Metacognition in the audiovisual McGurk illusion

Meijer, David; Noppeney, Uta

DOI:
10.1098/rstb.2022.0348

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Meijer, D & Noppeney, U 2023, 'Metacognition in the audiovisual McGurk illusion: perceptual and causal confidence', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 378, no. 1886, 20220348. https://doi.org/10.1098/rstb.2022.0348

PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Meijer D, Noppeney U. 2023 Metacognition in the audiovisual McGurk illusion: perceptual and causal confidence. *Phil. Trans. R. Soc. B* **378**: 20220348. https://doi.org/10.1098/rstb.2022.0348

Received: 17 March 2023
Accepted: 2 July 2023

One contribution of 16 to a theme issue 'Decision and control processes in multisensory perception'.

Subject Areas:
neuroscience, cognition

Keywords:
metacognition, audiovisual integration, Bayesian causal inference, McGurk illusion, confidence

Author for correspondence:
David Meijer
e-mail: meijerdavid1@gmail.com

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6751342.

THE ROYAL SOCIETY PUBLISHING

Metacognition in the audiovisual McGurk illusion: perceptual and causal confidence

David Meijer[1,2] and Uta Noppeney[1,3]

[1]Computational Neuroscience and Cognitive Robotics Centre, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
[2]Acoustics Research Institute, Austrian Academy of Sciences, Wohllebengasse 12-14, 1040, Wien, Austria
[3]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Kapittelweg 29, 6525 EN, Nijmegen, The Netherlands

DM, 0000-0002-7484-2783; UN, 0000-0002-7697-2290

Almost all decisions in everyday life rely on multiple sensory inputs that can come from common or independent causes. These situations invoke perceptual uncertainty about environmental properties and the signals' causal structure. Using the audiovisual McGurk illusion, this study investigated how observers formed perceptual and causal confidence judgements in information integration tasks under causal uncertainty. Observers were presented with spoken syllables, their corresponding articulatory lip movements or their congruent and McGurk combinations (e.g. auditory B/P with visual G/K). Observers reported their perceived auditory syllable, the causal structure and confidence for each judgement. Observers were more accurate and confident on congruent than unisensory trials. Their perceptual and causal confidence were tightly related over trials as predicted by the interactive nature of perceptual and causal inference. Further, observers assigned comparable perceptual and causal confidence to veridical 'G/K' percepts on audiovisual congruent trials and their causal and perceptual metamers on McGurk trials (i.e. illusory 'G/K' percepts). Thus, observers metacognitively evaluate the integrated audiovisual percept with limited access to the conflicting unisensory stimulus components on McGurk trials. Collectively, our results suggest that observers form meaningful perceptual and causal confidence judgements about multisensory scenes that are qualitatively consistent with principles of Bayesian causal inference.

This article is part of the theme issue 'Decision and control processes in multisensory perception'.

1. Introduction

Metacognition, the capacity to monitor one's own uncertainty, is important for adaptive behaviour [1–3]. In a noisy restaurant, if we feel unsure whether the waiter said 'beans' or 'greens', we may ask him to repeat it [4–6]. A wealth of research has shown that humans are able to meaningfully report their confidence about perceptual decisions [7–16]. They typically assign a higher confidence to their correct than incorrect decisions. Yet, research to date has almost exclusively focused on simple perceptual decisions based on single cues in one sensory modality, while more naturalistic scenarios such as a busy restaurant expose the brain to numerous signals that may come from common or independent sources. To communicate effectively with the waiter, the brain should integrate the waiter's speech signals selectively with his articulatory lip movements and segregate them from visual and auditory signals produced by other guests. Perception in complex audiovisual scenes thus relies inherently on solving the causal inference or binding problem [17–23].

Bayesian causal inference models (see also figure 1) address this computational challenge by explicitly modelling the potential causal structures that could have generated the sensory signals. In the case of common sources, signals are integrated weighted by their relative reliabilities into more precise estimates

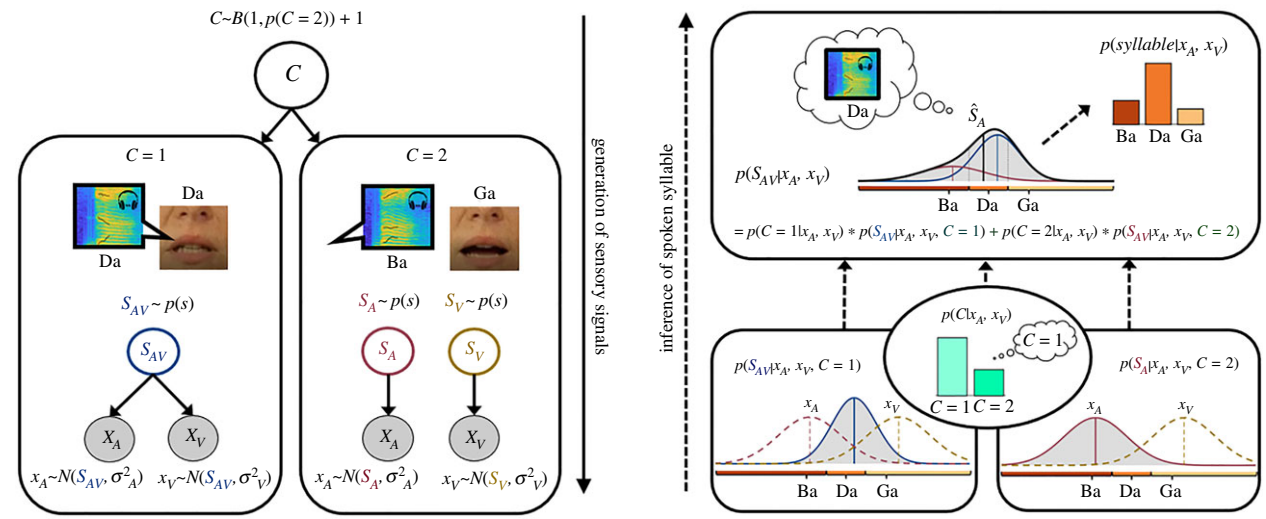

Figure 1. Bayesian causal inference model. (*a*) Generative model: the generative model of Bayesian causal inference explicitly models the potential causal structures (C = 1 or C = 2) that could have generated the auditory and visual signals. A common source (S_{AV}) generates an auditory Da speech signal (represented in the figure by a time-frequency spectrogram) and the corresponding articulatory lip movements. Alternatively, an auditory source (S_A) generates the Ba speech signal and an independent visual source (S_V) the lip movements articulating a Ga. The observed auditory and visual signals (x_A and x_V) are corrupted by independent noise. (*b*) Bayesian inference: the observer needs to make two closely related inferences based on the noisy auditory and visual signals: (i) perceptual inference about the spoken (i.e. auditory) syllable and (ii) causal inference about whether signals come from common or independent sources. In the case of common sources, signals should be fused weighted by their relative reliabilities into a more precise syllable percept (fusion estimate): $p(S_{AV}|x_A,x_V,C = 1) = w_A\widehat{S_A} + w_V\widehat{S_V}$ with $w_A = r_A/(r_A + r_V)$ and $r_A = 1/\sigma_A^2$. In the case of independent sources, the signal are segregated (segregation estimate). Hence, the phoneme percept should only depend on the speech signals: $p(S_A|x_A,x_V,C = 2)$. As the observer does not a-priori know whether signals come from common or independent sources, it needs to infer the causal structure from the noisy signals: $p(C|x_a,x_V)$. To account for observers' uncertainty about the signals' causal structure, a final percept is thought to be computed by averaging the fusion and the segregation estimates weighted by the posterior probabilities of common and independent sources. As a result, perceptual and causal estimates arise interactively in the inference process. Adapted from Körding *et al.* 2007 [18]. (Online version in colour.)

(i.e. fusion). In the case of independent sources, they are processed independently (i.e. segregation) [18,21]. As the brain does not know *a priori* whether signals come from common or independent sources, it needs to infer the causal structure from the noisy sensory signals themselves such as them occurring at the same time or location. To account for observers' uncertainty about the signal's causal structure, the brain is thought to compute a final estimate by combining the fusion and segregation estimates weighted by the posterior probabilities that signals come from common or independent sources. This decisional strategy is referred to as model averaging (for other decisional strategies see [24]).

Accumulating psychophysics and neuroimaging research has shown that human observers combine sensory signals consistent with the principles of Bayesian causal inference by dynamically encoding segregation, fusion and the final Bayesian causal inference estimates along the cortical hierarchy [25–29]. Yet, little is known about how observers monitor their uncertainties in multisensory environments, in which signals can come from common or separate sources. These more realistic scenarios require the brain to monitor two distinct, but intimately related forms of uncertainty [30]: causal and perceptual uncertainty. Causal uncertainty refers to observers' uncertainty about the environment's causal structure, i.e. about whether sensory signals come from common or independent sources. Perceptual uncertainty pertains to observers' reported perceptual estimate such as their perceived syllable extracted from auditory speech and/or visual facial movements. Causal and perceptual uncertainty interactively arise during perceptual inference and are corrupted by the same sensory noise [18,21]. At small intersensory discrepancies when auditory and visual signals are likely to come from one source, signals are fused into a unified more precise audiovisual percept associated with low causal and perceptual uncertainty. By contrast, at intermediate levels of audiovisual discrepancy, observers will be more uncertain about whether signals come from common or independent sources. According to Bayesian causal inference models, the brain would average the fusion and the segregation distributions approximately with equal weight leading to a broader posterior distribution and hence lower perceptual confidence (see figures 1 and 8 for illustration).

The relationship between causal and perceptual confidence can be studied by explicitly manipulating the discrepancy of the physical stimuli. Yet, even physically identical auditory and visual stimuli may elicit different perceptual and causal decisions because of internal and external noise. This inter-trial variability enables us to characterize the relationship between observers' causal and perceptual confidence over trials using dual tasks that combine causal and perceptual confidence reports.

