Multisensory decision-making: Spatiotemporal dynamics and computational principles



Yinan Cao

Brasenose College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Experimental Psychology

Trinity 2019

Multisensory decision-making: Spatiotemporal dynamics and computational principles

Yinan Cao

Brasenose College Trinity Term, 2019

A thesis submitted for the degree of Doctor of Philosophy in Experimental Psychology at the University of Oxford.

Abstract

Humans have evolved to fuse information across senses by weighing the reliability of each modality (e.g., a viewer relies more on the picture to follow a drama if the TV audio is faulty; Ernst & Banks, 2002) as well as to infer unseen causes of sensory events by inverting internal generative models of the world (causal inference or CI: "Was the film dubbed or not?"). These multisensory strategies resemble Bayesian inference, but the mechanism whereby the inference unfolds step-by-step in a biologically plausible fashion remains poorly understood. Moreover, the reliability-weighted fusion and CI strategies have complementary costs and benefits, but sometimes seem to describe human choices in experimental conditions that presumably only satisfy the assumption of each particular strategy (e.g., De Winkel, Katliar, & Bülthoff, 2015; Meijer, Veselič, Calafiore, & Noppeney, 2019).

In this thesis, I present converging evidence, from a series of human neuroimaging and psychophysics studies, suggesting that these rival accounts of multisensory inference can be reconciled by a time-resolved mechanism: The brain rapidly derives a fused sensory estimate for computational expediency and, later and if required, filtering out irrelevant signals based on the inferred sensory cause(s).

First, analysing time- and source-resolved human magnetoencephalographic data (**chapter 2**), I unveil a systematic spatiotemporal cascade of the relevant computations for multisensory inference, starting with early segregated unisensory representations, continuing with sensory fusion in parietal-temporal regions, and culminating as CI in the frontal lobe. This suggests that the distinct computations required for flexible multisensory behaviour (fusion and CI) coexist, but each dominates at different times and in distinct regions (**chapter 3**).

Furthermore, using behavioural paradigms such as speeded-judgement under deadline-induced time pressure as well as psychophysics leveraging sensory masking (**chapter 4**), I explicitly quantified how the unfolding of neural representations can be read out to drive behaviour. Collectively, slower versus faster behavioural choices reflect more of CI versus sensory fusion, respectively, indicating that the mapping from neural representational activity to behavioural choice corroborates the temporal hierarchy revealed in the neural data.

Table of Contents

Acknowledgements ..... *i*

Contribution ..... *ii*

Chapter 1. General introduction *1*

1.1 Describing and defining multisensory inference **1**
1.1.1 The multisensory brains in a multimodal world—Multisensory integration and structure inference 1
1.1.2 Some definitions and distinctions 5
1.1.3 Psychophysical theories of multisensory inference: A brief overview 7

1.2 Computational models of multisensory inference **10**
1.2.1 Probabilistic ideal-observer modelling for sensory fusion 11
1.2.2 Normative modelling for causal structure inference 13
1.2.3 Mixed findings for the best-fitting computational account 19
1.2.4 Bayesian computations using probabilistic population coding 22

1.3 The functional neuroanatomy of multisensory inference **26**
1.3.1 Reliability-weighted multisensory fusion in the brain 26
1.3.2 Multisensory causal structure inference in macroscopic brain signals 31
1.3.3 Reasoning, strategic arbitration, and multisensory conflict resolution in the brain 32
1.3.4 Potentials and limitations of MEG for characterising the computations of multisensory integration 35

1.4 Paradigms for unveiling the timecourse of decision formation **40**
1.4.1 Short overview 40
1.4.2 Variable (fixed)-duration task 42
1.4.3 Sensory masking disrupts short-term memory trace 43
1.4.4 Reaction-time (RT) paradigms 44
1.4.5 Weighting profile assessed with reverse correlation 46

Chapter 2. Spatiotemporal representational similarity analysis and INDSCAL of source MEG activity *48*

2.1 Representational similarity analysis (RSA) to MEG source-localised signals **48**
2.1.1 Diffeomorphic anatomical registration using exponentiated lie algebra (DARTEL) 48
2.1.2 Cross-validation for pre-whitened Euclidean distance 52

2.2 Multidimensional scaling with individual weighting (INDSCAL) **54**

Chapter 3. Spatiotemporal unfolding of neural representations underlying flexible multisensory inference *60*

Chapter abstract **60**

3.1 Chapter introduction **61**

3.2 Methods **63**
3.2.1 Participant details 63
3.2.2 Task design and stimuli 65
3.2.3 Experimental procedure and stimulus presentation 66
3.2.4 Analysis of behavioural data – Model fitting specifications 67
3.2.5 Analysis of behavioural data – Model comparison 69
3.2.6 Analysis of behavioural data – Sensory noise function 69
3.2.7 Neuroimaging data acquisition 71
3.2.8 Preprocessing of MEG data 71
3.2.9 Reconstruction of MEG sources 72
3.2.10 Representational similarity analysis of spatiotemporal MEG activity 73

3.3 Results **77**
3.3.1 Task summary 77
3.3.2 Modelling behaviour 78
3.3.3 Multisensory judgments follow Bayesian causal inference 79
3.3.4 Context-dependent cue weighting 81
3.3.5 MEG data reveal a spatiotemporal hierarchy of multisensory representations 88
3.3.6 Distinct markers of sensory integration in posterior and anterior brain regions 93
3.3.7 Complementary evidence for sensory fusion in parietal-temporal regions and causal inference in frontal regions 98
3.3.8 Prefrontal cortex drives flexible behaviour in conflicting environments 101
3.3.9 Additional control analyses ruling out task-based modulation as the sole determinant for neural effect 103

3.4 Chapter discussion **111**
3.4.1 Temporal hierarchy of multisensory computations 112
3.4.2 Graded context-dependent computations along parietal-frontal pathway 113
3.4.3 The frontal lobe in multisensory and domain-general inference 116

Chapter appendix I **118**

Chapter appendix II **129**

Chapter 4. Behavioural readout of temporal unfolding in multisensory inference ***140***

Chapter abstract **140**

4.1 Chapter introduction **140**

4.2 Behavioural Experiment I **142**
4.2.1 Methods 142
4.2.2 Results 146
4.2.2 Faster choices reflect higher tendency to fuse multisensory signals 150

4.3 Behavioural Experiment II **154**
4.3.1 Methods 155
4.3.2 Analyses and Results 158

4.4 Chapter discussion **160**

Chapter appendix **165**

Chapter 5. Conclusions and general discussion ***171***

5.1 Temporal sequence for multisensory inference **172**
5.1.1 Temporal unfolding of multisensory causal structure inference 172
5.1.2 The role of neural selective representation of sensory fusion 174

5.2 Parietal-frontal network for multisensory inference **177**
5.2.1 Posterior parietal cortex represents information in low-dimensional space 177
5.2.2 Choice predictability of neural activity in frontal vs. parietal cortex 178
5.2.3 Where is CI represented in the brain? 179

References ***182***

Acknowledgements

First and foremost, I must express my greatest gratitude to Chris Summerfield, Christoph Kayser and Charles Spence, the tremendous "3C" supervisory team for taking me under their wings, hosting me in their laboratories, and teaching me how to do science. Chris, your wisdom and your enthusiasm for science has been inspiring and influential, and I am grateful for the chance of being able to continue working with you in the near future as a postdoc. Christoph, thank you for all the extraordinary support, scientifically, mentally, and financially, that you have provided at every single step of this journey. Charles, thank you for showing me what it means by hardworking, and thank you for those multisensory dinners, and those candies that mysteriously showed up and disappeared day to day on the desk in the Tinbergen office.

Many thanks to past and present members of the Summerfield lab, Keno Juchems, Jan Balaguer, Andreas Jarvstad, Vickie Li, Fabrice Luyckx, Jacqueline Ding, Ron Dekker, Yasmin Meyer, Timo Flesch, Mira Dumbalska, Hannah Sheahan, Hannah Tickle, Andrew Saxe, and many others. Your kindness, support and good company have been a life-time treasure. Thanks to Dan Bang and Paul Muhle-Karbe for your career advice and kind guidance on job searching.

Collecting MEG data in Glasgow has been a tough but enjoyable experience. Prof. Joachim Gross was immensely supportive of my use of the computing resource. His generosity played an influential role in facilitating the computation-heavy data analyses requested by my neuroimaging project. Frances, I thank you for your excellent technical support, and I hope you liked the tea and wine that Bruno and I sent you. Drs. Robin Ince and Oliver Garrod shared many useful tips and guidance on analysis and programming, thank you! Liyu Cao, Cheng-yang Han, Ling-ling Hua, and Christoph Daube, thank you for being good friends when I was working there.

I thank my parents for everything you have done to support me. Doing a DPhil while family living 5,000 miles away on the other side of the world has been a tough few years but we're getting there slowly. In the coming years, I hope I can spend more time with you.

I thank Prof. Kia Nobre for being a great leader during the disastrous Tinbergen closure, Prof. Nick Yeung for acting as a remarkable Director of Graduate Studies, Dr. David Popplewell for being the best College Advisor that I can ever imagine, and Rebecca Cardus for your patience when helping with administration- and visa-related questions. I also thank the Clarendon Fund of Oxford University for financial support.

Bruno, you deserve a whole dedicated paragraph of my gratitude. This thesis would not have been possible without the support of you, as an insightful mentor, a tremendous supervisor, and a friend. Doing science with you during the past 7 years has been a luxury experience. I couldn't think of any other person on Earth who can dedicate so much time and effort discussing data with me at 3 AM UK time (of note, there is one-hour difference between Oxford and Marseille).

Contribution

This thesis presents the results of original experiments reported in **chapters 3** and **4**. I led the conceptualisation, design, analysis, and writing of all these.

In **chapter 3**, I present an experiment that was designed together with my thesis supervisors, Prof. Christopher Summerfield, Prof. Christoph Kayser, and Dr. Bruno L. Giordano. They also provided substantial feedback in the conceptualisation, design, analysis, and writing. I obtained additional help from Dr. Hame Park during data collection (who is also the co-author of the publication based on this experiment). This work has been published in the journal *Neuron*.

In **chapter 4**, I present two experiments that were designed together with Prof. Christopher Summerfield, Prof. Charles Spence, and Prof. Christoph Kayser. They also provided substantial feedback in the conceptualisation, design, analysis, and writing.

Chapter 1. General introduction

This thesis addresses neuro-computational mechanisms used by the human brain to support and organise inferential strategies for multisensory-guided action selection. I will begin the general introduction of the thesis by first describing two key aspects of flexible behaviour of biological systems in natural environments (multisensory integration and sensory structure inference; **section 1.1.1**), and then distinguishing mixed terminologies in the field of multisensory research (**section 1.1.2**).

Much has been established pertaining to the functional neuroanatomy of sensory integration (**section 1.3**) and Bayesian descriptions of multisensory behaviour (**section 1.2**). In this chapter, important psychophysical evidence implicating canonical principles of multisensory inference in the history of multisensory research (**section 1.1.3**), prominent computational models capitalising on Bayesian probability theory (**section 1.2.1**), and prominent experimental approaches to investigate temporal unfolding of perceptual decision (**section 1.4**) are reviewed, and key open questions concerning mechanisms of flexible multisensory inference in the human brain that are addressed in this thesis are highlighted (**section 1.2.3** and **section 1.3.2**).

1.1 Describing and defining multisensory inference

1.1.1 The multisensory brains in a multimodal world—Multisensory integration and structure inference

During sensory-guided action selection, biologically salient events are typically registered by more than one sensory modality, from communication signals

(Partan & Marler, 1999) to courtship displays (Ota, Gahr, & Soma, 2018). Bowerbirds are famous for their extraordinarily mesmerising multimodal mating behaviours (males wave their wings and produce vocalisations from deep in the throat); Peacock spiders (*Maratus volans*) have evolved to improvise arousing clapping sounds (using their body and legs) whilst coordinating elegant dance moves to court their partners; Movements of the lips, tongue or the eyes of humans convey complementary information that boosts speech intelligibility, and that is why we profit tremendously from visual cues arising from a speaker's face when communicating in challenging acoustic environments (e.g., in a noisy pub; Giordano et al., 2017); Pedestrians crossing a road may become more sensitive than usual to car horn sounds on a foggy day; Also, imagine trying to follow a drama on a broken television—if the TV audio is faulty, a viewer should rely more on the picture to follow the narrative, and vice versa. The mammalian brains are able to transform, seemingly effortlessly, the plethora of sensory inputs into an approximately veridical percept of the multisensory world. If each modality derives and transmits an independent sensory estimate, more accurate perceptual evaluations of the external world can be achieved through convergence, interaction and synthesis (integration) of those at least partially redundant or complementary multisensory estimates. Such integrative multisensory processes determine how humans and other animals perceive the variable and often ambiguous environments, and thus offer substantial survival value by allowing flexible decision-making and action selection in a constantly challenging world. **Multisensory integration** thus serves as a key aspect of coordinating flexible actions in response to complex sensory inputs, which per se underlies why biological brains efficiently function and operate the way they do.

One influential theory suggests that the brains of humans and other animals have evolved to implement a reliability-weighting fusion principle when judging sensory signals across senses (Alais & Burr, 2004; Angelaki, Gu, & DeAngelis, 2009; Ernst & Bülthoff, 2004; Raposo, Sheppard, Schrater, & Churchland, 2012). Probabilistic inference models have been able to describe such reliability-weighting multisensory behaviours at computational levels (**section 1.2.1**) and the neural substrates that are recruited for these computations have been revealed by human neuroimaging and non-human primate neurophysiology studies (**section 1.3.1**).

At the same time, flexible behaviour requires efficient representation of the richly structured knowledge of the world (Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Griffiths & Tenenbaum, 2005; Saxe, Tenenbaum, & Carey, 2005). The need for structural knowledge and the need to get it right echo the need for fast generalisation to un-experienced exemplars in future novel contexts (Behrens et al., 2018; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). A typical structural knowledge the brain needs to efficiently infer and represent is the **causal structures** underlying sensory events (Gershman, Norman, & Niv, 2015; Gershman, 2017; Körding et al., 2007).

Of particular interest are the ubiquitous examples of the computational challenges that **inference about causal structures** poses to the brains of biological systems in everyday multisensory behaviours. “Is the rattling noise coming from that crawling insect or from a rattlesnake that I have not yet seen?” A foraging chipmunk can approach her food more precisely by combining a concurrent sound with the sight of the moving insect, but only if the two inputs are indeed caused by the same source (e.g., her prey, not a dangerous predator). When hearing a “meowing” sound and

seeing a kitten walking by, we naturally wonder about whether it was one cat meowing whilst walking at the same time, or it was actually two different cats, one passed by the front of the house (i.e., the one we saw) whereas the other cat meowed in a nearby bush (the meow of which we heard)? In human daily activities, a crowd of faces along with competing conversations at a cocktail party make it a real challenge to identify who called our name.

At every instant of time, inferring about the causal structures of the sensory environments challenges the nervous systems since the systems have no direct access to the causal structures (typically unknown or *hidden*) underlying the observable signals themselves, but only to the "estimates" of these hidden structures that are derived from those "noisy" sensory inputs. Bayesian computations provide a normative framework for understanding the computational strategies used by observers to infer sensory causal structures via "inverting" a generative mental model of the world (**section 1.3.2**; Körding et al., 2007; Rohe & Noppeney, 2015a, 2015b; Odegaard, Wozny, & Shams, 2015; Odegaard, & Shams, 2016; Odegaard, Wozny, & Shams, 2016; Wozny, Beierholm, & Shams, 2011).

However, apart from the lack of a clear understanding of the neural underpinnings of multisensory causal inference (reviewed in **section 1.3.2**), the extent to which the Bayesian decision theories can inspire a test for the mechanistic role of different brain structures in implementing component processes of multisensory inference remains unclear. Admittedly, to gain an insight into what computations a biological system needs for goal-directed decision-making, it is useful to think of the decision as being decomposed into separable component processes that are essential for its appropriate commitment (Rangel, Camerer, & Montague, 2008). In one classic

study exemplifying this endeavour, a mechanistic account of the suboptimalities in human perceptual decision-making was postulated via identifying the origin of choice variability at each of the discrete stages of information processing (Drugowitsch, Wyart, Devauchelle, & Koechlin, 2016). One impediment toward understanding multisensory inference has been the lack of research mapping sensory representations predicted by Bayesian models onto the *full* spatiotemporal dynamics of neural activity.

These open questions motivate the projects of this thesis. Using neuroimaging method that is equipped to measure how neural signals unfold during a single multisensory choice (**chapters 2** and **3**), I will show that the neural representations underlying different component processes of flexible multisensory inference are organised in a temporal structure and across different brain regions. A systematic quantification of how these neural representations are read out for guiding behaviour will be discussed with psychophysical and modelling evidence drawn from two behavioural studies (**chapter 4**).

1.1.2 Some definitions and distinctions

When describing multisensory processes, distinct terms have been continuously preferred by different research groups in the multisensory community, such as "multimodal," "crossmodal," "heteromodal," "polymodal," "supramodal," "polysensory," "intersensory". As precisely stated by Stein (2012), "although unfortunate, this is probably not unusual for as diverse as an emerging field as ours…". Throughout this thesis, terminology consistency is maintained for the purpose of clarity and for promoting the proposal by Stein (2012). "Modality-specific," "multimodal" and "crossmodal" are used to describe stimulus inputs, cues,

tasks, settings or experimental conditions, whereas “unisensory” and “multisensory” are used to describe specifically the neuro-computational processes in response to the crossmodal events, tasks, settings or conditions, and at times more generally the resulting sensations, perceptions, and behaviours.

This thesis focuses on a spectrum of multisensory process that is not necessarily restricted to the sensory and perceptual stage(s) but potentially extends to the decisional levels, and/or occurs at the intersection between perceptual and decision stages. Of note, this thesis does not focus on “binding”, a special form of sensory integration describing multisensory object formation at the perceptual level (Atilgan et al., 2018; Bizley, Maddox, & Lee, 2016).

At the behavioural and modelling levels, multisensory integration and crossmodal cue combination are used as interchangeable concepts; and “fusion” is used at times to refer to a specific probabilistic model of reliability-weighted integration (see below: **section 1.2.1**). At the neural level, the term ‘integration’ is specifically used only when it can be distinguished from other types of multisensory processes at large. These processes are often pronounced as multisensory convergence (not within the main scope of this thesis) or interaction, since they may require that the individual senses maintain their own identities rather than being fused into a single entity. Examples for such processes include multisensory interactions in putatively unisensory cortices (Kayser & Logothetis, 2007; Lakatos, Chen, O’Connell, Mills, & Schroeder, 2007; Lewis & Noppeney, 2010; Schroeder & Foxe, 2005; Noesselt et al., 2007), via direct connectivity between sensory areas (Falchier, Clavagnier, Barone, & Kennedy, 2002), thalamo-cortical mechanisms (Lakatos et al., 2007), or top-down influences from higher-order regions (Macaluso & Driver, 2005). One example showing this distinction between integration and general multisensory

process can be found in the neural data of **chapter 3** ('bimodal' representation vs. fusion representation).

1.1.3 Psychophysical theories of multisensory inference: A brief overview

This section is designated to summarise a historical path of psychophysical investigations towards the modern views of two prominent fingerprints of context-dependent flexible multisensory behaviour: sensory reliability weighting and "unity" belief inference.

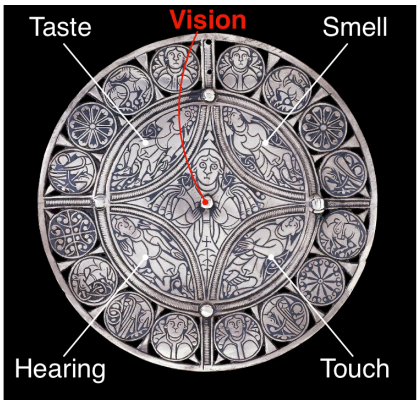


Figure 1.1. The Fuller Brooch (late Anglo-Saxon). Image source from the British Museum online database: https://www.britishmuseum.org/

Beyond reflecting the intellectual and technical sophistication of Anglo-Saxon taste and craftsmanship in the later ninth century, the Fuller Brooch, elegantly preserved in the British Museum, represents the earliest personifications of the five senses. Sight, perceived as the dominant sense in medieval times, commands the central position with staring eyes, and is flanked by figures representing hearing, touch, taste, and smell (see **Figure 1.1**). This centrality reflects a philosophical

tradition in which human vision is considered the dominant modality, a perspective that continued into modern-day psychology and neuroscience (Posner, Nissen, & Klein, 1976), and has been the subject of numerous empirical investigations (Bertelson & de Gelder, 2004; Jackson, 1953; Jousmäki & Hari, 1998; Pick, Warren, & Hay, 1969; Thomas, 1941; Warren & Cleaves, 1971; Witkin, Wapner, & Leventhal, 1952). Gibson (1933) showed that human volunteers perceive a straight rod as curved when they look at it through a prism that distorted its actual visual appearance. Rock and Victor (1964) demonstrated that when touch and sight provide discrepant information about shape and size, vision dominates the perceptual experience of the object (one critique has been that they tested this using only a single trial). A square object can be judged as smaller than its actual size when viewed through minifying lens; and it can be perceived as more rectangular when viewed through a narrowing lens (Rock & Harris, 1967). Numerous demonstrations of the rubber-hand illusion (Botvinick & Cohen, 1998; Durgin, Evans, Dunphy, Klostermann, & Simmons, 2007; Ehrsson, Spence, & Passingham, 2004) also confirm the visual capture of tactile processing as a robust mechanism in certain circumstances. Visual capture of audition has been widely found and formally tested using spatial ventriloquism paradigms in which participants' perceived location of a sound is usually attracted to the discrepant location of a concurrent visual stimulus (Bertelson, Vroomen, De Gelder, & Driver, 2000; Vroomen, Bertelson, & De Gelder, 2001). Visual dominance over audition is also an influential account for the McGurk and MacDonald (1976) illusion in which normal human participants report perceiving [da] for repeated utterances of the actual syllable [ba] dubbed on to lip movements articulating [ga]. Visual cues are also shown to dominate in non-human animal behaviours such as vocal communication in

monkeys (Partan & Marler, 1999), and food seeking in cows (Uetake & Kudo, 1994) and pigeons (Foree & LoLordo, 1973).

