The neural dynamics of hierarchical Bayesian causal inference in multisensory perception

Tim Rohe¹, Ann-Christine Ehlis¹,² & Uta Noppeney³

Transforming the barrage of sensory signals into a coherent multisensory percept relies on solving the binding problem – deciding whether signals come from a common cause and should be integrated or, instead, segregated. Human observers typically arbitrate between integration and segregation consistent with Bayesian Causal Inference, but the neural mechanisms remain poorly understood. Here, we presented people with audiovisual sequences that varied in the number of flashes and beeps, then combined Bayesian modelling and EEG representational similarity analyses. Our data suggest that the brain initially represents the number of flashes and beeps independently. Later, it computes their numbers by averaging the forced-fusion and segregation estimates weighted by the probabilities of common and independent cause models (i.e. model averaging). Crucially, prestimulus oscillatory alpha power and phase correlate with observers' prior beliefs about the world's causal structure that guide their arbitration between sensory integration and segregation.
In everyday life, the brain is constantly confronted with a myriad of sensory signals. Imagine you are skipping stones on a lake. Each time the stone bounces off the water's surface, you see the impact and hear a brief splash. Should you integrate or segregate signals from vision and audition to estimate how many times the stone hits the water's surface? Hierarchical Bayesian Causal Inference provides a rational strategy to arbitrate between information integration and segregation by explicitly modelling the underlying potential causal structures, i.e. whether visual impacts and splash sounds are caused by common or independent events1-2. Under the assumption of a common cause, signals are integrated weighted by their relative precisions (or reliabilities, i.e. the reciprocal of variance) into one single ‘forced-fusion’ numeric estimate. If, however, some splash sounds are caused by a stone hitting the water surface out of the observer’s sight (e.g. another person throwing a stone), audition and vision will provide conflicting information. In this segregation case, the brain needs to estimate the number of events independently for vision and audition. Importantly, the brain cannot directly access the causal structure of the world and audition. It needs to estimate the number of events independently for vision and audition. Crucially, prestimulus oscillatory alpha activity (i.e. alpha-, beta- and gamma-oscillations15-16) prior to stimulus onset may modulate the causal prior (i.e. observers’ belief about the world’s causal structure) or precision of sensory representations (e.g. visual variance) in early visual cortices and thereby in turn influence the outcome of the Bayesian Causal Inference. We combined psychophysics, computational modelling and EEG representational similarity analyses to characterize the neural dynamics of Bayesian Causal Inference in perception of audiovisual stimulus sequences.

Our data suggest that the brain initially represents the number of audiovisual flashes and beeps independently. Later, it computes their numbers by averaging the forced-fusion and segregation estimates weighted by the posterior probabilities of a common and independent causes. Crucially, prestimulus oscillatory alpha and gamma power as well as alpha phase correlate with observers’ causal priors. By resolving the computational operations of multisensory interactions in human neocortex in time, our study shows that the brain dynamically encodes and re-updates computational priors and multiple numeric estimates to perform hierarchical Bayesian Causal Inference.

Results
Experimental design and analysis. During the EEG recording, we presented 23 human observers with sequences of auditory beeps and visual flashes in a four (1 to 4 flashes) × four (1 to 4 beeps) factorial design (Fig. 1). Participants estimated and reported either the number of flashes or the number of beeps. We combined a general linear model (GLM) and a Bayesian modelling analysis to characterize the computations and neural mechanisms of how the brain combines information from vision and audition to estimate the number of auditory and visual stimuli.

Behavior – Audiovisual weight index and Bayesian modelling. Using a GLM (i.e. regression) approach, we computed a relative audiovisual weight index \( w_{AV} \) that quantifies the relative influence of the true number of beeps and flashes on participants’ numeric reports. The audiovisual weight index \( w_{AV} \) was analyzed as a function of numeric disparity between beeps and flashes (i.e. small \( s \) vs. large \( e \)) × task relevance (visual vs. auditory report). This audiovisual weight range differs from pure visual (90°) to pure auditory (0°) influence. As shown in Figs. 1c and 2a, observers’ reported number of beeps was mainly influenced by the true number of beeps and only slightly – but significantly – biased by the true number of flashes (circular mean \( w_{AV} = 3.871°, p < 0.001 \), one-sided randomization test on \( w_{AV} > 0° \); i.e. a visual bias on auditory perception15,18). Conversely, the reported number of flashes was biased by the true number of beeps (circular mean \( w_{AV} = 65.483°, p < 0.001 \), one-sided randomization test on \( w_{AV} < 90° \)), which is consistent with the well-known ‘sound-induced flash illusion8,17,18’. Yet, despite these significant biases operating from vision to audition and vice versa, observers did not fuse stimuli into one unified percept. Instead, the visual influence was stronger when the number of flashes was reported and the auditory influence was stronger when the number of
beeps was reported (effect of task on \( w_{AV} \), \( \text{LRTS} = 85.620, p < 0.001 \), randomization test of a likelihood-ratio test statistic (LRTS); Table 1). As a result, observers reported different perceived numbers of flashes and beeps for audiovisual stimuli with a numeric disparity. Thus, participants flexibly adjusted the weights according to the task-relevant sensory modality. Crucially, this difference between auditory and visual report increased significantly for large relative to small numeric disparities. In other words, audiovisual integration broke down for large numeric disparities, when auditory and visual stimuli were more likely to be caused by independent sources (significant interaction between task relevance and numeric disparity, \( \text{LRTS} = 1.761, p < 0.001 \); for analysis of response times, see Supplementary Notes and Supplementary Fig. 1).

Indeed, the model predictions in Fig. 1c show that this interaction between task relevance and numeric disparity is a key feature of Bayesian Causal Inference. As this behavioral profile can be accounted for neither by the classical forced-fusion model that assumes audiovisual stimuli are fused into one single estimate (i.e. common-source model) nor by the full-segregation model (i.e. independent-source model), the Bayesian Causal Inference model was the winning model for explaining the behavioral data based on formal Bayesian model comparison (Table 2). Further, the decisional function ‘model averaging’ outperformed ‘model selection’ and ‘probability matching’ at the group level (see Supplementary Table 1, consistent with\(^5\), but see\(^19\)). In the following, we will therefore focus selectively on Bayesian Causal Inference with model averaging.

**EEG – Conventional univariate ERP analysis.** Event-related potentials (ERPs) revealed the typical sequence of ERP components in response to audiovisual flashes and beeps (Fig. 1d), i.e. P1 (~50 ms), N1 (~100 ms), P2 (200 ms), N2 (280 ms) and P3 (~300 ms)\(^{20,21}\). In line with previous studies\(^{20,21}\), we observed early multisensory interactions in the classical ‘sound-induced flash illusion’ comparison (i.e. \( A_1V_1A_2 \) vs. \( A_1A_2 + V_1 \); Fig. 1e) over occipital electrodes starting at about ~70 ms (i.e. measured from the onset of the first flash-beep slot). Further, we observed a negative audiovisual-interaction component at 335–730 ms after stimulus onset. However, the current study did not focus on early multisensory interactions as evidenced in ERPs, but on the neural dynamics underlying Bayesian Causal Inference in perceptual decision-making.

**EEG – Multivariate decoding and audiovisual weight index.** To compute a neural audiovisual weight index \( w_{AV} \), we applied multivariate pattern analysis to single-trial EEG activity patterns.
(i.e. 64 electrodes) of 20 ms time intervals. We trained a support-vector regression model on EEG activity patterns independently at each time point of the audiovisual congruent conditions to establish a mapping between EEG activity patterns and number of audiovisual stimuli. We then generalized to the congruent and incongruent conditions (i.e. leave-one-run-out cross-validation).

First, we ensured that we could decode the stimulus number for congruent trials significantly better than chance. Indeed, the decoder was able to discriminate between, for instance, three and four flash-beeps nearly immediately after the presentation of the fourth flash-beep (Fig. 3a) and thus before the ERP traces, when averaged over parietal electrodes, started to diverge (Fig. 1d). Pooling over all four congruent conditions, we observed better than chance decoding accuracy from around 100 ms to 740 ms measured from the onset of the first flash-beep slot (Fig. 3b).

