration: Model-free and model-based analysis

Combining psychophysics, multivariate EEG pattern decoding and computational modelling, we next investigated the computational principles and neural dynamics underlying audiovisual integration of spatial representations with a model-free and a model-based analysis that was applied jointly to participants' behavioural and neural data.

The model-free analysis characterized audiovisual integration in terms of an audiovisual weight index w_{AV} that quantifies the influence of the true auditory and visual locations on (i) the perceived / reported auditory and visual spatial estimates (i.e., participants' localization responses; Fig. 1C) at the behavioural level and (ii) the spatial estimates decoded from EEG topographies independently for each post-stimulus time point at the neural level (Fig. 4). This audiovisual weight index ranges from pure visual (90°) to pure auditory (0°) influence. We performed the statistics on the behavioural and neural audiovisual weight indices using a 2 (visual reliability: high vs. low) x 2 (task-

relevance: auditory vs. visual report) x 2 (spatial disparity: ≤ 6.6° vs. > 6.6°) factorial design based on circular statistics [14].

The model-based analysis formally compared the following models as explanations for observers' behavioural and neural data: The (i) full segregation model(s) assume(s) that auditory and visual signals are processed independently and the observer reports the task-relevant spatial estimate irrespective of the spatial location of the other sensory modality (encircled in light blue in Fig 5A). (ii) The forced fusion model assumes that auditory and visual signals are integrated weighted by their reliabilities in a mandatory fashion irrespective of the environmental causal structure (encircled in yellow in Fig 5A). (iii) The Bayesian Causal Inference model computes a final auditory (or visual) spatial estimate by averaging the spatial estimates under forced fusion and full segregation assumptions weighted by the posterior probabilities of each causal structure (i.e., model averaging, encircled in dark blue; see supporting material Table S1 for other decision functions). For behavioural data the full segregation model fits the unisensory visual model to visual response trials and the unisensory auditory model to auditory response trials. For the EEG data, we also investigated the extent to which the brain maintains spatial representations of the task-irrelevant sensory signal. For this, we also fitted the unisensory visual (resp. auditory) model to the neural data. Hence, we fit five models to the neural data, but only three models to the behavioural data. Using a maximum likelihood procedure, we fitted the parameters (e.g., visual variances σ_{V1} , σ_{V2} for the two reliability levels) of each model individually to each participant's (i) behavioral localization responses and (ii) the spatial estimates decoded from EEG topographies independently at each time point (Fig. 5). Using Bayesian model comparison at the random effects level [20] we computed the exceedance probabilities of the different models for the behavioural and EEG data as a function of post-stimulus time.

Behaviour

The model-free analysis of the behavioural audiovisual weight index w_{AV} shows that observers integrated audiovisual signals weighted by their sensory reliabilities and task relevance (see Fig S1 for histograms of reported signal locations across all conditions). The relative influence of the visual signal on observers' perceived location was greater, when the location of the visual signal needed to be reported (main effect of task-relevance: $p < 0.001$). Observers thus flexibly adjusted the weights as a function of task relevance giving more emphasis to the sensory modality that needs to be reported. As a consequence, audiovisual signals were not fused into one unified percept; the reported auditory and visual locations differed for identical audiovisual signals. Critically, consistent with Bayesian Causal Inference this difference significantly increased for large (> 6.6°) relative to small (≤ 6.6°) spatial disparities. In other words, audiovisual integration broke down, when auditory and visual signals were far apart and more likely to be caused by independent sources. This significant interaction between task-relevance and spatial disparity ($p < 0.001$) is a qualitative profile that is characteristic for Bayesian Causal Inference.

Moreover, we observed significant two-way interactions between visual reliability and spatial disparity ($p = 0.0014$) as well as task-relevance ($p = 0.0002$). The effect of high vs. low visual reliability was stronger when the two signals were close in space and the auditory, i.e. less reliable signal needed to be reported. Again, this interaction is qualitatively expected for Bayesian Causal Inference, because the forced fusion model that integrates sensory signals weighted by their relative reliabilities is more influential for low spatial disparity trials, when it is likely that two signals come from a common source.

Indeed, formal Bayesian model comparison showed that the Bayesian Causal Inference model outperformed the full segregation and forced fusion models (85.6 $\pm$ 0.3 % variance explained, protected exceedance probability > 0.99; Table 2). Further, as shown in Fig 1D, the observed profile

of the audiovisual weight index w_{AV} is consistent with the profile predicted by Bayesian Causal Inference.
In summary, our model-free and model-based analysis of the behavioural data provide convergent evidence that human observers integrate audiovisual spatial signals weighted by their relative reliabilities at small spatial disparities but segregate signals at large spatial disparities when it is unlikely that they come from a common source.

EEG results - Temporal dynamics of audiovisual integration

To characterize the neural dynamics underlying integration of audiovisual signals into spatial representations, we combined EEG with multivariate pattern decoding. Based on a support-vector regression model trained on the EEG topographies of audiovisual spatially congruent trials [12,14,21], we decoded the spatial estimates of spatially congruent and disparate audiovisual signals at each post-stimulus time point. Consistent with our behavioural analysis we then applied a model-free and model-based analysis to those decoded 'spatial estimates'. Because the location of congruent audiovisual stimuli could be decoded better than chance only from 55 ms post-stimulus ($p < 0.001$), we report and interpret results limited to the time window from 55 ms to 700 ms (Fig S2).

Model-free EEG results

The model-free analysis investigated how visual reliability, task-relevance and spatial disparity influenced the decoded spatial locations across post-stimulus time using the audiovisual neural weight index w_{AV}. Our results show that sensory reliability significantly influenced the audiovisual neural weight index w_{AV} from 65 – 540 ms. As expected, the spatial representations were more strongly influenced by the true visual signal location, when the visual signal was reliable than unreliable (Fig 4A, Table 1). Moreover, consistent with our behavioural findings, we also observed a main effect of task-relevance between 190 – 700 ms (Fig 4B, Table 1). The decoded location was more strongly influenced by the visual signal when the visual modality was task relevant. Most importantly, we also observed a significant interaction between task relevance and spatial disparity from 350 – 435 ms. As discussed in the context of the behavioural results, this interaction is a qualitative profile characteristic for Bayesian Causal Inference: The brain integrates sensory signals at low spatial disparity (i.e. small difference for auditory vs. visual report), but computes different spatial estimates for auditory and visual signals at large spatial disparities (see Fig 4D, Table 1).
In addition to these key findings, we also observed a brief but pronounced main effect of spatial disparity at about 50 – 130 ms. In this time window, which coincides with the visual evoked N100 response, the decoded spatial estimate was dominated by the visual stimulus location. Moreover, we observed a small attractive effect of sound for small spatial disparity, but a repulsive effect for large spatial disparity as indicated by w_{AV} values above 90°. This profile suggests that early multisensory processing is already influenced by a spatial window of integration and the causal structure of the environment (Fig 4C, Table 1). Auditory stimuli affect the spatial representation of visual signals mainly when they close in space. However, because spatial disparity was inherently correlated with the eccentricity of the audiovisual signals by virtue of our factorial and spatially balanced design, these two effects cannot be fully dissociated. While signals were presented para-foveally or peripherally for small-disparity trials, they were presented always in the periphery for large-disparity trials.
For completeness, we also observed a significant interaction between spatial disparity and visual reliability between 45 – 100 ms and between 170 – 235 ms (Table 1). This interaction results from a larger spatial window of integration for stimuli with low versus high visual reliability. Basically, it is easier to determine whether two signals come from different sources when the visual input is reliable.

Finally, we asked when and how the neural audiovisual weights were related to the audiovisual weights that observers applied at the behavioural level. Hence, we computed the correlation between the values of the behavioural and neural weight indices w_{AV} separately for each time point. The Fisher z-transformed correlation coefficient fluctuated around chance level until about 100 ms. From 100 ms onwards it progressively increased over time, until it peaked and reached a plateau at about 350 ms (R = 0.92). As expected, this coincides with the time window where we observed a significant interaction between task-relevance and spatial disparity, i.e. the qualitative profile characteristic for Bayesian Causal Inference. After 500 ms, it then slowly decreased towards the end of the trial. Cluster permutation test confirmed, that the correlation between neural and behavioural weight indices w_{AV} was significantly better than chance, revealing two clusters between 175 – 550 ms (p = 0.0004) and 575 – 685 ms (p = 0.013). These results indicate that the neural representations expressed in EEG topographies are critical for guiding observers' responses.