A later upsurge of interest in a slightly more flexible hypothesis of 'modality-appropriateness' (Welch & Warren, 1980; Welch, DutionHurt, & Warren, 1986) was promoted to advance the theory of sensory dominance. This hypothesis posits that the most 'appropriate' modality for the task at hand (e.g., vision in spatial task versus hearing in temporal task, Shams, Kamitani, & Shimojo, 2000; Morein-Zamir, Soto-Faraco, & Kingstone, 2003; Recanzone, 2003) would dominate another for certain kinds of perceptual judgements. One of the earliest empirical demonstrations of auditory dominance over vision was by Sekuler et al. (1997) showing that moving balls appear to bounce when a simultaneous impact sound is presented. Somatosensation and haptics can also capture other modalities e.g., hearing (Caclin, Soto-Faraco, Kingstone, & Spence, 2002) or vision (van Beers, Wolpert, & Haggard, 2002).

However, the extent to which one modality dominates over another also scales with the uncertainty of the sensory estimate of a signal (i.e., dynamic modality re-weighting; Alais & Burr, 2004; Battaglia, Jacobs, & Aslin, 2003; Ernst & Banks, 2002), not just the precision of a modality defined using a broad concept of ecological validity such as "appropriateness". Probabilistic inference models in virtue of Bayesian decision theories are equipped to solve this challenge. These Bayesian models allow context-dependent flexibility in the sensory weights assigned to different modality-specific signals, and are able to describe a key fingerprint of flexible multisensory integration—sensory weighting by signal reliability—in a host of multimodal and sensorimotor situations (Battaglia et al., 2003; Gepshtein & Banks, 2003).

At the same time, it has long been realised in various psychophysical contexts that the strength and robustness of sensory integration can be influenced by subjective belief over a 'unitary' origin giving rise to sensory events (i.e., unity assumption; Welch & Warren, 1980; Chen & Spence, 2017). Spatiotemporal correspondence such as proximity of the spatial and/or temporal structures across individual signals serves as a bottom-up information the nervous systems use to form the unity belief (Cao, Giordano, Avanzini, & McAdams, 2016; Parise & Ernst, 2016; Parise, Spence, & Ernst, 2012; Roach, Heron, & McGraw, 2006; Wallace et al., 2004). In the following sections, probabilistic inference theories developed to systematically describe the reliability-weighting phenomenon (in a maximum-likelihood-estimation fusion model) and the unity belief inference (in a Bayesian structure inference model) will be reviewed and discussed.

1.2 Computational models of multisensory inference

In this section, some key computational theories of multisensory inference are reviewed with mathematical details and simulations of a Bayesian agent. I will also discuss some mixed results in the literature disagreeing upon the best-fitting computational account of multisensory behaviour in experimental settings initially designed to test for one model against another, which motivates our search in this thesis for a mechanistic account functionally reconciling these different models.

1.2.1 Probabilistic ideal-observer modelling for sensory fusion

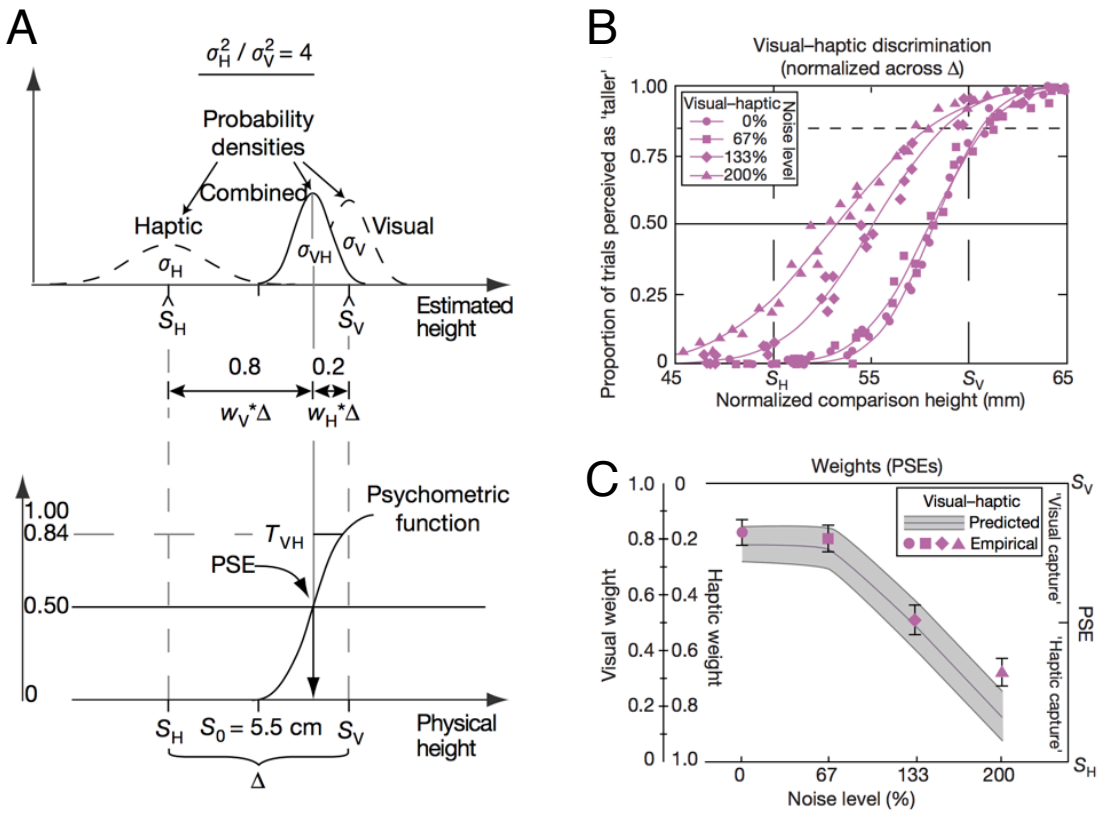


Figure 1.2. Schematic of the maximum-likelihood estimation (MLE) principle of reliability-weighted fusion. (**A**) Visually and haptically specified bar heights differ by a small amount (e.g., a few millimetres). Dashed Gaussians in the top panels represent probability density functions of the height estimate from vision and haptics, and solid gaussians represent probability densities for the combined estimate. The haptic variance is four times the visual variance. Thus, the combined probability density is shifted towards the visual estimate and its variance is 0.8 times the visual variance, as predicted by the fusion model. (**B**) Psychometric function for visual-haptic discrimination shifts leftward to the veridical haptic cue as visual noise increases. (**C**) Relative perceptual weights assigned to different sensory channels can be measured from psychometric functions in B and can be used to compare to those predicted by the fusion model (figures adapted from Ernst & Banks, 2002).

Generative probabilistic modelling for a cue-combination problem describes the way in which sensory inputs are generated by an underlying source, corrupted by noise processes, and how they are then observed and combined to make inference about a physical quantity. The mental representation of the noisy sensory inputs can be described as probability distributions whose inverse variance (1/variance)

represent the sensory 'reliabilities' of the different input signals. Formalised as a normative model of the world, the sensory fusion model describes how the sensory estimate of a physical quantity scales as a function of the relative reliabilities of different modalities. Ernst and Banks (2002) measured human participants' psychophysical judgement of the height of a simulated object, whilst requiring them to use sensory information from both touch (haptic inputs, "H" in the above schematic graph) and vision (visual inputs, "V"). The fusion model computes a posterior probability distribution (over the object's physical height), with the mean and variance of the posterior fully determined by a sum of the mean estimates of individual signals with each weighted by its relative reliability, and an inverse of the sum of the reliabilities of individual signals, respectively (**Figure 1.2C**). Computationally, this fusion strategy allows an agent to combine cues across different sensory modalities in proportion to their relative reliabilities, and it reflects a statistically-optimal use of crossmodal information by minimising uncertainty in the combined percept (Ernst & Bülthoff, 2004). Under the assumption of a common generative source underlying sensory events, the fusion model accounts for multisensory behaviours across a wide range of perceptual tasks and across many species (humans, rodents, non-human primates or even a type of fish without a cerebral cortex; Ernst & Banks, 2002; Fetsch, Turner, DeAngelis, & Angelaki, 2009; Raposo et al., 2012; Schumacher, de Perera, Thenert, & von der Emde, 2016).

1.2.2 Normative modelling for causal structure inference

The fusion theory (as described in **section 1.2.1** above), however, does not explain a typical class of flexible behaviour in situations where the causal structures of the sensory environment are uncertain and thus must be inferred *a posteriori*. This inferential process instead is addressed in causal inference models, whereby latent cause(s) underlying ambiguous sensory variables are inferred and an irrelevant information that likely arises from a separate cause gets down-weighted (Körding et al., 2007; Odegaard et al., 2015; Odegaard & Shams, 2016; Odegaard et al., 2016; Rohe & Noppeney, 2015a, 2015b; Wozny et al., 2010).

Bayesian causal models describe sensory representations in probabilistic terms, and incorporate rules that govern how a prior belief about the sensory causal structure is combined with incoming information to guide the combination of sensory cues.

Formulating multisensory inference using these frameworks (Körding et al., 2007; Wozny et al., 2010), I simulate a hierarchical Bayesian agent capitalising on distinct strategies of weighing and combining multisensory information (**Figure 1.3**). For this example, I consider a hypothetical task whereby human observers are asked to report their judgement of an auditory or visual feature (e.g., the spatial position of a sound burst or a visual flash) as task-relevant information. To quantify how the potential discrepancy between the audio and visual information influences participants' sensory judgement we combine, for each trial, the visual and auditory attributes independently (i.e., they could be either congruent or incongruent about their spatial location across different trials). I also manipulate the trial-by-trial reliability of the auditory signal by adding some background noise to it (i.e., the

higher the signal-to-noise ratio, the higher the auditory reliability relative to the visual signal).

The sensory estimate, *r,* is about a physical quantity in the external world. The respective unisensory estimates are described using Gaussian sensory likelihoods: $p(r_a|s_a)$ = N(s_a, σ_a^2) and $p(r_v|s_v)$ = N(s_v, σ_v^2), (**Eq. 1**), with *s* denoting the physical property of a stimulus, and $\sigma_a^2 (\sigma_v^2)$ the variance of the sensory likelihood distribution of unisensory auditory (visual) estimates, respectively. The sensory likelihood thus represents the probability of experiencing sensation *r* as a result of stimulus *s* occurring in the environment. On each trial, sensory estimate is draw from a Gaussian prior distribution over all possible values of *s* in the world, $p(s)$ = N(μ_p, σ_p^2), with mean μ_p and variance σ_p^2.

First, we assume that the reported sensory estimates are conditioned upon causal structure: either two independent causes (*c* = 2, **Figure 1.3A**, **Eq. 2**) or a single common cause (*c* = 1, **Figure 1.3B**, **Eq. 3**). Then the posterior probability of auditory or visual cue given a specific causal structure *c* and sensory observation *r*, is given by Bayes' rule:

(**Eq. 2**): $p(s_a|r_a, c = 2) = \frac{p(r_a|s_a)\, p(s)}{p(r_a)}$

$$p(s_v|r_v, c = 2) = \frac{p(r_v|s_v)\, p(s)}{p(r_v)}$$

(**Eq. 3**): $p(s|r_v, r_a, c = 1) = \frac{p(r_a|s)\, p(r_v|s)\, p(s)}{p(r_v, r_a)}$

Under the assumption of Gaussian distributions, the maximum-a-posteriori estimates are given by **Eq. 4** for segregation and by **Eq. 5** for fusion:

(**Eq. 4**): $\hat{S}_{a,c=2} = \frac{r_a/\sigma_a^2 + \mu_p/\sigma_p^2}{1/\sigma_a^2 + 1/\sigma_p^2}$, $\hat{S}_{v,c=2} = \frac{r_v/\sigma_v^2 + \mu_p/\sigma_p^2}{1/\sigma_v^2 + 1/\sigma_p^2}$

(**Eq. 5**): $\hat{S}_{a,c=1} = \hat{S}_{v,c=1} = \frac{r_a/\sigma_a^2 + r_v/\sigma_v^2 + \mu_p/\sigma_p^2}{1/\sigma_a^2 + 1/\sigma_v^2 + 1/\sigma_p^2}$

To perform causal structure inference, a Bayesian observer derives a probabilistic belief over multisensory causality (c = 1 for common cause, c = 2 for independent causes) to guide the estimation of the sensory feature. This belief (**Eq. 8** and **9**) is probabilistically determined by combining the sensory likelihoods with an integration tendency p_c, that is, an *a priori* belief in a common cause across the full experimental settings. Given the Gaussian assumption, the sensory likelihoods $p(r_v, r_a|c)$ of the visual and auditory estimates (r_v and r_a) under each causality assumption are:

(**Eq. 6**): $p(r_v, r_a|c=1) = \frac{1}{2\pi\sqrt{\sigma_v^2\sigma_a^2 + \sigma_v^2\sigma_p^2 + \sigma_a^2\sigma_p^2}} e^{-\frac{1}{2}\left[\frac{(r_v-r_a)^2\sigma_p^2 + (r_v-\mu_p)^2\sigma_a^2 + (r_a-\mu_p)^2\sigma_v^2}{\sigma_v^2\sigma_a^2 + \sigma_v^2\sigma_p^2 + \sigma_a^2\sigma_p^2}\right]}$

(**Eq. 7**): $p(r_v, r_a|c=2) = \frac{1}{2\pi\sqrt{(\sigma_v^2 + \sigma_p^2)(\sigma_a^2 + \sigma_p^2)}} e^{-\frac{1}{2}\left[\frac{(r_v-\mu_p)^2}{(\sigma_v^2 + \sigma_p^2)} + \frac{(r_a-\mu_p)^2}{(\sigma_a^2 + \sigma_p^2)}\right]}$

Given these sensory likelihoods and integration tendency, the posterior belief about the causal structure can be inferred using Bayes' rule:

(**Eq. 8**): $p(c=1|r_v, r_a) = \frac{p(r_v,r_a|c=1)p_c}{p(r_v,r_a|c=1)p_c + p(r_v,r_a|c=2)(1-p_c)}$

(**Eq. 9**): $p(c=2|r_v, r_a) = 1 - p(c=1|r_v, r_a)$

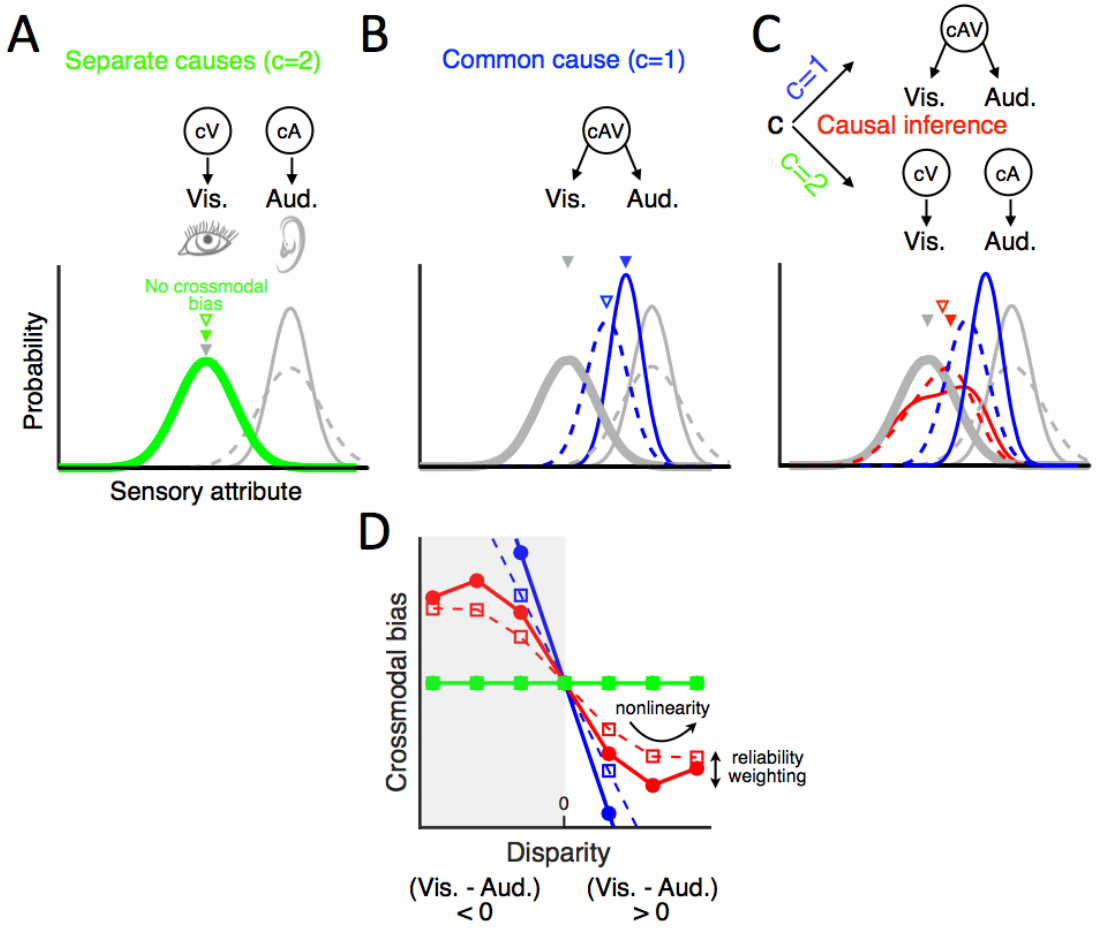


Figure 1.3. Computational models describing causal structure inference in a Bayesian agent. Schematic of different sensory causal structures giving rise to visual and acoustic stimuli. Top of **A** to **C**: inferred causality. Bottom: probability distribution of the perceived stimulus feature (e.g., stimulus position in space), and of the sensory estimate derived under different assumptions about the causal structure**.** Solid/dashed distributions indicate conditions with high/low auditory reliability. Upside-down triangle(s) represent the mean(s) of the distribution(s), with grey denoting task-relevant cue, and coloured denoting final estimates under different causal structures. **(A)** Assuming separate sources for two stimuli [cause (c)=2] leads to a sensory estimate that reflects the most likely stimulus in the task-relevant modality. **(B)** The assumption of a common source leads to the integration of both senses (reliability-weighted fusion). **(C)** With causal inference, the two hypotheses about the causal structures (c=1 or c=2) are combined probabilistically. The final Bayesian estimate combines the unisensory (task relevant) and the fused estimates, each weighted by its inferred probability and an *a priori* integration tendency. **(D)** Each candidate model predicts a unique relationship between crossmodal disparity (distinct visual vs. auditory stimulus positions are characterised by a large disparity) and bias (deviation of the final estimate from the true attribute). The shaded area corresponds to the example shown in **A** to **C**, i.e., visual stimulus is on the left side of the auditory stimulus.