We applied the same analysis approach as for behavioral responses to the audiovisual decoded numeric estimates and computed the neural audiovisual weight index \( w_{AV} \) which...
Fig. 2 Representational dissimilarity matrices (RDMs) for numeric reports and estimates of the BCI model. a Participants’ auditory and visual numeric reports (across-participants’ mean; n = 23) are plotted as a function of the true number of visual (#V) and auditory (#A) stimuli, separately for auditory (top) and visual report (bottom). The auditory reports are more strongly influenced by the true number of auditory stimuli, while the visual reports are more strongly influenced by the number of visual stimuli. Yet, crossmodal biases are also present. b RDM (across-participants’ mean) showing the absolute differences in participants’ numeric reports between all pairs of the 32 experimental conditions. The true number of bees (green) and flashes (red) for each condition is indicated by the number of dots. c RDMs (across-participants’ mean) computed from the numeric estimates and the estimate of the posterior probability of a common cause of the BCI model: The generative BCI model assumes that in case of a common cause (C = 1), the “true” number of audiovisual stimuli N_{AV} is drawn from a common numeric prior distribution leading to noisy auditory (n_A) and visual (n_V) inputs. In case of independent causes (C = 2), the “true” auditory (N_A) and visual (N_V) numbers of stimuli are drawn independently from the numeric prior distribution. To account for the causal uncertainty, the final BCI estimate of the auditory or visual stimulus number (N_A, or N_V, depending on whether auditory or visual modality is reported) is computed by combining the forced-fusion estimate of the auditory and visual stimuli (N^fl_{AV,C=1}) with the task-relevant unisensory visual (N^fl_{V,C=1}) or auditory estimates (N^fl_{A,C=1}), each weighted by the posterior probability of a common (C = 1) or independent (C = 2) causes, respectively (p(C|n_A,n_V)). Arrows indicate the influence of component estimates on the RDM of the final BCI estimate. The probability of independent causes (dashed) increases with larger numeric disparity and vice versa for the probability of a common cause (n.b. they sum to unity). Source data are provided as a Source Data file.

**Table 1** Effects of numeric disparity (D) and task relevance (T) on the behavioral and the neural audiovisual weight index w_{AV} computed from numeric estimates decoded from EEG activity patterns (i.e. significant clusters)

| Effect | LRTS | p    | Cluster # | Time (ms) | p    |
|--------|------|------|-----------|-----------|------|
| T      | 85.62 | <0.001 | 1         | 420–540   | <0.001|
| D      | 1.624 | <0.001 | 2         | 620–660   | 0.007 |
| T X D  | 1.761 | <0.001 | 2         | 400–480   | <0.001|
| T X D  | 1.761 | <0.001 | 2         | 560–620   | 0.039 |
| T X D  |       |       | 2         | 680–720   | 0.001 |

Note: p values are based on randomization tests using a circular log likelihood-ratio statistic (LRTS). Effects for the neural w_{AV} are cluster-based corrected for the multiple time points. Only significant (i.e. p < 0.05) clusters are reported, n = 23.

Table 1 quantified the relative auditory and visual influences on the decoded number of flashes and bees across poststimulus time (i.e. from 100 ms to 740 ms). We assessed how the neural audiovisual weight index was affected by numeric disparity between bees and flashes (i.e. small ≤ 1 vs. large ≥ 2) and task relevance (visual vs. auditory report) in a 2 × 2 repeated-measures analysis (Fig. 3c and Table 1). We observed that the neural influence was stronger for small relative to large numeric disparities from 400 to 480 ms poststimulus (i.e. effect of numeric disparity: 200–280 ms after the final flash-beep slot). Only when the numeric disparity was small and hence the two stimuli were likely to come from a common cause, auditory stimuli impacted the neural estimation of the number of flashes, which dominated the EEG activity patterns. Shortly later, i.e. 420–540 ms poststimulus, the influence of the auditory and visual stimuli on the decoded numeric estimate also depended on the sensory modality that needed to be reported (effect of task relevance; for additional effects see Table 1). The number of flashes influenced the decoded numeric estimates more strongly for visual report, whereas the number of bees influenced the decoded numeric estimates for auditory report. Crucially, at 560 ms and from 680 to 720 ms poststimulus, we observed a significant interaction between task relevance and numeric disparity, which is the key profile of Bayesian Causal Inference. As predicted by Bayesian Causal Inference (cf. Fig. 1c), the audiovisual weight indices for auditory and visual report were similar (i.e. integration) for small numeric disparity, but diverged (i.e. segregation) for large numeric disparities when it is unlikely that the flash and the beep sequences were generated by a common cause.

**EEG – Representational geometry of the numeric estimates and activity patterns.** Using representational similarity analysis, we compared the representational geometry of the full-segregation auditory or visual, forced-fusion and the final Bayesian Causal Inference (BCI) estimates with the representational geometry of observers’ numeric reports (Fig. 2) and EEG activity patterns across poststimulus time (Fig. 4). First, we estimated the representational dissimilarity matrices (RDMs) by computing the pairwise absolute distance between the BCI model’s four numeric estimates, i.e. (i) the forced-fusion, the full-segregation (ii) auditory and (iii) visual and (iv) the final BCI estimates as well as the posterior causal probability across all 32 conditions. As shown in Fig. 2c, the RDM for the forced-fusion estimate (N^fl_{AV,C=1}) was a weighted average of the RDMs of the full-segregation auditory (N^fl_{A,C=1}) and visual (N^fl_{V,C=1}) estimates. Further, because the auditory modality provides more precise temporal information (cf. Table 2) which is crucial for estimating the number of stimuli, the forced-fusion RDM is more similar to the auditory than the visual RDM. Finally, the RDM for the BCI estimate (i.e. N_{A} or N_{V}, depending on the sensory modality that needs to be reported) combined the forced-fusion estimate (N^fl_{AV,C=1}) with the task-relevant unisensory visual (N^fl_{V,C=1}) or auditory (N^fl_{A,C=1}) estimates (depending on report), weighted by the posterior probability of a common or separate causes, respectively (i.e. p(C = 1|n_A,n_V) or p(C = 2|n_A,n_V)). The probability of a common cause increased with smaller numeric disparity such that the influence of the forced-fusion estimate was greater for small numeric disparities. Figure 2b illustrates that the RDM computed from observers’ behavioral numeric reports was nearly identical to the BCI RDM. This match was confirmed statistically by a high correlation between the BCI RDM (i.e. N_{A} or N_{V}) and participants’ behavioral RDM (r = 0.878 ± 0.059, mean ± SEM, f_{22} = 59.806, p < 0.001, Cohen’s d = 12.471; one-sample t-test on Fisher’s z-transformed rank correlations against zero). Of course, this match between behavioral and BCI RDM was expected because the BCI RDM was computed from the predictions of the BCI model that well fit participants’ numeric reports (i.e. circular dependency; cf. Table 2).

Next, we characterized the neural dynamics of Bayesian Causal Inference by comparing the representational geometry obtained from EEG activity patterns across time with the representational geometries of (i) the forced-fusion, the full-segregation (ii) auditory and (iii) visual and (iv) the final BCI estimates. As shown in Fig. 4a, the RDMs obtained from EEG activity patterns significantly correlated with the unisensory auditory RDM (N^fl_{A,C=1}; significant cluster 60–740 ms, p < 0.001, one-sided cluster-based corrected randomization f_{22} test), the unisensory visual RDM (N^fl_{V,C=1}; cluster 100–720 ms, p < 0.001), the forced-fusion RDM (N^fl_{AV,C=1}; cluster 80–740 ms, p < 0.001) and the BCI RDM (N_{A} or N_{V}; significant cluster 80–740 ms, p < 0.001). In
short, the RDMs of EEG activity patterns correlated with multiple numeric estimates simultaneously. For the posterior probability of a common cause ($p(C=1|x_A, x_V)$), the correlation was weaker but significant in a later cluster (260–640 ms after stimulus onset, $p < 0.001$). The strong and sustained correlations of EEG RDMs and the RDMs of the four numeric estimates from the BCI model were expected because the four numeric estimates were highly correlated with one another. Hence, to account for these inherent correlations between these numeric estimates, we next computed the exceedance probability (i.e. the probability that the correlation with one numeric RDM was greater than that of any other RDMs) to determine which of the four numeric estimates was most strongly represented in the EEG activity patterns at a given time point (Fig. 4b). The exceedance probabilities showed that the EEG activity patterns predominantly encoded the unisensory visual estimate from 120 ms up to around 500 ms (i.e. 300 ms after the final flash-beep slot). This visual over auditory influence on EEG activity patterns at the scalp may be surprising, because the auditory sense exerts a stronger influence on observers’ reported numeric estimates (Fig. 1c) and provides more precise temporal information when estimated from observers’ numeric reports (i.e. $\sigma_A$ is smaller than $\sigma_V$ in Table 2). Potentially, the visual neural sources elicit EEG activity patterns in sensor space that are more informative about the number of events (see methods section for caveats and critical discussion of the decoding analysis). Indeed, additional multivariate decoding analyses of the unisensory auditory and visual conditions showed that the number of visual stimuli could be more accurately decoded from visual EEG activity patterns than the number of auditory stimuli from auditory EEG activity patterns (Supplementary Fig. 2). Potentially, this advantage for visual decoding under unisensory stimulation may further increase in an audiovisual context when the visual signal is task-relevant because of additional attentional amplification.