Model-based EEG results

In the model-based analysis we fitted five models to the decoded spatial estimates separately for each time point: (i) 'segregation auditory', (ii) 'segregation visual', (iii) 'full segregation', (iv) 'forced fusion' and (v) the 'Bayesian Causal Inference' model (Fig 5A). In contrast to the behavioural analysis, we also fitted the full segregation models separately for the auditory and visual estimates to examine when the auditory or visual estimates alone dominate the neural dynamics irrespective of task-relevance. The time course of the exceedance probabilities revealed a sequential pattern (Fig 5B): Initially, the 'segregation visual' model dominated until about 100 ms post stimulus. From 100 to about 200 ms the forced fusion model outperformed the other models indicating that spatial estimates are now influenced by both sensory modalities. From about 200 ms onwards the exceedance probability of the 'Bayesian Causal Inference' model progressively increased peaking at about 350 ms followed by a plateau until 500 ms. Thus, consistent with the model-free results, audiovisual interactions in line with Bayesian Causal Inference emerge relatively late at about 350 ms post-stimulus.

Discussion

Integrating information from vision and audition into a coherent representation of the space around us is critical for effective interactions with the world. This EEG study temporally resolved the neural dynamics that enables the brain to flexibly integrate auditory and visual signals into spatial representations in line with the predictions of Bayesian Causal Inference.

Auditory and visual senses code spatial location in different reference frames and representational formats. Vision provides spatial information in eye-centred, audition in head-centred references frames [22,23]. Furthermore, spatial location is directly coded in the retinotopic organization in primary visual cortex [24], while spatial location in audition is computed from sound latency and amplitude differences between the ears and coded in broad tuning functions in auditory cortices [25,26]. In order to merge spatial information from vision and audition the brain hence needs to establish coordinate mappings and/or transform spatial information into partially shared 'hybrid' reference frames as previously suggested by neurophysiological studies in non-human primates [25,27]. In the first step we therefore investigated the neural dynamics of spatial representations encoded in EEG topographies for unisensory auditory and visual signals using temporal generalization matrices. In vision spatial location is encoded initially at 55 ms in transient neural activity that then turn into temporally more stable representations from about 250 ms. In audition, the EEG topographies encode spatial location in neural representations that generalize across time from 100 ms onwards, but are again most pronounced at 250 ms. Moreover, spatial representations also generalize across sensory modalities starting from about 150 ms but culminating between 300 and 400 ms. These results suggest that unisensory auditory and visual spatial information are initially

coded in a modality-specific format, but are then transformed into a shared representational space in partly overlapping neural generators at about 300 ms.

Having characterized the neural dynamics of unisensory auditory and visual spatial representations across time, we next asked when and how the human brain combines spatial information from vision and audition into a coherent representation of space. Critically, the brain should integrate sensory signals only when they come from a common event, but segregate those from independent events [1,2]. Thus, multisensory perception inherently relies on the brain forming a causal model of the world. Spatial disparity is an important cue that informs the brain whether or not signals are generated by common events [10]. To investigate whether the brain arbitrates between sensory integration and segregation, we therefore presented observers with synchronous audiovisual signals that varied in their spatial disparity across trials. On each trial observers reported either the auditory or the visual location. As expected, a concurrent yet spatially disparate visual signal biased observers' perceived sound location towards the visual location – a phenomenon coined spatial ventriloquist illusion [15,28]. In line with the principles of reliability-weighted integration, this audiovisual spatial bias was stronger, when the visual signal was reliable. Most importantly, observers did not fuse audiovisual signals into one single percept, but reported different locations for auditory and visual signals in particular for large spatial disparities. In other words, we observed a significant interaction between spatial disparity and task-relevance for the behavioural audiovisual weight index. These behavioural results demonstrate that human observers arbitrate between sensory integration and segregation depending on spatial disparity, i.e. the likelihood of different causal structures of the world.

In our initial model-free analysis, we next investigated when the brain forms neural spatial representations that are influenced by reliability, task-relevance and spatial disparity as participant's behavioural reports. Our results show that visual reliability influenced how the brain weights and integrates audiovisual signals into spatial representations early from 65 ms onwards. By contrast, the task-relevance of the auditory or visual signals influenced the weighting and integration process only later from 190 ms. Most importantly, we observed a significant interaction between task-relevance and spatial disparity later at about 350 ms. Hence, at 350 ms the brain integrates sensory signals weighted by their bottom-up reliability and top-down task-relevance into spatial priority maps that take the likelihood of the different causal structures into account [15]. These spatial priority maps are behaviourally relevant for guiding spatial orienting and actions as indicated by the correlation of the neural and behavioural audiovisual weight indices, which progressively increase from 100 ms to culminate at about 300-400 ms. Our model-free analysis thus clearly demonstrates that the brain forms multisensory spatial priority maps that take the reliability and task-relevance of the spatial signals into account as well as the causal structure of the environment. The findings converge with a recent fMRI study suggesting a hierarchical organization of multisensory perception [14]: While low level auditory and visual areas predominantly encoded the unisensory auditory or visual locations [29–38], higher order visual areas and posterior parietal cortices combined audiovisual signals weighted by their sensory reliabilities [21,39–41]. Critically, only at the top of the hierarchy, in anterior parietal cortices, were spatial priority maps formed that weighted sensory signals by their bottom-up reliability and top-down task-relevance. Thus, the hierarchical organization observed in fMRI reflects the temporal dynamics of audiovisual integration observed with EEG. Collectively, behavioural and EEG results demonstrate that the brain does not fuse sensory signals into one unified percept, but gracefully transitions between sensory integration and segregation qualitatively in line with the principles of Bayesian Causal inference.

To test the predictions of Bayesian Causal Inference quantitatively we analysed the EEG data in a model-based fashion. Specifically, we fitted five models to observers' 'spatial locations' that were decoded from EEG topographies at each time point: the unisensory i. auditory and ii. visual model, iii. the full segregation audiovisual model, iv. the forced fusion audiovisual model and the v. Bayesian Causal Inference model (Fig 5A). As shown in Fig 5B, the time course of models' exceedance probability shows a multi-stage process. The full segregation visual model was the winning model for

the first 100 ms, the audiovisual forced fusion model dominated the time interval 100 - 250 ms and the Bayesian Causal Inference model outperformed from 250 ms onwards. Hence, our model-based analysis quantitatively corroborates the results of our model-free analysis. The hierarchical nature of Bayesian Causal Inference is reflected in the neural dynamics that is dominated successively by the full segregation, forced fusion and Bayesian Causal Inference model components. Yet, we need to emphasize that these results do not imply a purely sequential feed-forward architecture but only suggest that different computations are more prominent at different latencies in EEG topographies. Time-resolved decoding from EEG topographies has very limited abilities to dissociate multiple overlapping representations and will be dominated by the stronger neural signature and will mask weaker processes (e.g. also auditory full segregation model).

Collectively, our results demonstrate that multisensory perception relies on the dynamic encoding of multiple sensory estimates most likely via multiple feed-back loops across the cortical hierarchy. Only after 350 ms is a final perceptual estimate formed that takes the uncertainty about the world's causal structure into account and combines signals as predicted by Bayesian Causal Inference.

Methods

Participants

Sixteen right-handed participants gave informed consent to participate in the experiment. Three of those participants did not complete the entire experiment: two participants were excluded based on eye tracking results from the first day (the inclusion criterion was less than 10% of trials rejected because of eye blinks or saccades, see the Eye movement recording and analysis section for details) and one participant withdrew from the experiment. The remaining thirteen participants (7 females, mean age = 22.1; SD = 3.0) completed the three-day experiment and are thus included in the analysis. All participants had no history of neurological or psychiatric illnesses, normal or corrected-to-normal vision and normal hearing. The study was approved by the research ethics committee of the University of Birmingham and was conducted in accordance with the principles outlined in the Declaration of Helsinki.

Stimuli

The visual (V) stimulus was a cloud of 20 white dots (diameter = 0.43° visual angle, stimulus duration: 50 ms) sampled from a bivariate Gaussian distribution with vertical standard deviation of 2° and horizontal standard deviation of 2° or 12° visual angle presented on a dark grey background (67% contrast). Participants were told that the 20 dots were generated by one underlying source in the center of the cloud. The visual cloud of dots was presented at one of four possible locations along the azimuth (i.e., -10°, -3.3°, 3.3° or 10°).

The auditory (A) stimulus, was a 50 ms long burst of white-noise with 5 ms on/off ramp. Each auditory stimulus was delivered at 75 dB sound pressure level through one of four pairs of two vertically aligned loudspeakers placed above and below the monitor at four positions along the azimuth (i.e., -10°, -3.3°, 3.3° or 10°). The volumes of the 2 x 4 speakers were carefully calibrated across and within each pair to ensure that participants perceived the sounds as emanating from the horizontal midline of the monitor.