This psychophysics study characterized the relationship between perceptual and causal confidence in audiovisual syllable categorization. We presented human observers with spoken syllables (i.e. auditory phonemes), their corresponding articulatory facial movements (i.e. visemes), and their congruent (e.g. visual Ba and auditory Ba) and incongruent (e.g. visual Ga and auditory Ba) combinations. The incongruent phoneme-viseme pairs were designed to elicit veridical 'B/P' or illusory 'D/T' or 'G/K' auditory percepts (i.e. McGurk-MacDonald illusion) [31]. We refer to both 'D/T' and 'G/K' auditory percepts as illusory because the veridical auditory percept is a 'B/P' percept. On each audiovisual trial, observers

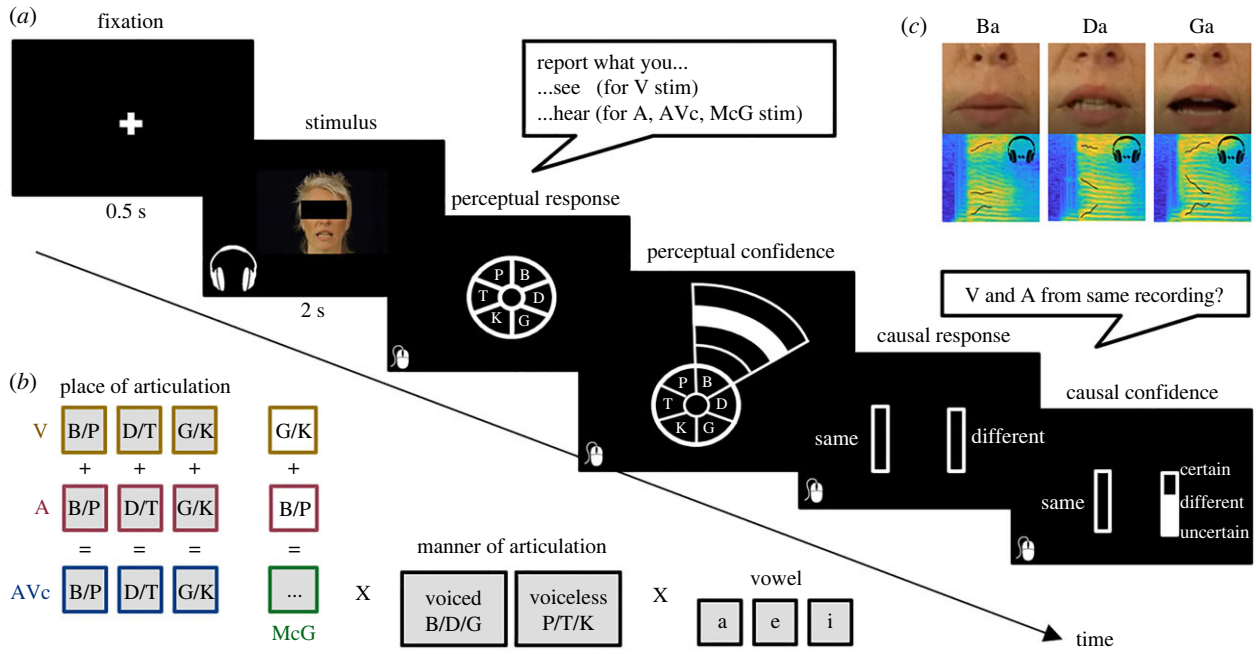

Figure 2. Experimental paradigm, example trial and stimuli. (*a*) Example trial: on each trial, observers are presented with a 2 s audio of a spoken syllable (e.g. Ba), a female face articulating a syllable (e.g. Ga) or their congruent or conflicting McGurk combinations. (Note: the eyes-covering black bar was added here for anonymization purposes, but it was not present during stimulus presentation.) After stimulus presentation, the six possible first-letter consonants were presented in a circle. Observers reported the syllable they heard on audio (A) and audiovisual (AV) trials and the syllable they saw on the visual (V) trials. They indicated their perceived first-letter consonant by moving their mouse over their preferred response option, after which a layered arc automatically appeared from which participants could select their perceptual confidence level on a 4-point scale (inner layer = low confidence = 1, outer layer = high confidence = 4). On AV trials, observers were then prompted to indicate whether the two auditory and visual stimulus components came from the 'same' or from two 'different' recordings together with their causal confidence by moving their mouse to a left or right bar. (*b*) Experimental paradigm and stimulus space: stimuli were videos, audios or their congruent and McGurk combinations. The stimuli differed along place of articulation (labial: B/P, dental: D/T and guttural: G/K), manner of articulation (voiced: B/D/G and unvoiced: P/T/K) and vowel (a, e, i). (*c*) Example stimuli: video frames of the articulatory positions of a Ba, Da and Ga stimuli and corresponding time-frequency spectrograms of the sounds with the first three formants in black. (Online version in colour.)

reported their perceived auditory phoneme, the signals' causal structure and their perceptual and causal confidence.

First, we assessed whether observers integrate audiovisual congruent and McGurk signals into percepts that are associated with greater confidence than their unisensory counterparts. Second, we characterized the complex relationship between causal and perceptual estimates and their associated confidence. Third, we selected audiovisual congruent and McGurk trials on which the sensory signals evoked causal and perceptual metamers, i.e. observers reach the same causal and perceptual decisional outcome. For instance, congruent (i.e. Da-Da) and McGurk (Ga-Ba) trials on which observers report a 'Da' percept and a common source are perceptual and causal metamers [30,32]. We assessed whether despite the same perceptual and causal outcomes observers may still be able to discriminate between them and assign greater levels of causal and perceptual confidence to the congruent than the McGurk trials.

2. Methods

(a) Participants

Fifteen participants were initially recruited, but one did not finish the study. Therefore, 14 participants were included in the analysis (two males, two left handers, mean age: 19.5, range: 18–22). The number of participants was a convenience sample. All participants gave written informed consent to participate in this psychophysics study, and they were compensated by means of study credits. Participants reported no history of psychiatric or neurological disorders, and no current use of any psychoactive medications. All had normal or corrected to normal vision and reported normal hearing. The study was approved by the research ethics committee of the University of Birmingham (ERN_11-0470AP4).

(b) Stimuli

Stimulus material was taken from close-up audiovisual recordings of a female actress' face on a dark background looking straight into the camera (figure 2*a*) and uttering the following 18 syllables: ba, be, bi, da, de, di, ga, ge, gi, pa, pe, pi, ta, te, ti, ka, ke and ki. In short, the recordings factorially combined six consonants (B, G, D, P, K, T) with three vowels (a, e, i). The six consonants can be organized into a two-dimensional space spanned by the dimensions of (i) place of articulation (i.e. production place along the vocal tract): B/P (labial), D/T (dental) and G/K (guttural); and (ii) voicing: unvoiced (P, T, K) and voiced (B, D, G). Audio and video were recorded with a camcorder (HVX 200 P; Panasonic). The video was acquired at 25 frames per second phase alternating line (PAL; 768 × 576 pixels); audio was acquired at 44.1 kHz (two channels). The recorded videos were edited (using PiTiVi 0.15.2) into 2000 ms long segments (50 frames) with the first articulatory movement starting at $t = 1$ second after the beginning of the movie (for further details see [33]). Each video started and finished with a neutral closed-lip view of the speaker's face.

We used the movies of 18 different syllables as congruent stimuli. We generated six McGurk stimuli by cross-dubbing the video and the audio-track from the B-vowel (auditory) + G-vowel (visual) and the P-vowel (auditory) + K-vowel (visual) stimuli for the three vowels (a, e, i). Further, we presented the

corresponding 18 auditory and visual components as unisensory stimuli (see also figure 2*b*).

Finally, we added a considerable amount of white noise to all auditory recordings over the full 2 s epoch length in order to increase the expected proportion of illusory percepts on McGurk trials [34,35]. The noise signal was sampled at random from a zero-centred normal distribution with a s.d. that was equal to 1/3 of the maximum amplitude of the speech signal. This resulted in a signal to noise ratio (SNR; computed with a 30 ms sliding window) that increased rapidly from −70 dB to −13 dB in the first 50 ms after syllable onset (±4 dB s.d. across the 18 auditory stimuli). Thereafter, the SNR gradually increased further to a maximum of 1.5 dB (±2.5 dB s.d.) at 150 ms after syllable onset.

(c) Experimental design and trial

The experiment included audiovisual (AV) congruent, AV McGurk trials, unisensory auditory (A) and visual (V) trials. On AV trials, observers reported the first-letter consonant that they heard and their perceived causal structure (common versus independent sources) together with their perceptual and causal confidence, respectively. On unisensory trials, observers reported their perceived first-letter consonant together with their perceptual confidence.

Each trial started with the presentation of a central fixation cross for 500 ms on a black background. Subsequently, the 2 s stimulus (A/V/AV) was presented. The woman's upper lip's screen position approximately matched the location of the fixation cross. On A-only trials, the fixation cross remained on screen during stimulus playback. After stimulus presentation, the six possible first-letter consonants were presented in a circle (figure 2*a*). Observers were instructed to report the syllable they heard on A and AV trials and the syllable that they saw on the V trials. Participants indicated their perceived first-letter consonant by moving their mouse over their preferred response option, after which a layered arc automatically appeared from which participants could select their confidence level on a 4-point scale (inner layer = low confidence, outer layer = high confidence). They indicated a response with a left-mouse click. Participants could still change their mind by moving their mouse to another letter until they had clicked. This procedure ensured that they provided their perceptual report and associated confidence simultaneously.

On AV trials only, participants were then prompted to indicate whether the two auditory and visual stimulus components came from the 'same' or from two 'different' recordings by moving their mouse to a left or right bar. The vertical mouse location indicated their causal confidence on a continuous scale (bottom = uncertain, top = certain). Again, participants were allowed to change their minds until they responded by left-mouse button click. Participants were instructed to focus on response accuracy and precision rather than speed. Furthermore, participants were encouraged to scale their confidence responses across trials such that they made use of all four levels for perceptual confidence and the entire certainty bar (i.e. from bottom to top) for causal confidence.

The experiment consisted of three 2 h sessions. It began with a short familiarization block that included 54 trials: 18 AV congruent trials, followed by 18 A-only trials and finally 18 V-only trials (each syllable appeared once). Before the beginning of each mini-block of 18 trials, the instructions appeared on the screen. For AV and A trials, these read 'report what you hear', whereas for V trials participants were instructed to 'report what you see'.