Crucially, from 450 ms poststimulus (i.e. 250 ms after the presentation of the final flash-beep; Fig. 4b), the EEG representational geometries progressively reflected the BCI estimate. Collectively, the model-based representational similarity analysis suggests that Bayesian Causal Inference evolves by dynamic encoding of multiple sensory estimates. First, the EEG activity is dominated by the numeric unisensory and forced-fusion estimates (i.e. $N_V(C=2)$, $N_A(C=2)$, and $N_{AV}(C=1)$) and later by the BCI estimate (i.e. $\hat{N}_A$ or $\hat{N}_V$) that takes into account the observers’ uncertainty about the world’s causal structure.

### Table 2 Results of the Bayesian model comparison of the Bayesian Causal Inference, the forced-fusion and the full-segregation model

| Model                      | $p_{\text{common}} \pm \sigma_p$ | $\sigma_p \pm \sigma_f$ | $\sigma_A \pm \sigma_V$ | $R^2$ | relBIC | $p_{\text{EP}}$ | % win |
|---------------------------|---------------------------------|-------------------------|-------------------------|-------|--------|-----------------|-------|
| Causal Inference (model averaging) | 0.42 ± 0.05                     | 2.26 ± 0.20             | 2.34 ± 0.29             | 0.53 ± 0.03 | 1.11 ± 0.23 | 0.874 ± 0.012 | 0     |
| Forced fusion             | -                               | 2.10 ± 0.22             | 4.38 ± 0.60             | 1.17 ± 0.04 | 1.45 ± 0.04 | 0.617 ± 0.016 | 8362.42 | 0     |
| Full segregation          | -                               | 2.15 ± 0.32             | 2.11 ± 0.32             | 0.55 ± 0.03 | 1.01 ± 0.11 | 0.846 ± 0.015 | 920.70 | 0     |

Note: $p_{\text{common}}$, causal prior; $\mu_p$, mean of the numeric prior; $\sigma_p$, standard deviation of the numeric prior; $\sigma_f$, standard deviation of the auditory likelihood; $\sigma_V$, standard deviation of the visual likelihood; $R^2$, coefficient of determination; relBIC, Bayesian information criterion at the group level, i.e. participant-specific BICs summed over all participants; BIC =$−$LL + $−0.5 \times \text{n}(\text{LL})$, LL =$−$log likelihood, $n =$ number of parameters, $n =$ number of data points) of a model relative to the Bayesian Causal Inference (“model averaging”) model (i.e. a smaller relBIC indicates that a model provides a better explanation of our data); $p_{\text{EP}}$, protected exceedance probability, i.e. the probability that a given model is more likely than any other model, beyond differences due to chance). % win, percentage of participants in which a model won the within-participant model comparison based on BIC.

**EEG – Effect of prestimulus oscillations on the causal prior.**

Previous research demonstrated that observers perceived a sound-induced flash illusion more often for low prestimulus alpha power and high beta as well as gamma power over occipital (i.e. visual) cortices. Within the framework of Bayesian Causal Inference, the occurrence of a sound-induced flash illusion may increase when visual precision is reduced or the causal prior (i.e. the probability of a common vs. independent causes, also known as binding tendency) is enhanced. We therefore investigated whether prestimulus oscillatory power (over occipital electrodes) alters participants’ multisensory perception as parameterized by the causal prior ($p_{\text{common}}$) or the precision of visual representations (i.e. the reciprocal of $\sigma_V$). For this, we sorted the trials into 10 deciles according to oscillatory power for each time and frequency point and re-fitted the causal prior or the precision of visual representations in the BCI model separately for each bin. Next, we computed the correlation of the causal prior or the visual precision with oscillatory power over deciles. This analysis showed that the causal prior correlated positively with gamma power ($p =$ 0.036, two-sided cluster-based corrected randomization $t_{23}$ test, starting at $−220$ ms prestimulus to stimulus onset) and negatively with alpha power ($p =$ 0.042, from $−320$ ms to $−100$ ms prestimulus, Fig. 5a, b). Figure 5c shows the weight index $w_{\text{AV}}$ computed from participants’ behavior for each decile and the corresponding model predictions. Both human and model behavior showed more audiovisual influences (i.e. $w_{\text{AV}}$ indices shifted towards 0.5) for high gamma power and low alpha power. Crucially, these audiovisual biases operated from vision to audition and vice versa (i.e. a bidirectional bias which cannot be modelled by changes in visual precision). Hence, prestimulus gamma and alpha oscillations tune how the brain arbitrates between sensory integration and segregation. High gamma and low alpha power prior to stimulus presentation increase the brain’s tendency to bind stimuli across the senses. For completeness, we did not observe any significant effect of oscillatory power on the visual precision (Supplementary Fig. 3A).

Given the prominent role of alpha oscillations in temporal binding in visual and multisensory perception, we next investigated whether the prestimulus alpha phase influenced the causal prior or visual precision. Using a similar sort-and-binning approach as for prestimulus power, we computed a circular-linear correlation between alpha phase and causal prior (or visual precision) over deciles as a function of prestimulus time. While there was again no significant effect of alpha phase on visual precision (Supplementary Fig. 3B), we observed a significant cluster from $−160$ ms to $−80$ ms prestimulus ($p = 0.015$, one-sided cluster-based corrected randomization $t_{23}$ test), where alpha phase correlated significantly with participants’ causal prior (Fig. 6a): trials with a specific alpha phase led to a higher causal prior than trials with an opposing alpha phase. Importantly, the relation between alpha phase and causal prior progressed consistently over time at alpha frequency (i.e. 10 Hz; Fig. 6c).

In support of this, a sinusoidal model in which the phase of an alpha oscillation modulated the causal prior outperformed a model that did not include a sinusoidal modulation from $−280$ ms to $−80$ ms prestimulus in 20 out of 23 participants (individual $F_{2,107}$ tests, $p < 0.05$, Fig. 6b; see Supplementary Fig. 4 for individual data from four representative participants).
However, the relation of alpha phase and causal prior was not consistent across participants ($z_{22} = 2.486, p = 0.082$, Raleigh test, Fig. 6b). These differences between participants are expected and may arise from differences in cortical folding and hence orientations of the underlying neural sources. To account for these differences across participants, we therefore aligned the alpha phase individually for each participant, such that the phase at the peak group effect at $-160$ ms prestimulus was consistent across participants (cf. Supplementary Fig. 5 for data without phase-alignment). Figure 6c, d shows that the alpha phase modulates the causal prior across nearly three cycles which is consistent across participants. Collectively, these results demonstrate that the power and phase of prestimulus alpha oscillations influence observers’ causal prior, which formally quantifies their...
apriori tendency to bind signals from audition and vision into a coherent percept.

**EEG – The relation of prior stimulus history, prestimulus alpha power and the causal prior.** Previous research has shown that prior stimulus history influences observers’ binding tendency\(^{24–26}\). For instance, prior congruent audiovisual speech stimuli increased observers’ tendency to bind incongruent audiovisual signals into illusionary McGurk percepts\(^{24}\). Hence, we investigated whether the numeric disparity of previous flash-beep stimuli (going back in history to five trials prior to stimulus onset) influenced observers’ causal prior on the current trial. Indeed, as shown in Fig. 7a, a 2 (numeric disparity: small vs. large) × 5 (stimulus order: 1, 2, 3, 4, 5 trials back) repeated-measures ANOVA revealed a significant main effect of numeric disparity \(F_{1,21} = 6.260, p = 0.021\), partial \(\eta^2 = 0.230\) and a significant interaction between numeric disparity and stimulus order \(F_{2,6,54.9} = 4.060, p = 0.015\), partial \(\eta^2 = 0.162\); Greenhouse-Geisser corrected df). Post-hoc tests for the effect of numeric disparity separately for specific stimulus order showed that the effect of numeric disparity was most pronounced for the first- and second-order previous stimulus (first order: \(t_{22} = 3.731, p = 0.001\), Cohen’s \(d = 0.778\); marginally significant second order: 

\[
t_{22} = 2.042, p = 0.053, \text{Cohen’s } d = 0.426; \text{two-sided paired } t\text{-tests) and tapered off with stimulus order.}
\]

