Unfolding of multisensory inference in the brain and behavior

Yinan Cao (yinan.cao@psy.ox.ac.uk)
University of Oxford, Walton Street, Oxford OX2 6AE, United Kingdom
Hame Park
Bielefeld University, 33615 Bielefeld, Germany
Bruno L. Giordano*
Centre National de la Recherche Scientifique and Aix-Marseille Université, Marseille, France
Christoph Kayser*
Bielefeld University, 33615 Bielefeld, Germany
Charles Spence*
University of Oxford, Walton Street, Oxford OX2 6GG, United Kingdom
Christopher Summerfield*
University of Oxford, Walton Street, Oxford OX2 6AE, United Kingdom
[* Equal contributions]

Abstract:

Human multisensory inference has recently been characterized as involving fusion, segregation, or a flexible arbitration between fusion and segregation by virtue of sensory causal inference (CI; see Rohe & Noppeney, 2015). Theoretical work suggests that this inference could be a monolithic process implemented in reciprocally-coupled neuronal assemblies (Zhang et al., 2019). An alternative view, however, is that the computations are structured in time, so that different processes dominate at different post-stimulus latencies. There is emerging neural evidence for this view (Aller & Noppeney, 2018; Cao et al., 2019). Furthermore, behavioral studies also suggested that fusion may be a rather automatic process, e.g., crossmodal biases tend to be stronger when participants respond faster or after acquiring only little sensory evidence (Noppeney et al., 2010). By contrast, CI requires additional processing time as it capitalizes on evaluating the degree of sensory discrepancy, maintaining beliefs over latent causes, and possibly exploring distinct decision strategies. Here, across three studies combining psychophysics, computational modelling, and representational similarity analysis (RSA) to source-resolved human magnetoencephalographic data, we show that multisensory inference unfolds in time, by rapidly deriving a fused sensory estimate for computational expediency and, later and if required, filtering out irrelevant signals based on the inferred sensory cause(s).

Keywords: multisensory inference; causal structure; magnetoencephalography (MEG).

Methods

An overview of the tasks and design factors for the 3 studies is illustrated in Fig. 1. In Study 1, 15 human participants made 4-alternative forced choice speeded judgements in an audiovisual rate categorization task (Fig. 1). The stimuli consisted of a temporal sequence of audiovisual pulses (flutter and flicker; duration of the entire sequence was 550 ms) presented at four possible repetition rates (9.1, 12.7, 16.4 or 20 Hz; i.e., number of pulses/s). In separate blocks, the participants were instructed to report either the auditory or the visual rate as "task-relevant" information, signaling their response with a button-press. To quantify how the discrepancy of crossmodal information influences behavior (Körding et al., 2007), we manipulated visual and auditory rates independently (i.e., they could either be congruent or incongruent across trials; Fig. 1 colormap). To quantify the reliability-dependent influence of one modality onto another, we varied the signal-to-noise ratio of the acoustic information. The paradigm thus comprised a factorial 4 (visual rates) by 4 (auditory rates) by 2 (auditory reliabilities) by 2 (task relevance) design. During this task, participants' brain activity was recorded using magnetoencephalography (MEG) that is equipped to measure how neural signals unfold during a single decision. There were 22 trials per condition.

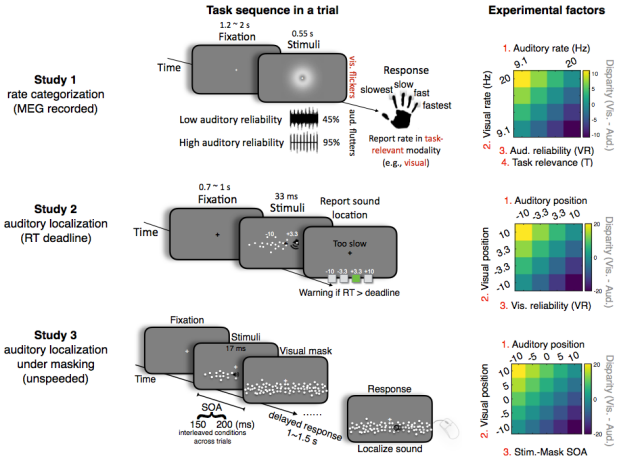

Fig. 1: Task structures and experimental manipulations (colormaps). See Methods for details.