The main task started after the familiarization block. Participants completed 16 blocks of 144 trials. Each block contained 36 V-only, 36 A-only and 36 AV-congruent trials (two repetitions of each of the 18 syllable), as well as 36 McGurk stimuli trials (six repetitions of each of the six McGurk stimuli sets). The AV-congruent and McGurk stimuli were randomly interleaved, whereas the unisensory trials were presented in separate mini-blocks.

(d) Experimental procedure

Participants were seated in a small dark room. They placed their chin on a chinrest that was positioned at a distance of 55 cm from a computer monitor (53 by 30 cm). Visual stimuli were shown at a frame-rate of 25 Hz in a central square of 20 by 20 cm, surrounded by a black background. Auditory stimuli were presented at 44.1 kHz by means of headphones (Sennheiser HD 280 Pro) at a comfortable listening volume (65 dB).

A Tobii EyeX eye tracker was used during the experiment to monitor whether participants focused their eye gaze on the centre of screen (±5 degrees) during stimulus presentation. The central area of focus corresponded approximately to the woman's upper lip. Participants who did not follow task instructions received corrective feedback. The eye-tracking data were not further analysed.

The experiment consisted of 2 h sessions that were performed on separate days, with maximally one week in between successive sessions. Participants had to finish the full experiment within two weeks. Participants were encouraged to take self-paced breaks between blocks within a session.

The experiments were run in Matlab 2014b using the Psychophysics Toolbox 3 [36,37] and the Tobii Eyex toolkit [38].

(e) Confidence normalization, statistical analyses and simulations

(i) Confidence normalization

The perceptual and causal confidence distributions varied substantially in mean and spread across participants. We therefore normalized the confidence distributions with the help of the cumulative distribution function. More specifically, each raw confidence value x was mapped onto a normalized confidence value equal to the corresponding value in the cumulative distribution. In cases with several identical raw confidence values, we assigned the average across their normalized confidence values to those:

$$\text{normalized confidence}(x) = \frac{1}{2}[\,p(X < x) + p(X \leq x)].$$

This transformation ensured that the mean overconfidence values in all participants was equal to 0.5 and that the confidence values were spread across the entire range. For instance, if a participant indicated maximal causal confidence (top of the bar) in 30% of the trials, the normalized causal confidence values for these trials would all be equal to the mean of 0.7 and 1 = 0.85 (see the electronic supplementary material, figure S1).

(ii) Statistical analysis

We employed generalized linear mixed models (GLMMs) that allowed us to incorporate fixed effects of interest as predictors and account for structure induced by subjects and stimuli by including those as random effects [39]. We started with the most comprehensive model and stepwise reduced the number of random effects until model estimation converged. Typically, our models incorporated subjects as random intercept and slope, and stimulus as intercept. For detailed specification of each model, i.e. its fixed effects predictors and random effects, please see the formulae in Wilkinson notation stated in each of the statistical results tables (electronic supplementary material). We used GLMMs with logit as link function and with a binomial distribution for binary outcomes such as 'correct versus incorrect' (i.e. accuracy) or C = 1 versus C = 2 (i.e. causal inference judgements.), a multinomial distribution for categorical outcomes

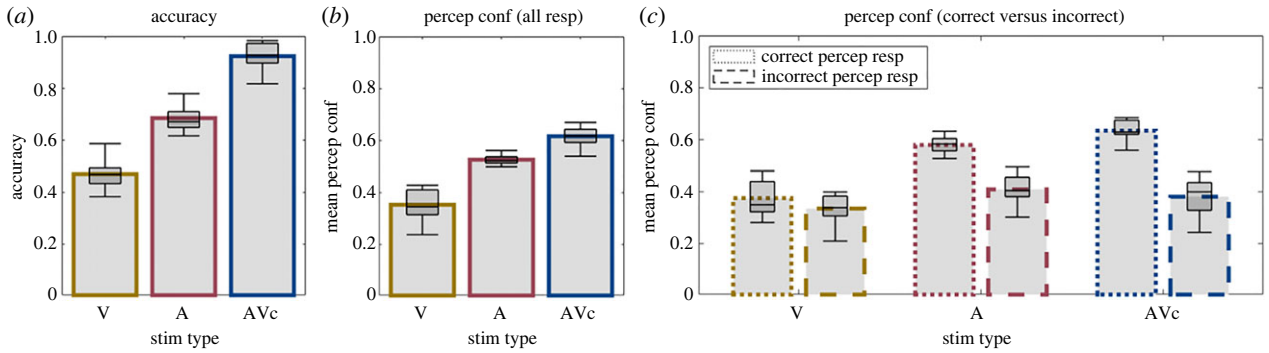

Figure 3. Perceptual accuracy and confidence for auditory (A), visual (V) and audiovisual congruent (AVc) stimuli. (*a*) Perceptual accuracy for V, A and AV congruent: perceptual accuracy is greater for AV congruent than unisensory A or V stimuli. (*b*) Perceptual confidence for V, A and AV congruent: perceptual confidence is greater for AV congruent than unisensory A or V stimuli. (*c*) Perceptual confidence for V, A and AV congruent separately for correct and incorrect syllable percepts: we observe an interaction between correct versus incorrect and sensory modality (A, V AV). In all figure components, bar plots show across subjects' means and box plots indicate across subjects' median response fraction with first and third quartiles and with whiskers indicating the minimum and maximum subject values. (Online version in colour.)

such as B/P, D/T versus G/K percepts and beta distributions for normalized confidence ratings (i.e. bounded to a range between 0 and 1).

The statistical analyses were performed in R v4.3.0 [40], using the glmmTMB and mclogit packages [41,42]. Based on the fitted models, we generated model predictions by computing marginal means of the response variables for each of the conditions, i.e. factor level combinations [43,44]. Details and results of each analysis can be found in the electronic supplementary material, tables (main).

Additional alternative models such as ordinal and ordered beta regression were used to analyse perceptual and causal confidence without the prior normalization described above [45,46]. These alternative analyses confirmed our main statistical results and are therefore not further discussed. Their details and results can be found in the electronic supplementary material, tables (alternative).

(iii) Simulations

To illustrate the relationship between perceptual and causal decisions as well as their corresponding confidence levels, we performed simulations based on the Bayesian causal inference model (for details see [18]). Originally, the Bayesian causal inference model has been developed to model spatial categorization responses in which continuous spatial estimates are mapped onto discrete spatial choices. Likewise, our simulations made the assumption that Ba, Da and Ga stimuli lie on a continuous abstract 'place of articulation' dimension that goes from labial (i.e. lip) 'Ba' to dental (i.e. 'teeth') 'Da' and finally to guttural (i.e. throat) 'Ga'. Further, we assumed that this dimension is shared across the visual and auditory senses. These continuous estimates are mapped onto 'Ba', 'Da' and 'Ga' categories via categorical perception. Auditory and visual senses provide information about the place of articulation via formants for different phonemes and articulatory movements (i.e. viseme). The second formant in the time-frequency spectrograms discriminates between Ba, Da and Ga (figure 2*c*). Likewise, the articulatory lip movements inform about the place of articulation.

Modelling audiovisual speech integration with a single shared audiovisual dimension (see also [47]) ignores complexities that arise from more structured inputs of phonemes and visemes that naturally live in a multidimensional space. For instance, our modelling approach ignores the dimension of voicing. Moreover, it ignores nonlinearities that may affect auditory and visual stimulus dimensions differently. An alternative way is to model audiovisual integration of phoneme-viseme pairs in a two-dimensional space in which the auditory stimulus only weakly activates visual information and vice versa [22]. Yet, such a two-dimensional space is agnostic about what the visual or auditory features refer to in physical space and what these cross-modal co-activations account for.

For the example shown in figure 1 and each of the four examples in figure 8, we sampled one auditory and one visual signal from $N(S_A = -1, \sigma_a^2 = 0.73)$ and $N(S_V = 1, \sigma_V^2 = 0.73)$ and we computed the likelihoods $\mathcal{L}(S; x_V)$ and $\mathcal{L}(S; x_A)$ (pink and brown dashed lines). Assuming a flat (i.e. uninformative) prior over the place of articulation dimension and a causal prior $P(C = 1) = 0.5$, we computed the posterior distribution:

$$P(S_A|x_A,x_V) = \sum_C P(S_{AV}|C,x_A,x_V)P(C|x_A,x_V).$$

This posterior distribution (solid black line) is a mixture of the full segregation $P(S_A|x_A,x_V,C = 2)$ (pink solid) and the fusion distributions $P(S_A|x_A,x_V,C = 1)$ (blue solid) weighted by the posterior probabilities over common and independent causes $P(C|x_A,x_V)$ (green bar plots). To obtain the discrete posterior probabilities over syllable categories 'B/P', 'D/T' and 'G/K' (orange bar plots), we integrated the continuous posterior probability distribution limited by the category boundaries that separate the three response choices (i.e. category boundaries 'Ba' versus 'Da' was set to −0.5 and 'Da' versus 'Ga' to 0.5). We present these simulation results to provide a qualitative explanation for the pattern of findings in our study. We note that for more complex abstract dimensions such as 'place of articulation' several assumptions of the Bayesian model may not fully hold, so that we refrain from formal quantitative modelling (e.g. Gaussian distributions, common decision dimension shared between auditory and visual senses, additional variability within categories etc. For related approaches see [48,49]).

3. Results

In the following, we present the key results organized in line with the results figures and electronic supplementary material, tables (main statistical analyses). The complete statistical results are presented only in the tables (see the electronic supplementary material).

(a) Performance accuracy and perceptual confidence in auditory, visual and audiovisual congruent conditions (figure 3)

Consistent with Bayesian models of multisensory perception observers integrated audiovisual signals into more precise estimates as indicated by their greater perceptual accuracy.