Now, we can proceed to describe how the inferred causal structure (**Eqs. 8** and **9**) can be used to obtain an estimate of the sensory attribute. Observers can potentially use different decision strategies in service of this arbitration between the sensory estimates under different latent causal structures given by **Eqs. 8** and **9**. Assuming a cost function (Ma, 2012) of minimising squared error of the responses, a model averaging (MA) strategy can be used by the observers (Körding et al., 2007). MA (**Eq. 10 below**) minimises the mean squared error of the final estimate by linearly combining the estimates derived from segregation and fusion (in **Eqs. 4** and **5**), each weighted by the inferred posterior probability over the respective causal structure:

(**Eq. 10**): $\hat{S}_a = p(c = 1|r_v, r_a)\hat{S}_{a,c=1} + p(c = 2|r_v, r_a)\hat{S}_{a,c=2}$

$$\hat{S}_v = p(c = 1|r_v, r_a)\hat{S}_{v,c=1} + p(c = 2|r_v, r_a)\hat{S}_{v,c=2}$$

This Bayesian causal model makes a distinct prediction for how the perceived sensory attribute varies with the experimental manipulation in this hypothetical task (crossmodal disparity, i.e., the difference between auditory and visual attributes), compared with the fusion model (**Figure 1.3D**). The fusion model predicts a crossmodal bias that grows linearly with disparity (cues are hence always combined) and scales with the relative reliability of the individual cues to reduce noise in the combined percept. By contrast, the causal inference model allows for an additional inference over sensory causality formulated by combining a sensory-driven estimate of signal probability (sensory likelihood) with an explicit tendency of sensory integration (Körding et al., 2007; Rohe & Noppeney, 2015b; Wozny et al., 2010). The inference model still predicts that the crossmodal bias scales with inter-sensory disparity and relative cue reliabilities (Rohe & Noppeney, 2015a), but the bias

diminishes for highly discrepant multisensory information that unlikely originates from a common source (i.e., a nonlinear dependency of the bias on disparity).

Observers can potentially rely on a more heuristic strategy by simply using probability matching (PM) stochastically over experimental trials. That is, with PM as decision strategy (**Eq. 11**) the observer aims to arbitrate between the fusion and segregation using a probabilistic rule, with the relative probability of each outcome matching the inferred probability of each causal scenario. This strategy was modelled using a stochastic selection criterion, γ, which was sampled uniformly between 0 and 1 independently on each trial:

(**Eq. 11**): $\hat{S}_a = \hat{S}_{a,c=1}\ if\ p(c = 1|r_v, r_a) > \gamma, \hat{S}_a = \hat{S}_{a,c=2}\ if\ p(c = 1|r_v, r_a) \leq\ \gamma$

$$\hat{S}_v = \hat{S}_{v,c=1}\ if\ p(c = 1|r_v, r_a) > \gamma, \hat{S}_v = \hat{S}_{v,c=2}\ if\ p(c = 1|r_v, r_a) \leq\ \gamma$$

With model selection (MS) as decision strategy an observer selects fusion (vs. segregation) as long as the inferred posterior probability of the respective causal structure "c = 1" (vs. "c = 2") exceeds 0.5. Practically, this corresponds to fixing γ in **Eq. 11** at 0.5. Model averaging, probability matching, and model selection strategies make qualitatively similar predictions about the nonlinear relationship between the crossmodal bias and crossmodal disparity, but differ in their predictions about the shape of the posterior distribution over sensory feature.

However, no matter how the inferred causal structures (**Eqs. 8** and **9**) are exploited to obtain an estimate of the sensory attribute, the different models under the assumption about different causal structures (common, separate, or uncertain) can be transformed to each other by restricting or setting free the *a priori* integration tendency p_c. Specifically, by restricting this parameter to 1 or 0, the causal inference process can be reduced to a fusion and segregation process, respectively, assuming

the other parameters (central prior, sensory variance, etc.) are held unchanged. This hierarchical model dependency has begun to receive increasing attention in recent computational modelling (Acerbi, Dokka, Angelaki, & Ma, 2018; Rohe & Noppeney, 2015a; Trommershauser et al., 2011).

1.2.3 Mixed findings for the best-fitting computational account

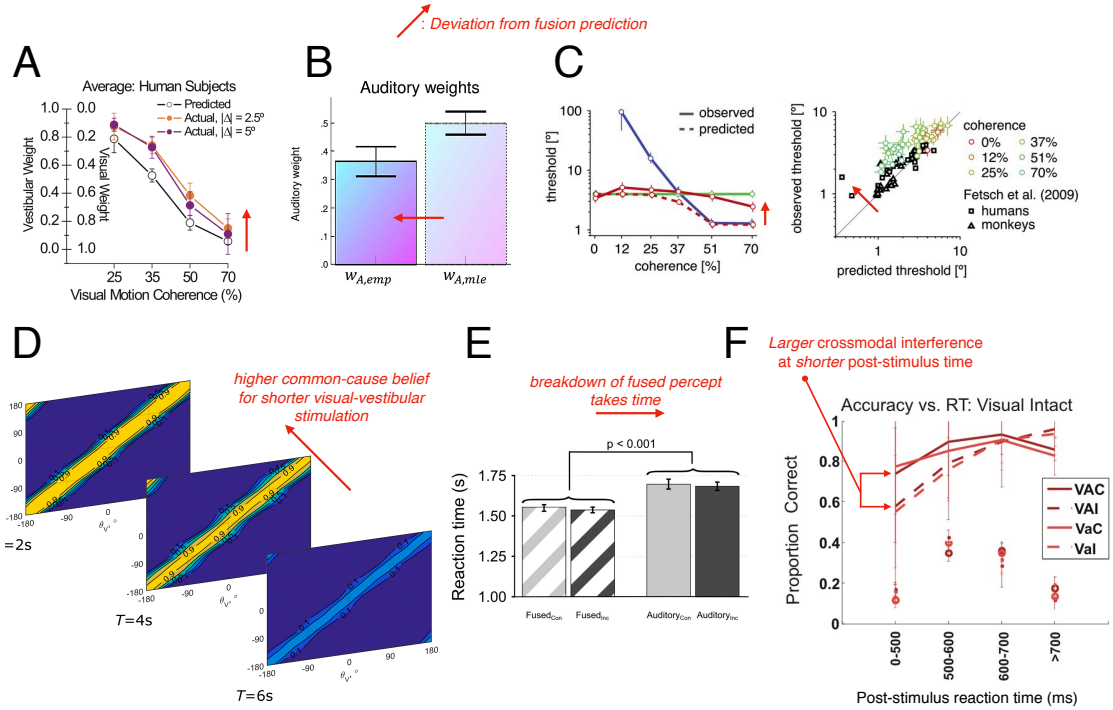


Figure 1.4. Psychophysical evidence implicating violations of fusion model (A, B, C), and time-dependent crossmodal influence on behaviour. (**A**) Fetsch et al. (2009): humans over-weigh vestibular information compared with what should be predicted by a fusion model. (**B**) Meijer et al. (2019): humans over-weigh visual information in spatial localisation. (**C**). Humans do not perform as well as a fusion model predicts in a speeded heading direction judgement task. (**D**) Shorter stimulation duration leads to higher interference of crossmodal information in visual-vestibular integration. (**E**) Conscious segregation of multimodal cues that usually tend to be fused is associated with significantly longer reaction times (RT). (**F**) Larger crossmodal interference is observed when human participants make faster decisions in a visual discrimination task in the presence of auditory distractor.

There have been mixed results about whether reliability-weighted fusion or Bayesian causal inference provides the best description of multisensory behaviour in

various experimental settings. Sometimes, these two models seem to unexpectedly explain human choices in experimental conditions that presumably only satisfy the assumption of each particular strategy (e.g., Butler, Campos, & Bülthoff, 2015; De Winkel et al., 2015; Meijer et al., 2019; Prsa, Gale, & Blanke, 2012; Rohe & Noppeney, 2018).

On one hand, a noticeable violation of the predictions of fusion has been documented in experiments that were originally designed to 'promote' sensory fusion across multisensory information. For instance, Meijer et al. (2019) pre-registered a psychophysical study in order to systematically (re)-test the fit of the MLE fusion model to an audiovisual spatial judgement task. Together with neural evidence supporting a deviation of the sensory weights estimated from fMRI BOLD activities from model predictions (Rohe & Noppeney, 2018), Meijer and colleagues identified a significant down-weighting of the auditory information during audiovisual integration (i.e., human participants tend to consistently over-weigh visual input; **Figure 1.4**). This result fits with an earlier observation in studies by Angelaki and colleagues (e.g., Fetsch et al., 2009), whereby significant over-weighting of one sensory modality (in this case, the vestibular modality in primates) was documented as evidence violating the prediction of a perfect reliability-weighted fusion (**Figure 1.4**). In the studies by Meijer et al. (2019) and Rohe and Noppeney (2018), one origin of this violation of fusion prediction might be related to the use of simulated spatial sound *a priori* promotes a detection of a slight mismatch between the spatial sound (presented via earphones) and spatial visual cue presented on a screen. In the study by Fetsch et al. (2009), the explanation could be that, in the case of heading discrimination, the extensive training given to the animals may have led to a sense of 'causal mismatch' between the visual optic flow (contributing to the feel of moving forward) and the

heading movement initiated by the robotic motion platform. In other words, causal inference leads to the observed violation of fusion. Sensory weighting may also differ from what a fusion model would have predicted in situations where a decision must be made in the presence of speed requirement (Drugowitsch, DeAngelis, Klier, Angelaki, & Pouget, 2014).

On the other hand, what seems more surprising is that in situations where a causal inference model should have been accounting for flexible behaviour, forced fusion is still observed even if the crossmodal disparity seems quite large (Butler et al., 2015; De Winkel et al., 2015; Prsa et al., 2012). Interestingly, two studies carried out by the same group of researchers (De Winkel et al., 2015; cf. De Winkel et al., 2017) found fusion explained behavioural data in one study whilst causal inference explained data in another using the same task but stimulations with longer durations. One plausible explanation is that fusion and causal inference constitute two component processes that co-exist in the brain (Rohe & Noppeney, 2015b), but dominate at different times. We are able to find some evidence supporting this possibility by looking at the actual data collected by De Winkel et al. (2017). By manipulating the duration of the visual-vestibular stimulation, De Winkel and colleagues (2017) estimated from participants' behavioural choices the probability of fusing crossmodal information as if they come from the same source (i.e., a posterior belief about common cause). What the researchers found was that the shorter the duration, the higher this fusion probability, indicating that the resolution of a crossmodal discrepancy depends on accumulating sufficient evidence about each sensory information, which per se requires additional processing time. A similar strand of findings has been found in psychophysical studies on audiovisual information integration. For example, crossmodal biases appear to be stronger when

participants respond faster (Noppeney, Ostwald, & Werner, 2010), and a conscious segregation of multimodal cues that usually tend to be fused requires additional time and effort (Gau & Noppeney, 2016).

These mixed findings partially motivate the projects in this thesis: Is there a mechanistic account that could reconcile these different computational theories? For instance, if these distinct computational strategies (fusion and CI) co-exist in the brain (Rohe & Noppeney, 2015b), do neural representations of the sensory information as predicted by these computational principles emerge simultaneously or sequentially, and in the same or across distinct brain regions? Crucially, what could be the neural coding scheme that organises them to guide flexible behaviour (**chapters 3** and **4**)?

1.2.4 Bayesian computations using probabilistic population coding

In this section, I will provide a summary of both theoretical and empirical studies using the framework of Poisson-like probabilistic population coding (PPC) to explain flexible multisensory integration (Deneve, Latham, & Pouget, 2001; Ma, Beck, Latham, & Pouget, 2006). I will begin by describing the basic principle of PPC and its relation to the probabilistic model of sensory fusion (**section 1.2.1**). I will then summarise how neurophysiology-driven studies have strived to modify existing models to accommodate neural data that seem to deviate from Poisson-like PPC in some subtle ways, and how theory-driven studies have strived to find neural basis to support the existing framework. In the end, I will introduce recent endeavour of extending PPC to causal structure inference.

One prediction of this PPC framework is that optimal Bayesian integration is readily achievable via a simple linear *unweighted* summation of the population patterns of activity elicited by the cues, assuming Poisson-like neuronal variability.

By ‘unweighted’, it simply means that there is no need whatsoever to modulate the population patterns of activity by any extra neural weight. Mathematically, this is easy to explain: the gain (peak amplitude) of the population activity profile already encodes the information about cue reliability of a sensory stimulus, and a linear summation of different population patterns of activity readily accounts for the reliability-weighted sensory integration in behaviour. This notion can be better understood by looking at the mathematical expression of the noise distribution p(***r***|s) for the population code ***r*** (a vector of spike counts, one for each neuron):

$$p(\boldsymbol{r}|s,g) = \prod_{i=1}^{n} \frac{1}{r_i!} e^{-f_i(s)} f_i(s)^{r_i} = (\prod_{i=1}^{n} \frac{1}{r_i!}) e^{-\sum_i f_i(s)}\, e^{\sum_i r_i log f_i(s)}$$

Where $f_i(s)$ is the tuning curve of the *i*th neuron within that population and takes the assumption of a Gaussian form: $f_i(s) = gA_i e^{-\frac{(s-s_{pref,i})^2}{2\sigma_{tc,i}^2}}$, in which g is a gain factor fully determined by stimulus cue reliability that might vary from trial to trial (Ma et al., 2006), A_i is the neuron’s firing amplitude at unit gain; $s_{pref,i}$ and $\sigma_{tc,i}$ denote the neuron’s preferred stimulus and the width of its tuning curve, respectively. Because it is reasonable to assume that $\sum_i f_i(s)$ is approximately independent of the stimulus *s*, we can replace it by a constant *K*. We find:

$$p(\boldsymbol{r}|s,g) = K(\prod_{i=1}^{n} \frac{1}{r_i!}) e^{\sum_i r_i \log(gA_i)} e^{-\sum_i r_i \frac{(s-s_{pref,i})^2}{2\sigma_{tc,i}^2}}$$

Using marginalisation to average out the random variable *g*, we find:

$$p(\boldsymbol{r}|s) = \int p(\boldsymbol{r}|s,g)p(g)dg = K(\prod_{i=1}^{n} \frac{1}{r_i!}) e^{-\sum_i r_i \frac{(s-s_{pref,i})^2}{2\sigma_{tc,i}^2}} \int e^{\sum_i r_i \log(gA_i)} p(g)dg$$

That is, the strength of the population activity readily contains the information about cue reliability, which per se is fully determined by the *g* parameter. It is obvious to see that taking the linear sum of any two population codes, ***r***1 and ***r***2, is equivalent to the optimal MLE solution given by **section 1.2.1** (see Ma et al., 2006).

However, the neural data from a series of neurophysiological studies do not seem to fully support this particular prediction. Morgan, DeAngelis, & Angelaki (2008) recorded from individual MSTd cells whilst non-human primates were presented with visual-vestibular heading stimuli. They found that although the linear combination rule as suggested by PPC provided a reasonably good fit to the neural data, the neural weights assigned to the neural responses to each unisensory input when predicting the neural responses a simultaneous combination of the two unisensory inputs varied considerably as a function of the unisensory cue reliability. This is at odds with the constant neural weight (presumably just fixed to 1) as proposed by the PPC. Furthermore, Fetsch, Pouget, DeAngelis, & Angelaki (2011) demonstrated that the neural re-weighting occurred even when stimulus reliability was varied randomly on a trial-by-trial basis.

Such a conflict with the PPC-predicted neural weights in empirical data (in MSTd) has inspired two parallel lines of thinking: **1)** perhaps a clever extension of the PPC is needed to explain why neural weights vary with cue reliability beyond the simple linear summation of population code, and **2)** maybe a better match between data and theory should be found in other brain structures, e.g., a downstream area to the MSTd, such as the parietal cortex (Hou, Zheng, Zhao, Pouget, & Gu, 2019). An extension of the standard PPC using the concept of *divisive normalisation* can account for why sensory cue reliability modulates the neural weight measured at the level of neuronal population (see Fetsch, DeAngelis, & Angelaki, 2013; Ohshiro,

Angelaki, & DeAngelis, 2011; Seilheimer, Rosenberg, & Angelaki, 2014). Changing the cue reliability of one particular sensory modality naturally affects the magnitude of the normalisation term acting upon the population of neurons inside the multisensory pool that upstream unisensory population code project to. This in turn suppresses the unisensory responses in a way that the neural weights reflect the modulated population gain of unisensory stimulation.

So far, there has been a growing interest in but extremely rare evidence on how Bayesian causal inference could be realised in Poisson-like PPC (Ma & Rahmati, 2013). The present view is that computations like causal inference requires a biologically unrealistically huge amount of operations to follow very specific combinations of their algorithmic rules (e.g., linear operations, quadratic operations, and divisive normalisation). It is not clear yet what solution could be there to guide a reconciliation or unification of theories. However, a search of canonical approximate inference algorithms such as sampling or variational methods could be one of the promising endeavours (Gershman & Beck, 2017). By contrast, alternative frameworks such as recurrent network models seem to be capable of recapitulating a wide variety of task performance as predicted by flexible Bayesian computations, and thus are treated as a useful tool for 'understanding' the inferential processes in the brain (Darlington, Beck, & Lisberger, 2018; Remington, Narain, Hosseini, & Jazayeri, 2018; Sohn, Narain, Meirhaeghe, & Jazayeri, 2019). Nonetheless, such approaches may suffer from one potential drawback: it could make it difficult to understand how exactly a network solves the task in every single step and how important each step is to the signature of task performance of interest. This is conceptually analogous to the case of training nearly any nonlinear neural networks to solve a computational problem in which certain component of the network (such as nonlinearity) does not

play a pivotal role in producing key features of the underlying computational process (such as learning dynamics), but some component does (e.g., network depth, see Saxe, McClelland, & Ganguli, 2019). For the purpose of making sense of biologically plausible approximations to flexible inference such as causal structure inference, such a "lesion" approach to network architecture could be a promising direction.

1.3 The functional neuroanatomy of multisensory inference

1.3.1 Reliability-weighted multisensory fusion in the brain

Only a handful of human functional imaging and non-human primate neurophysiology studies have strived to shed light upon the neural substrates of reliability-weighted fusion across modalities. Converging evidence from these studies has implicated the superior temporal and posterior parietal regions as being critical not only in representing a fused multisensory estimate, but also in signalling trial-by-trial neuronal weighting of sensory cues according to their reliability.

A first set of evidence came from single-unit studies on non-human primates. Using a visual-vestibular heading discrimination task, Gu, Angelaki, and DeAngelis (2008) related neural activity recorded from a higher-order sensory region of macaque monkeys (the dorsal portion of the medial superior temporal area, MSTd) to behaviour by converting spike activities into perceptual choices via simulating the behaviour of an ideal observer. In this analysis, the simulated model observer effectively performs the discrimination task by comparing distributions of firing-rate distributions recorded in response to different stimuli. This procedure resulted in neurometric curves (**Figure 1.5**) that quantified the sensitivity and

pattern of cue weighting in a single neuron that remarkably resembled the patterns in the behavioural data. These findings indicate that neurons of this higher-order sensory cortex show close parallels with behaviour with respect to both the sensitivity enhancement (upper panels) and crossmodal cue weighting (lower panels of **Figure 1.5**; see Fetsch et al., 2013), as predicted by the reliability-weighted fusion principle.

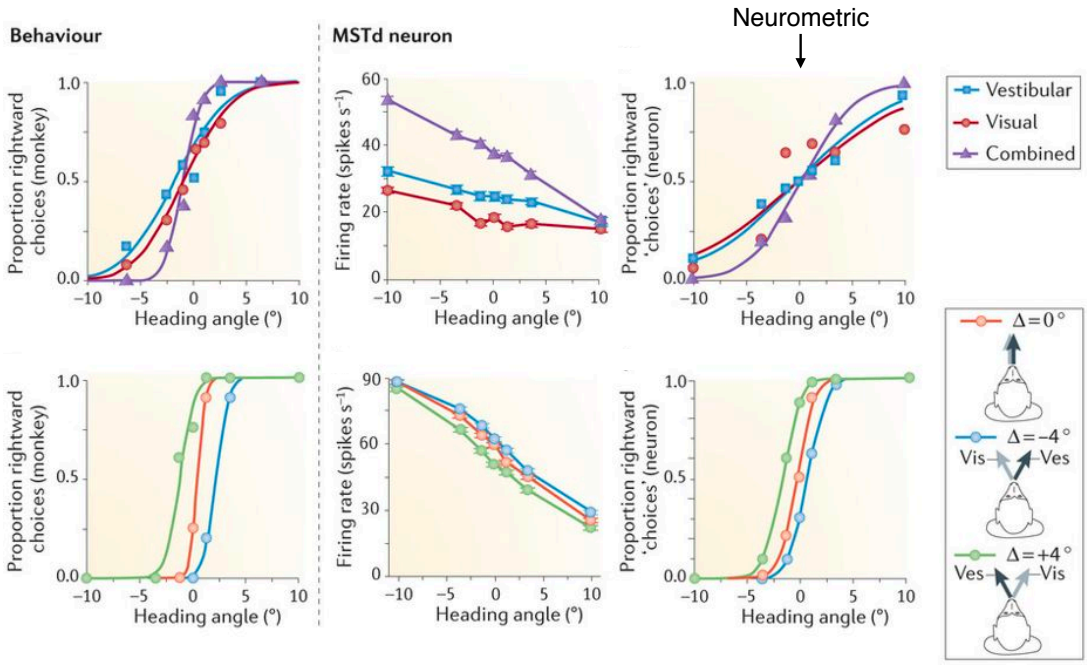


Figure 1.5. Neural correlates of reliability-weighted sensory fusion in neurons of the MSTd in non-human primates. Behavioural data (psychometric functions) for a single session are plotted showing the proportion of rightward choices as a function of signed heading angle. The firing rate responses (tuning curves) of a single example neuron from the dorsal medial superior temporal area (MSTd). The firing rates were converted into simulated choices by an ideal observer using receiver operating characteristic (ROC) analysis. The resulting 'neurometric' functions quantify the sensitivity of the neuron to small changes in heading during the vestibular (blue), visual (red) and combined (purple) conditions. In a separate study (Fetsch, Pouget, DeAngelis, & Angelaki, 2011), the cues were placed in conflict to test for reliability-based cue weighting (lower panels). Figures adapted from Fetsch, DeAngelis, and Angelaki (2013).