Our results so far suggest that previous stimulus history (i.e. numeric disparity of previous trials) and pre-stimulus alpha power predict observers’ tendency to bind audiovisual signals (Fig. 7b). This raises the intriguing question whether alpha power mediates the effect of previous stimulus history. For instance, given the well-established role of pre-stimulus alpha oscillations in visual perception\(^{27–30}\) and attention\(^{31}\), one may argue that alpha power is adjusted according to observers’ causal expectations based on prior stimulus history. Contrary to this conjecture, the numeric disparity of previous stimuli did not significantly predict alpha power (Fig. 7c; all clusters \(p > 0.05\); Bayes factors provided substantial evidence in favor of a null effect, Supplementary Fig. 8). However, we observed a marginally significant interaction between numeric disparity of the previous trial and alpha power on observers’ causal prior in two clusters from \(-340\) to \(-240\) ms, \(p = 0.069\) and from \(-220\) to \(-120\) ms \(p = 0.096\); two-sided cluster-based corrected randomization \(t_{22}\) test, Fig. 7d, top panel). The correlation between alpha power and the observers’ causal prior was prominent when prior numeric disparity was small (cluster from \(-480\) to \(-80\) ms, \(p = 0.006\)), but not significant when previous numeric disparity was large (i.e. all clusters \(p > 0.05\)). In summary, alpha power did not mediate, but...
to some extent (i.e. only marginally significant) moderated the effect of stimulus history on observers’ causal prior, i.e. their tendency to bind audiovisual signals (Fig. 7b).

**Discussion**

To form a coherent percept of the world, the human brain needs to integrate signals arising from a common cause, but segregate signals from independent causes. Perception thus relies crucially on inferring the world’s causal structure. To characterize the neural dynamics of how the brain solves this binding problem, we presented participants with sequences of beeps and flashes that varied in their numeric disparity.

Behaviorally, the number of beeps biased observers’ perceived number of flashes – a phenomenon coined sound-induced flash illusion. Conversely, the number of flashes biased observers’ perceived number of beeps, albeit only to a small degree. This asymmetry of crossmodal biases operating from vision to audition and vice versa can be attributed to the smaller precision of vision for temporal estimation, which is consistent with forced-fusion models of reliability-weighted integration (and Bayesian Causal Inference models, cf. Table 2). Crucially, as predicted by Bayesian Causal Inference, participants did not fully fuse auditory and visual stimuli into one unified percept, but they reported different numeric estimates for the flash and beep components of numerically disparate flash-beep stimuli. Moreover, audiovisual integration and crossmodal biases decreased for large numeric disparities, when the flash and beep sequences were unlikely to arise from a common cause. Thus, observers flexibly arbitrated between audiovisual integration and segregation depending on the probabilities of the underlying causal structures as predicted by Bayesian Causal Inference (see Fig. 2c).

At the neural level, our univariate and multivariate EEG analyses revealed that the computations and neural processes of multisensory interactions and Bayesian Causal Inference dynamically evolve poststimulus. Initially, the univariate ERP analyses revealed an early audiovisual interaction effect starting at about 70 ms poststimulus that is related to the visual P1 component and...
Thus, causal inference starts immediately after stimulus poststimulus (i.e. 200 ms after the sound-induced visual input and may thereby facilitate the emergence of the enhance the excitability in visual cortices and the salience of the fi

10 NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-019-09664-2 | www.nature.com/naturecommunications

ARTICLE NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-019-09664-2

presentation based on numeric disparity and influences early audiovisual interactions and biases as quantified by the neural weight index. However, only relatively late, starting at about 200–300 ms and peaking at 400 ms after the onset of the final stimulus slot, does the brain compute numeric estimates consistent with Bayesian Causal Inference by averaging the forced-fusion estimate with the task-relevant unisensory estimate weighted by the posterior probabilities of common and independent causal structures (i.e. model averaging). The exceedance probability of the hierarchical Bayesian Causal Inference estimate steadily rises until its peak, where it outperforms all other numeric estimates in accounting for the representational geometries obtained from EEG activity patterns (i.e. exceedance probability = 1). Likewise, the relative audiovisual weight index $w_{AV}$ revealed a task-relevance by numeric-disparity interaction at similar latencies as the characteristic qualitative profile for Bayesian Causal Inference.

This dynamic evolution of neural representations dovetails nicely with a hierarchical organization of Bayesian Causal Inference that has recently been shown in fMRI research\(^\text{10,11}\). Low-level sensory areas represented sensory estimates mainly under the assumption of separate causes, whereas posterior parietal areas integrated the signals weighted by their sensory precision under the assumption of a common cause. Only at the top of the cortical hierarchy, in anterior parietal areas, did the brain form a final Bayesian Causal Inference estimate that takes into account the observers’ uncertainty about the signals’ causal structure. Collectively, fMRI and EEG research jointly suggest that computations involving unisensory estimates rely on lower-level regions at earlier latencies, while Bayesian Causal Inference estimates that take into account the world’s causal structure arise later in higher-level cortical regions. Previous fMRI research implicated prefrontal cortices in the computations of the causal structure\(^\text{24,33}\), which may in turn inform the integration processes in parietal and temporal cortices\(^\text{34}\).
A recent neural network model with a feedforward architecture by Cuppini et al.\textsuperscript{35,36} suggests that this explicit causal inference relies on a higher convergence layer, while the audiovisual biases in numeric estimates may be mediated via direct connectivity between auditory and visual layers and emerge from spatio-temporal receptive fields in auditory and visual processing. In contrast to such a feed-forward architecture, we generally observed a mixture of multiple representations that were concurrently expressed in EEG activity patterns, even though different numeric estimates dominated neural processing at different poststimulus latencies. Therefore, we suggest that Bayesian Causal Inference is iteratively computed via multiple feed-back loops across the cortical hierarchy whereby numeric estimates as well as causal inferences are recurrently updated as the brain accumulates knowledge about the causal structure and sources in the environment\textsuperscript{12,13}.

In Bayesian inference, prior knowledge and expectations are crucial to guide the perceptual interpretation of the noisy sensory inputs\textsuperscript{37}. Multisensory perception in particular relies on a so-called causal prior that quantifies observers’ prior beliefs about the world’s causal structure\textsuperscript{1,2}. A ‘high’ causal prior (i.e. the belief that signals come from a common cause) influences multisensory perception by increasing observers’ tendency to bind audiovisual signals irrespective of the signals’ instantaneous intersensory congruency\textsuperscript{24}. In the current study, we investigated whether the neural activity before stimulus onset is related to observers’ causal prior. Indeed, low prestimulus alpha power and high gamma power were associated with a high causal prior, i.e. they increased participants’ tendency to integrate audiovisual stimuli. Accumulating research has shown effects of prestimulus alpha power on perceptual decisions such as detection threshold, decisional biases or perceptual awareness\textsuperscript{27,28,38}. Further, low alpha power was also shown to increase the occurrence of the sound-induced flash illusion\textsuperscript{16} (though see Keil et al.\textsuperscript{15} for an effect in beta power). In our study, low prestimulus alpha power predicted a larger causal prior leading to stronger bidirectional interactions between audition and vision and audiovisual biases (see Fig. 5c, audiovisual weight index $w_{AV}$). These enhanced audiovisual interactions might be explained by a tonic increase in cortical excitability for states of low alpha oscillatory power and associated high gamma power (though see Yuval-Greenberg et al.\textsuperscript{39} for a cautionary note). Moreover, if peaks and troughs of alpha oscillatory activity are modulated asymmetrically\textsuperscript{40}, low alpha power may also induce larger alpha troughs, thereby extending the temporal windows where gamma bursts and audiovisual interactions can occur\textsuperscript{41}. Indeed, our results show that observers’ causal prior depends not only on the tonic level of alpha power, but also on its phase. Prestimulus alpha phase may thus influence audiovisual
binding by defining the optimal time window in which neural processing can interact across auditory, visual and association areas, thereby modulating the temporal parsing of audiovisual signals into one unified percept23,42.

Next, we investigated whether the fluctuations in alpha power may enable observers to adopt dynamically to the statistical structure of the sensory inputs. Previous research has shown that prior exposure to congruent signals increases observers’ tendency to integrate sensory signals, while exposure to incongruent signals enhances their tendency to process signals independently (24,25), but see26). In the current study, we also observed that previous low numeric disparity trials predicted a greater causal prior or tendency to bind audiovisual signals into a coherent percept. Surprisingly, however, the numeric disparity of previous audiovisual stimuli did not significantly influence alpha power. It only modulated the effect of alpha power on the causal prior (i.e. a marginally significant interaction between alpha power and prior numeric disparity). More specifically, alpha power correlated with observers’ causal prior mainly when previous stimuli were of low rather than large numeric disparity.