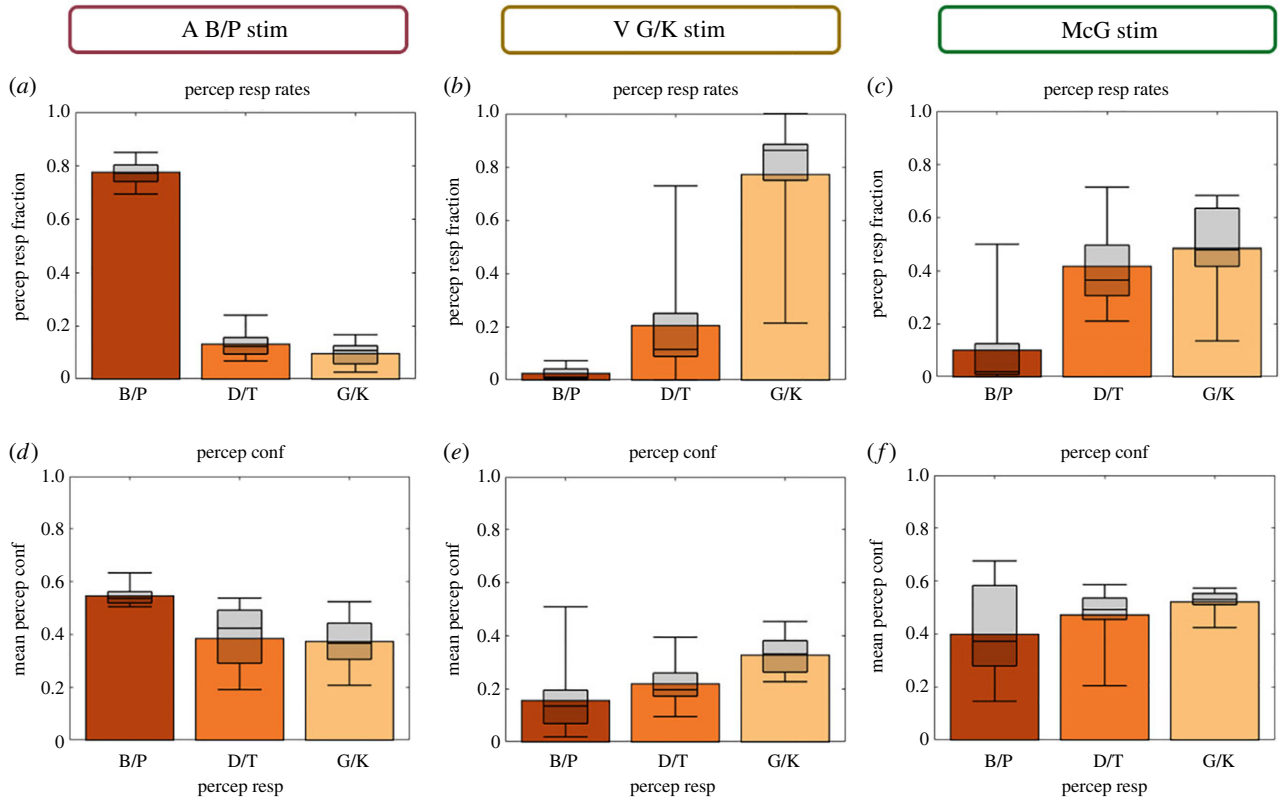

Figure 4. Perceptual response fractions (*a–c*) and perceptual confidence (*d–f*) for auditory B/P, visual G/K and McGurk stimuli. (*a*) Response fractions over 'B/P', 'D/T' and G/K' percepts (top) and associated perceptual confidence (bottom) separately for auditory B/P stimulus (*a,d*), visual G/K stimulus (*b,e*) and McGurk stimulus (i.e. auditory B/P and visual G/K; (*c,f*)). In all figure components, bar plots show across subjects' means and box plots indicate across subjects' median response fraction with first and third quartiles and with whiskers indicating the minimum and maximum subject values. (Online version in colour.)

Moreover, because the variance of the posterior distribution of the audiovisual distribution is smaller or equal to that of either unisensory posterior distribution, we would expect participants to be more confident on audiovisual congruent than unisensory conditions. Indeed, in line with these predictions, we observed an increase in perceptual accuracy and confidence for the audiovisual congruent relative to the unisensory conditions (figure 3*a*,*b*; electronic supplementary material, table S1A-B). Performance accuracy increased significantly for audiovisual congruent relative to both auditory and visual conditions. Likewise, perceptual confidence increased significantly for audiovisual congruent relative to visual (but not auditory) conditions. The electronic supplementary material, figure S3 further shows that this response pattern for mean perceptual confidence is highly consistent across subjects.

Observers were sensitive to their own accuracy as indicated by a significant increase in perceptual confidence for correct relative to incorrect responses (figure 3*c*; electronic supplementary material, table S1B). This metacognitive sensitivity was significantly greater for congruent than auditory and in particular than visual conditions (i.e. significant interaction between correct/incorrect $\times$ sensory modality). This effect can be explained by the differences in performance accuracy across sensory modalities. At first sight, it may be surprising that observers were 50% accurate and hence significantly better than the chance level of 16.67% (i.e. 1/6 possible response options) on the visual first-letter categorization task, but showed no metacognitive sensitivity. In other words, they were unable to discriminate between their correct and wrong visual decisions despite showing better than chance performance. This metacognitive insensitivity arises from the fact that the six response options are spanned by the place of articulation (e.g. B versus D versus G) and the voicing (i.e. B versus P) dimensions. While the visual modality is very informative about the place of articulation, it is uninformative about the voicing dimension. Hence, the vast majority of errors are 'voicing errors' such as misclassifying a Ba as a Pa stimulus. As shown in the electronic supplementary material, figure S2, observers were at chance (i.e. approximately 50% correct) when discriminating between voiced (e.g. Ba) and the corresponding unvoiced (e.g. Pa) visual stimuli. When observers perform a task at chance, it is impossible for them to metacognitively discriminate between correct and incorrect responses. So, observers' incapacity to discriminate between voiced and unvoiced syllables based on the visual input alone explains their metacognitive insensitivity for unisensory visual trials (see also the electronic supplementary material, figure S2).

(b) Perceptual decisions and confidence on unisensory and McGurk trials (figure 4)

The bar plots in figure 4 show how the brain combines auditory 'B/P' and visual 'G/K' information on McGurk conflict trials together with observers' associated confidence levels. The perceptual accuracy for place of articulation (n.b. pooled over voicing) is comparable for unisensory auditory B/P and visual G/K stimuli. We assessed this statistically using a multinomial mixed effects model in which we classified 'B/P' responses as 'correct' for auditory B/P stimuli and as 'other' for visual G/K stimuli. Conversely, we labelled the 'G/K' responses as 'correct' for visual G/K stimuli and as

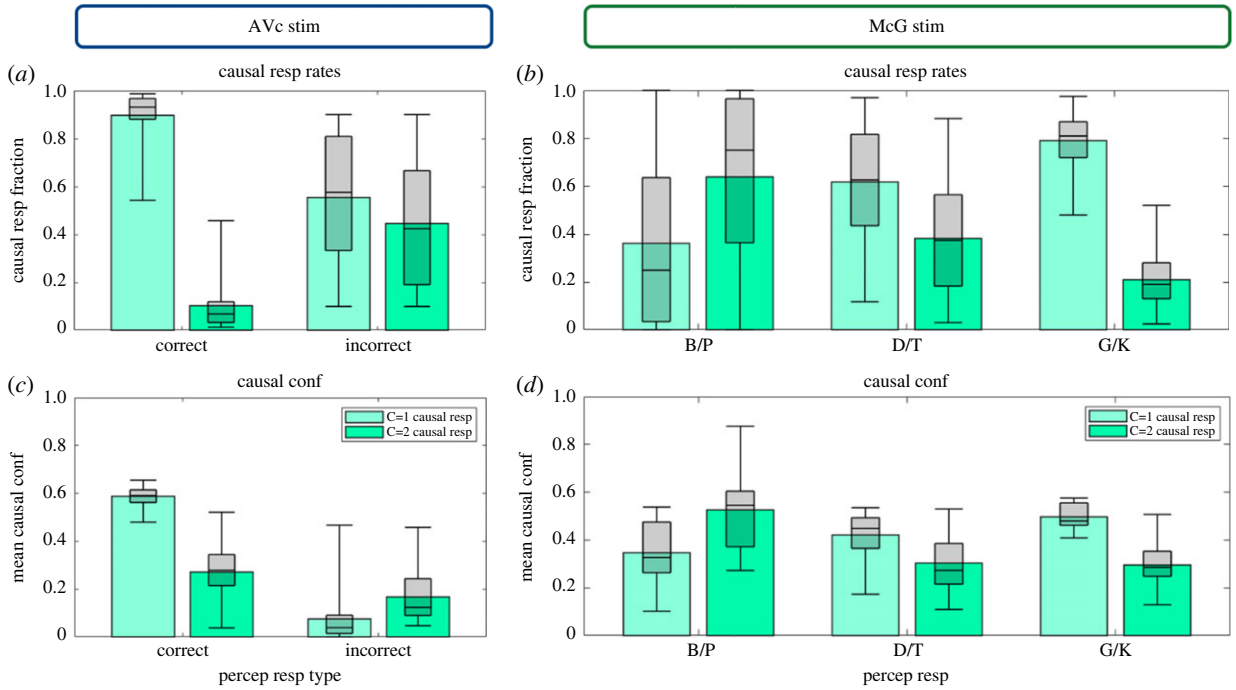

Figure 5. Causal response fractions (*a*,*b*) and causal confidence (*c*,*d*) for audiovisual congruent and McGurk stimuli. (*a*,*c*) Response fractions over 'common' (C = 1) and 'separate' (C = 2) cause responses (*a*) and the associated causal confidence (*c*) separately for audiovisual congruent trials on which observers categorized the syllable correctly or incorrectly. (*b*,*d*) Response fractions over 'common' and 'separate' cause responses (*b*) and the associated causal confidence (*d*) separately for McGurk trials with 'B/P', 'D/T' and 'G/K' percepts. In all figure components, bar plots show across subjects' means and box plots indicate across subjects' median response fraction with first and third quartiles and with whiskers indicating the minimum and maximum subject values. (Online version in colour.)

'other' for auditory B/P stimuli. The 'D/T' responses were labelled as 'D/T' responses for both auditory and visual stimuli. This analysis revealed no significant differences in response probabilities for 'correct' versus 'D/T' responses or for 'correct' versus 'other' responses between auditory and visual stimuli (see the electronic supplementary material, table S2A). Critically, however, the (predicted) response probabilities for the 'G/K' category (i.e. 'other') on auditory B/P trials is four times higher than the response probability for 'B/P' category (i.e. 'other') on visual G/K trials (electronic supplementary material, table S2A). On 8% of the trials an auditory B/P stimulus is perceived as 'D/T' and on 4% of the trials it is perceived as 'G/K'. By contrast, a visual G/K stimulus is perceived as a 'D/T' on 12% of the trials, but it is almost never perceived as a 'B/P' syllable (1%). This difference between auditory and visual response probabilities explains that McGurk stimuli are mainly perceived as 'D/T' and 'G/K', i.e. perceptual categories that are consistent with perceptual interpretations of both the visual G/K and the auditory B/P stimulus (electronic supplementary material, table S2B; i.e. significant 'intercept' effects for 'D/T' and for 'G/K' responses relative to 'B/P' responses).