Some theoretical studies have recently strived to shed light on the weighted combination rule observed in primate MSTd neurons at cellular and/or circuit levels. It has been shown that sensory fusion may be implemented through frugal computations—e.g., operated in feedforward network processes (Ma et al., 2006; Ohshiro et al., 2011, 2017). The architecture of this simple network-level computation is characterised by a set of unisensory inputs converging onto another topographically-aligned layer of multisensory neurons in a feedforward fashion. The network-predicted weights for the neural combination rule turn out to be depending on the coherence of the visual optical flow in a manner similar to that exhibited by the MSTd neurons during the visual-vestibular cue-integration task (Fetsch et al., 2011; Ohshiro et al., 2011, 2017).

The neural implementation of reliability-weighted fusion, however, highly likely depends on a much broader network of brain regions, rather than a single highly localised cortical structure such as the MSTd. To test for a causal link between MSTd activity and the behavioural readout of reliability-weighted fusion, in a follow-up study capitalising on a similar behavioural paradigm as in the research by Gu et al. (2008) and Fetsch et al. (2011), Gu, DeAngelis and Angelaki (2012) used micro-stimulation and reversible inactivation techniques to perturb brain activity in the primate MSTd whilst monitoring behavioural performance during multisensory choices. Their results indeed support a causal link between MSTd neurons and multisensory heading perception. But they also suggest that other brain structures must be playing a similar role (e.g., ventral intraparietal area: Chen, DeAngelis, & Angelaki, 2013). Specifically, in spite of a significant increase in the discrimination thresholds, which clearly indicates the causal role of MSTd in heading perception, the tested animals were still able to combine visual and vestibular cues “near-optimally”

as evidenced in the behavioural readouts, even in the presence of inactivation of large portions of MSTd neurons bilaterally. One conclusion drawn upon these pieces of evidence is that sensory integration relies on a distributed network of multiple brain structures and potentially the functional interplay between them (Bizley, Jones, & Town, 2016; Kayser & Shams, 2015).

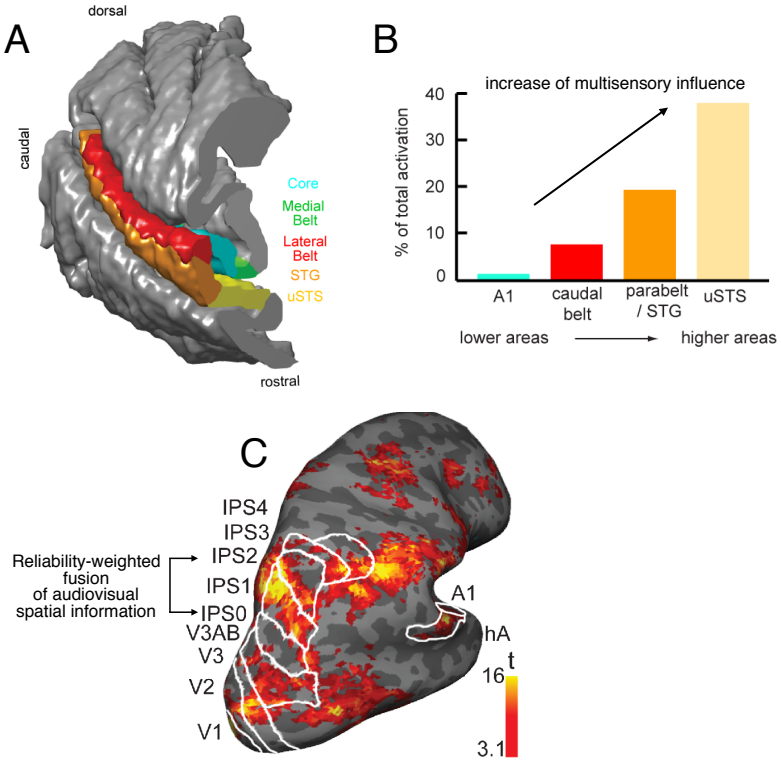


Figure 1.6. A distributed and topographical organisation of sensory integration across modalities (audiovisual integration across the three panels here; A and B for non-human primate and C for human). (**A**) Three-dimensional rendering of a segment of a primate brain. Different structures investigated in fMRI experiments are colour-coded and comprise classical auditory cortex (core and belt) as well as auditory association cortex (parabelt) and general association cortex (STS). (**B**) A graded strength of visual influence along the auditory hierarchy, as measured by the fraction (%) of BOLD response to visual stimulation relative to the sum of BOLD responses to all three conditions (visual-only, auditory-only, and combined audiovisual). Visual influence over the auditory hierarchy increases from lower to higher areas. (**C**) Along the dorsal visual hierarchy, neural representation of the sensory encoding as predicted by a reliability-weighted fusion model can be decoded from BOLD activity of the posterior regions of the intraparietal sulcus in humans (A and B are adapted from: Kayser, Petkov, Remedios, & Logothetis, 2012; C: Rohe & Noppeney, 2015b).

Another brain structure, the posterior parietal cortex, has been highlighted in human non-invasive imaging studies as playing a functional role in reliability-weighted sensory fusion (Helbig et al., 2012; Rohe & Noppeney, 2015, 2016). Of particular interest are the neuroimaging studies by Noppeney and colleagues, in which the sensory representations as predicted by the reliability-weighted fusion model can be decoded from brain activities over the posterior parietal regions. In these regions, multisensory fusion processes seem to occur in a manner as if merging the sensory evidence about space representation provided by different modalities, weighted by their relative reliabilities, but regardless of how likely it is that these actually relate to the same object or originate from distinct sources.

Of note, distinctions must be drawn between brain structures that appear to code neural activity more in the frame of sensory convergence and/or interactions, such as early sensory (and subcortical) structures, and regions that are more directly in the frame of subserving multisensory fusion according to cue reliability. Based on convergent neural evidence gleaned from studies in the past decade, there seems to be three things that we have already known concerning the neural substrates of multisensory fusion. In what follows, I summarise these three points:

1) Multisensory influences already exist at early processing stages, but they progressively increase in higher areas (Bizley & King, 2009; Bizley, Nodal, Bajo, Nelken, & King, 2007; Dahl, Logothetis, & Kayser, 2009; Kayser et al., 2007, 2012; Lewis & Noppeney, 2010; Schroeder & Foxe, 2005);

2) The neural underpinnings of reliability-weighted sensory fusion, however, are more likely to be confined to higher-order association cortices than those of sensory integration and/or convergence defined in a general sense, and to the

timescale of 100-ms to 250-ms post-stimulus onset (Aller & Noppeney, 2019; Boyle, Kayser, & Kayser, 2017; although nuances in latencies possibly depend on the nature of different stimuli and/or tasks under investigation);

3) Those earliest (on the scale of faster than 50-ms post-stimulus onset) multisensory interactions (mediated via colliculo-pulvinar pathways, Cappe, Rouiller, & Barone, 2009; Hackett et al., 2007; or via direct cortico-cortical pathways connecting primary sensory cortices, Cappe & Barone, 2005; Falchier et al., 2002), although not necessarily exhibit the characteristics of fusion according to reliability, could instead potentially be capitalised upon for the build-up of the neural representation of a fully fused sensory estimate.

1.3.2 Multisensory causal structure inference in macroscopic brain signals

In contrast to the existing evidence for when and where reliability-weighted fusion might be coded in the brain (as reviewed in **section 1.3.1** above), the neural theories concerned with sensory causal inference had remained largely unexplored until very recently. This served as one of the motivations underlying the human neuroimaging project of this thesis (**chapter 3**). At least for the neural representation of space, regions more anterior in the intraparietal sulcus (in contrast to the more posterior regions encoding sensory information as predicted by fusion) represented the spatial estimate as predicted by the causal inference model, hence appeared to adapt their neurally-coded sensory representation based on the likely causal origin (Rohe & Noppeney, 2015b). Nevertheless, there remain many open questions. One of the starkest is with regards to mechanism: Is the anterior patch of the parietal cortex

the only prominent cortical centre encoding sensory information as predicted by causal inference? If there could be a network of brain structures encoding multisensory signals for causal inference, which of them also directly drives the flexible behaviour exhibiting nonlinear weighting of disparate crossmodal information? What is the appropriate, physiologically plausible mechanism underlying the timecourse of multisensory causal inference in the brain?

1.3.3 Reasoning, strategic arbitration, and multisensory conflict resolution in the brain

Although neural evidence concerning particularly how the brains infer "multisensory" causal structures has been rare, one could gain some valuable insights from past studies on the neural basis of human reasoning, causal learning, rule abstraction, and multisensory conflict detection and resolution.

Several seminal studies on the neural mechanisms of inductive reasoning and/or probabilistic rule representation have strongly implicated the frontopolar region and lateral prefrontal structures as playing a functional role (Boorman, Behrens, Woolrich, & Rushworth, 2009; Donoso, Collins, & Koechlin, 2014). In Donoso et al. (2014), for instance, human volunteers actively monitor a currently executed inferential strategy as well as maintaining two or three alternative strategies in a decision-making task whereby a number displayed on a computer screen needs to be assigned to one of several candidate outcome options. Unbeknownst to participants, these input-outcome contingencies are not fixed but changing across different 'episodes' (containing a variable number of trials) and must be acquired through representing the reward-based structure of the task through trial and error. A

computational model combining reinforcement-learning algorithms with certain features designed to track inferential strategy switching was proposed by Donoso et al. (2014), and the model-predicted behavioural components such as value representations, strategic rejection/confirmation signals were then mapped to brain activities. Consistent with past work revealing value signals for alternative actions in the frontopolar cortex (Boorman et al., 2009), the rostro-lateral PFC regions were identified by Donoso et al. as representing the reliability associated with different inferential strategies and arbitrating between them to support reasoning and guide flexible behaviour.

The lateral PFC has also been implicated in causal structure representation and learning (Tomov, Dorfman, & Gershman, 2018). Mapping to whole-brain neural activities a full posterior probability distribution over causal structures as predicted by a causal structure learning model, Tomov et al. (2018) were able to identify a key network of brain structures including the lateral PFC and the frontal pole that support structural learning. These neural underpinnings remarkably overlap with a functional hierarchy of the lateral PFC as guiding the acquisition of "abstract" rules underlying complex stimulus–outcome relations (Badre, Kayser, D'Esposito, 2010; Frank & Badre, 2012). Causal structure inference may relate to hidden structure representation in a more generalised context, implicating a domain-general neural coding scheme of rule abstraction.

In the context of multisensory perception, the lateral PFC—the inferior frontal gyri/sulci in particular—has been highlighted as a key brain structure responsible for evaluating sensory discordance (Adam & Noppeney, 2010; Calvert, 2001; Hein et al., 2007; Noppeney et al., 2010). These studies consistently identified the lateral PFC as signalling a neural signature of conflict resolution and response selection, i.e., neural

activations (as measured by fMRI BOLD signals) in response to incongruent audiovisual inputs (e.g., a hammer impact sound combined with a picture of a musical instrument) are significantly greater than those in response to congruent inputs. Moreover, in virtue of connectivity approach to fMRI data, the left inferior frontal sulcus, has been implicated in the role of accumulating multisensory evidence via weighting its functional links to unisensory cortical structures (e.g., occipital and temporal regions).

Furthermore, the lateral PFC has been highlighted in forming beliefs over inferential states of the sensory environment, based on expectations or prior experience, in face of uncertainty about noisy sensory inputs (Gau & Noppeney, 2016; Kayser & Kayser, 2018; Noppeney et al., 2010). Together with recent causal findings in primate neurophysiology—e.g., the activity in the ventrolateral PFC appears to carry over the representations of past decision-related signals and bias the representations of subsequent decisions in a trial-by-trial fashion (Tsunada, Cohen, & Gold, 2019)—the lateral PFC acts as a strong candidate structure for integrating different types of predictive information, such as a prior over stimulus attributes or statistical correspondence across modalities, with bottom-up sensory evidence.

Surprisingly, however, the role of the prefrontal structures in multisensory processes has long been at least in part dominated by the view that the PFC acts as a "convergence zone" for multisensory information (Jones & Powell, 1970), with the ventrolateral PFC receiving afferent projections from auditory and visual cortices and association areas implementing sensory fusion (Romanski, 2007, 2012; Sugihara, Diltz, Averbeck, & Romanski, 2006). Despite these fruitful endeavours systematically revealing multisensory neurons in the frontal cortex, there is still an

open question about how flexible behaviour capitalises upon the presence of these multisensory neurons.

One step forward is to have a closer look at the neural representational activities that carry diagnostic information about what computations these frontal structures need for the efficient multisensory inference. The MEG project in **chapter 3** strives to follow this endeavour. Another open question is concerned with how neural representations are read out to guide behaviour. Recent work has emphasised the importance of mapping neural code to behaviour because not all spatiotemporal features of neural population activity are directly relevant for driving behaviour (Panzeri, Harvey, Piasini, Latham, & Fellin, 2017). The studies in **chapter 4** are dedicated to address this question.

1.3.4 Potentials and limitations of MEG for characterising the computations of multisensory integration

First and foremost, it is important to realise that any methodological approach or brain imaging modality, when implemented in isolation, provides only limited insights about the mechanism of any neuro-computational process. Any approaches have certain conditions under which they will produce spurious results, and these limitations should be borne in mind when choosing a method and when analysing the data. Comprehensive reviews about the inherent interpretational ambiguities of using functional neuroimaging methods for characterising multisensory computations can be found elsewhere (Kayser et al., 2012; Noppeney, 2012). In this section, I will briefly introduce a few key insights provided by these reviews. I will mainly focus on explaining the potentials and limitations of MEG as an imaging modality when

approaching a difficult neuro-computational problem such as multisensory causal inference.

A succinct but certainly excellent overview highlighting the pros and cons of different experimental methods is given by Panzeri and colleagues (2015):

"Current experimental methods emphasise different dimensions. Multi-neuron recordings resolving individual neurons at high temporal resolution provide information about both spatial and temporal dimensions and their non-separable intersection (Berényi et al., 2013). Imaging techniques such as two-photon imaging resolve individual neurons in space but often lack the fine temporal resolution to inform about spike timing (Kerr & Denk, 2008). Measures of mass activity at high temporal resolution (LFP, ECoG, EEG, MEG) cannot resolve individual neurons (and sometimes even brain regions), but capture temporal activity patterns that are coordinated across multiple neurons (da Silva, 2013)."

In this thesis, I adopted MEG (**chapter 3**) to investigate predictions from Bayesian models of multisensory integration. These models make fine-grained predictions about the mental representations of external stimulus. One key motivation of the thesis is to map these model predictions onto the temporal dynamics of neural activity, and so requires a technique with high temporal resolution.

Mass signals such as MEG have the potential of enabling comparisons and/or synergies of methods and/or findings across different species

Measurements of neural mass signals such as MEG share a closer physiological origin with local field potentials (LFP; they are both caused by post-synaptic (dendritic) sources of current in the brain) than BOLD haemodynamic response (measured using functional magnetic resonance imaging). A major gulf we have to bridge is the one between the techniques typically used to measure brain activity in the human and in non-human primates. Numerous important studies in the neurophysiological basis of multisensory processing have been capitalising on analysis of neural mass signals such as LFP activity (Atilgan et al., 2018;

Chandrasekaran, Lemus, & Ghazanfar, 2013; Ghazanfar, Maier, Hoffman, & Logothetis, 2005; Kayser, Petkov, & Logothetis, 2008; Perrodin, Kayser, Logothetis, & Petkov, 2015). Neural mass signals with high temporal resolution measured noninvasively in humans no doubt provide a parsimonious way of moving towards reconciliation or unification of cross-species findings.

Beamforming enables investigation of widespread cortical activation in a cognitive paradigm

Source reconstruction technique such as beamforming or multiple sparse priors (MSP) approach mitigates more or less the ill-posed inverse problem in MEG source reconstruction. A low-dimensional subspace of the experimentally measured activity usually suffices to explain the population dynamics underlying information processing. This means that the input data usually have a rank (e.g., independent components) that is less than the number of whole-brain MEG sensors. Going from this low-dimensional subspace to an accurate source reconstruction on a high-resolution grid (e.g., tens of thousands of grid points inside the brain) is ill-posed. Beamforming technique mitigates this issue accompanied by the inverse problem of source estimation by imposing biophysical constraints to the estimation procedure. Beamformer essentially designs a spatial filter for each source of interest in the brain that aims to achieve unity passband at that particular location whilst assigning zero weights to all other sources. Using beamforming rather than MSP has been shown to be more powerful to infer sources for subtle sensor-level effects (Hunt, 2011).

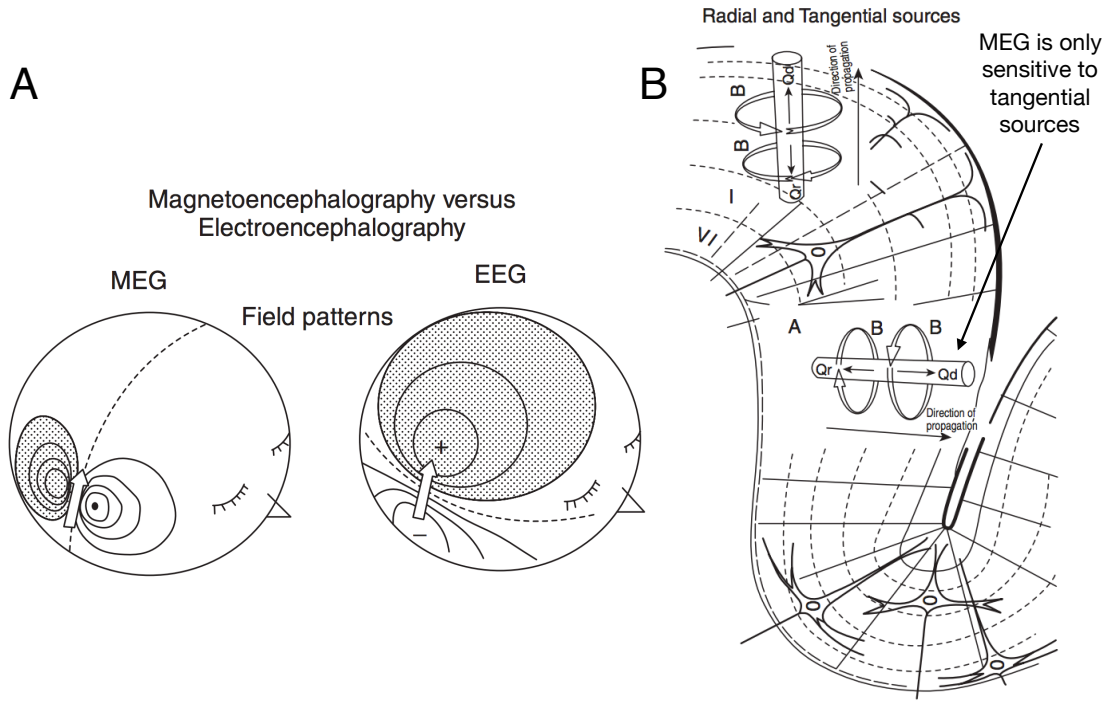


Figure 1.7. Biophysical difference between magnetic and electric fields. (**A**) Field pattern of M/EEG generated by a current dipole model. (**B**) The radial sources at the top of a gyrus are not detectable by MEG. Adapted from Hansen, Kringelbach, and Salmelin (2010).

On one hand, a key advantage of MEG over EEG in source estimation is its less sensitivity to the influence of inhomogeneities in the conductivity of different tissues (e.g., caused by factors such as white matter anisotropy and skull sinuses, Wolters et al., 2006). These tissue inhomogeneities smear out the electric potential much more than the magnetic field (see **Figure 1.7A**). On the other hand, MEG is only sensitive to tangential sources (whilst EEG reflects both tangential and radial sources; **Figure 1.7B**), but this also contributes to a remarkable simplification of the 'inverse problem'. Recent theories suggest that it is beneficial to develop advanced beamforming techniques for MEG and EEG fusion in the same participants to improve localisation of deep sources. There has been no consensus on the capacity of MEG for localising deep brain structures. Although my thesis does not focus on subcortical structures, these deep sources might be playing equally important role as

cortical structures in subserving multisensory processing (King, 2013). Historically, another limitation of the beamformer approach is that it has difficulty resolving distant sources that are highly correlated with one another. Recent theory suggests that such a caveat is mitigated in advanced analysis that applies beamforming to single-trial data: the correlation between sources is lower at the single-trial level than at the trial-averaged level. The MEG source reconstruction in **chapter 3** in the present thesis adopted this approach.