Collectively, our results show that observers’ causal prior dynamically adapts to the statistical structure of the world (i.e. previous audiovisual numeric disparity), but that these adaptation processes are not mediated by fluctuations in alpha power. Instead, spontaneous (i.e. as yet unexplained by stimulus history) fluctuations in prestimulus gamma and alpha power as well as alpha phase correlated with observers’ causal prior. Alpha power, phase and frequency (i.e. speed)43–45 together with gamma power may thus dynamically set the functional neural system into states that facilitate or inhibit interactions across brain regions46 and temporal parsing of audiovisual signals into common percepts23,41.

In conclusion, to our knowledge this is the first study that resolves the neural computations of hierarchical Bayesian Causal Inference in time (see also47,48). We show that prestimulus oscillatory alpha power and phase correlates with the brain’s causal prior as a binding tendency that guides how the brain dynamically arbitrates between sensory integration and segregation (see49,50 for related studies showing that top-down predictions may be furnished via alpha/beta oscillations). Initially, about 70 ms after stimulus presentation, we observed non-specific audiovisual interactions, which may increase the bottom-up salience of sensory signals. Our multivariate analyses suggested that unsensory numeric estimates initially dominated the EEG activity pattern. Only later, from about 200–400 ms after the final stimulus slot, EEG signals encoded the Bayesian Causal Inference estimates that combine the forced-fusion and task-relevant segregation estimates weighted by the probabilities of common and independent cause models (i.e. model averaging). Thus, consistent with the notion of predictive coding, the brain may accumulate evidence concurrently about i. auditory (or visual) numeric estimates and ii. the underlying causal structure (i.e. whether auditory and visual signals come from common or independent sources) over several hundred milliseconds via recurrent message passing across the cortical hierarchy to compute Bayesian Causal Inference estimates13,51. By resolving the computational operations of multisensory interactions in human neocortex in time, our study reveals the hierarchical nature of multisensory perception. It shows that the brain dynamically encodes and re-computational priors and multiple numeric estimates to perform hierarchical Bayesian Causal Inference.

Methods

Participants. After giving written informed consent, 24 healthy volunteers participated in the EEG study based on previous calculations of statistical power. One participant did not attend the interview session and was excluded. Thus, data from 23 participants were analyzed (10 female; mean age 36.0 years, range 25–61 years). Participants were screened for current or former psychiatric disorders (as verified by the structured clinical interview for DSM-IV disorders, SCID-I, German version), cardiovascular disorders, diabetes and neurological disorders. One participant reported an asymptomatic arteriosclerotic malformation. Because behavioral and EEG was inconspicuous, the participant was included. All participants had normal or corrected-to-normal vision and audition. The study was approved by the human research review committee of the Medical Faculty of the University of Tübingen and at the University Hospital Tübingen (approval number 728/2014BO2).

Stimuli. The flash-beep paradigm was an adaptation of previous “sound-induced flash illusion” paradigms27,28. The visual flash was a circle presented in the center of the screen on a black background (i.e. 100% contrast; Fig. 1a) briefly for one frame (i.e. 16.7 ms, as defined by the monitor refresh rate of 60 Hz). The maximum grayscale value (i.e. white) of the circle was at radius 4.5° with smoothed inner and outer borders by defining the grayscale values of circles of smaller and larger radius by a Gaussian of 0.9° STD visual angle. The auditory beep was a pure tone (2000 Hz, ~70 dB SPL) of 27 ms duration including a 10 ms linear on/off ramp. Multiple visual flashes and auditory beeps were presented sequentially at a fixed SOA of 66.6 ms (see below).

Experimental design. In the flash-beep paradigm, participants were presented with a sequence of i. one, two, three or four flashes and ii. one, two, three or four beeps (Fig. 1a). On each trial, the number of flashes and beeps were independently sampled from one to four leading to four levels of absolute numeric disparity (i.e. zero congruence; three = maximal level of disparity; Fig. 1b). Each flash and/or beep was presented sequentially in fixed temporal slots that started at 0, 66.7, 133, 200 ms. The temporal slots were filled up sequentially. For instance, if the number of beeps was three, they were presented at 0, 66.6, 133 and 200 ms, while the fourth slot was left empty. Hence, if the same number of flashes and beeps were presented on a particular trial, flashes and beeps were presented in synchrony. On numerically disparate trials, the ‘surplus’ beeps (or flashes) were added in the subsequent fixed time slots (e.g. in case of 2 flashes and 3 beeps: we present 2 flash-beeps at 0 and 66.6 ms in synchrony and a single beep at 133 ms).

Across experimental runs, we instructed participants to selectively report either the number of flashes or beeps and to ignore the stimuli in the task-irrelevant modality. Hence, the 4 × 4 × 2 factorial design manipulated (i) the number of visual flashes (i.e. one, two, three or four), (ii) the number of auditory beeps (i.e. one, two, three or four) and (iii) the task relevance (auditory- vs. visual-selective report) yielding 32 conditions in total (Fig. 1b). For analyses, we reorganized trials based on their absolute numeric disparity (|#A - #V| ≤ 1: small numeric disparity; |#A - #V| ≥ 2: large numeric disparity). Thus, we analyzed the data in a 2 (task relevance: visual vs. auditory report) × 2 (numeric disparity) factorial design. Please note that for model estimation, a dual-task design combining auditory and visual report on the same trials may have been preferable. However, the dual-task demands would render the interpretation of the neural processes ambiguous, so that in our experimental participants reported either their auditory or visual percept on each trial.

The duration of a flash-beep sequence was determined by the number of sequentially presented flash and/or beep stimuli (see above for the definition of temporal slots). Irrespective of the number of flashes and/or beeps, a response screen was presented 750 ms after the onset of the first flash and beep for a maximum duration of 2.5 s instructing participants to report their perceived number of flashes (or beeps) as accurately as possible by pushing one of four buttons. The order of buttons was counterbalanced across runs to decorrelate motor responses from numeric reports. On half of the runs, the buttons from left to right corresponded to one to four stimuli; on the other half, they corresponded to four to one. After a participant’s response, the next trial started after an inter-trial interval of 1–1.75 s.

In every experimental run, each of the 16 conditions was presented 10 times. Participants completed 4 runs of auditory- and 4 runs of visual-selective report in a counterbalanced fashion (except for one participant performing 5 runs of auditory and 3 of visual report). Further, each participant completed two uni-sensory runs with visual or auditory stimuli only (i.e. 4 uni-sensory conditions presented 40 times per run) from which we computed the difference wave (see below). Before the actual experiment, participants completed 56 practice trials.

Experimental setup. Psychtoolbox 3.0922 (www.psychtoolbox.org) running under MATLAB R2016a (MathWorks) presented audiovisual stimuli and sent trigger pulses to the EEG recording system. Auditory stimuli were presented at ~70 dB SPL via two loudspeakers (Logitech Z130) positioned on each side of the monitor. Visual stimuli were presented on an LCD screen with a 60 Hz refresh rate (EIZO FlexScan S2202W). Button presses were recorded using a standard keyboard. Participants were seated in front of the monitor and loudspeakers at a distance of 85 cm in an electrically shielded, sound-attenuated room.

Overview of GLM and Bayesian modelling analysis for behavioral data. To characterize how human observers arbitrate between sensory integration and
segregation, we developed a general linear model (GLM)-based and a Bayesian modelling analysis approach. The GLM-based analysis computed a relative weight index \( w_{AV} \) which quantified the relative influence of the auditory and the visual numeric stimuli on observers’ auditory and visual behavioral numeric reports. This GLM-based analysis allowed us to reveal audiovisual weight profiles in our 2 (numeric disparity) \times 2 (task relevance) factorial design that are qualitatively in line with the principles of Bayesian Causal Inference (Fig. 1c).

The Bayesian modelling analysis fitted the full-segregation, the forced-fusion and the Bayesian Causal Inference (BCI) model to the behavioral numeric reports with different decision functions. We then used Bayesian model comparison to determine the model that is the best explanation for observers’ behavioral data (Table 2 and Supplementary Table 1).

Behavior - GLM-based analysis for reported number of stimuli. We quantified the influence of the true number of auditory and visual stimuli on the reported (behavioral) auditory or visual numeric estimates using a linear regression model\(^{11}\).