The large fraction of 'G/K' perceptual responses on McGurk trials is consistent with previous studies of the McGurk illusion [34,35]. It can be explained by the additional noise that we added to the auditory component signal (see methods section) to decrease its reliability. As a result, the auditory signal received a lower weight when observers integrate auditory and visual signals during perception. However, it is important to note that observers did not simply go with the visual signal and ignored the sound on trials with a 'G/K' percept. Instead, they integrated audiovisual signals even on those trials with visual dominant 'G/K' percept as evidence by their voicing classification accuracy. The substantial increase in accuracy on the voicing dimension for McGurk relative to visual-only trials demonstrates that observers relied on both visual and auditory information even on trials with a visual dominant 'G/K' percept (see the electronic supplementary material, figure S2).

Overall, perceptual confidence is significantly higher for McGurk than for unisensory auditory or visual stimuli. Observers were thus more confident on conflicting McGurk trials than on unisensory trials. Moreover, this increase in perceptual confidence for McGurk relative to unisensory trials was particularly pronounced for 'D/T' and 'G/K' relative to 'B/P' perceptual outcomes (i.e. significant interactions between stimulus type and perceptual outcome; electronic supplementary material, table 2C). Similarly, observers' perceptual confidence was higher on McGurk trials with 'G/K' outcomes, i.e. perceptual interpretations that integrate audiovisual information, than with 'B/P' outcomes where audiovisual information was successfully segregated (electronic supplementary material, table S2D). Collectively, these results suggest that observers are more confident even on conflicting McGurk trials when they integrate audiovisual signals into a 'G/K' percept than on unisensory trials or on McGurk trials on which they perceive separate causes (i.e. 'different' sources).

(c) The relationship between perceptual and causal inference (figure 5)

Perceptual and causal decisions are intimately related in the inference process and susceptible to shared sensory noise. These dependencies explain that correct common cause 'C = 1' responses on congruent trials go together with correct perceptual responses (i.e. significant effect of trial correctness on causal response fraction; figure 5*a*; electronic supplementary material, table S3A). Likewise, observers associated a greater

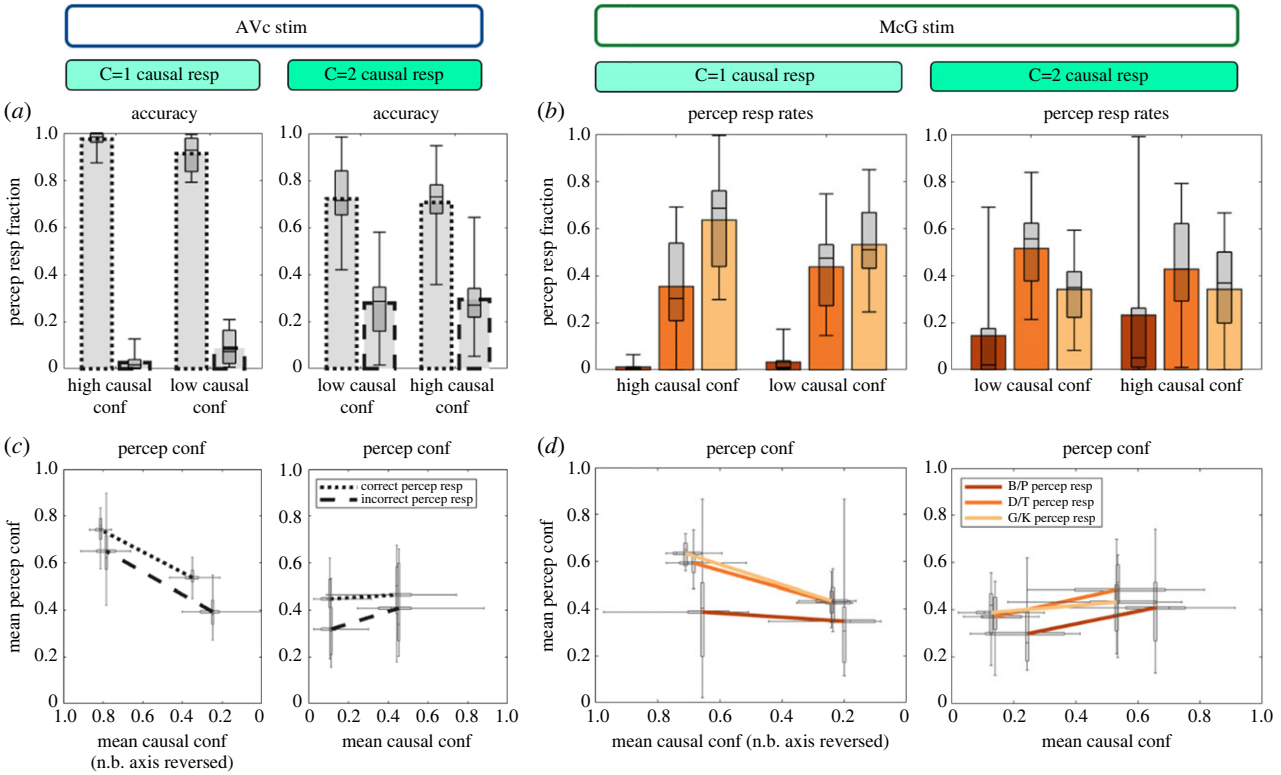

Figure 6. Perceptual and causal response fractions and their associated perceptual and causal confidence. (*a*,*c*) Fraction of correct and incorrect perceptual responses (*a*) as well as their associated causal and perceptual confidence (*c*) separately for 'common' (C = 1) and 'separate' (C = 2) cause response trials. (*b*,*d*) Fraction of 'B/P', 'D/T' and 'G/K' percepts (*b*) as well as their associated causal and perceptual confidence (*d*) separately for 'common' and 'separate' cause response trials. In all figure components, bar plots show across subjects' means and box plots indicate across subjects' median response fraction with first and third quartiles and with whiskers indicating the minimum and maximum subject values. Likewise, the crosses in the line plots (*c*,*d*) indicate first and third quartiles and the longer cross whiskers the minimum and maximum subject values. (Online version in colour.)

causal confidence to correct common cause responses mainly when they categorized the first letter correctly (i.e. significant interaction in causal confidence for common 'C = 1' versus independent 'C = 2' cause responses x correct versus incorrect; electronic supplementary material, table S3C; figure 5*c*).

The close relationship between causal and perceptual inference is also manifest in the McGurk trials (figure 5*b*,*d*; electronic supplementary material, table S3B,D). The fraction of common cause responses is directly associated with observers' perceptual categorization responses, increasing from 'B/P' to 'D/T' and 'G/K' responses (i.e. significant main effect of perceptual category on causal response fractions; electronic supplementary material, table S3B; figure 5*b*). Likewise, the causal confidence for common cause 'C = 1' responses was greatest for 'G/K' percepts, while the causal confidence for independent cause 'C = 2' responses peaked for 'B/P' percepts (i.e. significant interactions between causal and perceptual outcomes on causal confidence; electronic supplementary material, table S3D; figure 5*d*). In other words, as expected from Bayesian causal inference (figure 1), the integrated percept that is conditional on a common cause receives more weight (over the segregated percept) when a common cause is deemed more probable (as expressed by greater causal confidence with 'C = 1' responses), thereby increasing the 'D/T' and 'G/K' response probabilities (also see the electronic supplementary material, figure S4). Vice versa, larger causal confidence with 'C = 2' responses lead to higher weights for the segregated percept, thus increasing the probability of a 'B/P' percept.

(d) The relationship between perceptual and causal confidence (figure 6)

Our dual-task design enables us to characterize the relationship between causal and perceptual confidence based on inter-trial variability. To visualize this, we sorted observers' perceptual decisions and confidence according to their causal decision and confidence on those trials (median split per response category for each subject, although note that such a median split was unnecessary for the statistical analysis). Consistent with Bayesian causal inference models, common source decisions and high causal confidence were associated with high perceptual accuracy (figure 6*a*) and perceptual confidence (figure 6*c*). Statistically, this observation is supported by a significant interaction between causal decision and causal confidence on observers' perceptual accuracy. Moreover, we observed a significant main effect of causal confidence (electronic supplementary material, tables S4A,C) as well as a significant interaction between perceptual accuracy and causal response on perceptual confidence. Observers' perceptual confidence increased with their causal confidence in particular for common cause responses, but less so for independent cause responses. Thus, when observers made a wrong response the correlation between their perceptual and causal confidence was attenuated most likely because of random guesses (see below).