Study 2 and 3 replaced the rate stimuli with spatial stimuli in order to test the generalizability of our hypothesis while omitting the modality-specific task switch (i.e., auditory localization task only). Study 2 (N = 22) required participants to make speeded judgements under time pressure (controlled by an adaptive reaction-time (RT) deadline), while Study 3 (N = 12) leveraged a masking paradigm in order to test the extent to which interrupting the short-term memory trace of

unisensory information at different points in time would hinder CI differently. The auditory spatial signal (33 ms; white-noise burst convolved with head-related impulse response) was onset-synchronized with a visual signal (33 ms; 25 white dots on a grey background). The visual dots were drawn from a 2D Gaussian distribution centered on one of the four locations and with variable horizontal standard deviation controlling for the visual reliability about the estimation of the cloud origin (see Fig. 1; each participant completed ≥ 1792 trials in total, i.e., ~ 60 trials per condition). In the masking study, participants used a mouse cursor to localize the transient auditory signal (17 ms) in the presence of a synchronized visual signal subsequently masked by a large canvas of dots. The stimulus-mask SOA varied trial-by-trial randomly across 2 conditions: 150 ms vs. 200 ms. Unisensory localization (auditory and visual) trials were interleaved within the multisensory trials in Study 3.

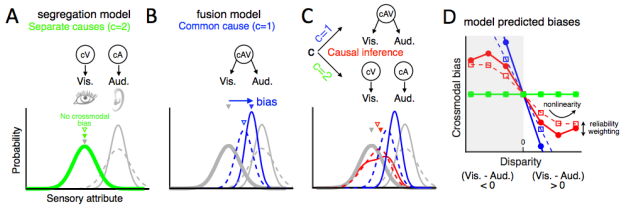

Fig. 2: Schematic of different sensory causal structures giving rise to visual and acoustic stimuli. Top of (A) to (C): inferred causality. Bottom: probability distribution of the perceived stimulus feature, and of the sensory estimate derived under different assumptions about the causal structure. (D) Each candidate model predicts a unique relationship between crossmodal disparity (distinct visual vs. auditory rates are characterized by a large disparity) and bias (deviation of the final estimate from the true attribute).

Results

General modelling framework: We compared the predictions of three classes of models concerning participants' behavior. Each model encodes probability distributions over sensory signals and incorporates rules that govern how a prior belief about the sensory causal structure is combined with incoming information to judge the event rate in the task-relevant modality (Fig. 2). These models make distinct predictions about how the perceived event rate varies with experimental manipulations (crossmodal disparity, i.e., the difference between auditory and visual rates; cue reliability; and task relevance). One key behavioral variable is the level of crossmodal bias, i.e., the extent to which judgements about the relevant modality are biased by the irrelevant modality, and how this bias varies with disparity. The segregation model proposes that sensory estimates are fully independent and predicts no crossmodal bias. The fusion model instead predicts a bias that grows linearly with disparity, because relevant and irrelevant sensory signals are fused irrespective of their congruency. Finally, the inference model allows for an additional inference about sensory causality, i.e., that observers allow for some signals to be fused and some to be segregated, and that fusion is more likely for signals having a similar rate (Körding et al., 2007). This inference model predicts that the bias increases with disparity and relative cue reliabilities, but critically, it predicts that the growth rate of bias should diminish for highly discrepant information that is unlikely to originate from a common source, i.e., reflecting a nonlinear dependency of bias on disparity, in contrast to the fusion model predicting a linear dependency.