In this regression model, the reported number of stimuli was predicted by the true number of auditory and visual stimuli separately in the four conditions in the 2 (numeric disparity) \times 2 (task relevance) factorial design. The auditory \((\beta_A)\) and visual \((\beta_V)\) parameter estimates quantified the influence of the experimentally defined auditory and visual stimuli on the perceived number of stimuli for a particular condition. To obtain a relative audiovisual weight index \( w_{AV} \), we computed the four-quadrant inverse tangents of the auditory \((\beta_A)\) and visual \((\beta_V)\) parameters estimates for each of the four conditions (i.e. \( w_{AV} = \arctan(\beta_A, \beta_V) \)). An audiovisual weight index \( w_{AV} \approx 90^\circ \) indicates purely visual and \( w_{AV} \approx 0^\circ \) purely auditory influence on the reported/decoded number of stimuli.

We performed the statistics on the behavioral audiovisual weight indices using a two (auditory vs. visual report) \times 2 (large vs. small numeric disparity) factorial design based on a likelihood-ratio test statistic (LRTS) for circular measures\(^{39}\). 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 (i.e. main and interaction effects) and the reduced model that obtained a model not allowing any mean differences between conditions. To refrain from making any parametric assumptions, we evaluated the effects of task relevance, numeric disparity and their interaction in the factorial design using randomization tests (5000 randomizations)\(^{34}\). To account for the within-participant repeated-measures design at the second random-effects level, randomizations were performed within each participant. For the main effects of numeric disparity and task relevance, we varied random numbers within the levels of the non-test factor\(^{30}\). For tests of the numeric-disparity \times task-relevance interactions, we randomized the simple main effects (i.e. (A1B1, A2B2) and (A1B2, A2B1)) which are exchangeable under the null-hypothesis of no interaction\(^{56}\). To test deviations of simple main effects (i.e. (A1B1, A2B2) and (A1B2, A2B1)) which are exchangeable under the null-hypothesis of no interaction\(^{56}\)

\[
N_{AV} = \left\{ \begin{array}{ll} N_{AV,C1} & \text{if } P(C = 1|s_A, s_V) > 0.5 \\
N_{AV,C2} & \text{if } P(C = 1|s_A, s_V) \leq 0.5 
\end{array} \right.
\]

According to the ‘model selection’ strategy, the brain reports the numeric estimate selectively from the more likely causal structure (Equation (6) only for \( N_{AV} \)).

Thus, Bayesian Causal Inference formally requires three numeric estimates \( (N_{AV,C1}, N_{AV,C2}, N_{AV,C3}) \) which are combined into a final estimate \( N_{AV} \) or \( N_{AV} \), depending on which sensory modality is task-relevant) according to one of the three decision functions.

We evaluated whether and how participants integrate auditory and visual stimuli based on their auditory and visual numeric reports by comparing (i) the full-segregation model that estimates stimulus number independently for vision and audition (i.e. formally, the BCI model with a fixed \( p_{common} = 0 \)), (ii) the forced-fusion model that integrates auditory and visual stimuli in a mandatory fashion (i.e. formally, the BCI model with a fixed \( p_{common} = 1 \)) and (iii) the BCI model (i.e. model averaging: Table 2). Because the decisional strategy of ‘model averaging’ outperformed the other decision functions (Equations (4)–(7)) based on Bayesian model comparison at the group level (Supplementary Table 1), the main report and analysis of the neural data focus on model averaging.

To arbitrate between the full-segregation, forced-fusion and BCI models, we fitted each model to participants’ numeric reports (Table 2) based on the predicted distributions of the auditory (i.e. the marginal distributions: \( p(N_{AV}|N_{AV,C1}) \) and visual (i.e. \( p(N_{AV}|N_{AV,C2}) \)) numeric estimates that were obtained by marginalizing over the internal variables \( s_A \) and \( s_V \) that are not accessible to the experimenter (for further details of the fitting procedure, see Kording et al.\(^{31}\)). These distributions were generated by simulating \( s_A \) and \( s_V \) at 5000 times (i.e. continuous variables sampled from Gaussian distributions) for each of the 32 conditions and inferring \( N_{AV} \) and \( N_{AV} \) from Equations (1)-(5). To link the continuous distributions \( p(N_{AV}|N_{AV,C1}) \) and \( p(N_{AV}|N_{AV,C2}) \) to participants’ categorical auditory or visual numeric reports (i.e. from (1,2,3,4)), we assumed that participants selected the button that is closest to \( N_{AV} \) or \( N_{AV} \) and binned \( N_{AV} \) and \( N_{AV} \) accordingly into a four-bin histogram. From these predicted multinomial distributions (i.e. one for each of the 32 conditions; auditory and visual numeric reports were linked to \( N_{AV} \) and \( N_{AV} \), respectively), we computed the log likelihood of participants’ numeric reports and summed the log likelihoods across conditions. To obtain maximum likelihood estimates for the five parameters of the models \( p_{common} \mu_A \phi_A, \phi_V, \sigma_A, \sigma_V \); formally, the forced-fusion and full-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 R2013b). This optimization algorithm was initialized with 10 different parameter settings that were defined based on a priori grid-search.

We report the results (across-participants’ mean and standard error) of the parameter setting with the highest log likelihood across these initializations (Table 2 and Supplementary Table 1). This fitting procedure was applied individually to each participant’s data set for the Bayesian Causal Inference (with three different decision functions), the forced-fusion and the full-segregation models. The model fit was assessed by Nagelkerke’s coefficient of determination\(^{33}\).
using a null model of random guesses of stimulus number 1–4 with equal probability 0.25. 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. The BIC depends on both model complexity and model fit. We performed Bayesian model comparison at the random effects group level as implemented in SPM12 to obtain the protected exceedance probability (the probability that a given model is more likely than any other model, beyond differences due to chance) for the candidate models.

To generate predictions for the audiovisual weight index based on the Bayesian Causal Inference model (with model averaging), we simulated new \( x_{AV} \) and \( x_{V} \) for 10000 trials for each of the 32 conditions using the fitted BCI model parameters of each participant. For each condition, we extracted the audio-unisensory (\( N_{VA} - C \)), ii. unisensory auditory (\( N_{VI} - C \)), estimates, iii. forced-fusion (\( N_{AV} - C \)), iv. final BCI auditory numeric estimate (\( N_{V} \) or \( N_{I} \) depending on whether the auditory or visual modality was task-relevant) and v. posterior probability estimate of each causal structure (\( p(C = 1 | x_{AV} ) \)). Next, we used the mode of the resulting (kernel-density estimated) distributions for each condition and participant to compute the model predictions for the audiovisual weight index \( w_{AV} \) (Figs. 1c, 5c) and the RDMs (see multivariate EEG analysis, Fig. 2c).

EEG – Data acquisition and preprocessing

EEG signals were recorded from 64 active electrodes positioned in an extended 10–20 montage using electrode caps (actiCap, Brain Products, Gilching, Germany) and two 32 channel DC amplifiers (BrainAmp, Brain Products). Electrodes were referenced to FCz using AFz as ground during recording. Signals were digitized at 1000 Hz with a high-pass filter of 0.1 Hz. Electrode impedances were kept below 25 kOhm.

Preprocessing of EEG data was performed using Brainstorm 3.42 running on Matlab R2015b. EEG data were band-pass filtered (0.25–45 Hz for the main EEG analyses). Eye blinks were automatically detected using data from the FP1 electrode (i.e. a blink was detected if the band-pass (1.5–15 Hz) filtered EEG signal exceeded two times the STDev of the minimum duration between two consecutive blinks; was 800 ms). Signal-space projectors (SSPs) were created from band-pass filtered (1.5–15 Hz) 400 ms segments centered on detected blinks. The first spatial component of the SSPs was then used to correct blink artifacts in continuous EEG data. Further, all data were visually inspected for artifacts from blinks (i.e. residual blink artifacts after correction using SSPs), saccades, motion, electrode drifts or jumps and contaminated segments were discarded from further analysis (on average 6.4 ± 0.9 % SEM of all trials discarded). Finally, EEG data were referenced to the average of left and right mastoid electrodes and downsampled to 200 Hz. For analysis of event-related potentials (ERPs) and decoding analyses (see below), all EEG data were normalized with a 200 ms prestimulus baseline and were analyzed from 100 ms before stimulus onset up to 750 ms after stimulus onset, when the response screen was presented.

EEG – Preprocessing for multivariate analyses

Single-trial EEG data from the 64 electrodes were binned in time windows of 20 ms. Hence, given a sampling rate of 200 Hz, each 20 ms time window included four temporal sampling points. 64-electrode trials (for each time sample) were concatenated across the four sampling points within each bin resulting in a spatio-temporal EEG activity pattern of 256 features. EEG activity patterns were z scored to control for mean differences between the first sampling point. The first sampling point in the 20 ms time window was taken as the window’s time point in all analyses.