Limitations of non-invasive functional imaging for characterising sub-voxel-level representations of multisensory computations

Macroscopic neural activities reflect summary statistics aggregating neural responses across multiple neurons or multiple populations of neurons. Although the analyses in **chapter 3** adopted multivariate approaches to neural data, which essentially quantify the representational pattern activities across the voxels inside a spatiotemporal searchlight, the various interpretational limitations of mass-univariate statistical approaches that have been well-studied still apply to these multivariate pattern approaches.

Imagine a voxel or a searchlight identified by fMRI-BOLD or beamformed-MEG activities that seem to reflect a multisensory-fusion computation in which audiovisual features appear to be *integrated* into a single quantity. This result may inspire different interpretations with lots of subtitles if there is no additional constraint. For instance, that can arise either from one neuron sensitive to both auditory and visual features, or from greatly overlapping (or spatially proximate) but distinct populations that are sensitive to the auditory feature and visual feature independently. Similarly, auditory segregation can arise (potentially) because all

neurons sensitive to an auditory feature have the specific orientation to which MEG is most sensitive; whereas fusion can potentially arise because the number of auditory segregation vs. visual segregation neurons happens to match perfectly the reliability of the auditory and visual signals. In this case, the mass signals will reveal fusion even if it is a mere arithmetic average of the neural signals coming from the entire neuronal population. However, what is much more difficult to explain with these sub-voxel averaging processes at the macroscopic level is the task-modulated sensory bias as characterised by a causal inference process (**chapter 3**). In this sense, the results presented in this thesis are part of an emerging picture that it is critically important to unify evidence from different experimental species by drawing upon different methodological tools when approaching a difficult neuro-computational problem such as multisensory causal inference (human fMRI/EEG: Rohe & Noppeney, 2015b; Aller & Noppeney, 2019; human MEG: **chapter 3** and Cao et al., 2019; non-human primate electrophysiology: Fang, Li, Qi, Li, Sigman, & Wang, 2019).

1.4 Paradigms for unveiling the timecourse of decision formation

1.4.1 Short overview

The Bayesian framework outlined above (**section 1.2**) does not directly permit a strong inference about the mechanisms by which an integrated multisensory percept is calculated in the brain. We know that multisensory integration constitutes an instance of Bayes-optimal coding mechanisms when faced with sensory uncertainty; we know that humans and other animals seem to use precision weighting for sensory integration in a manner as if they perform Bayesian inference; we know that the

multisensory brains flexibly arbitrate between strategies of integrating versus segregating sensory evidence via resolving causal relations between multiple sensory features; we also know that the nervous systems (the lateral/medial prefrontal cortex in particular) are able to track, represent probabilistic summary statistics, e.g., posterior (orbitofrontal cortex (OFC): Chan, Niv, & Norman, 2016; lateral prefrontal cortex and OFC, Tomov et al., 2018), prior (insula and OFC: Vilares, Howard, Fernandes, Gottfried, & Körding, 2012; OFC: O'Reilly, Jbabdi, Rushworth, & Behrens, 2013). However, we also know that a tracking of overall Bayesian states may entail high processing costs due to large state spaces (Koechlin & Hyafil, 2007; Ma & Rahmati, 2013). As the number of dimensions increases, the computational demands of storing and updating increase combinatorially (Diaconis & Freedman, 1986), a phenomenon known as the "curse of dimensionality". This poses several challenges as to: **1)** whether the brain actually *represents* sensory estimates pertaining to different potential causes as formulated within the Bayesian models in an explicit neural coding scheme, and how the arbitration between these, as pronounced in behavioural readouts, is achieved in the nervous systems; and **2)** how best to characterise the brain's temporal dynamics and spatial hierarchy that transform, seemingly effortlessly, the plethora of sensory inputs into a percept of the multisensory world.

It seems critical to notice that a mechanistic account should provide detailed understanding about what the discrete information processing stages are that intervene between sensory encoding and choice behaviour, and how they are ordered in time and across the cortical hierarchy.

In this section, past experimental paradigms that have been used to investigate the unfolding of sensory-guided decision in time are summarised. I outline different

methods addressing the question of how discrete component processes of an information processing system are ordered to come online.

1.4.2 Variable (fixed)-duration task

When we talk about timecourse of decision formation, it is relevant to mention an additional line of work: sequential sampling models and their application to random-dot motion direction discrimination (Gold & Shadlen, 2007; Shadlen & Kiani, 2013). These models give decision-making researchers a chance to quantify the time needed to make a sensory-guided decision. Sequential sampling models have provided a dominant theoretical framework guiding computational and neurophysiological investigations of perceptual decision-making. Whilst these models come in many forms that can make similar predictions of choice behaviour despite invoking fundamentally different mechanisms, they share the basic principle that decisions are formed by accumulating sensory evidence to a decision bound.

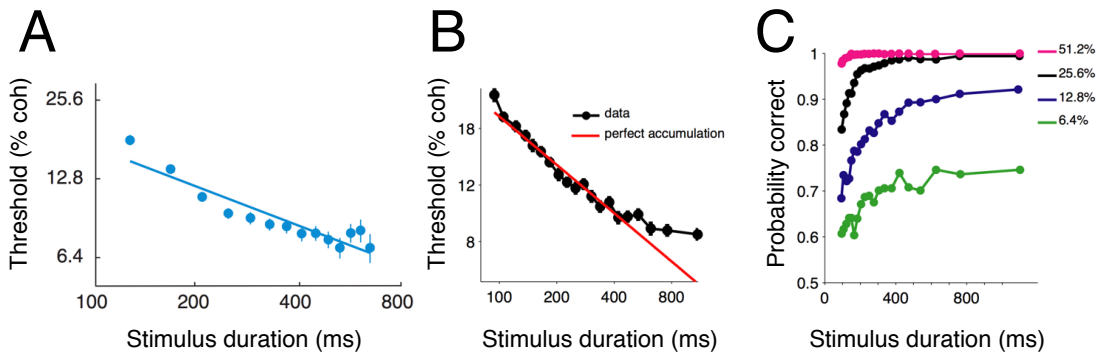


Figure 1.8. Quantifying timecourses of perceptual decision-making through manipulating stimulus duration. (**A**) Decision accuracy of visual motion-direction discrimination improves as a function of motion-viewing duration. Psychophysical threshold is defined as the motion coherence supporting 82% correct. (**B**) Behavioural discrimination thresholds deviate from perfect accumulation (red line) for longer viewing duration. (**C**) Ceiling improvement in perceptual accuracy for long motion viewing duration (presumably longer than the time at which a decision is already formed). Probability of correct response is plotted as a function of stimulus duration for four strength levels of sensory evidence determined by different motion

coherence. A is reproduced from Gold and Shadlen (2007); B and C are reproduced from Kiani, Hanks, and Shadlen (2008).

By systematically varying the duration of stimulus presentation, researchers have characterised neural correlates of the unfolding of decision process in time (**Figure 1.8**). For a random-dot motion direction discrimination task in which motion viewing time is determined experimentally and varied trial-to-trial unbeknownst to the observers, discrimination performance improves as a function of the square root of the viewing duration. This is a result that is expected if at each successive moment in time the brain acquires an independent sample of noisy evidence. More importantly, instead of exhibiting a perfectly linear function for the threshold improvement, participants' performance reaches a plateau (**Figures 1.8B** and **1.8C**). This can be explained by an optional stopping criterion adopted by the observer: the evidence accumulation process is terminated before the end of a sensory stimulus. Using other perceptual decision tasks and/or stimuli, a similar phenomenon has been found: observers show little or no improvement with further prolonged viewing duration (Ludwig, Gilchrist, McSorley, & Baddeley, 2005; Uchida, Kepecs, & Mainen, 2006). Hence, the fixed-duration paradigm can inform the experimenters about approximately when a decision maker initiates, persists or ends the decision process.

1.4.3 Sensory masking disrupts short-term memory trace

In addition, sensory stimuli are sometimes backward masked at different post-stimulus latencies (Ratcliff & Rouder, 2000). This is to limit to different extents the persistency of the neural representation of the stimulus in short-term memory (Smith & Ratcliff, 2009). The RT distributions found in such masking paradigms resemble

those obtained from tasks with response-terminated stimuli (see next **section 1.4.3** below). These distributions also suggest no evidence of increasing skewness at low stimulus discriminability, as would be expected if the decision process were driven by a decaying perceptual trace. The most plausible interpretation of this finding is that the drift rate in the decision process depends on a durable representation of the stimulus stored in short-term memory traces, which preserve the information during the course of forming perceptual decision. Backward masking has also been used in multisensory integration to facilitate an understanding of the precise timing of discrete stages by which the influence of one modality (e.g., audition) onto another (vision) comes into play (Chen & Spence, 2011).

1.4.4 Reaction-time (RT) paradigms

1.4.4.1 Sub-second decision under deadline-induced time pressure

Deadline RT paradigms are particularly useful for studying the timecourse of decision-making (perceptual decision: Noppeney et al., 2010; value-based decision: Blangero & Kelley, 2017). Previous studies have also shown that perceptual decisions are biased in favour of certain choice alternatives when the time available to view and/or to respond to the stimulus is limited (Dekel & Sagi, 2019; Diederich & Busemeyer, 2006; Feng, Holmes, Rorie, & Newsome, 2009; Murphy, Boonstra, & Nieuwenhuis, 2016; Stanford, Shankar, Massoglia, Costello, & Salinas, 2010; Whiteley & Sahani, 2008). As a result, subtle changes in choice patterns captured at different post-stimulus reaction times (RTs) help to dissect the component processes underlying the modulations of the inferential strategy during the formation of a decision. RT paradigms serve as powerful tools for dissecting into the strategic computations, expressed as different behavioural readouts, captured at different processing stages (**chapter 4**).

1.4.4.2 Multisensory decision under speeded response

The RT paradigm has also been applied to multisensory settings (Drugowitsch et al., 2014; Drugowitsch, DeAngelis, Angelaki, & Pouget, 2015; Noppeney et al., 2010). Using an RT version of audiovisual object categorisation task, Noppeney et al. (2010) revealed temporal dynamics of multisensory decision-making. Humans made more incorrect judgement about visual object category (larger crossmodal bias induced by an incongruent auditory object) for shorter RTs, because of a presumably higher belief near stimulus onset about common cause underlying temporally concomitant crossmodal signals.

Moreover, a modelling study by Drugowitsch et al. (2014) offers two critical insights. First, it is important to take into account both behavioural choice and reaction time when modelling multisensory inference. Drugowitsch et al. (2014) demonstrate that the multisensory benefit (enhancement of perceptual sensitivity) in an RT task is worse than predicted by the MLE model (Ernst & Banks, 2012; Angelaki et al., 2009), which has only been tested in non-speeded tasks (unlimited viewing time or sufficiently long stimulus duration with no pressure to response fast). Second, some of their test participants made less accurate but much faster choices when they were presented with both visual and vestibular information, compared to when only visual information was available. This may be specific to the vestibular-visual modalities used (vestibular alone induces more rapidly responses), but at the same time emphasises the need to carefully take into account the imbalance of information processing latencies between different modalities. This particular point supports the need to combine the RT and fixed-duration (with masking) paradigms to understand the unfolding of multisensory decision formation (**Figure 1.9**).

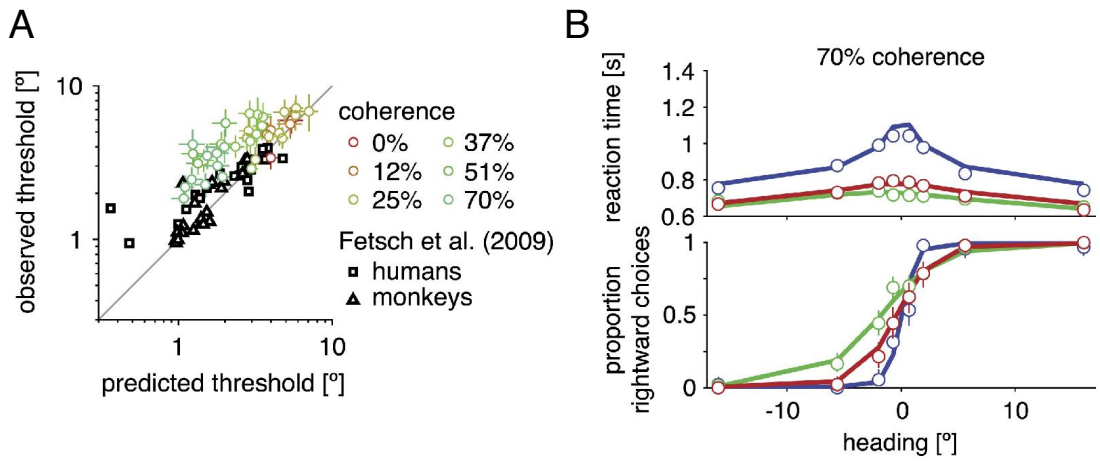


Figure 1.9. Sensory weighting between modalities changes as a function of RT. (**A**) Observed vs. predicted discrimination thresholds for multisensory condition with congruent visual-vestibular information about heading direction. Direct comparison between the thresholds estimated from the RT task in Drugowitsch et al. (2014) and those estimated from a previous study using a fixed-duration version of the same task (Fetsch et al., 2009). (**B**) Discrimination performance and reaction times in RT-version of visual-vestibular heading discrimination. Reproduced from Drugowitsch et al. (2014).

1.4.5 Weighting profile assessed with reverse correlation

To quantify at which time point an input signal contributes most to the final choice, some studies use deliberately constructed stimuli, usually stochastic in nature, with time-varying fluctuations in the choice-predictive power in different stimulus segments. For instance, in a paradigm that involved estimating which of two streams contained more pulses, Raposo et al. (2012) and Sheppard, Raposo, and Churchland (2013) quantified the ‘‘excess rate’’ defined as the difference in the mean window-averaged event rates between the selected trials that preceded high- versus low-rate decisions. ungrammatical consistently observed that most human and rodent subjects are influenced by sensory evidence throughout the trial duration regardless of whether stimuli are reliable or unreliable. Similar method has been widely used to quantify motion energy dynamics and its choice predictability reflected in temporally flexible

weighting process in both humans and monkeys (Levi, Yates, Huk, & Katz, 2018; Yates, Park, Katz, Pillow, & Huk, 2017). Multisensory integration has also begun to adopt such reverse-correlation tools (Parise & Ernst, 2016, 2017) although not yet for addressing timecourse of multisensory processing.

Chapter 2. Spatiotemporal representational similarity analysis and INDSCAL of source MEG activity

This chapter describes important concepts underlying the multivariate analysis techniques used in **chapter 3** (MEG study). An advanced diffeomorphic anatomical co-registration method that is equipped to improve the spatial specificity of MEG source-space RSA is described in **section 2.1.1**; Enhancement of the reliability of dissimilarity pattern using cross-validation distance and multivariate noise normalisation is discussed in **section 2.1.2**. Finally, a multidimensional-scaling method with individual weighting is described in **section 2.2**.

2.1 Representational similarity analysis (RSA) to MEG source-localised signals

2.1.1 Diffeomorphic anatomical registration using exponentiated lie algebra (DARTEL)

It has been well-known that source reconstruction for MEG data, i.e., inferring the current sources that generate the multichannel magnetic fields measured at the sensors, is a complex and 'ill-posed' problem. It is ill-posed because the number of different combinations of candidate sources that could have potentially generated the sensor-space signals (Baillet, Mosher & Leahy, 2001). An alternative approach to source reconstruction that circumvents the ill-posed bioelectromagnetic inverse

problem is beamforming (Hillebrand & Barnes, 2005; van Veen, Van Drongelen, Yuchtman, & Suzuki, 1997). A beamformer is an optimal spatial filter that reconstructs the contribution of a dipole within a grid specifying candidate locations (i.e., source points) whilst blocking the contributions of all other source points. Therefore, beamforming techniques do not need to specify a priori the number of active sources.

A challenge for beamforming-based MEG source reconstruction, however, is spatial blurring of source estimation. Apart from several frequently implicated origins of the spatial blurring such as correlated signals between distant source points (Hadjipapas, Hillebrand, Holliday, Singh, & Barnes, 2005) as well as depth-dependent resolvability of the inference problem (Hillebrand & Barnes, 2002), it is important to note that mis-alignment of statistical maps (e.g., RSA correlation maps) obtained from the brains of different participants also contributes to the spatial blurring of source estimation at the step of group-level inference. For MVPA and RSA, decoding and correlation analyses are typically performed in 'native' neural space based on an isotropic source-projection grid prepared for each participant by resampling their individual T1-weighted anatomical images (Haxby et al., 2001). For statistical inference, these native-space statistical maps must be co-registered to a common brain template. The extent to which such co-registration steps introduce spatial distortion for the inference maps presumably depends on the anatomical similarity between the cohort of participants in a study and the brain template used for co-registration (e.g., MNI/ICBM templates).

A clever solution to address these limitations is to create a common brain template using the anatomical images of the participants themselves. This idea constitutes the basic intuition underlying DARTEL, a diffeomorphic algorithm for

image deformation that essentially preserves the topology of the wrapped image by virtue of estimating a "flow field" (Ashburner, 2007; see also Klein et al., 2009; Yassa & Stark, 2009). DARTEL strikingly improves the spatial specificity of the anatomical template prepared for group-level inference, implies better alignment of the statistical maps across individual brains, and achieves more accurate anatomical labelling of group-level inference results (Giordano et al., 2018; Thornton, Weaverdyck, Mildner, & Tamir, 2019). Since the DARTEL space is truly specific to the group of participants of interest, it avoids the dependence on external atlases that might lack universal representativeness. A demonstration for the stark improvement of spatial specificity of DARTEL against a rigid-body registration is shown in **Figure 2.1**.

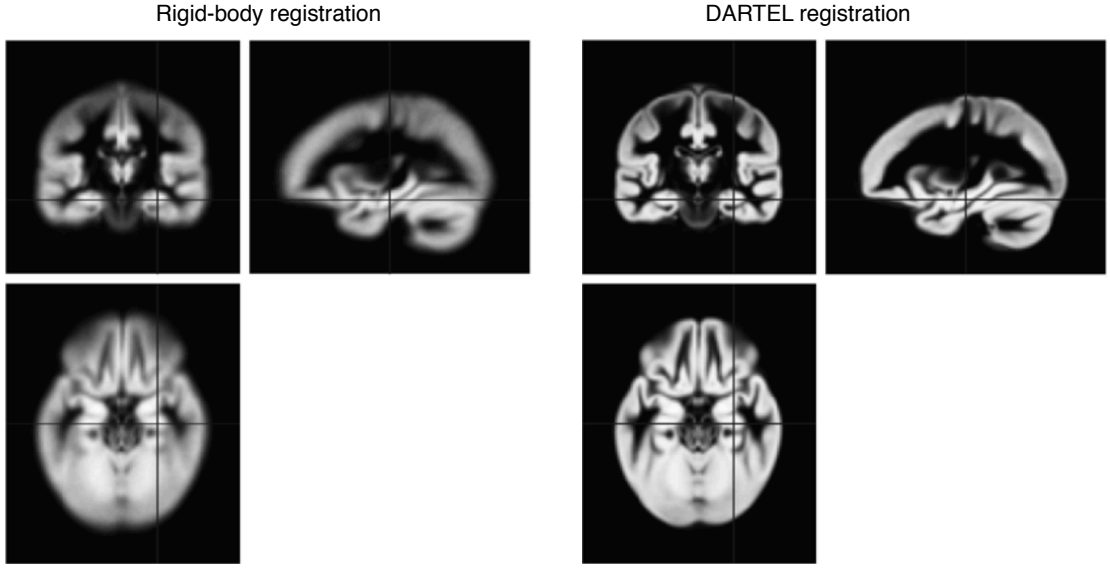


Figure 2.1. Enhancing spatial specificity of inference brain template using DARTEL. Comparison between an average template (grey-matter images) obtained from initial rigid-body registration (left) and an DARTEL template obtained through 18 iterations of the diffeomorphic algorithm for image deformation (right). Figures adapted from Ashburner (2007).

In practice, DARTEL has recently been implemented in a MEG RSA study (Cao, Summerfield, Park, Giordano, & Kayser, 2019; also see **chapter 3**) and a MEG-fMRI fusion RSA study (Giordano et al., 2018; **Figure 2.2**). DARTEL approach improves the spatial specificity and accuracy of MEG RSA results and facilitates the comparison between multimodal imaging techniques. As shown in **Figure 2.2** (left panel: "Categories dominance"), MEG RSA based on DARTEL reveals in the exact same region (of the right auditory cortex)—as fMRI does for the same cohort of study participants—the encoding of sensory prediction of a category emotion-encoding model describing similarity-rating behaviour (see Giordano et al., 2018). However, since fMRI signals are likely biased towards a particular latency of the evoked response, the fMRI RSA fails to detect the neural representation of a rival model that also describes behaviour (right panel: "Dimensions dominance"), which emerged at a later post-stimulus time and in fact well captured by the MEG RSA. DARTEL hence endows the MEG RSA with spatiotemporal specificity to reveal the unfolding of neural representations of sensory information.