Experimental design and procedure

In a spatial ventriloquist paradigm, participants were presented with synchronous, spatially congruent or disparate visual and auditory signals (Fig 1A). On each trial, visual and auditory locations were independently sampled from four possible locations along the azimuth (i.e., -10°, -3.3°, 3.3° or 10°) leading to four levels of spatial discrepancy (i.e., 0°, 6.6°, 13.3° or 20°). In addition, we manipulated the reliability of the visual signal by setting the horizontal standard deviation of the Gaussian cloud to 2° (high reliability) or 14° (low reliability) visual angle. In an inter-sensory selective-attention paradigm, participants reported either their auditory or visual perceived signal location, and ignored signals in the other modality. For the visual modality, they were asked to determine the location of the center of the visual cloud of dots. Hence, the 4 x 4 x 2 x 2 factorial design manipulated (1) the location of the visual stimulus ({-10°, -3.3°, 3.3°, 10°}, i.e., the mean of the Gaussian) (2) the location of the auditory stimulus ({-10°, -3.3°, 3.3°, 10°}) (3) the reliability of the visual signal ({2°,14°}, STD of the Gaussian) and (4) task-relevance (auditory- / visual-selective report) resulting in 64 conditions (Fig 1B). To characterize the computational principles of multisensory integration, we reorganized these conditions into a two (visual reliability: high vs. low) x two (task-relevance: auditory vs. visual report) x two (spatial disparity: ≤ 6.6° vs. > 6.6°) factorial design for the statistical analysis of the behavioural and EEG data. In addition, we included 4 (locations: 10°, -3.3°, 3.3° or 10°) x 2 (visual reliability: high, low) unisensory visual conditions and 4 (locations: 10°, -3.3°, 3.3° or 10°) unisensory auditory conditions.

On each trial, synchronous audiovisual, unisensory visual or unisensory auditory signals were presented for 50 ms, followed by a response cue 1000 ms after stimulus onset (Fig 1B). The response was cued by a central pure tone (1000 Hz) and a blue colour change of the fixation cross presented synchronously for 100 ms. Participants were instructed to withhold their response and avoid blinking until the presentation of the cue. They fixated a central cross throughout the entire experiment. The next stimulus was presented after variable response interval of 2.6-3.1 s.

Stimuli and conditions were presented in a pseudo-randomized fashion. The stimulus type (bisensory vs. unisensory) and task-relevance (auditory vs. visual) was held constant within a run of 128 trials. This yielded four run types: audiovisual with auditory report, audiovisual with visual report, auditory with auditory report and visual with visual report. The task relevance of the given run was displayed to the participant at the beginning of the run. Further, across runs we counterbalance the response hand (i.e. left vs. right hand). The order of the runs was counterbalanced across participants. All conditions within a run were presented an equal number of times. Each participant completed 60 runs leading to 7680 trials in total (3840 auditory and 3840 visual localization tasks, i.e. 96 trials for each of the 76 conditions were included in total; apart from the 4 unisensory auditory conditions that included 192 trials). The runs were performed across three days with 20 runs per day. Each day was started with a brief practice run.

Experimental set up

Stimuli were presented using Psychtoolbox version 3.0.11 [42] (http://psychtoolbox.org/) under MATLAB R2014a (MathWorks Inc.) on a desktop PC running Windows 7. Visual stimuli were presented via a gamma-corrected 30” LCD monitor with a resolution of 2560 x 1600 pixel at a frame rate of 60 Hz. Auditory stimuli were presented at a sampling rate of 44.1 kHz via 8 external speakers (Multimedia) and an ASUS Xonar DSX sound card. Exact audiovisual onset timing was confirmed by recording visual and auditory signals concurrently with a photo-diode and a microphone.

Participants rested their head on a chin rest at a distance of 475 mm from the monitor and at a height that matched participants’ ears to the horizontal midline of the monitor.

Participants responded by pressing one of four response buttons on a USB keypad with their index, middle, ring and little finger respectively.

Eye movement recording and analysis

To address potential concerns that results were confounded by eye movements, we recorded participants' eye movements. Eye recordings were calibrated in the recommended field of view (32° horizontally and 24° vertically) for the EyeLink 1000 Plus system with the desktop mount at a sampling rate of 2000 Hz. Eye position data were on-line parsed into events (saccade, fixation, eye blink) using the EyeLink 1000 Plus software. The 'cognitive configuration' was used for saccade detection (velocity threshold = 30°/sec, acceleration threshold = 8000°/sec^2, motion threshold = 0.15°) with an additional criterion of radial amplitude larger than 1°. Individual trials were rejected if saccades or eye blinks were detected from -100 to 700 ms post-stimulus.

EEG data acquisition

Continuous EEG signals were recorded from 64 channels using Ag/AgCl active electrodes arranged in 10-20 layout (ActiCap, Brain Products GmbH, Gilching, Germany) at a sampling rate of 1000 Hz, referenced at FCz. Channel impedances were kept below 10kΩ.

EEG pre-processing

Pre-processing was performed with the FieldTrip toolbox [43] (http://www.fieldtriptoolbox.org/). For the decoding analysis raw data were high pass filtered at 0.1 Hz, re-referenced to average reference, and low pass filtered at 120 Hz. Trials were extracted with 100 ms pre-stimulus and 700 ms post-stimulus period, baseline corrected, temporally smoothed with a 20 ms moving window and down-sampled to 200 Hz (note, that a 20 ms moving average is comparable to a Finite Impulse Response (FIR) filter with a cut-off frequency of 50 Hz). Trials containing artefacts were rejected based on visual inspection. Furthermore, trials were rejected if they included (i) eye blinks, (ii) saccades, or when (iii) the distance between eye fixation and the central fixation cross exceeded 2 degrees, or (iv) participants responded prior to the response cue or (v) there was no response. For event related potentials (ERPs; Fig 2A, S3A) the pre-processing was identical to the decoding analysis, except that a 45 Hz low-pass filter was applied without additional temporal smoothing with a temporal moving window. Grand average ERPs were computed by averaging all trials for each condition first within each participant, then across participants.

EEG multivariate pattern analysis

For the multivariate pattern analyses we computed ERPs by averaging over sets of 8 randomly assigned individual trials from the same condition. To characterize the temporal dynamics of the spatial representations (see below) we trained linear support vector regression models (SVR, LIBSVM [44], https://www.csie.ntu.edu.tw/~cjlin/libsvm/) to learn the mapping from ERP topography patterns of the i. unisensory auditory, ii. unisensory visual or iii. audiovisual congruent conditions to external spatial locations separately for each time point over the course of the trial (Fig 2, S3). All SVR models were trained and evaluated in a 12-fold stratified cross-validation (12 ERPs/fold) procedure with default hyper parameters ($C = 1$, $\nu = 0.5$). The specific training and generalization procedures were adjusted to the scientific questions (see below for details)

Overview of behavioural and EEG analysis

Combining psychophysics, computational modelling and EEG we addressed two questions: First, focusing selectively on the unisensory auditory and visual conditions we investigated when spatial representations are formed that generalize across auditory and visual modalities. Second, focusing on the audiovisual conditions we investigated when and how human observers integrate audiovisual

signals into spatial representations that take into account the observer's uncertainty about the world's causal structure consistent with Bayesian Causal Inference. In the following we will describe the analysis approaches to address these two questions in turn.

Shared and distinct neural representations of space across vision and audition

First, we investigated how the brain forms spatial representations in either audition or vision using the so-called temporal generalization method. Here, the SVR model is trained at time point t_1 to learn e.g. the mapping from visual ERP topography to external spatial location. This learnt mapping is then used to predict spatial locations from ERP topographies across all other time points. The decoding accuracy as quantified by the Pearson correlation coefficient between the true and decoded stimulus locations is entered into a training time x generalization time matrix. The generalization ability across time illustrates the similarity of EEG topographies sustaining the encoding features and has been proposed to assess the stability of neural representations [19].

Second, to examine whether and when neural representations are formed that are shared across vision and audition, we generalized not only to ERP topographies across time from the same sensory modality, but also from the other sensory modality (i.e. from vision to audition and vice versa). This cross-sensory generalization reveals neural representations that are shared across sensory modalities.

To assess whether decoding accuracies and circular correlations were better than chance we entered the Fisher z-transformed Pearson correlation coefficients into a between subjects Monte-Carlo permutation test using the one-sample t-statistic with 5000 permutations ([45], as implemented in the FieldTrip toolbox). To correct for multiple comparisons within the two-dimensional (i.e. time x time) data, cluster-level inference was used based on the maximum of the summed t-values within each cluster ('maxsum') with a cluster defining threshold of $p < 0.05$ and a two-tailed p-value was computed.

Computational principles of audiovisual integration: Model-free and model-based analysis

To characterize how human observers integrate auditory and visual signals into spatial representations at the behavioural and neural levels, we developed a model-free and a model-based analysis approach that we applied jointly to both (i) reported auditory and visual spatial estimates (i.e., participants' behavioural localization responses) and (ii) the decoded neural spatial estimates [12,14]. The neural spatial estimates were obtained by training a SVR model on the audiovisual congruent trials to learn the mapping from ERP topography to external spatial locations. This learnt mapping at time t was then used to decode the spatial location from the ERP topographies of the spatially congruent and incongruent audiovisual conditions at time t.