Overview of EEG analyses

We characterized the neural processes underlying multisensory integration by combining several analysis approaches:

1. Univariate EEG analysis: We identified multisensory integration by testing for audiovisual interactions focusing on the classical ‘sound-induced flash illusion’ conditions, where one flash is presented together with two beeps.

2. Multivariate EEG analysis and neural audiovisual weight index \( w_{AV} \):
   We computed the audiovisual weight index \( w_{AV} \) which quantifies the relative influence of the true number of auditory and visual stimuli on the ‘internal’ numeric estimates decoded from EEG activity patterns using support-vector regression (SVR) models (libSVM 3.20) to map the numeric estimates to stimulus numbers.

3. Multivariate EEG analysis and Bayesian Causal Inference model: We assessed how the numeric estimates obtained from the BCI model, i.e. the unisensory auditory and visual full-segregation, the forced-fusion and the Bayesian Causal Inference estimates (i.e. based on model averaging) are dynamically encoded in EEG activity pattern across post-stimulus time using representational dissimilarity analyses. In supplementary analyses, we also directly decoded the numeric estimates from EEG activity patterns using support-vector regression or canonical correlation analyses (Supplementary Methods and Supplementary Fig. 6).

4. Pre-stimulus EEG activity and parameters of the Bayesian Causal Inference model: We investigated whether the power or phase of brain oscillations as measured by EEG before the stimulus onset correlates with the causal prior or the visual precision parameters of the Bayesian Causal Inference model (selectively refitted to trials binned according to their oscillatory power or phase).

EEG – Univariate analysis of audiovisual interactions

To assess basic components in ERPs and early audiovisual interactions, we averaged trial-wise EEG activity from the ERPs for audiovisual conditions. We then averaged the ERPs across centro-parietal electrodes (i.e. Cz, CP1, CP2, P1, Pz, P2; Fig. 1d) or occipital electrodes (i.e. O1, O2, Oz, PO3, POz, PO4; Fig. 1e). To analyze early audiovisual interactions as reported for the sound-induced flash illusion, we computed the difference between conditions and the corresponding unisensory conditions (i.e. \( A_{AV} A_{V} \); \( A_{AV} A_{I} \); \( A_{VA} A_{V} \); \( A_{VA} A_{I} \)). However, the auditory and visual trials were acquired in separate unisensory runs and may therefore differ in attentional and cognitive context. Further, our experimental design did not include null trials to account for anticipatory effects around stimulus onset. Thus, we compared the difference between the BCI model predicted and unisensory audiovisual interactions need to be interpreted with caution. To test whether the difference wave deviated from zero at the group level, we used a non-parametric randomization test (5000 randomizations) in which we flipped the sign of the individual difference waves and computed two-sided one-sample t tests as a test statistic. To correct for multiple comparisons across the sampling points, we used cluster-based corrections with the sum of the t values across a cluster as cluster-level statistic and an auxiliary cluster-defining threshold of t = 2 for each time point.

EEG – Multivariate GLM-based analysis, decoding accuracy and audiovisual weight index

For each 20 ms time window, we trained linear support-vector regression (SVR) models (libSVM 3.20) to learn the mapping from spatial–temporal audiovisual activity patterns to the number of stimulus trials (ambient audio–visually congruent conditions (including conditions of auditory and visual report) from all but one run. The SVR’s parameters (C and \( \nu \) were optimized using a grid search within each cross-validation fold (i.e. nested cross-validation). Before training the SVR models, we recoded the stimulus numbers as labels to the range of \(-1,1\) (i.e. \(-1 = 1 \text{ stimulus}; 0.33 = 2 \text{ stimulus}; 0.33 = 3 \text{ stimulus}; 1 = 4 \text{ stimulus})). This learnt mapping from EEG activity patterns to external number of stimulus was then used to decode the number of stimuli from spatio-temporal EEG activity patterns of the audiovisual congruent and incongruent audiovisual conditions of the remaining run. In a leave-one-run-out cross-validation scheme, the training-test procedure was repeated for all runs. To account for SNR differences across runs, predicted stimulus numbers were \( z \)-scored within each run. The decoded stimulus numbers for the congruent and incongruent conditions were used to assess i. decoding accuracy based on congruent trials only and ii. to compute the audiovisual weight index \( w_{AV} \) in subsequent GLM-based analysis approaches (see below).

First, we computed decoding accuracy based selectively on the audiovisual congruent conditions. We decoded stimulus numbers 1–4 at all time points even though the distinctions between high flash and/or beep numbers (i.e. three vs. four) was only possible at later time points. Hence, as expected, the decoder was able to discriminate between higher stimulus numbers (e.g. three vs. four stimuli) only after about 250 ms (Fig. 3a). Next, we evaluated the decoder’s accuracy in terms of the Pearson correlation between true and decoded stimulus number selectively in audiovisual congruent conditions (Fig. 3b). We tested whether individual Fisher’s z-transformed correlation coefficients were larger than zero at the group level using a one-sided test (i.e. using a one-sample t test). The resulting correlation coefficient was compared across conditions (in 5000 randomizations) and a cluster-based correction for multiple comparisons across time intervals (as applied to difference waves, see above; cluster-level statistic: sum of the t values in a cluster; auxiliary cluster defining threshold t = 2). Second, we quantified the influence of the true number of auditory and visual stimuli on the decoded (neural) auditory or visual numeric estimates in a GLM-based analysis approach that was equivalent to our behavioral analysis. In a linear regression model, the decoded number of stimulus was predicted by the true number of auditory and visual stimuli separately for the four conditions in the 2 (numeric disparity) × 2 (task relevance) factorial design (see behavioral analysis for further details). Statistical analysis was also equivalent to the behavioral analysis with the exception that we accounted for multiple comparisons across time using a cluster-based correction (cluster-level statistic: sum of the LRTS values in a cluster; auxiliary clustering defining threshold LRTS = 2). Unless otherwise stated, all results are reported at \( p < 0.05 \) corrected for multiple comparisons across the audiovisual weight index \( w_{AV} \) (Figs. 3c, 4c). We computed the means’ bootstrapped confidence intervals (1000 bootstraps).

EEG – Multivariate analyses of BCI estimates

To characterize the neural dynamics of Bayesian Causal Inference, we next investigated whether and when the four numeric estimates of the Bayesian Causal Inference (BCI) model are represented in EEG activity patterns using support-vector regression (i.e. similar to a previous fMRI study), canonical correlation analysis and representational similarity analysis (RSA). Because these three analysis approaches yield comparable results, we focus in the main manuscript on the RSA (see Supplementary Methods and Supplementary Fig. 6).

To define the representational dissimilarity matrices (RDMs) for the RSA, we computed the pairwise absolute distance between the BCI model’s four numeric estimates, i.e. i. unisensory visual (\( N_{VA} - C \)), ii. unisensory auditory (\( N_{VI} - C \)), iii. forced-fusion (\( N_{AV} - C \)), iv. final BCI auditory numeric estimate

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-019-09664-2 | www.nature.com/naturecommunications
**EGG - The effect of stimulus presentation oscillations on the causal prior and the visual precision parameters**

We investigated whether stimulus presentation oscillatory power or phase over occipital electrodes (i.e. O1, O2, Oz, P7, P8, Pz) is related to the brain's prior binding tendency as quantified by the causal prior (i.e. $P_{\text{common}}$) or the precision of the visual representation (i.e. $\sigma_c$). We computed the band-power across trials in the frequency range of interest for each participant (based on Equation (8)). We then correlated the causal prior across deciles for each participant with $P_{\text{common}}$ and $\sigma_c$. For each participant, we fitted a linear regression model to the data across deciles for each condition (i.e. visual or auditory presentation). We then computed the correlation coefficient for each decile and assessed the significance using a one-sample t-test. We then compared the correlation coefficients across conditions separately for previous small and large numeric disparities.