Study 1---Temporal hierarchy of multisensory inference

A CI model formulated with a free probabilistic belief of common cause (p_c) explained the data better than did models that did not incorporate the inference of latent cause(s) (i.e., segregation and reliability-weighted fusion; group-level Bayesian (BIC) and corrected Akaike Information Criterion (AICc) relative to CI ≥ 468 and 547, respectively). We further examined why CI outperforms the other models in describing the behavioral responses, using an alternative analysis. Specifically, we quantified crossmodal bias, defined as the deviation of participants' response from the actual task-relevant rate (Fig. 3A), and used a general linear model (GLM) to predict how the magnitude (i.e., absolute value) of this bias depended on the contextual factors: task, reliability, their interaction, as well as disparity (Fig. 3B; all effects were assessed using maximum-statistics permutation controlling for multiple comparisons, family-wise error FWE = 0.05). Importantly, we included an effect of squared disparity in this model to capture whether the bias scales nonlinearly with disparity, as predicted by CI, or simply follows a linear dependency, as predicted by sensory fusion. A reliability-weighted cue combination is captured by the interaction between task and reliability rather than by the main effect of reliability. This is because reliability was manipulated only for the acoustic signal, which, under reliability-based cue weighting, would result in different biases for the two tasks. Indeed, this GLM revealed no main effects of task ($t(14) = 1.84$, mean $\beta = 0.097$, SEM = 0.053) and auditory reliability ($t(14) = 1.91$, mean $\beta = 0.043$, SEM = 0.022), but a significant interaction between task relevance and auditory reliability ($t(14) = -6.36$, mean $\beta = -0.16$, SEM = 0.025). Lastly, the GLM revealed a significantly negative effect of squared disparity ($t(14) = -9.28$, mean $\beta = -0.21$, SEM = 0.022), confirming a reduction of the bias growth rate for larger disparities (i.e., nonlinear scaling) as suggested by CI.

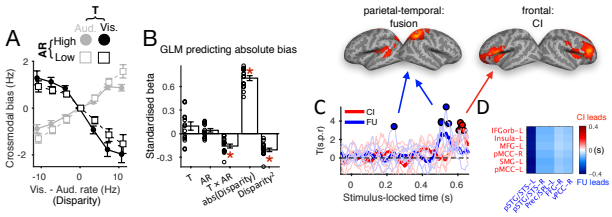

Fig. 3: (A) Crossmodal bias, reflecting the disparity-dependent influence of the task-irrelevant cue. (B) GLM quantifying the influence of task (T; visual task minus auditory task), auditory reliability (AR; low minus high), and the linear and quadratic effects of the absolute disparity on the absolute bias. (C-D) RSA model

encoding timecourses for fusion and CI, and the relevant brain regions encoding each model (see also Cao et al., 2019).

At the neural level, using spatiotemporal RSA to source-localized MEG, we reveal that the distinct computations required for flexible multisensory perception (segregation, fusion and CI) coexist, but each dominates at different points in time and in distinct regions. The initially segregated unisensory signals are fused in temporal and parietal lobes (Fig. 3C). However, this fused information gives way to more flexible representations formed under multisensory CI in the frontal cortex (Fig. 3D).

Study 2---Stronger fusion in faster decisions

Behavior: A general linear model (GLM) was used to predict how the magnitude (i.e., absolute value) of the bias depended on RT level (4 bins), visual reliability, their interaction, as well as crossmodal disparity (Fig. 4B; all effects were assessed using maximum-statistics permutation controlling for multiple comparisons, family-wise error FWE = 0.05). Similar to Cao et al. (2019), we included an effect of squared disparity in this model to capture whether the bias scales nonlinearly with disparity, as predicted by Bayesian models of CI ($t(21) = -3.99$, mean $\beta = -0.083$, SEM = 0.021). The bias was dependent on visual reliability, indicating reliability-weighted influence of the task-irrelevant visual cue ($t(21) = -9.59$, mean $\beta = -0.37$, SEM = 0.039). Importantly, this GLM revealed a significant effect of RT ($t(21) = -6.72$, mean $\beta = -0.194$, SEM = 0.029), and a significant interaction between RT and linear dependency of the bias on disparity ($t(21) = -5.03$, mean $\beta = -0.13$, SEM = 0.026). The visual bias significantly increases with faster RTs, and the rate of bias growth with respect to disparity ramps towards fusion for faster RTs (Fig. 4B).

Computational modeling: We first fit the CI model to each participant's data by varying freely only 1 model parameter across the 4 RT bins. Model comparison shows that both a decreasing probabilistic belief of common cause p_c and decreasing auditory noise over time could similarly capture the time-varying decreasing crossmodal bias (see Fig. 4C). It might seem tricky to interpret why p_c decreases over time within a trial since such a parameter has been commonly assumed as a long-term statistical 'prior'. Here, we show that a novel probabilistic model updating across time the posterior belief of common cause provides naturally a parsimonious account for the observed decreasing p_c (Fig. 4D). This model simulates 5,000 trajectories of the common-cause posterior for each experimental condition (equation below: where $\vec{X}_{vis,1\ldots t}$ and $\vec{X}_{aud,1\ldots t}$ encapsulates the overall visual and auditory evidence accumulated until time t, respectively; c is the hidden cause underlying audio-visual signals: c=1 indicates common cause).