Model-free analysis

Regression model to compute audiovisual weight index w_{AV}

In the 'model free' analysis approach we quantified the influence of the auditory and visual signals on the reported (behavioural) or decoded (neural) spatial estimates using a linear regression model. In this regression model, the perceived/decoded spatial locations were predicted by the true auditory and visual spatial locations for each of the eight conditions in the 2 (visual reliability: high vs. low) x 2 (task-relevance: auditory vs. visual report) x 2 (spatial disparity: ≤ 6.6° vs. > 6.6°) factorial design (Fig 1A). Hence, the regression model included 16 regressors in total, i.e., 8 (conditions) x 2 (true auditory or visual spatial locations). The auditory ($ß_A$) and visual ($ß_V$) parameter estimates quantified the influence of auditory and visual signals on the perceived/decoded signal location for a particular condition. To obtain a summary index for the relative audiovisual weights, we computed the relative audiovisual weight (w_{AV}) as the four-quadrant inverse tangent of the visual ($ß_V$) and

auditory ($ß_A$) parameter estimates from the regression model for each of the eight conditions based on participants' behavioural responses (Fig 1C) and their decoded spatial estimates across the time course of the trial (Fig 4A-D). If the reported/decoded estimate is dominated purely by the visual signal, w_{AV} is 90°. For pure auditory dominance, it is 0°.

Permutation of circular indices w_{AV} for behavioural and neural data

We performed the statistics on the behavioural and neural audiovisual weight indices using a two (auditory vs. visual report) x two (high vs. low visual reliability) x two (large vs. small spatial disparity) factorial design based on the likelihood ratio statistics for circular measures (LRTS) [46]. Similar to an analysis of variance for linear data, LRTS computes the difference in log-likelihood functions for the full model that allows differences in the mean locations of circular measures between conditions and the reduced null model that does not model any mean differences between conditions. LRTS were computed separately for the main effects and interactions. To refrain from making any parametric assumptions, we evaluated the main effects of visual reliability, task-relevance, spatial disparity and their interactions in the factorial design using permutation testing of the LRTS. Permutations were constrained to occur within each participant. For the main effects of visual reliability, task-relevance and spatial disparity, w_{AV} values were permuted within the levels of the non-tested factors. For tests of the two-way interactions, we permuted the simple main effects of the two factors of interest within the levels of the third factor [47]. For tests of the three-way interaction, values were freely permuted across all conditions [48]. All tests were based on 5000 random permutations and a right-tailed p-value was computed.
To assess the similarity between behavioural and neural audiovisual weights (w_{AV}), we computed the circular correlation coefficient (as implemented in the CircStat toolbox [49]) between the 8 neural and 8 behavioural w_{AV}s from our 2 (high vs. low visual reliability) x 2 (auditory vs. visual report) x 2 (large vs small spatial disparity) factorial design separately for each time point.
To correct for multiple comparisons within the one (i.e. time) data, cluster-level inference was used based on the maximum of the summed t-values within each cluster ('maxsum') with a cluster defining threshold of $p < 0.05$ and a two-tailed p-value was computed.

Model-based analysis

Bayesian Causal Inference model

In the 'model based' analysis approach we formally investigated whether participants integrate auditory and visual signals into spatial representation quantitatively in line with the predictions of Bayesian Causal Inference. Details of the Bayesian Causal Inference model of audiovisual perception can be found in [2,11,13]. Briefly, the generative model of Bayesian Causal Inference assumes that common ($C = 1$) or independent ($C = 2$) causes are sampled from a binomial distribution defined the common cause prior P_{common}. For a common source, the 'true' location S_{AV} is drawn from the spatial prior distribution $N(\mu_{AV}, \sigma_P)$. For two independent causes, the 'true' auditory (S_A) and visual (S_V) locations are drawn independently from this spatial prior distribution. For the spatial prior distribution, we assumed a central bias (i.e., $\mu = 0$). We introduced sensory noise by drawing x_A and x_V independently from normal distributions centered on the true auditory (resp. visual) locations with parameters σ_A^2 (resp. σ_V^2). Thus, the generative model included the following free parameters: the common source prior p_{common}, the spatial prior variance σ_P^2, the auditory variance σ_A^2 and the two visual variances σ_V^2 corresponding to the two visual reliability levels. The posterior probability of the underlying causal structure can be inferred by combining the common-source prior with the sensory evidence according to Bayes rule:

$$p(C = 1|x_A, x_V) = \frac{p(x_A, x_V|C=1)p_{common}}{p(x_A, x_V)} \quad (1)$$

In the case of a common source (C = 1), the optimal estimate of the audiovisual location is a reliability-weighted average of the auditory and visual percepts and the spatial prior.

$$\hat{S}_{AV,C=1} = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_V}{\sigma_V^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} + \frac{1}{\sigma_P^2}} \tag{2}$$

In the case of independent sources (C = 2), the optimal estimates of the auditory and visual signal locations (for the auditory and visual location report, respectively) are independent from each other.

$$\hat{S}_{A,C=2} = \frac{\frac{x_A}{\sigma_A^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_P^2}}, \quad \hat{S}_{V,C=2} = \frac{\frac{x_V}{\sigma_V^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_V^2} + \frac{1}{\sigma_P^2}} \tag{3}$$

To provide a final estimate of the auditory and visual locations, the brain can combine the estimates under the two causal structures using various decision functions such as 'model averaging', 'model selection' and 'probability matching' [11]. In the main paper, we present results using 'model averaging' as the decision function which was associated with the highest model evidence and exceedance probability at the group level (see supporting Table S1). According to the 'model averaging' strategy, the brain combines the integrated forced fusion spatial estimate with the segregated, task-relevant unisensory (i.e., either auditory or visual) spatial estimates weighted in proportion to the posterior probability of the underlying causal structures.

$$\hat{S}_A = p(C=1|x_A, x_V)\, \hat{S}_{AV,C=1} + (1 - p(C=1|x_A, x_V))\hat{S}_{A,C=2} \tag{4}$$

$$\hat{S}_V = p(C=1|x_A, x_V)\, \hat{S}_{AV,C=1} + (1 - p(C=1|x_A, x_V))\hat{S}_{V,C=2} \tag{5}$$

Model fitting to behavioural and neural spatial estimates and Bayesian Model Comparison

At the behavioural level, we compared three models based on individual's behavioural localization responses: (i) the observers may process and report auditory and visual signals independently (i.e., the audiovisual full segregation model, equation (3)). (ii) They may integrate auditory and visual in a mandatory fashion irrespective of spatial disparity (i.e., the forced fusion model, equation (2)). (iii) The observer may perform Bayesian Causal Inference, i.e., combine estimates from the forced fusion and the task-relevant estimate from the full segregation model depending on the probability of the underlying causal structures (equation (4) and (5), i.e., model averaging, for other decision functions see supporting Table S1). At the neural level, we also investigated whether the sensory signals are represented irrespective of whether they are task-relevant and reported. Hence, at the neural level we fitted two additional models: (iv) the auditory full-segregation model and (v) the visual full segregation model.

To arbitrate between those three (behavior) or five (neural) models, we fitted each model individually to a participant's behavioural localization responses (or neural response decoded from EEG topography) based on the predicted distributions of the auditory spatial estimates (i.e., $p(\hat{S}_A|S_A,S_V)$) and the visual spatial estimates (i.e., $p(\hat{S}_V|S_A,S_V)$). These distributions were obtained by marginalizing over the internal variables x_A and x_V that are not accessible to the experimenter (for further details of the fitting procedure see [2]). These distributions were generated by simulating x_A and x_V 10000 times for each of the 64 conditions and inferring $\hat{S}_A$ and $\hat{S}_V$ from equations (1)-(5). To link $p(\hat{S}_A|S_A,S_V)$ and $p(\hat{S}_V|S_A,S_V)$ to participants' auditory and visual discrete localization responses at the behavioural level, we assumed that participants selected the button that is closest to $\hat{S}_A$ or $\hat{S}_V$ and binned the $\hat{S}_A$ and $\hat{S}_V$ accordingly into a histogram (with four bins corresponding to the four buttons). Thus, we obtained a histogram of predicted auditory or visual localization responses for each condition and participant. Based on these histograms we computed the probability of a participant's counts of localization responses using the multinomial distribution (see [2]). This gives

the likelihood of the model given participants' response data. Assuming independence of conditions, we summed the log likelihoods across conditions.
Using the same procedure, we also fitted each model individually to a participant's decoded neural spatial estimates at each time point after binning the neural spatial estimates decoded from each ERP topography across time based on their distance from the four true locations (i.e., -10°, -3.3°, 3.3° or 10°) into four spatial bins.
To obtain maximum likelihood estimates for the parameters of the models (p_{common}, σ_P, σ_A, σ_{V1} - σ_{V2} for the two levels of visual reliability; formally, the forced fusion and segregation models assume p_{common} = 1 or = 0, respectively), we used a non-linear simplex optimization algorithm as implemented in Matlab's fminsearch function (Matlab R2016a). This optimization algorithm was initialized with a parameter setting that obtained the highest log likelihood in a prior grid search.
The model fit for behavioural and neural data (i.e. at each time point) was assessed by the coefficient of determination R^2 [50] defined as