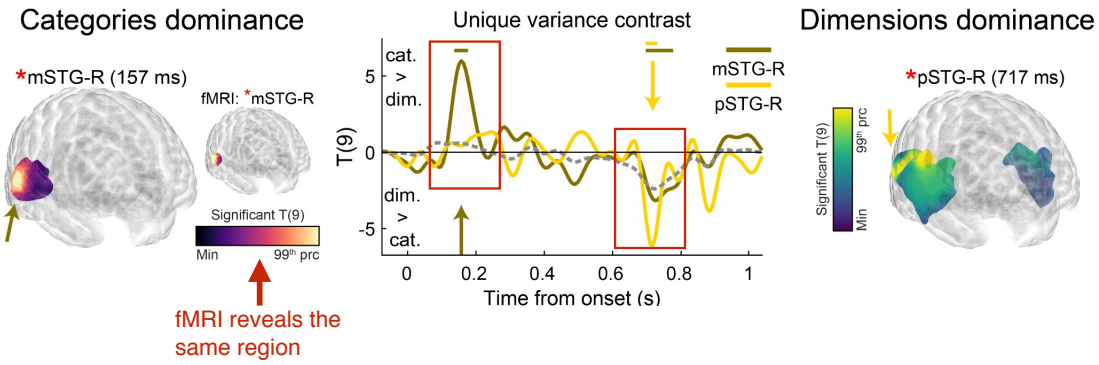


Figure 2.2. A MEG-fMRI fusion study directly comparing RSA to MEG and fMRI data collected from the same cohort of participants (Giordano et al., 2018). This work investigates the neural representations of two prominent models in the literature of emotion processing ("Categories" vs. "Dimensions"). Whilst fMRI reveals in the same region (of the right auditory cortex), as MEG does, the encoding

of sensory prediction of the category model, it fails to detect the neural representation of the other model, which emerges at a later time post-stimulus, likely because fMRI is biased towards a particular latency of the evoked response (rightmost panel). The middle panel shows the timecourses of model encoding contrast extracted from the brain regions preferentially dominated by the neural representations of the sensory predictions under each model.

2.1.2 Cross-validation for pre-whitened Euclidean distance

Representational similarity analysis (RSA) quantifies the statistical association between the pairwise dissimilarity of multivariate brain activity in different conditions (MEG representational dissimilarity matrix, "RDM" hereafter) and model RDMs that quantify hypotheses about the nature of the neural representations (here the computational models; Kriegeskorte, Mur, & Bandettini, 2008; Walther et al., 2016). Euclidean distance is one important measure of pattern dissimilarity because of its superior interpretability as compared with other dissimilarity measures such as correlation-based distances (Diedrichsen & Kriegeskorte, 2017; Walther et al., 2016). Cross-validation and multivariate noise normalisation (pre-whitening) are two independent methods for improving the reliability of the Euclidean dissimilarity measure.

Cross-validation. Since the Euclidean distance is non-negative by definition, any condition-nonspecific noise would bias positively the estimation of Euclidean distance. As noise level increases, the estimated dissimilarity pattern deviates more and more from the true activity patterns that are specific to the experimental conditions or stimuli (**Figure 2.3** left). Given intrinsic noise in any neural data, such condition-nonspecific neural distances would hinder the reliability of the dissimilarity measures and in turn the interpretability and sensitivity of RSA. One solution to this challenge is to 'cancel out' these condition-nonspecific neural distances by splitting the original multi-run neural data into *M* partitions and projecting the pattern vectors

of one partition onto those of the rest *M*-1 partitions. As shown in a simulation (**Figure 2.3** right), this cross-validation approach presumably cancels out noise, because the noise components of the pattern vectors are orthogonal across random partitions.

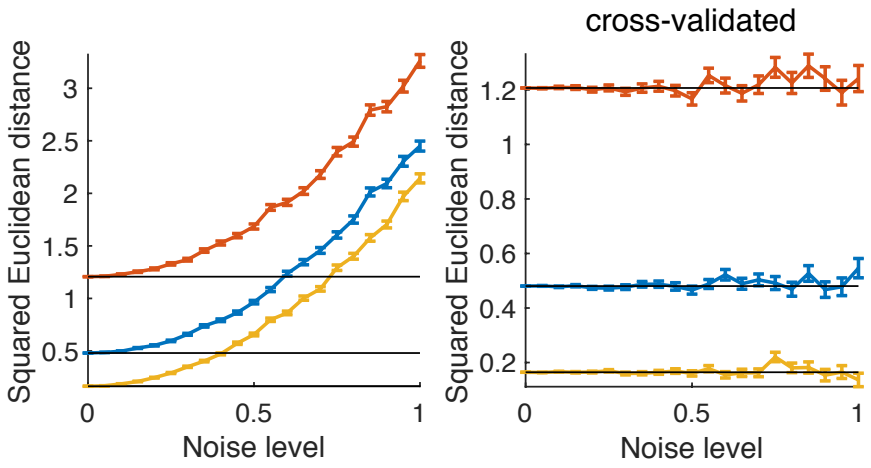


Figure 2.3. Cross-validation recovers true pattern distance from added noise. Simulation specifies three conditions whereby sensory attributes = log([1, 2, 3]), and neural responses are simulated by adding Gaussian noise with different variability levels (SD of Gaussian = 0 to 1 with incremental step of 0.05). The true condition-specific distances (squared Euclidean) are plotted as the horizontal black lines. Non-cross-validated distances are affected by noise and deviate more from the ground truth as the noise level increases (left panel), whereas cross-validation eliminates the effect of noise and preserves the true dissimilarity structure (right panel). Lines (error bars) represent the means (SEMs) across 1,000 simulations.

Multivariate noise normalisation (pre-whitening). It has been well-studied that noise normalisation improves reliability of dissimilarity measures (not only Euclidean distances but also correlation-based distances, such as Pearson distance; see Guggenmos, Sterzer, & Cichy, 2018). The "multivariate" nature of this noise normalisation lies in using a variance-covariance structure to not only suppress sources with high error variance (e.g., some sensors or voxels are characterised by higher intrinsic noise than others), but also to model explicitly the noise covariance between sources (e.g., in MEG the activities of spatially and temporally nearby

sources are likely more correlated than others). As a result, the errors of different measurement sources (e.g., sensors, time, beamformed voxels, etc.) are rendered approximately independent of each other (see also **chapter 3** Appendix II). Of note, this pre-whitening transforms Euclidean distance to Mahalanobis distance. The cross-validation on Mahalanobis distances, also known as the "crossnobis estimator" has proven to be a preferable unbiased measure of representational dissimilarity (Carlin & Kriegeskorte, 2017; Diedrichsen & Kriegeskorte, 2017; Diedrichsen, Provost, & Zareamoghaddam, 2016; Walther et al., 2016).

2.2 Multidimensional scaling with individual weighting (INDSCAL)

As a multivariate technique, multidimensional scaling (MDS) is a fruitful tool for understanding the structural relationships among psychophysical or neural representations elicited by different stimuli or experimental conditions. The fundamental objective of MDS is to unveil relationships among a set of stimuli or conditions by projecting them to a low-dimensional space so that the distances among the stimuli or conditions reflect their relative dissimilarities. In psychology and neuroscience, researchers usually strive to construct an MDS space with 'meaningful' coordinate axes (i.e., dimensions) so that the representational geometry is directly interpretable. For instance, the meaning of an MDS dimension should imply the conveying of an experimental manipulation (e.g., the sensory modality by which a stimulus is registered) or stimulus attribute (e.g., loudness of a sound/brightness of a Gabor grating).

However, a distortion of the underlying representational structure caused by averaging the representational dissimilarity matrices (RDM) of different individual participants, a known issue in psychophysics and cognitive science, has often (if not always) been overlooked in cognitive neuroscience. In a seminal paper by Ashby Maddox, & Lee (1994), titled "*On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model*", a systematic analysis of the problem of collapsing dissimilarity data across individuals was demonstrated through two simulation studies. In the first study, Ashby et al. constructed a set of nine objects differing in two physical dimensions (e.g., size and contrast for visual objects), and then added Gaussian noise to each of these dimensions to simulate a set of noisy psychological spaces of 50 different human observers who were asked to rate the dissimilarity among different pairs of the 9 objects. Next, for each observer, the perceived dissimilarity of each pair of stimuli was modelled using distinct distance models assuming different exponential factors in a Minkowski metric (see Ashby et al. 1994 for details). The researchers finally averaged the dissimilarity data of individual observers and used model recovery to test for the effect of structure distortion as a result of such an averaging procedure. Their results are summarized in the following figure (**Figure 2.4**; data adapted from their Tables 1 and 2). The data averaging across participants elicited a fundamental change of the representational structure in the psychological space, and as a consequence, the correct model generating the data could not be identified as the best fit to the data (**Figure 2.4**). A similar conclusion was obtained from their second study on a set of simulated identification data, whereby Ashby et al. showed strong evidence of the distortion effect of collapsing individual data. The true model ('decision-bound model') that

generated the data was shown to fit the averaged data strikingly worse than a 'similarity-choice model'.

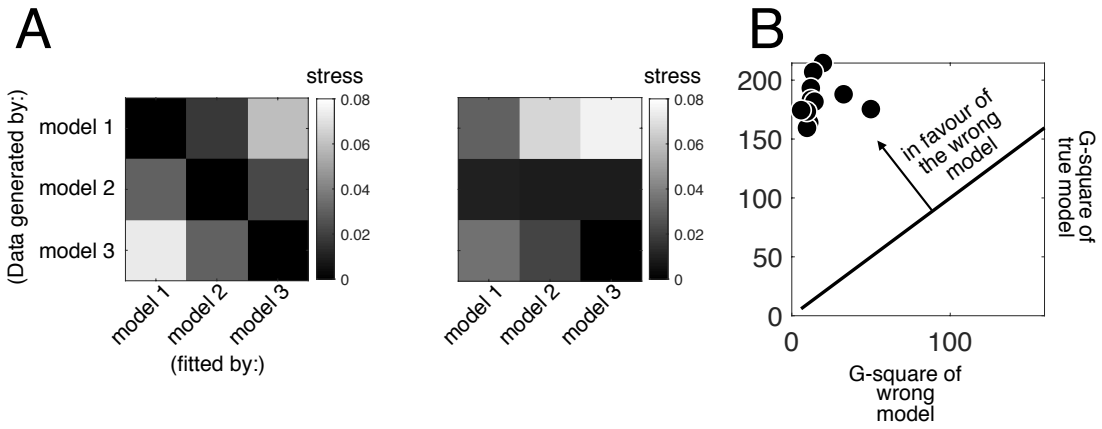


Figure 2.4. Distortion of representational structure by similarity/identification data averaging. (A) Ground-truth mean stress (a measure of goodness-of-fit of a model to data; the smaller the better) obtained by modelling single-participant data (left panel), in comparison with model recovery using group-averaged similarity rating data (right panel). Apparently, the data generated by the true model (particularly model 1 and 2) cannot be best recovered by their respective source models (right panel). (**B**) In a simulation for a stimulus-identification experiment, the true model that actually generated the data ('decision-bound model') turns out to provide a worse fit to the group-averaged data than an alternative wrong model ('similarity-choice model'). G-square was computed as -2 times log-likelihood, thus the lower the better a model fit to data. All figure panels here are produced using the original statistics reported in Tables 1 and 2 of Ashby et al. (1994).

As the paper concluded, "…*The problem apparently is that averaging across subjects changes the underlying psychological structure of the data. Thus, it would be extremely dangerous to fit an MDS model to data that have been averaged across a large group of subjects and then to claim that the resulting MDS solution is a valid representation of the underlying psychological space*."

Individual Differences Scaling (INDSCAL) has been known to address this issue. The key advantage of INDSCAL is expressed as its capacity of minimising the danger of distorting representational structure and its enhancement of the interpretability of the dimensions revealed by the scaling technique.

Here, the basic concept and algorithm underlying INDSCAL are introduced, and an example directly comparing the advantage of INDSCAL using individual dissimilarity data vs. MDS approach to averaged data is discussed. In particular, a visualisation of how INDSCAL recovers the interpretability of MDS dimensions will be presented in this section.

The core concept of INDSCAL concerns explicitly capturing the individual differences (among participants) in the 'weighting' of different psychological or neural dimensions among dissimilarity data. Assuming a set of R dimensions underlying N stimuli, the subjective representation of the distance between the *j*th and *k*th stimuli for the *i*th participant can be derived using the mapping **L** from a weighted Euclidean distance measure d_{jk} to the subjective representation s_{jk}. By taking into account the inter-individual differences in the dimensional weights $w_s^{(i)}$, a weighted version of the distance metric can be computed as follows:

$$d_{jk}^{(i)} = \sqrt{\sum_{s=1}^{r} w_s^{(i)} (x_{js} - x_{ks})^2}$$

Here, INDSCAL modelling is performed in R using the "smacof" package (de Leeuw, & Mair, 2009). In order to demonstrate the extent to which data averaging reduces the interpretability of MDS dimensions and how INDSCAL addresses the issue, I use the neural data of the MEG experiment (**chapter 3**). The dataset contains neural responses to pulsatile audiovisual stimuli (sound flutters combined with visual flickers) with congruent or incongruent pulsing rates across the modalities. For clarity, I focus on comparing INDSCAL against MDS for group-averaged data using the neural data localised to the same brain region for all participants (MNI coordinates: x = 14, y = -84, z = 8) of right primary visual cortex (Calc-R). In this MEG experiment, participants were asked to choose their perceived auditory or visual rate of the

compound audiovisual stimuli from 4 possible rate levels (slowest to fastest). For each participant, a neural RDM was constructed by computing the cross-validated (5-folds across trials) neural distances among the multivariate source-localised MEG responses to different stimuli presented under different experimental conditions (e.g., auditory vs. visual report, different signal-to-noise ratio of the auditory flutter; see **chapter 3** for further details).

As shown in **Figure 2.5A**, the INDSCAL solution (metric; in order to be comparable with **Figure 2.5B**) reveals two dimensional structures for interpreting the neural representation of the audiovisual stimuli in V1. Each symbol represents the neural response to a particular stimulus under a given condition (explained by the colourmap and legends shown on the left). A macro-structure unveiled as four separable clusters is mainly varying along the vertical axis, which can be easily interpreted as the dimension of the visual rate of the stimuli. This makes lots of sense since the neural data are from the early visual cortex. Moreover, within each of the four clusters there is a micro-structure representing the auditory rate of different stimuli, and these neural representations about the auditory rate are all positively correlated with the horizontal dimension revealed by INDSCAL. INDSCAL hence provides two underlying dimensions for the interpretation of the neural representational geometry.

By contrast, applying commonly-used MDS method to group-averaged neural RDMs distorts the micro-structure within each visual-rate cluster: By inspecting **Figure 2.5**, one can notice that dimension 1 (x-axis) is no long interpretable because for some clusters it still positively aligns with the auditory rate, but for some other cluster it becomes negatively correlated with the auditory rate (marked by the ellipses). As shown in **Figure 2.5B**, manipulating different MDS optimisation

criterion (squared stress, Sammon's criterion, or strain) does not change this result. Hence, this reduced interpretability of scaling dimensions is caused by averaging individual neural RDMs.

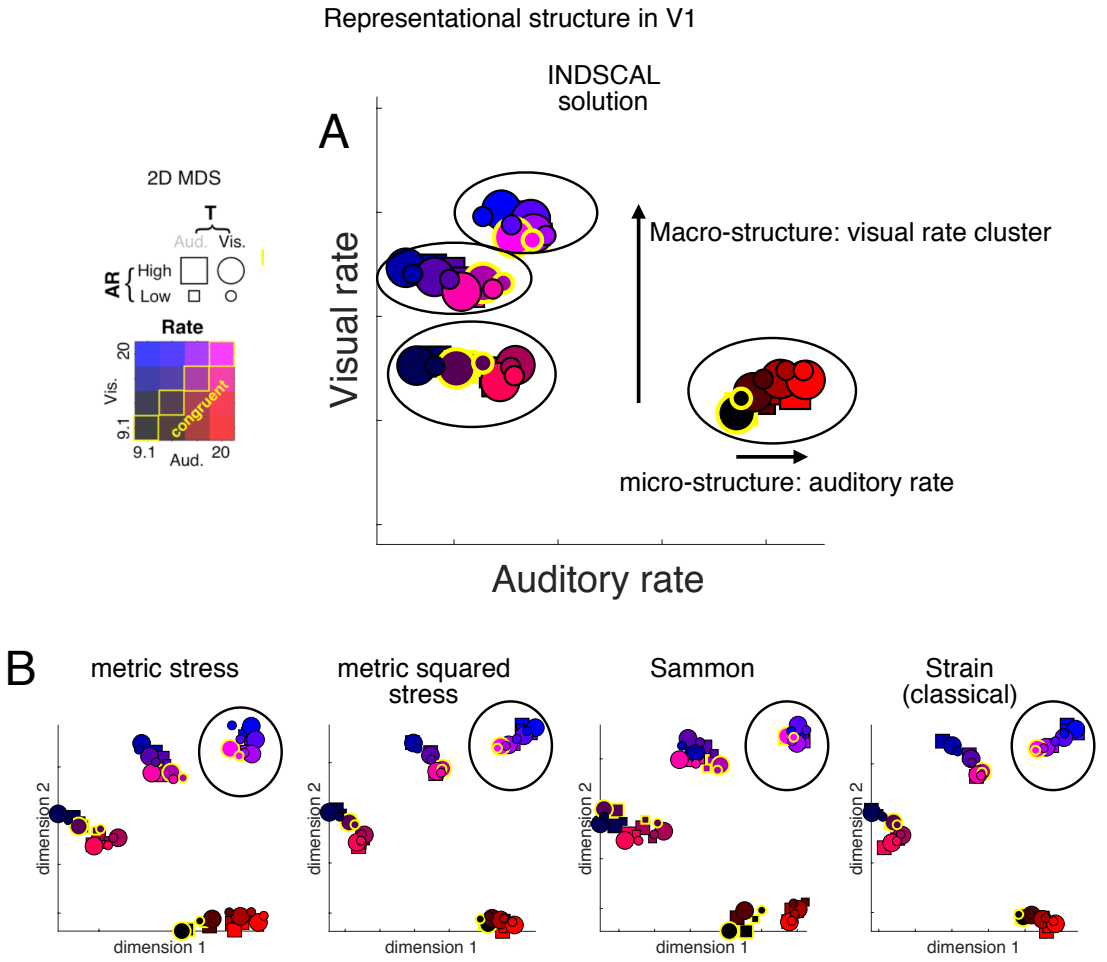


Figure 2.5. INDSCAL supports interpretability of scaling dimensions whereas MDS on group-averaged neural RDMs fails to do so. (**A**) 2D space revealed by INDSCAL approach to MEG neural distances (V1 data). Colourmap indicates the combination of auditory (x-axis) and visual (y-axis) rates for each stimulus under a given experimental condition (squares: 'Task' is to report auditory rate, circles: 'Task' is to report visual rate; big vs. small: high vs. low auditory reliability (high signal-to-noise ratio)). (**B**) MDS (using different goodness-of-fit criterion) on group-averaged neural RDMs. Dimension 1 is not interpretable because for some clusters dimension 1 positively aligns with the auditory rate (see the arrangement of colour on the map in **A**), but for other cluster it becomes negatively correlated with the auditory rate (marked by the ellipses).

Chapter 3. Spatiotemporal unfolding of neural representations underlying flexible multisensory inference

Chapter abstract

The main contribution of this chapter is to describe a model-based magnetoencephalography (MEG) study that seeks to reconcile previous different computational accounts of multisensory decision-making, by combining psychophysics, Bayesian modelling and multivariate MEG techniques. *How does the brain arbitrate between the expediency of fusing multimodal signals (for better precision) and the imperative to perform causal structure inference (for lower bias) in service of optimal behaviour?* This chapter provides a plausible account that bridges the frontal cortex functions, adaptive behaviour and multisensory computations at distinct timescales. When combining information across different senses humans need to flexibly select cues of a common origin whilst avoiding distraction from irrelevant inputs. The brain could solve this challenge using a hierarchical principle, by deriving rapidly a fused sensory estimate for computational expediency and, later and if required, filtering out irrelevant signals based on the inferred sensory cause(s). Analysing time- and source-resolved human magnetoencephalographic data we unveil a systematic spatiotemporal cascade of the relevant computations, starting with early segregated unisensory representations, continuing with sensory fusion in parietal-temporal regions and culminating as causal inference in the frontal lobe. Our results reconcile previous computational accounts of multisensory perception by showing that prefrontal cortex guides flexible integrative behaviour based on candidate representations established in sensory and association

cortices, thereby framing multisensory integration in the generalised context of adaptive behaviour.