$$p(c=1|\vec{X}_{vis,1\ldots t},\vec{X}_{aud,1\ldots t}) = \frac{p(x_{vis,t},x_{aud,t}|c=1)p(c=1|\vec{X}_{vis,1\ldots t-1},\vec{X}_{aud,1\ldots t-1})}{\sum_c p(x_{vis,t},x_{aud,t}|c)p(c|\vec{X}_{vis,1\ldots t-1},\vec{X}_{aud,1\ldots t-1})}$$

A free parameter q sets a threshold which the noisy posterior trajectories first cross, thus resulting in simulated RTs and multinomial choice probabilities. The agent tends to fuse audio-visual signals when hitting the q threshold, while it tends to segregate signals when hitting the (1-q) threshold (example trajectories shown in Fig. 4D left). Each participant's data were binned every 100 ms from 0 to 1500 ms post-stimulus onset, and a non-decision time parameter captured sensorimotor processing time. A key parameter in this model is the initial p_c (a common-cause belief held before accumulating audio-visual evidence encoded in working memory traces). As shown in simulations (Fig. 4D middle), for values of the initial $p_c > 0.25$ (congruency rate in current task), fitting the CI model independently to the data at each time point would result in a decreasing trajectory of p_c, thus explaining qualitatively the observed data in Study 2. This dynamic model also predicts that for a decreasing trajectory of p_c, the initial common-cause belief should be higher than the task statistics of 0.25. By fitting this model to each participant's RT and choice data, we indeed observed that the mean for the estimated initial $p_c = 0.75$ (SEM = 0.031 across participants; Fig. 4D right).

Study 3---Stronger fusion at earlier mask latency

As mentioned above, an alternative explanation for Study 2's results is that the slower decisions were solely characterized by a more precise sensory representation for the auditory signal (sensory refinement over time). This possibility was tested in Study 3 (masking paradigm). Of note, the mechanism of model progression (fusion progresses to CI) and this 'sensory refinement' account would make distinct predictions in the masking task. Specifically, only the former predicts an increase of audiovisual fusion for shorter mask latency. This is because an inference about the sensory causal structure of the world requires the brain to maintain at least partial access to the representations of individual cues beyond the fused representation. By analyzing the probability of fusion [i.e., percent bias: (response – auditory)/(visual – auditory), a significantly stronger fusion was found at the shorter mask SOA (Fig. 5A). This effect mainly originated from the conditions with smaller audiovisual disparities, which fits with the observations in a previous work (Körding et al., 2007; Fig. 5B). The inference about independent sensory causes at the relatively longer SOA (200 ms) likely promoted the repulsive (negative) bias in the opposite direction to the task-irrelevant visual signal. Importantly, the unisensory representations (quantified using localization precision in the interleaved unisensory trials) did not change significantly across SOA, thus unlikely played a role in inducing the time-varying effect of fusion (Fig. 5C).

In sum, the brain appears to arbitrate between the expediency of sensory fusion (for better precision) and the imperative to perform causal structure inference (for lower bias), by orchestrating the two processes in a temporal sequence and along a cerebral hierarchy to guide flexible behavior.

Acknowledgments

This project was funded by the European Research Council (to CK ERC-2014-CoG; grant No 646657), UK's Biotechnology and Biological Sciences Research Council (grant BB/M009742/1 to BLG), and a Human Brain Project award to C. Summerfield. YC was funded by the Clarendon Fellowship of Oxford University.

References

Aller, M., & Noppeney, U. (2019). To integrate or not to integrate: Temporal dynamics of hierarchical Bayesian causal inference. *PLoS Biol., 17*(4), e3000210.

Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal inference in the multisensory brain. *Neuron*.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS One*, *2*(9), e943.

Noppeney, U., Ostwald, D., & Werner, S. (2010). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *J. Neurosci.*, *30*, 7434–7446.

Rohe, T., & Noppeney, U. (2015). Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS Biol., 13*, e1002073.

Zhang, W. H., Wang, H., Chen, A., Gu, Y., Lee, T. S., Wong, K. M., & Wu, S. (2019). Complementary congruent and opposite neurons achieve concurrent multisensory integration and segregation. *eLife, 8*, e43753.