$$\mathrm{R}^2 = 1 - \exp(-\frac{2}{n}(\mathrm{l}(\widehat{ß}) - \mathrm{l}(0)))$$

where $\mathrm{l}(\widehat{ß})$ and l(0) denote the log likelihoods of the fitted and the null model, respectively, and n is the number of data points. For the null model, we assumed that an observer randomly chooses one of the four response options, i.e. we assumed a discrete uniform distribution with a probability of 0.25. As in our case the Bayesian Causal Inference model's responses were discretized to relate them to the four discrete response options, the coefficient of determination was scaled (i.e., divided) by the maximum coefficient (cf. Nagelkerke [50]) defined as

$$\max(\mathrm{R}^2) = 1 - \exp(\frac{2}{\mathrm{n}}\mathrm{l}(0))$$

To identify the optimal model for explaining participants' data, we compared the candidate models using the Bayesian information criterion (BIC) as an approximation to the model evidence [51]. The BIC depends on both model complexity and model fit. We performed Bayesian model selection [20] at the group level as implemented in SPM8 [52] to obtain the protected exceedance probability for the candidate models.

Figures

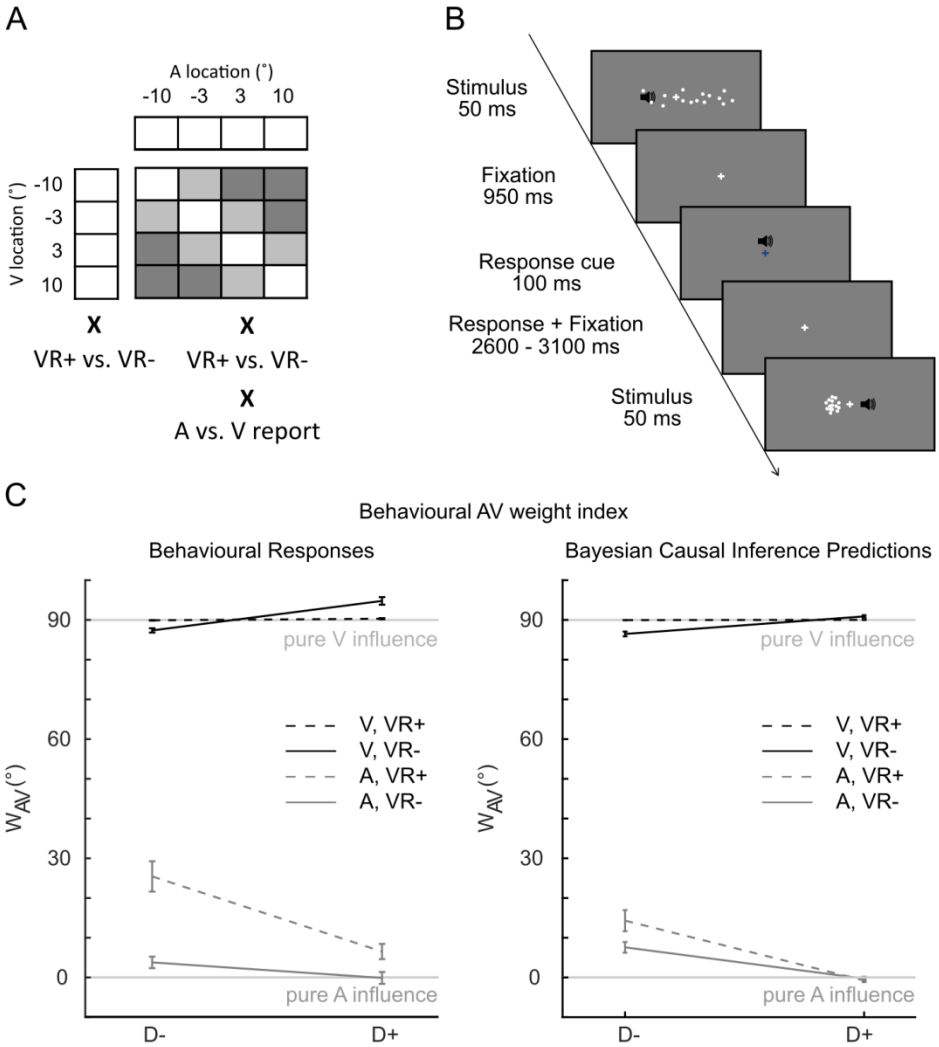

Figure 1 Experimental design, example trial, behavioural and predicted audiovisual weights (w_{AV})(A) Experimental design. In a 4 x 4 x 2 x 2 factorial design, the experiment manipulated i. the location of the visual (V) signal (-10˚, -3.3˚, 3.3˚, and 10˚), ii. the location of the auditory (A) signal (-10˚, -3.3˚, 3.3˚, and 10˚), iii. the reliability of the visual signal (high [VR+] versus low [VR-] , as defined by the spread of the visual cloud), and iv. task relevance (auditory versus visual report). In addition, we included unisensory auditory and visual VR+ and VR- trials. **(B)** Time course of an example trial. **(C)** Behavioural audiovisual weight index w_{AV} estimated from behavioural responses (left) and from the predictions of the Bayesian Causal Inference model (right; across-participants circular mean ± 68% CI; n = 13). The audiovisual weight index w_{AV} is shown as a function of (i) visual reliability: high [VR+] versus low [VR-], (ii) task relevance: auditory [A] versus visual [V] report, (iii) audiovisual spatial disparity: small [≦6.6; D-] versus large [>6.6; D+],