3.1 Chapter introduction

We experience the world via multiple sensory modalities. Where information arrives simultaneously in two modalities with differing reliability, the most precise estimates are formed when signals are combined in proportion to their relative reliability. For example, imagine trying to follow a drama on a broken television. If the TV audio is faulty, a viewer should rely more on the picture to follow the narrative, and vice versa. One influential theory suggests that the brains of humans and other animals have evolved to implement this reliability-weighting principle when judging sensory signals (Alais & Burr, 2004; Angelaki et al., 2009; Ernst & Bülthoff, 2004; Raposo et al., 2012). A challenge for the nervous system, however, is that sensory signals should only be fused when they originate from a common source. For instance, if there is a chance that a film is dubbed, combining information about lip movements with speech will render the dialogue difficult to understand. To meet this challenge, the brain must infer the probability that sensory signals share a common cause (or, to continue the example, that the film was dubbed or not). There is evidence from psychophysics that our brain indeed carries out causal inference (CI) to achieve behavioural flexibility during multisensory integration (Körding et al., 2007; Kayser & Shams, 2015). For example, when localising auditory and visual signals we tend to fuse these when they likely originate from nearby sources, but not when they originate from disparate locations, suggesting that the probability of fusion is determined by high-level inference over the probable cause(s) of sensation (De

Corte et al., 2018; Körding et al., 2007; Rohe & Noppeney, 2015a, 2015b; Wozny et al., 2010).

Reliability-weighted fusion and CI have complementary costs and benefits. Fusion may allow rapid inference through frugal computations, for example implemented in feedforward circuits (Ma et al., 2006; Ohshiro et al., 2011, 2017), and serves as a good rule of thumb for many circumstances that give rise to correlated multimodal signals (Parise & Ernst, 2016). CI permits adaptive behaviour, but may be slower and more computationally costly, as it requires inference to be carried out over potential states of the world (Kayser & Shams, 2015). We do not understand how the brain arbitrates between the expediency of fusing multimodal signals and the imperative to perform CI in service of optimal perception. One possibility is that the brain hedges its bets by both computing a rapid fused estimate and, later and where required, inferring the likely cause(s) of multimodal signals. This prediction can be evaluated using time-resolved neuroimaging methods, such as magnetoencephalography (MEG), that are equipped to measure how neural signals unfold during a single decision.

Here, combining a multivariate analysis approach to MEG data with computational modelling of behaviour, we asked where in the brain and when neural signals predicted by models of sensory fusion and CI emerge during perception. In line with past results, we predicted that fused estimates would be emerging rapidly and in parietal-temporal association cortices (Beauchamp et al., 2004a; Boyle et al., 2017; Calvert, 2001; Macaluso & Driver, 2005). We further reasoned that CI would rely at least in part on the frontal cortex, a structure that is thought to subserve causal reasoning and sensory conflict resolution across a wide range of tasks (Donoso et al.,

2014; Giordano et al., 2017; Koechlin & Summerfield, 2007; Noppeney et al., 2010; Tomov et al., 2018).

Using a multisensory rate-categorisation paradigm, we show that flexible multisensory behaviour is best described by a Bayesian causal inference model. We also find that the neural representations of sensory fusion and inference unfold hierarchically in time and across brain regions. This comprises a cascade from early unisensory encoding in primary sensory cortices, to reliability-weighted fusion in parietal-temporal cortices, and to CI primarily in the frontal lobe. We find that neural representations within the dorsomedial and ventrolateral PFC are directly predictive of categorical choices, and that the vlPFC subserves a particular behavioural benefit of inferring sensory causes to minimise perceptual bias in discrepant crossmodal contexts. Our results reconcile previous rival computational models of multisensory integration (see **chapter 1**. General Introduction: **Figure 1.2**), by showing that distinct computational strategies are orchestrated as a temporal sequence and along a parietal-frontal network. These results also suggest that the neurocomputational mechanisms underlying flexible multisensory perception can be understood in a more general framework of causal reasoning that subserves adaptive behaviour in ambiguous sensory environments.

3.2 Methods

3.2.1 Participant details

Sixteen right-handed adults (8 females; aged 19 to 35 years, mean = 24.9, SD = 4.8) participated in this study. All reported normal hearing and vision, were briefed

on the nature and goal of this study, and received financial compensation for their participation. The study was conducted in accordance with the Declaration of Helsinki and was approved by the local ethics committee (College of Science and Engineering, University of Glasgow). Written informed consent was obtained from all participants prior to the study. The head movement of one male participant during MEG acquisition was excessive (see **section 3.2.8 Preprocessing of MEG data**) thus the data were not included in the analysis. All results are reported for an N of 15.

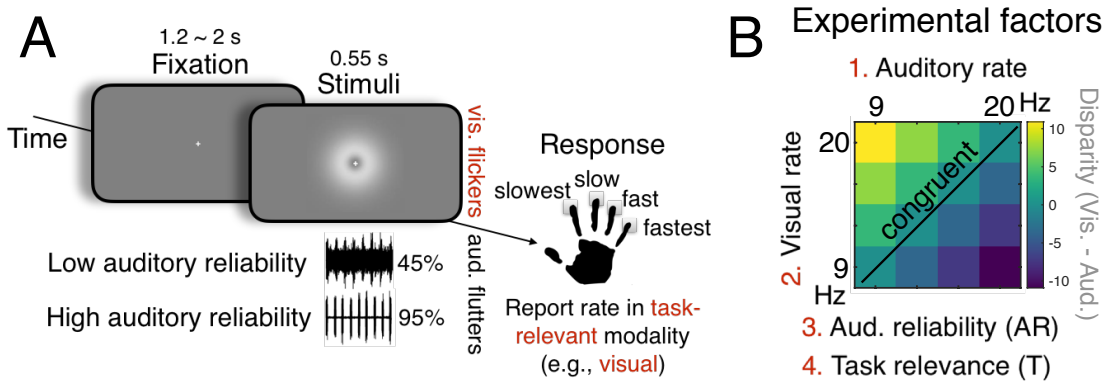


Figure 3.1. Trial structure and experimental manipulations for the audiovisual rate-categorisation task. (**A**) In separate blocks, human participants reported either the auditory or visual rate as task-relevant information in a 4-choice speeded categorisation task. Participants were passively exposed to unisensory stimuli (auditory flutter/visual clicker were presented in separate exposure phases) in order to establish the knowledge of rate categories prior to the main experiment. The panel illustrates the structure of a multisensory trial where the task-relevant modality is vision. (**B**) The experimental design featured 4 orthogonal factors: visual and auditory rates, auditory reliability, and task-relevant modality (visual versus auditory; block design). Signed crossmodal disparity (visual rate minus auditory rate) is visualised using a colourmap.

3.2.2 Task design and stimuli

Participants categorised the temporal rate of stochastic sequences of brief visual and auditory pulses (flicker/flutter; see Roach et al., 2006 for conceptually similar stimuli), presented either in unisensory or multisensory conditions (see **Figure 3.1**). In each block participants were instructed to report either the auditory or visual rate as the task-relevant information. They were asked to respond as accurately and as quickly as possible and to think of the visual and auditory pulses as originating either from independent sources or from a common generative process. Such a design allows for an analysis of participants' flexible strategies concerning when and how to use multisensory cues for a judgement based on their task relevance. The pulse-trains (full duration = 550 ms) consisted of sequential 16.7-ms visual flickers (Gaussian annulus) and auditory flutters (amplitude-modulated white noise) with congruent or incongruent rates (each being one of four possible levels: 9.1, 12.7, 16.4 or 20 Hz, i.e., 5, 7, 9 or 11 events presented in 550 ms). The first/last auditory flutter was always onset-synchronised with the first/last visual flicker, while intermediate pulses were temporally jittered. The inter-event intervals (IEIs; time from one event's offset to the onset of a subsequent event) of a perfectly periodic event sequence were perturbed using a zero-mean Gaussian jitter with a standard deviation equal to the 10% of the periodic-sequence IEI. The jittered visual events were further rounded to the nearest possible video flips and had a minimal IEI of 2 visual frames to avoid flicker fusion (event-onset-asynchrony $\geq$ 50 ms). The jitter was independent for the two modalities rendering the auditory and visual pulses temporally asynchronous regardless of the overall congruency in auditory and visual rates.

The visual stimulus was a difference of two concentric Gaussians (standard deviation of 3^{o} and 1.5^{o} respectively; resulting in a Gaussian *annulus* that made it

possible to present a constant central fixation cross – white, diameter = 0.4°– throughout the entire experiment). The auditory stimuli were noise bursts generated by modulating (square wave) the amplitude of a 550-ms white noise with 3-ms onset-/offset cosine ramp. We manipulated the modulation depth – either 95% or 45% of the peak amplitude – resulting in two levels of auditory reliability. These sound stimuli were then convolved with a head-related transfer function (elevation = 0°, azimuth = 0°; PKU-IOA HRTF database, Qu et al., 2009) to promote perceived co-localisation of the sound with the visual stimuli projected to an external screen. The brightness of the visual flicker was jittered on a trial-by-trial basis to de-correlate the visual contrast from sound intensity. Across trials, the auditory and visual rates, auditory reliability and the task-relevant modality were manipulated factorially in a 4 (visual rates) by 4 (auditory rates) by 2 (auditory reliabilities) by 2 (task relevance) design, resulting in 64 multisensory conditions.

3.2.3 Experimental procedure and stimulus presentation

Each run of MEG data collection (5 minutes) contained all conditions (64 multisensory and 12 unisensory). Within each run the auditory and visual tasks were separated into two task-specific blocks starting with on-screen instructions about the relevant modality for upcoming trials. Unisensory and multisensory conditions were interleaved within each task-specific block. The auditory block contained 32 multisensory trials plus 8 auditory trials (4 rates × 2 reliabilities), whereas the visual block consisted of 32 multisensory trials plus 4 visual trials (4 rates). The order of auditory and visual blocks was counterbalanced across runs. Prior to the experiment, participants were passively exposed to 64 trials of unisensory stimuli (visual and

auditory separately) with repeatedly increasing rates (slowest to fastest, repeated 8 times) and were instructed to memorise the 4 rate categories for use in the following main task. No feedback on correctness or speed was provided during the main experiment.

Stimulus presentation was controlled in Matlab (Mathworks, Inc, Natick, USA) using the Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). Visual stimuli were projected through a DLP projector (Panasonic D7700) onto a non-translucent screen at 1280 × 720 pixels at 60 fps covering a field of view of 25^{o} × 19^{o}. The viewing distance was 180 cm. Acoustic flutters (sampling rate = 48 kHz; depth = 16 bits; 80 dB SPL root-mean-square value, measured with a Brüel & Kjær Type 2205 sound-level meter, A-weighting) were presented through Etymotic ER-30 insert tubephones. We used inverse filtering methods to eliminate the spectral distortion of the sound stimuli induced by the frequency response of the tubephone system (Giordano et al., 2018; for the effect of this filtering technique on the tubephone frequency response function, see supplemental **Figure S3.1**).

3.2.4 Analysis of behavioural data – Model fitting specifications

To estimate the best-fitting model parameters, for each participant we implemented an optimisation search that maximised the log-likelihood of each model given the participant's data. Suppose the counts of the four choices in a given condition are N_i, with $i = \{1, 2, 3, 4\}$. Then the log-likelihood of a model that predicts the response probabilities p_i associated with each of the 4 choices is given by:

(**Eq. 1**): $LL(model\ |\{N_i\}) = \ln C + \sum_{i=1}^{4} N_i \ln p_i$

where C denotes the multinomial coefficient, i.e., $C = (\sum_{i=1}^{4} N_i)!/(\prod_{i=1}^{4} N_i!)$. Maximising **Eq. 1** is equivalent to maximising $\sum_{i=1}^{4} N_i \ln p_i$ because $\ln C$ is a constant. However, $\ln C$ must be estimated for calculating the generalised coefficient of determination (see **Analysis of behavioural data – Model comparison** below). In practice, C can be written as a product of a series of binomial coefficients:

(**Eq. 2**):

$$C = \binom{N_1}{N_1}\binom{N_1+N_2}{N_2}\binom{N_1+N_2+N_3}{N_3}\binom{N_1+N_2+N_3+N_4}{N_4}$$

For large N_i, the logarithm of the factorial function (**Eq. 2**) can be approximated using Stirling's formula (MacKay, 2003):

(**Eq. 3**): $\ln \frac{a!}{b!(a-b)!} \approx b \ln \frac{a}{b} + (a-b) \ln \frac{a}{a-b}$

Parameter estimation relied on the Bayesian Adaptive Direct Search (BADS; Acerbi & Ma, 2017) maximising the sum of log-likelihoods across all 64 multisensory conditions. To avoid local optima, for each model we generated a grid of 500 random parameter guesses as starting values. Monte Carlo sampling (N = 20,000) was used to generate the model-predicted distributions of rate estimates given by **Eqs. 4-11** in **chapter 1**. We obtained discrete categorisation probabilities by binning the continuous distributions to the closest stimulus rate among the 4 levels used in the experiment (similar to Körding et al., 2007; Rohe & Noppeney, 2015b).

3.2.5 Analysis of behavioural data – Model comparison

We used Bayesian random-effects model comparison to determine the model that best explains the data at the group level using both Schwarz's Bayesian Information Criterion (BIC) and the corrected Akaike Information Criterion (AICc) (Rigoux, Stephan, Friston, & Daunizeau, 2014). BIC = $-2LL + k \times \ln(n)$, AICc = $-2LL + 2k + 2k(k+1)/(n-k-1)$, corrected for sample size (Cavanaugh, 1997); where LL denotes the log-likelihood, k the number of free parameters, n the total number of data points, and ln the natural logarithm. Log model evidence was obtained for each participant and model by multiplying the BIC or AICc by -½. We then calculated each model's posterior frequency and protected exceedance probability using the variational Bayesian analysis (VBA) toolbox (Daunizeau, Adam, & Rigoux, 2014; summarised in **Table S3.1**). We also report the models' goodness-of-fit using generalised coefficient of determination R^2 (Nagelkerke, 1991):

(**Eq. 4**): $R^2 = \left\{1 - e^{\left(-\frac{2}{n}\{LL(\hat{\beta}) - LL(0)\}\right)}\right\} / R^2_{max}$

where $LL(\hat{\beta})$ and $LL(0)$ denote the log-likelihood of the fitted and a 'null' model respectively. The null model describes a chance-level observer with response probability = 0.25 over 4 choices. $R^2_{max} = 1 - e^{\left(-\frac{2}{n}\{LL(0)\}\right)}$ is a scaling factor proposed by Nagelkerke (1991).

3.2.6 Analysis of behavioural data – Sensory noise function

We considered that the sensory noise (σ^2) in the above sensory likelihood functions might depend on the rate, r, in a power-law fashion (Nieder & Miller, 2003). To determine the precise nature of this dependency, we compared different

functions describing the rate-dependent noise using the data from unisensory trials. We started from the following general form:

(**Eq. 5**): $\sigma_r^2 = \gamma + p \cdot r^k$

with γ denoting a baseline (i.e., rate independent) noise, r being the rate, p and k a scaling factor and power coefficient, respectively. For practical purposes we re-parameterised the power function as follows [here r_1 (σ_1) and r_4 (σ_4) denote the lowest and highest rates (SD of noise), respectively]:

(**Eq. 6**): $\sigma_r^2 = \sigma_1^2 + (r^k - r_1^k)\,(\sigma_4^2 - \sigma_1^2)/(r_4^k - r_1^k)$

and specifically considered four candidate models describing how parameters change with modality and reliability:

(**Model 1**): parameters being both modality-specific and reliability-specific;

(**Model 2**): parameters being modality-specific but reliability-independent;

(**Model 3**): parameters being both modality-independent and reliability-independent;

(**Model 4**): constant noise across rates, but being both modality-specific and reliability-specific.

We compared these models in capturing the rate dependency of noise using cross-validation on unisensory trials (i.e., partitioning the 22 runs in 5 folds with alternating runs across folds). Model 2 outperformed the others, as demonstrated by an exceedance probability P_{exc} = 0.98 (protected P_{exc} = 0.53) and a mean posterior model probability = 0.47 (SD = 0.056; see **Figure S3.7** for model details).

3.2.7 Neuroimaging data acquisition

MEG data were acquired using a 248-magnetometers whole-head MEG system (MAGNES 3600 WH, 4-D Neuroimaging) at a sampling rate of 1017.25 Hz with participants seated upright. The positions of five coils marking fiducial landmarks on the head of the participants were acquired at the beginning and end of each run. Runs associated with excessive head movements, MEG noise or containing reference-channel jumps were discarded. For each participant we selected the acceptable 22 noise- and artifact-free runs with the smallest average head movement. Across these, the maximum change in head position was on average 3.9 mm (SD = 1.12 mm). To facilitate source analysis, we acquired for each participant the digitised head shape and the locations of the five coils, as well as a whole-brain, high-resolution, structural T1-weighted MP-RAGE image (192 sagittal slices, 256 × 256 matrix size, 1 mm^3 voxel size; Siemens 3T Trio scanner; 32-channel head coil).

3.2.8 Preprocessing of MEG data

All analyses were carried out in Matlab using SPM12 (Wellcome Trust, London), Fieldtrip (Oostenveld et al., 2011) and custom code. Signal preprocessing was initially carried out on unsegmented MEG data from each run. Infrequent SQUID jumps (observed in 2.3% of the channels, on average) were repaired using piecewise cubic polynomial interpolation. For each participant, we then removed those channels that consistently deviated from the median spectrum (shared variance < 25%) on at least 25% of the runs (number of removed channels = 8.4 on average; STD = 2.2). Environmental magnetic noise was removed using regression based on principal components of reference channels. Both the MEG and reference data were then

filtered using a forward-reverse 70 Hz FIR low-pass (-40 dB at 72.5 Hz), a 0.2 Hz elliptic high-pass (-40 dB at 0.1 Hz) and a 50 Hz FIR notch filter (-40 dB at 50 ± 1Hz), and were subsequently re-sampled to 150 Hz. Residual magnetic noise was once more removed using the same regression approach. ECG and EOG artifacts were removed using independent component analysis ("runica" in fieldtrip, 30 components) and were identified based on their time course and topographies (Hipp & Siegel, 2013).

MEG data from each run were then segmented into trials. Given the condition-wise differences in reaction times (**Figure S3.2**), we aligned the MEG trial-wise segmentations not only to stimulus onset (stimulus-locked window = -0.1 to 0.7 s from stimulus onset) but also to trial-by-trial response onset (response-locked window = -0.7 to 0.1 s from response onset). Segmented MEG data were then corrected on a block-by-block basis to avoid motor contamination of the response on the brain activities specific to each condition, which are our primary interests in subsequent analyses. To this purpose, and independently for each run, the motor-related signals were approximated by averaging MEG data across trials of different experimental conditions but having the same button press (N of conditions = 76, ensuring the same finger response could occur in many distinct conditions) and were finally subtracted from the single-trial MEG data.

3.2.9 Reconstruction of MEG sources

For MEG source analysis, we prepared for each participant a native-space grid of 3.5-mm resolution by re-sampling a group-level anatomical template in native space. The group anatomical template was based on Diffeomorphic Anatomical

Registration Through Exponentiated Lie algebra (DARTEL, Ashburner, 2007). A group-level mask was then created by considering non-cerebellum template voxels associated with a grey-matter probability > 0.1, and was back-deformed to the native space of each individual. The native space grid for each individual finally considered 6-connected voxels associated with a participant-specific grey-matter probability > 0.25. Depth-normalised lead fields for each participant were then computed based on a single-shell conductor model and the source-projection filters was derived for each block using a linearly-constrained-minimum-variance (LCMV) beamformer (regularisation = 5%). Data were projected onto the dipole orientation of maximum variance across runs.

3.2.10 Representational similarity analysis of spatiotemporal MEG activity

Representational similarity analysis (RSA) quantifies the statistical association between the pairwise dissimilarity of multivariate brain activity in different conditions (MEG representational dissimilarity matrix RDM) and model RDMs that quantify hypotheses about the nature of the neural representations (here the computational models; Kriegeskorte et al., 2008; Walther et al., 2016). We implemented a whole-brain MEG RSA (Giordano et al., 2018) to assess the local encoding of four different types of rate estimates in the MEG activity as predicted by the candidate models fit to behavioural data. The MEG RDMs were computed in native source space of each participant within a spatiotemporal searchlight (10-mm spatial radius, 80-ms temporal extent, and 40-ms temporal overlap between consecutive searchlights). Importantly, we retained the temporal structure within each searchlight, rather than averaging the MEG signal within it. Searchlight-specific

RDMs quantified the cross-validated Mahalanobis distances between condition-specific spatiotemporal response patterns (Walther et al., 2016).