Likewise, on McGurk trials, observers' causal inference outcome and causal confidence were closely related with their perceptual outcome and perceptual confidence. As

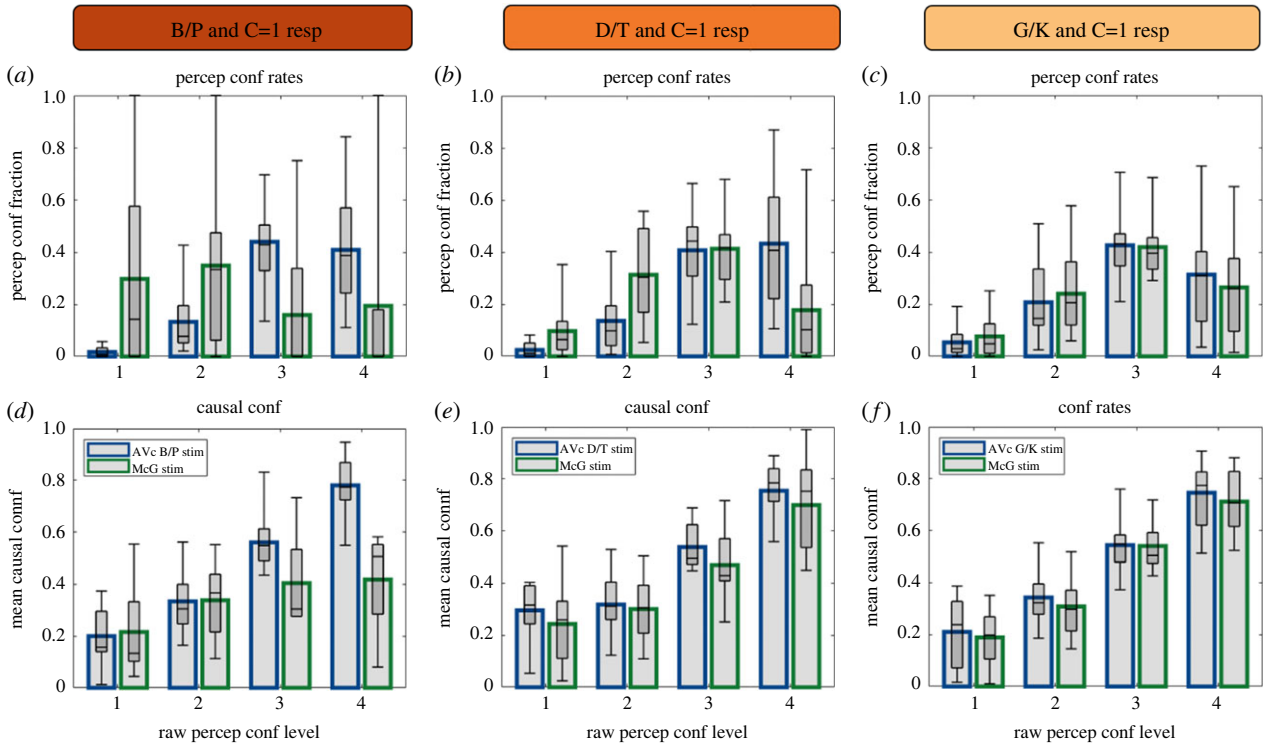

Figure 7. Perceptual and causal metamers. (*a–c*) Response fractions over the four perceptual confidence levels for congruent stimuli with correct perceptual response and C = 1 causal response (blue) and McGurk stimuli with correct perceptual voicing response and C = 1 causal response (green) separately for 'B/P' (*a*), 'D/T' (*b*), and 'G/K' (*c*) percepts. (*d–f*) Causal confidence for the four perceptual confidence levels for congruent (blue) and McGurk (green) trials (as in *a–c*) separately for 'B/P' (*d*), 'D/T' (*e*) and 'G/K' (*f*) percepts. In all figure components, bar plots show across subjects' means and box plots indicate across subjects' median response fraction with first and third quartiles and with whiskers indicating the minimum and maximum subject values. (Online version in colour.)

predicted by the Bayesian causal inference model, illusory 'D/T' and 'G/K' percepts on McGurk trials were significantly more frequent for common cause responses with high relative to low causal confidence (i.e. significant interaction between causal decision and causal confidence on perceptual response fractions 'D/T' versus 'B/P' and 'G/K' versus 'B/P'; electronic supplementary material, table S4B; figure 6*b*). By contrast, 'B/P' percepts were observed mainly on trials with independent cause decision and high level of causal confidence (electronic supplementary material, table S4B; figure 6*b*).

On McGurk trials, perceptual confidence was positively correlated with causal confidence (significant main effect of causal confidence on perceptual confidence; electronic supplementary material, table S4D; figure 6*d*). However, this main effect arose from a complex three-way interaction between perceptual outcome (e.g. 'G/K' versus 'B/P'), causal outcome (C = 1' versus 'C = 2) and causal confidence (electronic supplementary material, table S4D; figure 6*d*). As shown in figure 6*d*, perceptual and causal confidence were closely related for all trials apart from those with 'B/P' percepts with 'C = 1', i.e. wrong, causal responses—most likely because those erroneous responses reflect random guesses during lapses of attention.

In summary, both congruent and McGurk trials demonstrate that causal inference outcome and causal confidence implicitly affect observers' perceptual choices and confidence.

(e) Causal and perceptual metamers (figure 7)

The results presented so far show that on a large percentage of McGurk trials observers integrated the auditory B/P and the visual G/K stimulus components into an auditory 'D/T' or 'G/K' percept and report a common cause (C = 1). On these trials, observers thus integrated audiovisual McGurk signals into perceptual and causal metamers of the corresponding D/T and G/K congruent trials with correct perceptual and causal responses. This raises the critical question of whether, despite identical perceptual and causal decisions, observers were metacognitively aware that the 'D/T' and 'G/K' percepts on McGurk trials rely on incongruent audiovisual information and hence report lower perceptual and causal confidence relative to their congruent metamers.

To address this question, we categorized congruent trials (with correct syllable response and 'C = 1' causal response) and McGurk trials (with correct voicing response and 'C = 1' causal response) according to their perceptual confidence levels separately for 'B/P' (figure 7*a*,*d*), 'D/T' (figure 7*b*,*e*) and 'G/K' (figure 7*c*,*f*) percepts. For instance, figure 7, left column, shows the response probabilities for the four different confidence levels as a fraction of all trials with a 'B/P' percept and 'C = 1' causal outcome in the top row and the associated causal confidence ratings in the bottom row (blue = AVc B/P stimulus; green = McGurk stimulus). We assessed statistically whether the mean perceptual confidence (e.g. averaged across all AVc B/P trials with correct perceptual and C = 1 responses) differed between AV congruent and McGurk stimuli. These statistical analyses were performed separately for each perceptual category, i.e. separately for 'B/P' (figure 7*a*; electronic supplementary material, table S5A), 'D/T' (figure 7*b*; electronic supplementary material, table S5B) and 'G/K' (figure 7*c*; electronic supplementary material, table S5C) percepts. We observed significantly lower perceptual confidence on McGurk trials with 'B/P' and 'D/T' percepts compared to

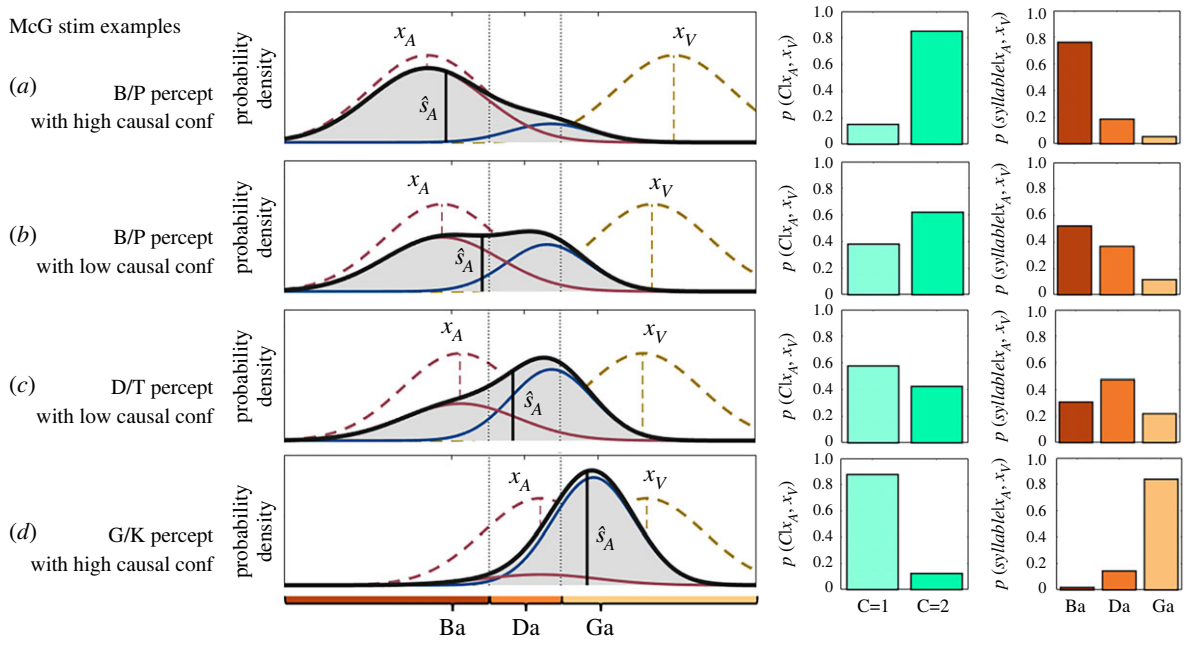

Figure 8. Simulation results of the Bayesian causal Inference model. Each row shows the simulation results for a particular noisy auditory and visual signal pair that have been sampled from the generative model (see methods for details). The noisy auditory and visual signals were set to (*a*) $x_A = -1.45$; $x_V = 2.20$; (*b*) $x_A = -1.23$; $x_V = 1.88$; (*c*) $x_A = -0.95$; $x_V = 1.76$; (*d*) $x_A = 0.23$; $x_V = 1.80$. For each sampled audiovisual signal pair, we show the likelihoods $\mathcal{L}(S; x_V)$ and $\mathcal{L}(S; x_A)$ (pink and brown dashed lines), the posterior distribution (solid black line) as a mixture of the full segregation $P(S_A|x_A,x_V,C = 2)$ (pink solid) and the fusion distributions $P(S_A|x_A,x_V,C = 1)$ (blue solid) weighted by the posterior probabilities over common and independent causes $P(C|x_A,x_V)$ (green bar plots). The discrete posterior probabilities over 'B/P', 'D/T' and 'G/K' perceptual categories (orange bar plots) were obtained by integrating over the continuous posterior probability distribution between the respective category boundaries. (Online version in colour.)

their corresponding AV congruent trials (electronic supplementary material, table S5A-B) suggesting that observers could metacognitively discriminate between AV congruent and McGurk trials despite identical perceptual outcome. Importantly, however, the perceptual confidence was not statistically different between McGurk trials with 'G/K' percepts and their congruent metamers (electronic supplementary material, table S5C). Likewise, Akaike information criterion (AIC) and Bayesian information criterion (BIC) did not provide consistent evidence for a difference in perceptual confidence between AV congruent and McGurk trials with 'G/K' percept.