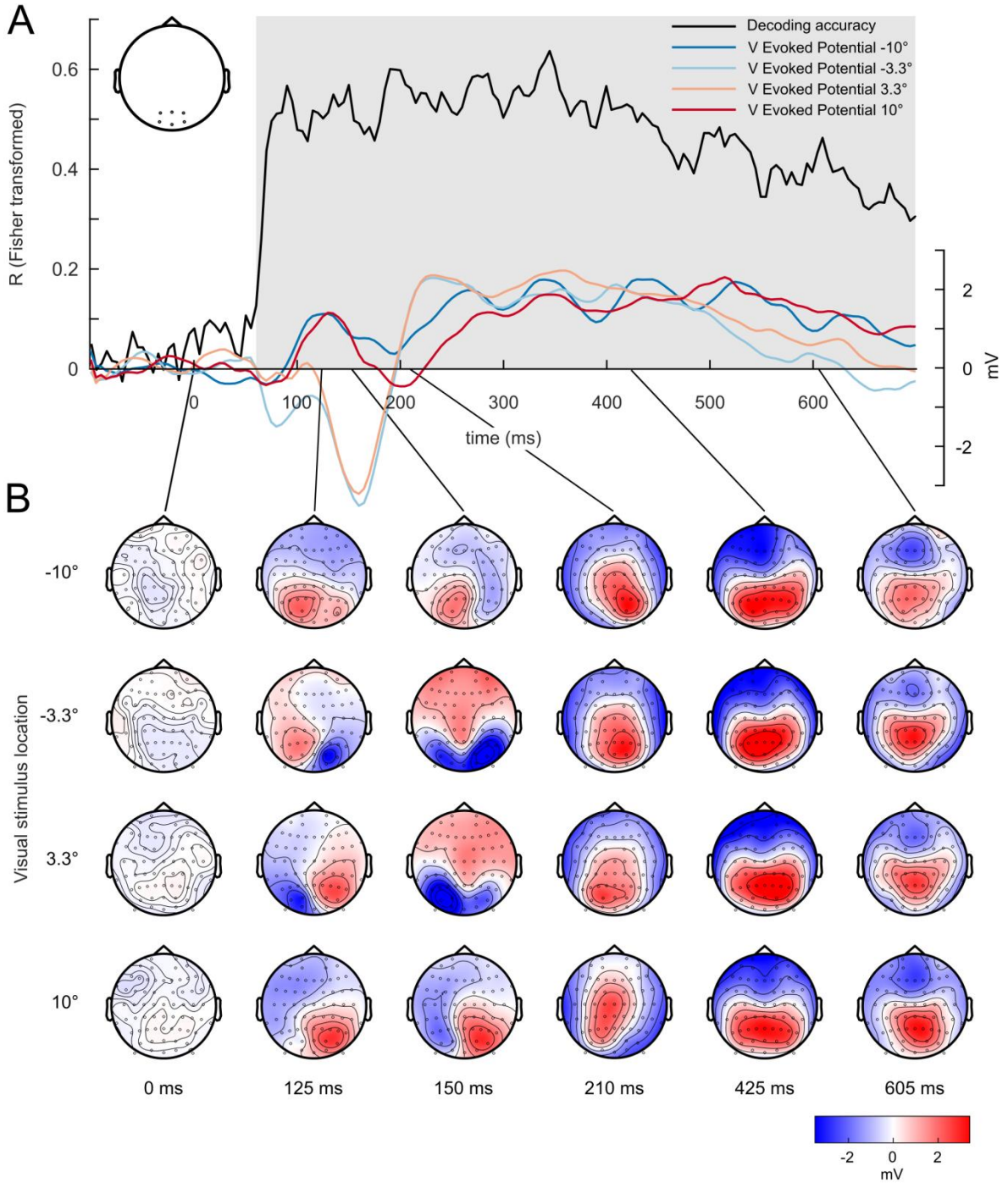

Figure 2 Time resolved decoding of visual stimulus location. (A) Time course of decoding accuracy (i.e. Pearson correlation between true and predicted visual stimulus locations, black line) and the grand average evoked potentials for the unisensory visual signals at -10°, -3.3°, +3.3°, 10° degree, averaged over occipital channels. Shaded grey area indicates the time window where the decoding accuracy is significantly better than chance. The inset shows the electrodes on which the evoked responses are based. **(B)** Grand average topographies for the unisensory visual signals at -10°, -3.3°, 3.3°, 10° degree at the given time points.

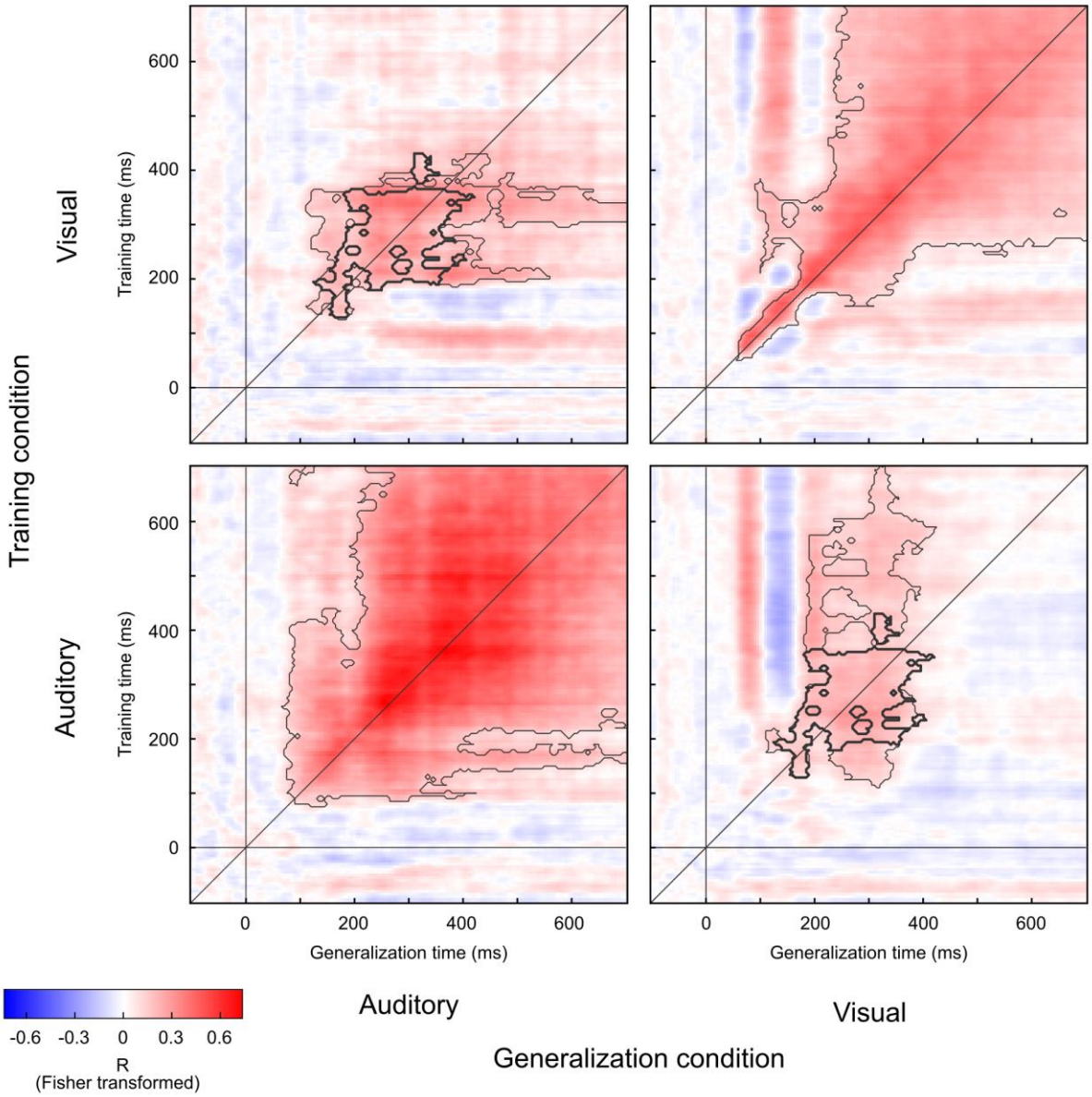

Figure 3 Temporal generalization matrices within and across auditory and visual senses. Each temporal generalization matrix shows the decoding accuracy for each training (y axis) and testing (x axis) time point. We factorially manipulated the training data (auditory vs. visual stimulation) and testing data (auditory vs. visual stimulation). The grey line along the diagonal indicates where the training time is equal to the testing time (i.e. the time resolved decoding accuracies). Horizontal and vertical grey lines indicate the stimulus onset. The thin black lines encircle clusters with decoding accuracies that were significantly better than chance at $p < 0.05$ corrected for multiple comparisons. The thick grey lines encircle the clusters with decoding accuracies that were significantly better than chance jointly for both i. auditory to visual and ii. visual to auditory cross-temporal generalization at $p < 0.05$ corrected for multiple comparisons.