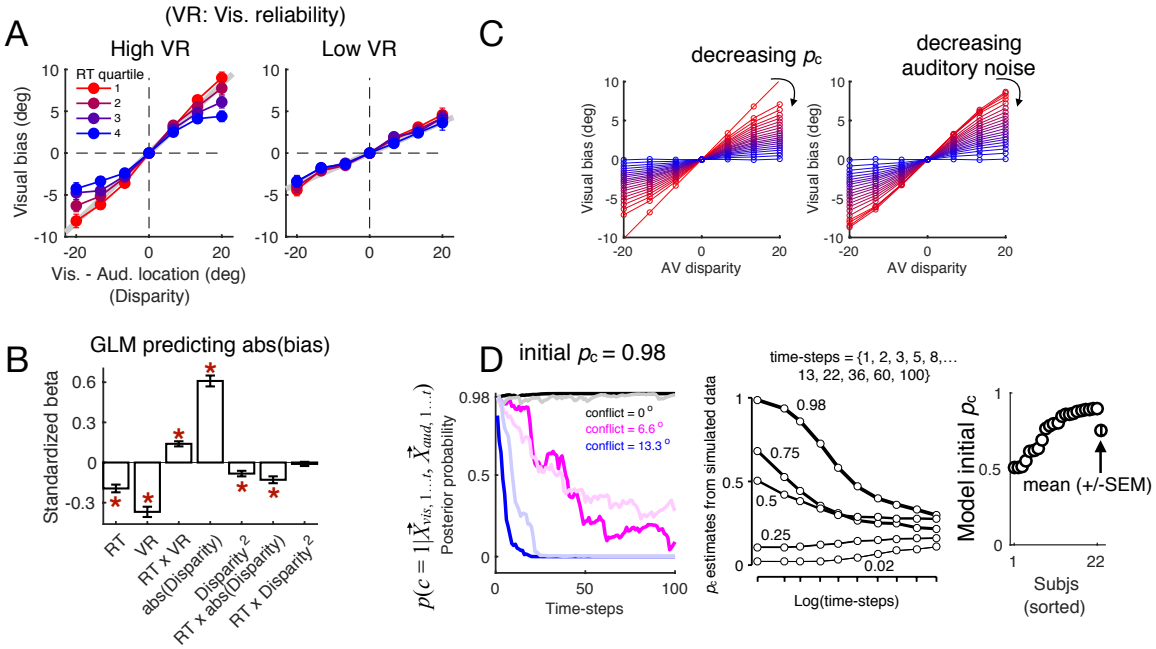

Fig. 4: (A) Crossmodal bias varies with reaction time (RT), disparity, and reliability. (B) GLM predicting bias magnitude. (C) Both a decreasing probabilistic belief of common cause p_c and decreasing auditory noise over time could similarly capture the time-varying decrease in crossmodal bias. (D) Posterior updating in the dynamic probabilistic model. Estimated initial p_c across participants (N = 22; sorted).

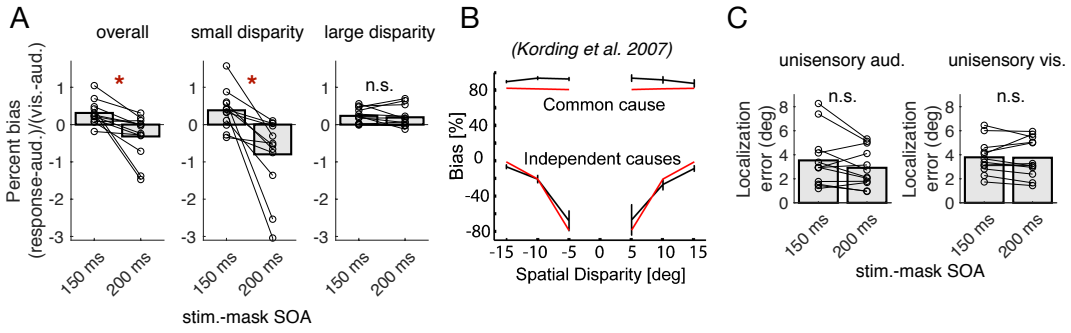

Fig. 5: (A) Percent bias as a function of stimulus-mask onset SOA. (B) Repulsion effect (figure adapted from Körding et al. 2007 Figure 3b). (C) Non-significant effect of stimulus-mask SOA on unisensory precision measured in interleaved unisensory trials.