Next, we assessed whether observers assigned lower causal confidence levels to McGurk trials than congruent stimuli, when we account for differences in perceptual confidence levels by including perceptual confidence as a regressor in our GLMMs (electronic supplementary material, table S5D–F). Again, we observed a lower causal confidence for McGurk relative to congruent trials only for 'B/P' and 'D/T' percepts (electronic supplementary material, table S5D-E; n.b. expressed by the significant interaction between perceptual confidence and McGurk versus AV congruent trials). By contrast, neither the main effect of stimulus type (i.e. McGurk versus AV congruent) nor its interaction with perceptual confidence was significant for trials with 'G/K' percepts (electronic supplementary material, table S5F). These null results were further corroborated by formal Bayesian model comparison. Both AIC and BIC jointly provided evidence that AV congruent and McGurk trials did not significantly differ in their causal confidence (i.e. the more parsimonious model without the predictor stimulus type was a better fit to the data; electronic supplementary material, table S5F). These results suggest that observers regularly integrate conflicting signals from McGurk stimuli into auditory 'G/K' percepts that are associated with comparable perceptual and causal confidence as their metameres from G/K congruent stimuli. On those subsets of McGurk trials observers are no longer metacognitively aware of the conflicting 'B/P' phoneme and 'G/K' viseme stimuli.

4. Discussion

This study investigated how human observers form confidence judgements when presented with spoken syllables, articulatory lip movements or their congruent and McGurk combinations. In such multisensory information integration tasks, observers need to monitor two intimately related sorts of uncertainty: perceptual uncertainty about environmental properties (e.g. syllable's first letter) and causal uncertainty about whether signals come from common or independent sources. Our results demonstrate that human observers form meaningful perceptual and causal confidence judgements that are qualitatively in line with the principles of Bayesian causal inference.

A wealth of research has shown that human observers integrate audiovisual signals from common sources weighted by their relative reliabilities into more precise percepts [50–53]. Sensory integration reduces observers' uncertainty about the current state of the world. In our study, auditory and visual senses provide both redundant and complementary information about syllables [54]. The auditory sense facilitates the discrimination between voiced and unvoiced consonants (e.g. 'B/G/D' versus 'P/K/T') that is left ambiguous by the visual sense alone. Further, speech signals and the articulatory lip movements together inform about the place of articulation (e.g. 'B/P' versus 'D/T' versus 'G/K').

Unsurprisingly, observers benefit substantially from audiovisual integration. They show superior syllable categorization accuracy and higher perceptual confidence on audiovisual congruent relative to unisensory trials (figure 3). Only on less than 10% of the audiovisual congruent trials did observers miscategorize the syllables. As evidence for observers' metacognitive sensitivity, they assigned lower levels of confidence to perceptual categorization errors relative to their correct responses [11]. Moreover, these perceptual categorization errors also revealed a close relationship between perceptual and causal inference. Observers' causal accuracy and confidence were greater, when observers categorized the syllable correctly (figure 5*a*,*c*). Conversely, observers' perceptual accuracy and confidence were greater for trials with high than low causal confidence (figure 6*a*,*c*). As we will discuss in greater detail below, this positive relationship between perceptual and causal accuracy resp. confidence is consistent with Bayesian causal inference models, in which perceptual and causal inference arise interactively and are susceptible to shared sensory noise [18,21].

McGurk trials provide additional insights into the formation of confidence judgements by introducing a small inter-sensory conflict along the 'place of articulation' dimension, i.e. by combining an auditory B/P with a visual G/K signal. Unisensory auditory B/P stimuli were predominantly perceived as 'B/P', but in approximately 20% of the trials as 'D/T' or 'G/K'. Unisensory visual G/K signals were mainly perceived as 'G/K' and in 20% of the trials as 'D/T', but nearly never as 'B/P'. This inter-sensory difference in the distribution over perceptual categories explains that McGurk combinations are mainly integrated into 'D/T' and 'G/K' percepts that are possible perceptual explanations for both auditory B/P and visual G/K signals (figure 4). Moreover, the perceptual outcome on a McGurk trial is characteristically influenced by observers' causal inference on that trial. Consistent with Bayesian causal inference models, observers perceive the first letter of the auditory syllable as a 'D/T' or 'G/K' particularly, when they infer that auditory and visual signals come from common sources and hence integrate them (figures 5 and 8). The proportion of visual biased 'G/K' percepts and the perceptual confidence increases even further for common source judgements with high relative to low causal confidence (figure 6*b*,*d*). Conversely, observers reported a veridical 'B/P' percept, i.e. a percept unbiassed by the conflicting visual G/K signal, when they inferred that auditory and visual signals come from independent sources. Again, this proportion of 'B/P' percepts increased for high relative to low causal confidence. Moreover, for both common and independent source responses, perceptual and causal confidence were positively related: a greater causal confidence was generally associated with a greater perceptual confidence. In short, McGurk trials replicated the tight relationship between perceptual and causal inference/confidence over trials that we already observed for congruent trials.

As shown in our simulations (figure 8), this intimate relationship naturally arises in Bayesian causal inference models, because perceptual and causal inference are based on the same auditory and visual inputs that vary across trials because of sensory noise. Thus, when noisy auditory and visual signals are close together along the abstract 'place of articulation dimension', observers are likely to infer a common source and integrate audiovisual signals into a visual biased 'G/K' phoneme (figure 8 bottom row). By contrast, an auditory dominant 'B/P' percept arises only, when the probability of independent sources is very high (figure 8, top row).

While the Bayesian causal inference model can qualitatively explain the relationship between perceptual and causal decisions resp. confidence, our results cannot dissociate whether the brain forms Bayesian or approximate confidence estimates when exposed to multiple sensory signals under causal uncertainty. In these situations, the posterior probability distribution turns bimodal. Perceptual confidence may be related to a variety of quantities [55]. For instance, observers' confidence judgements may reflect the posterior probability at the particular perceptual estimate or the entropy of the full bimodal posterior probability distribution [10,56]. Alternatively, because observers perceive auditory speech signals categorically as 'B/P', 'D/T' or 'G/K', their perceptual confidence may be related to the posterior probabilities over the discrete response options rather than a continuous posterior distribution over a hypothesized place of articulation dimension. Further, in the discrete case, it is unclear whether observers' perceptual confidence reflects Bayesian confidence, i.e. the posterior probability that a decision is correct or some other quantity. For example, in a three alternative visual categorization task observers' perceptual confidence has recently been shown to reflect the difference in posterior probability between the two most likely options [57]. Because in our experimental paradigm observers made a choice among six options that were arranged in a two-dimensional 'place of articulation' x 'manner of articulation' space, it is likely that observers formed approximate or simple heuristic confidence judgements. Future research combining psychophysics with formal quantitative modelling is needed to dissociate between these different strategies to form confidence judgements.

McGurk trials provide critical insights into whether observers metacognitively monitor only the final integrated percept or whether they access the unisensory signals and underlying inference processes. An early intriguing study by Hillis *et al*. (2002) [32] has previously suggested that observers lose access to individual cues after integrating them within but not between the senses. In an oddity judgement task, conflicting visual- but not visuohaptic-cues were fused into perceptual metamers that were indistinguishable from the standard percepts derived from congruent cues. Following this rationale, we selected McGurk trials on which observers integrated the conflicting audiovisual signals into illusory 'D/T' and 'G/K' percepts and perceived a common cause. We then compared those to their corresponding perceptual and causal metamers of congruent trials, i.e. congruent audiovisual signals that elicited veridical 'D/T' and 'G/K' percepts and were perceived as coming from a common source (figure 7). We reasoned that if observers move beyond the integrated percept and retain access to their unisensory ingredients, they should assign lower confidence to the conflicting McGurk signals than their congruent counterparts [30]. Contrary to this conjecture, we observed closely matched perceptual and causal confidence ratings between congruent and McGurk stimuli with 'G/K' percepts and common cause 'C = 1' responses (figure 7; electronic supplementary material, table S5). Thus, observers obtained comparable confidence levels for perceptual and causal metamers that were unaffected by the underlying true causal structure, at least for visual biased

'G/K' percepts with common source judgements. These results suggest that observers metacognitively monitor mainly the final posterior distribution to form confidence judgements. When they integrate audiovisual signals into 'G/K' responses and infer a common source, they do not seem to have access to the unisensory signals or the true inter-sensory conflict. Future research needs to assess whether these findings generalize to other sets of McGurk stimuli. For instance, a previous online study found a significant reduction in perceptual confidence for McGurk trials with 'Da' or 'Na' percept relative to the corresponding congruent stimuli [58]. Potentially, by adding noise to auditory component signals in both congruent and McGurk trials our study may have made it more difficult for participants to discriminate metacognitively between congruent and McGurk stimuli.

In conclusion, our results show that observers form meaningful causal and perceptual confidence estimates. Consistent with the principles of Bayesian causal inference, these two forms of uncertainty are closely related over trials with higher causal confidence typically associated with higher perceptual confidence. Further, when a common source of the sensory signals is inferred, confidence is directly informed by the final integrated percept with no or only very limited access to the unisensory signals and their true causal structure.

Ethics. The study was approved by the research ethics committee of the University of Birmingham (ERN_11-0470AP4). All participants provided written informed consent to participate in the study.

Data accessibility. Unfortunately, we cannot make the raw data available because participants did not indicate explicitly in the ethics consent form that they approve of their data to be shared. We have updated our consent forms for future data collections.

Summarized data for all conditions are provided in the electronic supplementary materials (see tables of statistical analyses) [59].

Authors' contributions. D.M.: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing—original draft, writing—review and editing; U.N.: conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing—original draft, writing—review and editing.

Both authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This research was funded by the ERC starting grant ('multsens'). D.M. is currently supported by the Austrian Science Fund (FWF, ZK-66, 'Dynamates').

Acknowledgements. We thank Sonal Patel for help with data acquisition.