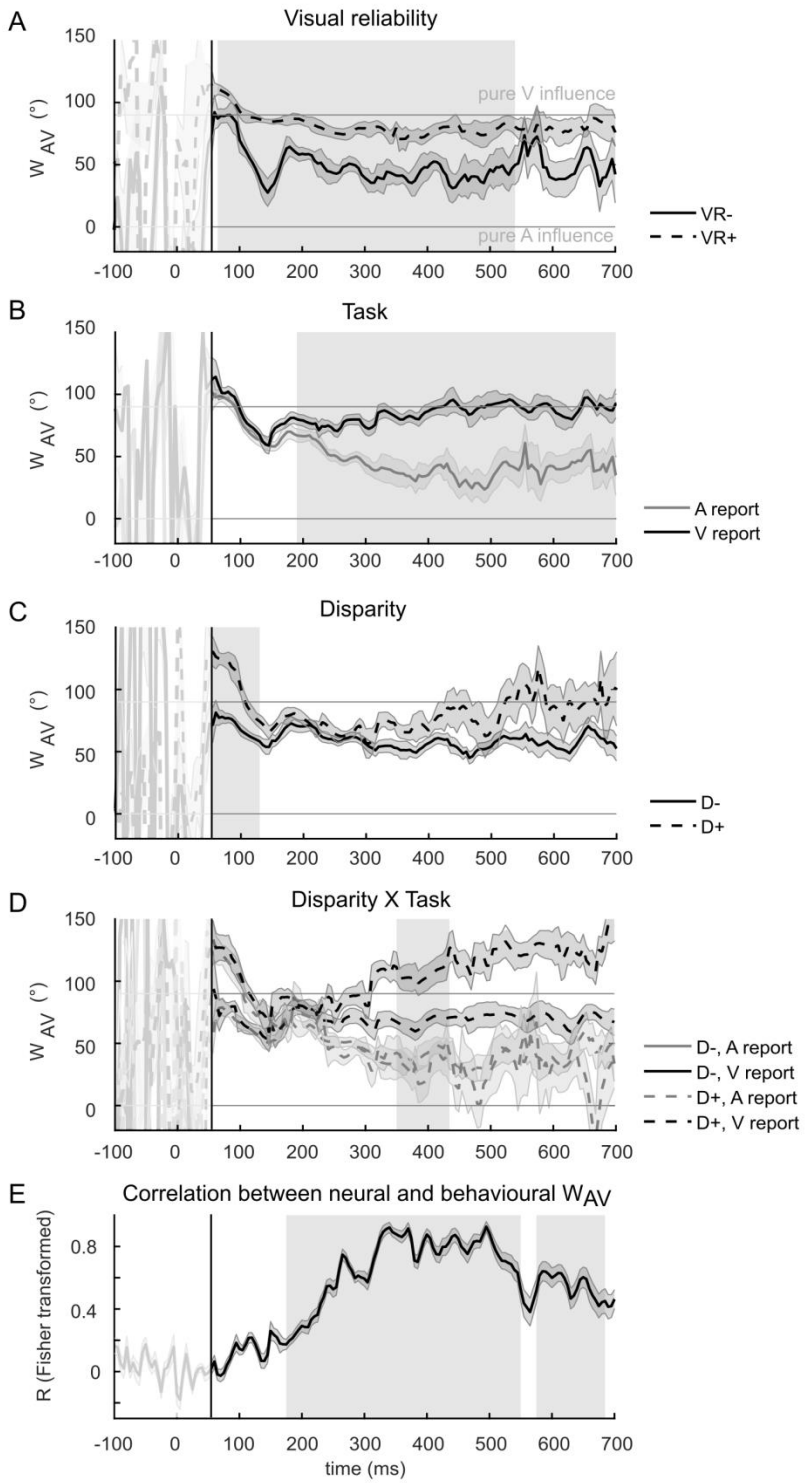

Figure 4 'Model free' EEG analysis results. The neural audiovisual weight index w_{AV} (across-participants' circular mean ± 68% CI; n = 13). Neural w_{AV} as a function of time is shown for **(A)** visual reliability: high [VR+] versus low [VR-], **(B)** task relevance: auditory [A] versus visual [V] report, **(C)** audiovisual spatial disparity: small [≦6.6; D-] versus large [>6.6; D+], **(D)** the interaction between task relevance x disparity. Shaded grey areas indicate the time windows where the main effect of **(A)** visual reliability, **(B)** task relevance, **(C)** audiovisual spatial disparity or **(D)** the interaction between task relevance x disparity on w_{AV} was statistically significant at $p < 0.05$ corrected for multiple comparisons across time. **(E)** Time course of circular-circular correlation (across-participants mean after Fisher z-transformation ± 68% CI; n = 13) between the neural- and the equivalent behavioural weight index w_{AV} (shaded if $p < 0.05$ corrected for multiple comparisons). The early time window

until 55 ms (delimited by black vertical line) is shaded in white, because the decoding accuracy was not greater than chance and hence the neural weight index w_{AV} is variable and uninterpretable in this window.

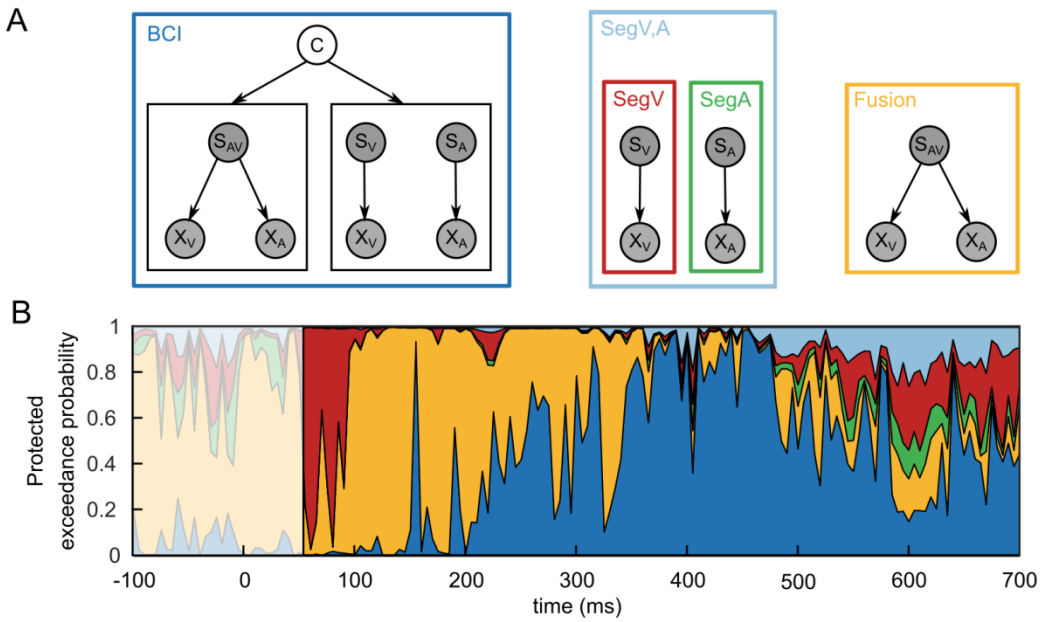

Figure 5 'Model based' EEG analysis results. (A) Schematics of the models that were fitted to observers' behavioural and neural responses: 'segregation unisensory auditory' (SegA, green), 'segregation unisensory visual' (SegV, red), 'full audiovisual segregation' (SegV,A, light blue), 'forced fusion' (Fusion, yellow) and 'Bayesian Causal Inference' model (with model averaging, BCI, dark blue). **(B)** Time course of the protected exceedance probabilities of the five models. The early time window until 55 ms (delimited by black vertical line) is shaded in white, because the decoding accuracy was not greater than chance and hence the model fit is not interpretable in this window.

Tables

	Behavioural	Neural
VR	p = 0.61	65 – 540 ms: p = 0.0002*
TR	p = 0.0002*	190 – 700 ms: p = 0.0002*
D	p = 0.77	50 – 130 ms: p = 0.0042*
VR x TR	p = 0.0002*	n.s.
VR x D	p = 0.0014*	45 – 100 ms: p = 0.0008* 170 – 235 ms: p = 0.016*
TR x D	p = 0.0002*	350 – 435 ms: p = 0.0182*
VR x TR x D	p = 0.79	n.s.
VR in A	p = 0.0002*	not tested
VR in V	p = 0.43	not tested
VR in D -	p = 0.47	115 – 170 ms: p = 0.0076* 245 - 480 ms: p = 0.0008*
VR in D +	p = 0.92	50 – 385 ms: p = 0.0002*
TR in D -	p = 0.0002*	265 – 310 ms: p = 0.039* 320 – 370 ms: p = 0.03* 380 – 530 ms: p = 0.001* 575 – 630 ms: p = 0.025*
TR in D +	p = 0.0002*	235 – 700 ms: p = 0.0002*

Table 1 Statistical significance of main, interaction and simple main effects for the behavioural and neural audiovisual weight indices (w_{AV}) ('model free' approach). TR = task relevance (visual, V or auditory, A, report), VR = visual reliability (high or low), D = disparity (small or large). Asterisks (*) denote significant results ($p < 0.05$, corrected at the cluster level for multiple comparisons), n.s. = not significant ($p \geqq 0.05$)

	p_c	σ_p	σ_A	σ_{V1}	σ_{V2}	R^2	$relBIC_{group}$	PEP
BCI	0.15 ± 0.04	36.4 ± 11.0	4.4 ± 0.2	0.3 ± 0.15	3.5 ± 0.24	0.86 ± 0.003	0	0.9992
Fus	-	71.5 ± 7.4	9.7 ± 0.3	7.9 ± 0.3	9.8 ± 0.35	$0.46 \pm 5*10^4$	$-3.73*10^4$	$2.9*10^{-4}$
SegV,A	-	42.7 ± 15.2	4.5 ± 0.3	0.4 ± 0.16	3.5 ± 0.23	0.85 ± 0.004	-865.1	$5.3*10^{-4}$

Table 2 Model parameters (across-subjects mean ± SEM) of the computational models fit to observers' behavioural localizations reports. R^2 = coefficient of determination, $relBIC_{group}$ = group level relative BIC, PEP = protected exceedance probability. BCI = Bayesian Causal Inference Model; Fus = Fusion Model; SegV,A = Full segregation audiovisual model