**EEG - The relation of prior stimulus history, prestimulus alpha power and the causal prior.**

To investigate whether the numeric disparity of prior stimulus history influences observers' causal prior, we sorted the participants into two groups (small and large numeric disparity) based on their performance in a previous small numeric disparity task. We then computed the causal prior across deciles for each participant (based on Equation (8)) and assessed the significance using a one-sample t-test. We then compared the correlation coefficients across conditions separately for previous small and large numeric disparities. We then compared the causal prior across deciles for each participant with $P_{\text{common}}$ and $\sigma_c$. For each participant, we fitted a linear regression model to the data across deciles for each condition (i.e. visual or auditory presentation). We then computed the correlation coefficient for each decile and assessed the significance using a one-sample t-test. We then compared the correlation coefficients across conditions separately for previous small and large numeric disparities. We then compared the causal prior across deciles for each participant with $P_{\text{common}}$ and $\sigma_c$. For each participant, we fitted a linear regression model to the data across deciles for each condition (i.e. visual or auditory presentation). We then computed the correlation coefficient for each decile and assessed the significance using a one-sample t-test. We then compared the correlation coefficients across conditions separately for previous small and large numeric disparities.
Counted twice for A

Putative audiovisual interactions in EEG have been suggested to emerge because of interactions involving only once for the multisensory condition (i.e. AV). Likewise, early interaction analyses. Moreover, interactions computed as AV vs. A can be counted twice for the sum of the unisensory conditions (i.e. A + V), but be involved only once for the multisensory condition (i.e. AV). Likewise, early putative audiovisual interactions in EEG have been suggested to emerge because of anticipatory ERP effects that precede all stimulus presentations and are therefore counted twice for A + V, but only once for AV (see68). Therefore, multisensory interactions should optimally be computed including ‘null events’ to account for non-specific expectation effects (i.e. AV + Null vs. A + V). Further, in our study unisensory and multisensory conditions may differ in attentional context, because auditory and visual conditions were performed in separate experimental runs, where either auditory or visual information were task-relevant. Collectively, these factors need to be taken into account when interpreting audiovisual interactions in EEG (or MRI) responses in our and other studies.

Multivariate decoding: The EEG activity patterns measured across 64 scalp electrodes represent a superposition of activity generated by various neural sources located for instance in auditory, visual or higher-order association areas. The extent to which auditory or visual information can be decoded from EEG activity pattern therefore depends inherently not only on how information is neurally encoded by the ‘neuronal generators’ in source space, but also on how these neural activities are expressed and superimposed in sensor space (i.e. as measured by scalp electrodes). For example, the number of auditory beeps is perceptually more precisely represented than the number of flashes (based on observers’ behavioral reports, Table 2), suggesting that the brain encodes the timing and number of events with a greater precision in audition than vision. Nevertheless, supplementary analyses in sensor space revealed that the number of unisensory flashes can be more accurately decoded from EEG activity patterns than the number of unisensory beeps (Supplementary Fig. 2). These discrepancies between precision (or accuracy) measured at the behavioral/perceptual level and EEG decoding accuracy at the sensor level may result from differences in neural encoding in source space or how these neural activities are expressed in sensor space (e.g. source orientation, superposition etc.). Potentially, the greater decorrelation of visual numeric information may contribute to the visual bias we observed for the audiovisual weight index wAV (i.e., wAV > 45°; Fig. 3c) and the dominance of the visual numeric estimates in our decoding analysis based on the estimate of the Bayesian Causal Inference model (Eq. 4a).

In the analysis of the audiovisual weight index wAV, we trained the support-vector regression model on the audiovisual congruent conditions pooled over task relevance to ensure that the decoder was based on activity patterns generated by sources related to auditory, visual and audiovisual integration processes. Moreover, this approach ensures that the selectivity of task relevance on the audiovisual weight index wAV cannot be attributed to differences in the decoding model (see71 for a related discussion). In a supplementary analysis, we also trained the support-vector regression models separately for visual and auditory report and obtained comparable results (Supplementary Fig. 7) suggesting that our results are immune to this particular choice of the decoding approach. While the univariate interaction analysis (see above) cannot identify linear response combinations, this multivariate decoding analysis cannot exclude the possibility that auditory and visual stimuli jointly influence EEG activity pattern even though auditory and visual signals are not integrated at the single neuron level.

In our second multivariate analysis approach, we decoded (directly: support-vector regression, canonical correlation analysis; or indirectly: representational similarity analysis) the numeric estimates of the Bayesian Causal Inference model from EEG activity patterns and then computed the exceedance probability that one numeric estimate was more likely encoded than any other one. The decoding approaches using support-vector regression, canonical correlation analysis and representational similarity analysis provided comparable results indicating that our results are robust to the specific decoding approach (Fig. 4 and Supplementary Fig. 6). However, given the caveats discussed above (e.g. superposition of EEG activity patterns) and the high correlation between the different numeric estimates in the model, it is likely that multiple numeric estimates are concurrently represented in the brain even if the exceedance probability is high for only one particular numeric estimate.

Data availability statement
The raw behavioral and EEG datasets generated and analyzed in the current study are available in a G-Node repository2, [https://doi.org/g-node.org/ee65189d139c984a94867679425224497/]. The source data underlying Figs. 1–7, Tables 1–2 and Supplementary Figs. 1–8 and Supplementary Table 1 are provided as a Source Data file in the same G-Node repository.

Code availability
The Matlab code to fit the Bayesian Causal Inference model to the behavioral data is available in a G-Node repository3, [https://doi.org/g-node.org/ee65189d139c984a94867679425224497/]. Custom Matlab code for the analyses of EEG data are available from the corresponding author on reasonable request.

Received: 31 July 2018 Accepted: 14 March 2019
Published online: 23 April 2019

References
1. Kording, K. P. et al. Causal inference in multisensory perception. PLoS ONE 2, e0000943 (2007).
2. Shams, L. & Beierholm, U. R. Causal inference in perception. Trends. Cogn. Sci. 14, 425–432 (2010).
3. Alais, D. & Burr, D. The ventriloquist effect results from near-optimal bimodal integration. Curr. Biol. 14, 257–262 (2004).
4. Ernst, M. O. & Banks, M. S. Multisensory Organ. Trends. Cogn. Sci. 7, 194–202 (2003).
5. Shams, L., Ma, W. J. & Beierholm, U. Sound-induced flash illusion as an optimal percept. Neuron 16, 1923–1927 (2002).
6. Shams, L., Kamitani, Y. & Shimojo, S. What you see is what you hear. Nature 408, 788 (2000).
7. Alais, D. & Burr, D. The ventriloquist effect results from near-optimal bimodal integration. Curr. Biol. 14, 257–262 (2004).
8. Shams, L., Ma, W. J. & Beierholm, U. Sound-induced flash illusion as an optimal percept. Neuron 16, 1923–1927 (2002).
9. Shams, L., Kamitani, Y. & Shimojo, S. What you see is what you hear. Nature 408, 788 (2000).
10. Wallace, M. T. et al. Unifying multisensory signals across time and space. Exp. Brain Res. 158, 252 (2005).
11. Rohe, T. & Noppeney, U. Cortical hierarchies perform Bayesian causal inference in multisensory perception. PLoS Biol. 13, e1002073 (2015).
12. Rohe, T. & Noppeney, U. Distinct computational principles govern multisensory integration in primary sensory and association cortices. Curr. Biol. 26, 509–514 (2016).
13. Friston, K. The free-energy principle: a unified brain theory? Nat. Rev. Neurosci. 11, 127 (2010).
14. Bastos, A. M. et al. Canonical microcircuits for predictive coding. Neuron 76, 695–711 (2012).
15. Kell, J., Müller, N., Hartmann, T. & Weise, N. Prestimulus beta power and phase synchrony influence the sound-induced flash illusion. Cereb. Cortex 24, 1278–1288 (2013).
16. Lange, J., Oostenveld, R. & Fries, P. Reduced occipital alpha power indexes enhanced excitability rather than improved visual perception. J. Neurosci. 33, 3212–3220 (2013).
17. Andersen, T. S., Tippana, K. & Sams, M. Factors influencing audiovisual fission and fusion illusions. Brain. Res. Cogn. Brain. Res. 21, 301–308 (2004).
18. Wozny, D. R., Beierholm, U. R. & Shams, L. Human trimodal perception follows optimal statistical inference. J. Vis. 8, 24–24 (2008).
19. Wozny, D. R., Beierholm, U. R. & Shams, L. Probability matching as a computational strategy used in perception. PLoS Comput. Biol. 6, https://doi.org/10.1371/journal.pcbi.1000871 (2010).
20. Mishra, J., Martinez, A., Sejnowski, T. J. & Hillyard, S. A. Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. J. Neurosci. 27, 4120–4131 (2007).
21. Wozny, D. R., Beierholm, U. R. & Shams, L. Probability matching as a computational strategy used in perception. PLoS Comput. Biol. 6, https://doi.org/10.1371/journal.pcbi.1000871 (2010).
22. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 4 (2008).
23. VanRullen, R. Perceptual cycles. Trends. Cogn. Sci. 20, 723–735 (2016).
24. Gau, R. & Noppeney, U. How prior expectations shape multisensory perception. Neuroimage 124, 876–886 (2016).
25. Nahorna, O., Berthommier, F. & Schwartz, J.-L. Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect. J. Acoust. Soc. Am. 137, 362–377 (2015).