References

1. Pouget A, Drugowitsch J, Kepecs A. 2016 Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374. (doi:10.1038/nn.4240)
2. Knill DC, Pouget A. 2004 The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719. (doi:10.1016/j.tins.2004.10.007)
3. Ma WJ, Jazayeri M. 2014 Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* **37**, 205–220. (doi:10.1146/annurev-neuro-071013-014017)
4. Fleming SM, Dolan RJ, Frith CD. 2012 Metacognition: computation, biology and function. *Phil. Trans. R. Soc. B* **367**, 1280–1286. (doi:10.1098/rstb.2012.0021)
5. Flavell JH. 1979 Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry. *Am. Psychol.* **34**, 906–911. (doi:10.1037/0003-066x.34.10.906)
6. Yeung N, Summerfield C. 2012 Metacognition in human decision-making: confidence and error monitoring. *Phil. Trans. R. Soc. B* **367**, 1310–1321. (doi:10.1098/rstb.2011.0416)
7. Overgaard M, Sandberg K. 2012 Kinds of access: different methods for report reveal different kinds of metacognitive access. *Phil. Trans. R. Soc. B* **367**, 1287–1296. (doi:10.1098/rstb.2011.0425)
8. Fleming SM, Daw ND. 2017 Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114. (doi:10.1037/rev0000045)
9. Fleming SM, Maniscalco B, Ko Y, Amendi N, Ro T, Lau H. 2015 Action-specific disruption of perceptual confidence. *Psychol. Sci.* **26**, 89–98. (doi:10.1177/0956797614557697)
10. Aitchison L, Bang D, Bahrami B, Latham PE. 2015 Doubly Bayesian analysis of confidence in perceptual decision-making. *PLoS Comput. Biol.* **11**, e1004519. (doi:10.1371/journal.pcbi.1004519)
11. Fleming SM, Lau HC. 2014 How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443. (doi:10.3389/fnhum.2014.00443)
12. Maniscalco B, Lau H. 2012 A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Conscious Cogn.* **21**, 422–430. (doi:10.1016/j.concog.2011.09.021)
13. de Gardelle V, Le Corre F, Mamassian P. 2016 Confidence as a common currency between vision and audition. *PLoS ONE* **11**, e0147901. (doi:10.1371/journal.pone.0147901)
14. de Gardelle V, Mamassian P. 2014 Does confidence use a common currency across two visual tasks? *Psychol. Sci.* **25**, 1286–1288. (doi:10.1177/0956797614528956)
15. Mamassian P. 2016 Visual confidence. *Annu. Rev. Vis. Sci.* **2**, 459–481. (doi:10.1146/annurev-vision-111815-114630)
16. Grimaldi P, Lau H, Basso MA. 2015 There are things that we know that we know, and there are things that we do not know we do not know: confidence in decision-making. *Neurosci. Biobehav. Rev.* **55**, 88–97. (doi:10.1016/j.neubiorev.2015.04.006)
17. Shams L, Beierholm U. 2022 Bayesian causal inference: a unifying neuroscience theory. *Neurosci. Biobehav. Rev.* **137**, 104619. (doi:10.1016/j.neubiorev.2022.104619)
18. Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L. 2007 Causal inference in multisensory perception. *PLoS ONE* **2**, e943. (doi:10.1371/journal.pone.0000943)
19. Noppeney U. 2021 Solving the causal inference problem. *Trends Cogn. Sci.* **25**, 1013–1014. (doi:10.1016/j.tics.2021.09.004)
20. Noppeney U. 2021 Perceptual inference, learning, and attention in a multisensory world. *Annu. Rev. Neurosci.* **44**, 449–473. (doi:10.1146/annurev-neuro-100120-085519)
21. Rohe T, Noppeney U. 2015 Sensory reliability shapes perceptual inference via two mechanisms. *J. Vis.* **15**, 22. (doi:10.1167/15.5.22)
22. Magnotti JF, Beauchamp MS. 2017 A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLoS Comput. Biol.* **13**, e1005229. (doi:10.1371/journal.pcbi.1005229)
23. Magnotti JF, Ma WJ, Beauchamp MS. 2013 Causal inference of asynchronous audiovisual speech. *Front. Psychol.* **4**, 798. (doi:10.3389/fpsyg.2013.00798)
24. Wozny DR, Beierholm UR, Shams L. 2010 Probability matching as a computational strategy used in perception. *PLoS Comput. Biol.* **6**, e1000871. (doi:10.1371/journal.pcbi.1000871)
25. Rohe T, Ehlis AC, Noppeney U. 2019 The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. *Nat. Commun.* **10**, 1907. (doi:10.1038/s41467-019-09664-2)
26. Rohe T, Noppeney U. 2016 Distinct computational principles govern multisensory integration in primary sensory and association cortices. *Curr. Biol.* **26**, 509–514. (doi:10.1016/j.cub.2015.12.056)
27. Rohe T, Noppeney U. 2015 Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS Biol.* **13**, e1002073. (doi:10.1371/journal.pbio.1002073)

28. Aller M, Noppeney U. 2019 To integrate or not to integrate: temporal dynamics of hierarchical Bayesian causal inference. *PLoS Biol.* **17**, e3000210. (doi:10.1371/journal.pbio.3000210)
29. Ferrari A, Noppeney U. 2021 Attention controls multisensory perception via two distinct mechanisms at different levels of the cortical hierarchy. *PLoS Biol.* **19**, e3001465. (doi:10.1371/journal.pbio.3001465)
30. Deroy O, Spence C, Noppeney U. 2016 Metacognition in multisensory perception. *Trends Cogn. Sci.* **20**, 736–747. (doi:10.1016/j.tics.2016.08.006)
31. McGurk H, Macdonald J. 1976 Hearing lips and seeing voices. *Nature* **264**, 691–811. (doi:10.1038/264746a0)
32. Hillis JM, Ernst MO, Banks MS, Landy MS. 2002 Combining sensory information: mandatory fusion within, but not between, senses. *Science (N.Y.)* **298**, 1627–1630. (doi:10.1126/science.1075396)
33. Gau R, Noppeney U. 2016 How prior expectations shape multisensory perception. *Neuroimage* **124**, 876–886. (doi:10.1016/j.neuroimage.2015.09.045)
34. Tiippana K, Hayes E, Möttönen R, Kraus N, Sams M. 2010 *The McGurk effect at various auditory signal-to-noise ratios in American and Finnish listeners*. In *Proc. of the Int. Conf. on Auditory-Visual Speech Processing*, *September 30–October 3, 2010*, pp. 166–169. Hakone, Kanagawa, Japan: International Speech Communication Association (ISCA). (https://isca-speech.org/archive)
35. Tiippana K, Vainio M, Tiainen M. 2011 Audiovisual speech perception: acoustic and visual phonetic features contributing to the McGurk effect. *I-Perception* **2**, 768. (doi:10.1068/ic768)
36. Brainard DH. 1997 The psychophysics toolbox. *Spat. Vis.* **10**, 433–436. (doi:10.1163/156856897X00357)
37. Pelli DG. 1997 The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–442. (doi:10.1163/156856897X00366)
38. Gibaldi A, Vanegas M, Bex PJ, Maiello G. 2017 Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. *Behav. Res. Methods* **49**, 923–946. (doi:10.3758/s13428-016-0762-9)
39. Brauer M, Curtin JJ. 2018 Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychol. Methods* **23**, 389–411. (doi:10.1037/met0000159)
40. R Core Team. 2023 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See https://www.Rproject.org/.
41. Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Maechler M, Bolker BM. 2017 glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J.* **9**, 378–400. (doi:10.32614/RJ-2017-066)
42. Elff M. 2022 mclogit: multinomial logit models, with or without random effects or overdispersion. R package version 0.9.6. See https://CRAN.R-project.org/package=mclogit.
43. Lenth R. 2023 emmeans: estimated marginal means, aka least-squares means. R package version 1.8.6. See https://CRAN.R-project.org/package=emmeans.
44. Lüdecke D. 2018 ggeffects: tidy data frames of marginal effects from regression models. *J. Open Source Softw.* **3**, 772. (doi:10.21105/joss.00772)
45. Christensen RHB. 2022 Ordinal – regression models for ordinal data. R package version 2022.11–16. See https://CRAN.R-project.org/package=ordinal.
46. Kubinec R. 2022 Ordered beta regression: a parsimonious, well-fitting model for continuous data with lower and upper bounds. *Polit. Analysis* 1–18. (doi:10.1017/pan.2022.20)
47. Lindborg A, Andersen TS. 2021 Bayesian binding and fusion models explain illusion and enhancement effects in audiovisual speech perception. *PLoS ONE* **16**, e0246986. (doi:10.1371/journal.pone.0246986)
48. Ma WJ, Zhou X, Ross LA, Foxe JJ, Parra LC. 2009 Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS ONE* **4**, e4638. (doi:10.1371/journal.pone.0004638)
49. Bejjanki VR, Clayards M, Knill DC, Aslin RN. 2011 Cue integration in categorical tasks: insights from audio-visual speech perception. *PLoS ONE* **6**, e19812. (doi:10.1371/journal.pone.0019812)
50. Ernst MO, Banks MS. 2002 Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433. (doi:10.1038/415429a)
51. Alais D, Burr D. 2004 Ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* **14**, 257–262. (doi:10.1016/S0960-9822(04)00043-0)
52. Meijer D, Noppeney U. 2020 Computational models of multisensory integration. In *Multisensory perception: from laboratory to clinic* (eds K Sathian, VS Ramachandran), pp. 113–133. New York, NY: Academic Press. (doi:10.1016/B978-0-12-812492-5.00005-X)
53. Meijer D, Veselič S, Calafiore C, Noppeney U. 2019 Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation. *Cortex* **119**, 74–88. (doi:10.1016/j.cortex.2019.03.026)
54. Ernst MO, Bülthoff HH. 2004 Merging the senses into a robust percept. *Trends Cogn. Sci.* **8**, 162–169. (doi:10.1016/j.tics.2004.02.002)
55. Peters MAK. 2022 Confidence in decision-making. *Oxford research encyclopedia of neuroscience*. Accessed 17 March 2023. (doi:10.1093/acrefore/9780190264086.013.371)
56. Adler WT, Ma WJ. 2018 Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS Comput. Biol.* **14,** e1006572. (doi:10.1371/journal.pcbi.1006572)
57. Li HH, Ma WJ. 2020 Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nat. Commun.* **11**, 2004. (doi:10.1038/s41467-020-15581-6)
58. Kimmet F, Pedersen S, Cardenas V, Rubiera C, Johnson G, Sans A, Baldwin M, Odegaard B. 2023 Metacognition and causal inference in audiovisual speech. *Multisensory Res.* **36**, 289–311. (doi:10.1163/22134808-bja10094)
59. Meijer D, Noppeney U. 2023 Metacognition in the audiovisual McGurk illusion: perceptual and causal confidence. Figshare. (doi:10.6084/m9.figshare.c.6751342)