References

1. Shams L, Beierholm UR. Causal inference in perception. Trends Cogn Sci. Elsevier Ltd; 2010;14: 425–432. doi:10.1016/j.tics.2010.07.001
2. Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L. Causal inference in multisensory perception. PLoS One. 2007;2: e943. doi:10.1371/journal.pone.0000943
3. Alais D, Burr D. Ventriloquist effect results from rear-optimal bimodal integration. Curr Biol. 2004;14: 257–262. doi:10.1016/S0960-9822(04)00043-0
4. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002;415: 429–433. doi:10.1038/415429a
5. von Helmholtz H. Handbuch der physiologischen Optic (Vol. 3). Leipzig, Germany: Leopold Voss; 1867.
6. Fetsch CR, Turner AH, DeAngelis GC, Angelaki DE. Dynamic reweighting of visual and vestibular cues during self-motion perception. J Neurosci. 2009;29: 15601–15612. doi:10.1523/JNEUROSCI.2574-09.2009
7. Butler JS, Smith ST, Campos JL, Bülthoff HH. Bayesian integration of visual and vestibular signals for heading. J Vis. 2010;10: 23. doi:10.1167/10.11.23
8. van Beers RJ, Wolpert DM, Haggard P. When feeling is more important than seeing in sensorimotor adaptation. Curr Biol. 2002;12: 834–837. Available: http://www.ncbi.nlm.nih.gov/pubmed/12015120
9. Roach NW, Heron J, McGraw P V. Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. Proceedings Biol Sci. 2006;273: 2159–2168. doi:10.1098/rspb.2006.3578
10. Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA. Unifying multisensory signals across time and space. Exp Brain Res. 2004;158: 252–258. doi:10.1007/s00221-004-1899-9
11. Wozny DR, Beierholm UR, Shams L. Probability matching as a computational strategy used in perception. PLoS Comput Biol. 2010;6. doi:10.1371/journal.pcbi.1000871
12. Rohe T, Noppeney U. Cortical hierarchies perform Bayesian causal inference in multisensory perception. PLoS Biol. 2015;13: e1002073. doi:10.1371/journal.pbio.1002073
13. Rohe T, Noppeney U. Sensory reliability shapes perceptual inference via two mechanisms. J Vis. 2015;15: 22. doi:10.1167/15.5.22
14. Rohe T, Noppeney U. Distinct computational principles govern multisensory integration in primary sensory and association cortices. Curr Biol. Elsevier Ltd; 2016;1: 1–6. doi:10.1016/j.cub.2015.12.056
15. Bonath B, Noesselt T, Martinez A, Mishra J, Schwiecker K, Heinze H-J, et al. Neural basis of the ventriloquist illusion. Curr Biol. 2007;17: 1697–703. doi:10.1016/j.cub.2007.08.050
16. Friston K. A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci. 2005;360: 815–836. doi:10.1098/rstb.2005.1622
17. Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci. 1999;2: 79–87. doi:10.1038/4580
18. Talsma D. Predictive coding and multisensory integration: an attentional account of the multisensory mind. Front Integr Neurosci. 2015;09: 19. doi:10.3389/fnint.2015.00019
19. King J-R, Dehaene S. Characterizing the dynamics of mental representations: the temporal generalization method. Trends Cogn Sci. Elsevier Ltd; 2014;18: 203–10. doi:10.1016/j.tics.2014.01.002
20. Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies - Revisited. Neuroimage. Elsevier Inc.; 2014;84: 971–985. doi:10.1016/j.neuroimage.2013.08.065
21. Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE. Neural correlates of reliability-based cue weighting during multisensory integration. Nat Neurosci. 2012;15: 146–54. doi:10.1038/nn.2983

22. Grothe B, Pecka M, McAlpine D. Mechanisms of sound localization in mammals. Physiol Rev. 2010;90: 983–1012. doi:10.1152/physrev.00026.2009
23. Gardner JL, Merriam EP, Movshon JA, Heeger DJ. Maps of Visual Space in Human Occipital Cortex Are Retinotopic, Not Spatiotopic. J Neurosci. 2008;28: 3988–3999. doi:10.1523/JNEUROSCI.5476-07.2008
24. Wandell BA, Dumoulin SO, Brewer AA. Visual field maps in human cortex. Neuron. 2007. pp. 366–383. doi:10.1016/j.neuron.2007.10.012
25. Werner-Reiss U, Groh JM. A rate code for sound azimuth in monkey auditory cortex: implications for human neuroimaging studies. J Neurosci. 2008;28: 3747–3758. doi:10.1523/JNEUROSCI.5044-07.2008
26. Ortiz-Rios M, Azevedo FAC, Kuśmierek P, Balla DZ, Munk MH, Keliris GA, et al. Widespread and Opponent fMRI Signals Represent Sound Location in Macaque Auditory Cortex. Neuron. 2017;93: 971–983. doi:10.1016/j.neuron.2017.01.013
27. Mullette-Gillman OA, Cohen YE, Groh JM. Eye-centered, head-centered, and complex coding of visual and auditory targets in the intraparietal sulcus. J Neurophysiol. 2005;94: 2331–52. doi:10.1152/jn.00021.2005
28. Bertelson P, Radeau M. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. Percept Psychophys. 1981;29: 578–584. Available: http://www.ncbi.nlm.nih.gov/pubmed/7279586
29. Lakatos P, Chen C-M, O'Connell MN, Mills A, Schroeder CE. Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron. 2007;53: 279–292. doi:10.1016/j.neuron.2006.12.011
30. Kayser C, Petkov CI, Augath M, Logothetis NK. Functional Imaging Reveals Visual Modulation of Specific Fields in Auditory Cortex. J Neurosci. 2007;27: 1824–1835. doi:10.1523/JNEUROSCI.4737-06.2007
31. Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ. Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. Brain Res Cogn Brain Res. 2002;14: 115–28. Available: http://www.ncbi.nlm.nih.gov/pubmed/12063135
32. Noesselt T, Rieger JW, Schoenfeld MA, Kanowski M, Hinrichs H, Heinze H-J, et al. Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. J Neurosci. 2007;27: 11431–11441. doi:10.1523/JNEUROSCI.2252-07.2007
33. Lewis R, Noppeney U. Audiovisual Synchrony Improves Motion Discrimination via Enhanced Connectivity between Early Visual and Auditory Areas. J Neurosci. 2010;30: 12329–12339. doi:10.1523/JNEUROSCI.5745-09.2010
34. Werner S, Noppeney U. Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. J Neurosci. 2010;30: 2662–2675. doi:10.1523/JNEUROSCI.5091-09.2010
35. Lee H, Noppeney U. Temporal prediction errors in visual and auditory cortices [Internet]. Current Biology. 2014. pp. R309–R310. doi:10.1016/j.cub.2014.02.007
36. Schroeder CE, Foxe J. Multisensory contributions to low-level, "unisensory" processing. Current Opinion in Neurobiology. 2005. pp. 454–458. doi:10.1016/j.conb.2005.06.008
37. Atilgan H, Town SM, Wood KC, Jones GP, Maddox RK, Lee AKC, et al. Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding. Neuron. 2018;97: 640–655.e4. doi:10.1016/j.neuron.2017.12.034
38. Butler JS, Foxe JJ, Fiebelkorn IC, Mercier MR, Molholm S. Multisensory representation of frequency across audition and touch: high density electrical mapping reveals early sensory-perceptual coupling. J Neurosci. 2012;32: 15338–44. doi:10.1523/JNEUROSCI.1796-12.2012
39. Fetsch CR, DeAngelis GC, Angelaki DE. Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. Nat Rev Neurosci. Nature Publishing

Group; 2013;14: 429–42. doi:10.1038/nrn3503
40. Morgan ML, Deangelis GC, Angelaki DE. Multisensory integration in macaque visual cortex depends on cue reliability. Neuron. 2008;59: 662–73. doi:10.1016/j.neuron.2008.06.024
41. Nardo D, Santangelo V, Macaluso E. Spatial orienting in complex audiovisual environments. Hum Brain Mapp. 2014;35: 1597–1614. doi:10.1002/hbm.22276
42. Brainard DH. The Psychophysics Toolbox. Spat Vis. 1997;10: 433–436. doi:10.1163/156856897X00357
43. Oostenveld R, Fries P, Maris E, Schoffelen JM. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci. 2011;2011: 156869. doi:10.1155/2011/156869
44. Chang C-C, Lin C-J. LIBSVM : a library for support vector machines. ACM Trans Intell Syst Technol. 2011;2: 1–27. doi:10.1145/1961189.1961199
45. Nichols T, Holmes A. Nonparametric Permutation Tests for Functional Neuroimaging. Human Brain Function: Second Edition. 2003. pp. 887–910. doi:10.1016/B978-012264841-0/50048-2
46. Anderson CM, Wu CFJ. Measuring Location Effects from Factorial Experiments with a Directional Response. Int Stat Rev. 1995;63: 345–363. Available: http://www.jstor.org/stable/1403484%5Cnhttp://about.jstor.org/terms
47. Edgington ES, Onghena P. Randomization Tests [Internet]. 4th ed. Chapman & Hall/CRC; 2007. Available: https://www.crcpress.com/Randomization-Tests-Fourth-Edition/Edgington-Edgington-Onghena/p/book/9781584885894
48. Gonzalez L, Manly BFJ. Analysis of variance by randomization with small data sets. Environmetrics. 1998;9: 53–65. doi:10.1002/(SICI)1099-095X(199801/02)9:1<53::AID-ENV285>3.0.CO;2-#
49. Berens P. CircStat: A MATLAB toolbox for circular statistics. J Stat Softw. 2009;31: 1–21. doi:10.1002/wics.10
50. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78: 691–692. doi:10.1093/biomet/78.3.691
51. Kass RE, Raftery AE. Bayes Factors. J Am Stat Assoc. 1995;90: 773–795. doi:10.1080/01621459.1995.10476572
52. Friston K, Holmes AP, Worsley KJ, Poline J-P, Frith CD, Frackowiak RSJ. Statistical parametric maps in functional imaging: A general linear approach. Hum Brain Mapp. 1994;2: 189–210. doi:10.1002/hbm.460020402