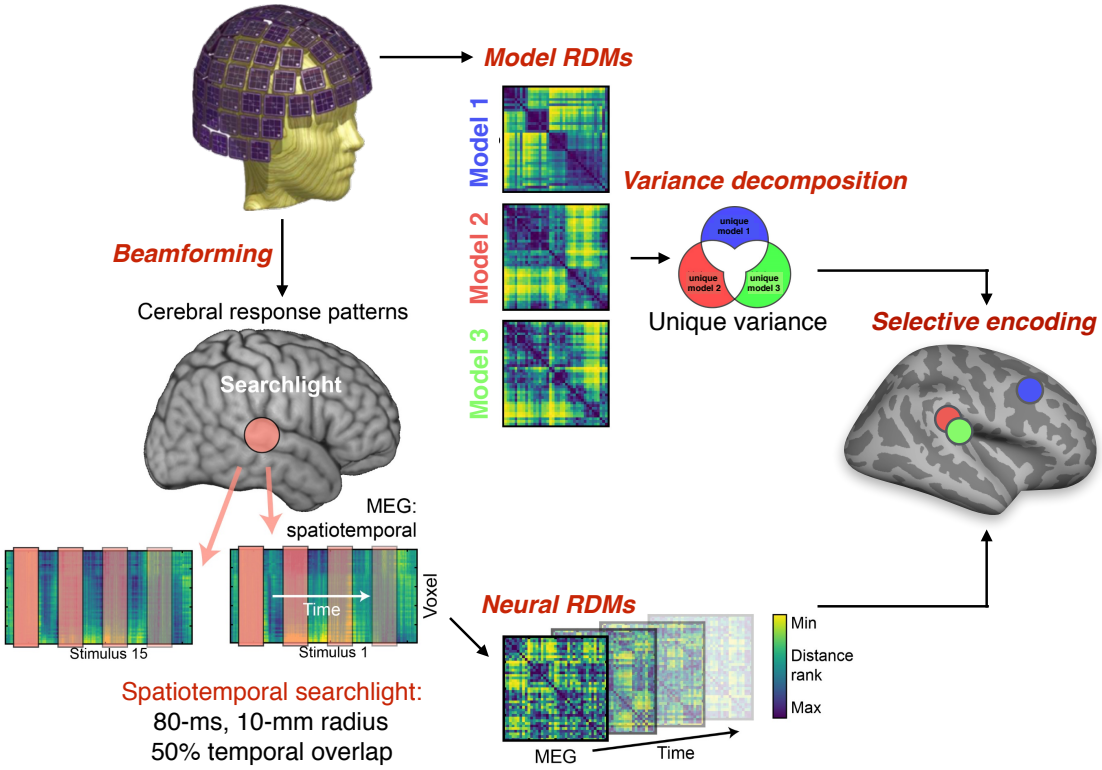


Figure 3.2. A schematic diagram showing the analysis pipeline of MEG spatiotemporal RSA. Multi-channel MEG signals are beamformed to localise the sources of neural activity using individual participant's anatomical MR structure. A whole-brain searchlight (80-ms long, 10-mm radius) captures the multivariate distance between the spatiotemporal (time and beamformed 'voxels') neural activity patterns in response to different experimental conditions (or stimuli). Neural representational dissimilarity matrices are constructed by computing the pairwise multivariate distances across all possible combinations of 2 conditions. RSA provides a systematic approach to quantify the statistical associations between the spatiotemporal neural RDMs and a set of fully orthogonalized model RDMs (which capture the unique variance in each model that is unexplained by any other model RDMs). This reveals the selective encoding of the sensory representations as predicted by each unique model describing a hypothesised computational nature of a given neural representation. This variance decomposition is achieved through deriving the semi-partial rank correlation between the neural RDM and a particular model RDM in the presence of all other model RDMs. A portion of this figure was adapted from Giordano et al. (2018).

More specifically, we: **1.** partitioned the 22 runs into 5 folds (alternating consecutive runs across folds); **2.** whitened within-searchlight data for each fold independently using the residuals of a GLM predicting condition-specific response patterns; and **3.** computed the cross-validated Euclidean distance of condition-specific whitened data based on the covariance between pairs of cross-validation folds (average of covariance between the 10 possible pairs of the 5 folds). To construct model RDMs, we first obtained the model-predicted rate estimates for each condition and participant. For each condition, the model-predicted rate was obtained by averaging 20,000 posterior rate estimates derived using the best-fitting model parameters. The model RDMs were then constructed by taking the pairwise absolute difference between the model-predicted rate estimates for each pair of the 64 conditions (equivalent to the Euclidean distance between condition-specific rate estimates).

We used the Spearman rank correlation between MEG and model RDM to quantify the MEG representation of the model-predicted rate estimates. The RSA correlation maps were computed in native space, Fisher Z transformed, were then transformed to the DARTEL group space (Gaussian smoothing FWHM = 8 mm), and finally assembled in T statistics for group-level inference. Importantly, the rate predictions from the different models and the associated model RDMs were not independent of each other. As such, a significant RSA correlation with a given model RDM is itself not a strong proof of selective representation because it might be the by-product of a correlated model. To address the key question of selective model representation, we used variance partitioning (see e.g., Hebart et al., 2018; Seibold & McPhee, 1979) to assess the neural encoding of the unique variance ("selective encoding", hereafter) of each model after partialling out all other model RDMs using

rank regression (semi-partial correlation). Note that for fusion we partialled out the CI RDM but not the SV and SA RDMs, because fusion by nature is a linear combination of SA and SV, while CI also builds on either SA or SV depending on the task, and hence partialling out all other models would be overly stringent in comparison to the selective encoding tests for the other models (group averaged percent variance retained in the CI, fusion, SA and SV RDMs after partialling out all other model RDMs = 73.7, 80.9%, 75.7 and 80.0%, respectively; SEM ≤ 4%). Note that a focus on semi-partial correlations for establishing selective model encoding does not automatically lead to the observation of segregated networks for the encoding of different models, as multiple predictors can exhibit significant partial correlations (see e.g., spatiotemporal overlap of selective encoding of SA and SV models in temporal cortex, and overlap of SA and CI model encoding in cingulate cortex in **Figure 3.5A**).

Significance testing relied on a permutation-based random-effects (RFX) framework (shuffling condition labels; 10,000 permutations for each test), in conjunction with a spatiotemporal cluster-mass enhancement of the group-level statistics (parametric cluster-forming threshold of T(14) = 1.76 for 1-sided inference, Maris and Oostenveld, 2007). Correction for multiple comparisons relied, for each model RDM independently, on the maximum-statistics across the entire analysis space (see below; FWE = 0.05; 1-sided inference for both correlations and semi-partial correlations). Importantly, while correlations were tested across all RSA time windows and grey-matter DARTEL-space voxels, semi-partial correlations were tested only within a mask of significant correlations. In other words, we adopted a two-step approach that initially assessed the non-selective encoding of a particular model, and subsequently tested for selective encoding (Giordano et al., 2018), which

mitigated false-positives in a strict way. This RSA was carried out twice, once using the data aligned to stimulus onset, and once aligned to the behavioural response onset. Region-of-interests (ROIs) were identified as the DARTEL coordinates of local MEG model-encoding peaks on the spatiotemporal T(s.p.r) maps and are reported using MNI coordinates in **Table 3.1**.

3.3 Results

3.3.1 Task summary

Fifteen human volunteers participated in an audiovisual rate categorisation task (four-choice speeded judgement; **Figure 3.1A**) whilst their brain activity was measured using magnetoencephalography (MEG). In separate blocks participants were instructed to report either the auditory or visual rate as task-relevant information and signalled their response with a button press. Unlike paradigms that focus on sensory fusion by requiring explicit use of both sensory modalities (Ernst & Bülthoff, 2004), our task permits the analysis of individual and flexible strategies in processing multisensory cues based on their task relevance. To quantify how the discrepancy of crossmodal information influences behaviour (Körding et al., 2007), we manipulated visual and auditory rates independently (i.e., they could be either congruent or incongruent across trials; **Figure 3.1B**). To quantify the reliability-dependent influence of one modality onto another (Ernst & Bülthoff, 2004), 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.

3.3.2 Modelling behaviour

We compared the predictions of three classes of models about participants' behaviour. 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. Specifically, we considered: **1.** a model of 'sensory segregation', **2.** a model for reliability-weighted 'sensory fusion' (Ernst & Bülthoff, 2004), and **3.** Bayesian models of multisensory 'causal inference' (Körding et al., 2007; Wozny et al., 2010). See **Figure 1.3** of **chapter 1** General Introduction for a schematic illustrating the differences in sensory estimate given by these 3 classes of computational models.

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 behavioural 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. This model does, however, predict that bias will scale with the reliability of individual cues. 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 that are similar in rate (Körding et al., 2007; Rohe & Noppeney, 2015b; Wozny et al., 2010). This inference model predicts that the bias increases with disparity and relative cue reliabilities (Rohe & Noppeney, 2015a), but

critically, it predicts that the 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). Hence, among these candidates, only the inference model reflects the behavioural flexibility to exploit multisensory information when of benefit, and to otherwise avoid distraction from cues that likely have an independent origin.

3.3.3 Multisensory judgments follow Bayesian causal inference

We determined which candidate model best accounts for participants' behaviour. We maximised each model's log-likelihood of explaining individual participant's responses across all conditions and derived the best-fitting model parameters (**Table S1** for model comparison; **Table S3.2** for parameter estimates). A Bayesian causal inference model formulated with a free probabilistic belief of common cause (p_c) and with a model-averaging decision strategy, "CI (MA)", explained the data better than models not incorporating the inference of latent cause(s) (i.e., segregation and reliability-weighted fusion, **Figure 3.3A**; group-level Bayesian (BIC) and corrected Akaike Information Criterion (AICc) relative to CI (MA) $\geq$ 468 and 547, respectively). Model averaging here refers to a decision function that averages the fused and the task-relevant segregated sensory estimates, each weighted by their inferred probabilities. The CI (MA) also fared best in random-effects model comparison (**Figure 3.3A;** protected exceedance probability P_{exc} = 0.967 and 0.989 using BIC and AICc model evidence, i.e., the probability corrected for chance level that a model is more likely than any others in describing data). Other variants of the CI model with alternative decision strategies ("probability matching"

and "model selection", see **Methods**) provided worse fit than the CI model with model-averaging strategy (protected $P_{exc} < 0.0062$). Across participants, the average coefficient of determination R^2 of this best CI model was 0.87 (SEM = 0.0078; **Figure 3.3B**). Quantifying CI model performance as a function of disparity further emphasised that participants' behaviour was best explained by causal inference for discrepant contexts (**Figure 3.3C**).

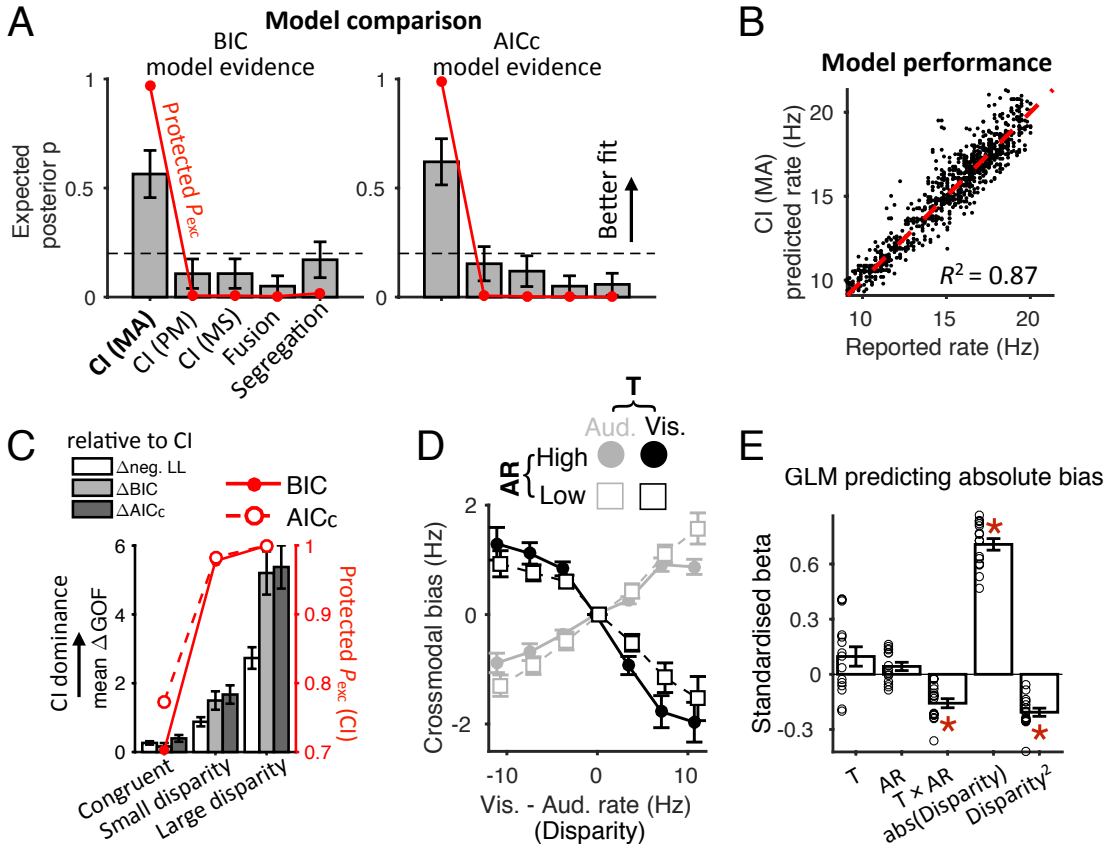


Figure 3.3. Behavioural data analysis. **(A)** Model comparison: CI = Bayesian causal inference, with decision strategy of MA = model averaging, PM = probability matching, or MS = model selection. Fusion: a linear combination of audio-visual information following a belief in a common cause; Segregation: focus on a single modality following a belief in separate causes. Red curves = protected exceedance probability P_{exc} (probability that a given model is more likely than any other models); grey bars = expected posterior probabilities; error bars = standard deviation (SD) of posterior probabilities; dashed lines = chance level (p = 0.2). **(B)** Rate estimates as predicted by the winning CI model in comparison with observed (trial averaged) rates. Each dot represents individual a condition of one participant. R^2 = generalised coefficient of determination, mean across participants. **(C)** Model comparison (CI (MA) vs. Fusion vs. Segregation) as a function of disparity confirms the leading role of CI, specifically at high disparities. CI dominance is measured as the mean

goodness-of-fit (GOF) difference (log-likelihood, BIC, AICc; mean of Fusion - CI and Segregation - CI), and the protected P_{exc} of CI (against Fusion and Segregation, using both BIC and AICc model evidence, same as in **C**). Congruent: disparity = 0 (16 conditions), small disparity = 3.64 Hz (24 conditions), large disparity ≥ 7.27 Hz (24 conditions). **(D)** Crossmodal bias, reflecting the disparity-dependent influence of the task-irrelevant stimulus. Reported rates are grouped by task (T) and auditory reliability (AR). Disparity is signed: visual minus auditory rate. **(E)** Standardised regression coefficients quantifying the influence of task (T; visual task minus auditory task), auditory reliability (AR; low minus high), and the linear and quadratic effect of the absolute disparity on the absolute bias (see **Methods**). Red asterisk *= significant (family-wise error corrected $p \leq 0.05$); error bars = ± 1 SEM of the group average, N = 15; circles = participant-specific betas. See also **Tables S3.1-S3.2** and **Figure S3.2D**.

3.3.4 Context-dependent cue weighting

We further examined why causal inference outperforms the other models in describing the behavioural responses, using an alternative analysis. Specifically, we quantified crossmodal bias, defined as the deviation of participants' response from the actual task-relevant rate (**Figure 3.3D**), 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 and squared disparity (**Figure 3.3E**; all effects were assessed using maximum-statistics permutation controlling for multiple comparisons, family-wise error FWE = 0.05).

We quantified the magnitude of the crossmodal bias as the absolute deviation of the reported rate from the task-relevant rate in each trial (**Figure 3.3D and 3.3E**). To correct for a baseline bias caused by a potential response tendency towards intermediate rates, we adjusted the task-relevant rate using the mean reported rate derived from congruent trials using an established method (Rohe & Noppeney, 2015a). Such a central tendency would make a rate estimate appear to be attracted towards the signal in the other modality (even if it might not be the case), particularly

for conditions with large crossmodal disparities involving the lowest and/or highest rates. However, this baseline bias can be measured and compensated for by replacing the task-relevant physical rate in each condition with the trial-averaged perceived rate in the respective congruent conditions. This adjustment was carried out independently for each participant, and for each task and reliability level. We used a general linear model (GLM) to predict the magnitude of the bias using the experimental factors and specific interactions of these (**Figure 3.3E**):

(Eq. 7): abs(crossmodal bias) $\sim \beta_0 + \beta_1$ **T** $+ \beta_2$ **AR** $+ \beta_3$ **(T** $\times$ **AR)** $+ \beta_4$ **abs(Disparity)** $+ \beta_5$ **Disparity**^{2}

where T = task relevance, AR = auditory reliability. GLM predictors were Z-scored to obtain comparable effect sizes. Significance testing relied on a permutation procedure shuffling the condition labels and we corrected for multiple comparisons using maximum-statistics (10,000 permutations; family-wise error FWE = 0.05 across GLM effects). The results of this analysis were also validated in a separate analysis that considered a more complex GLM with additional higher-order effects (**Figure S3.2D**), or a GLM with additional cubic effect of disparity (resulting for all participants in an increase of BIC relative to the original GLM; average BIC increase = 2.9; SEM = 0.31; see **Figure 3.4** below).

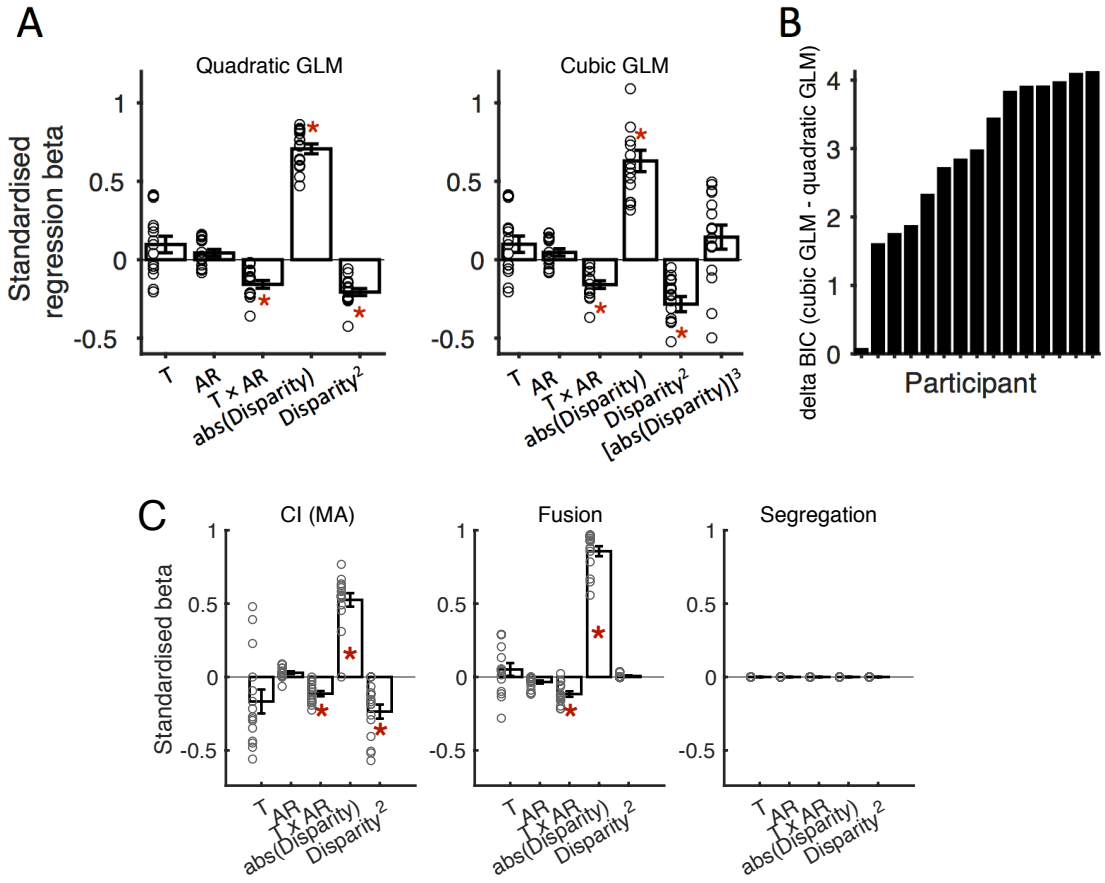


Figure 3.4. Nonlinear dependence of crossmodal bias on disparity. (**A**) Adding a cubic term to the GLM does not change the significance and overall size of the effects (red asterisk = FWE-corrected *p*-value ≤ 0.05). The group-averaged cubic coefficient is not significantly different than 0 (FWE-corrected *p*-value = 0.084). (**B**) For all of the 15 participants adding the cubic term to the GLM results in higher BIC values than for the original quadratic GLM without the cubic term, indicating that the quadratic GLM is preferable (a lower BIC of a model indicates a better goodness-of-fit to data). (**C**) GLMs of abs(crossmodal bias) using rate estimates predicted by each computational model fitted to each participant's behavioural responses. These models correspond to the ones for model comparison in Figure 3.3A and only CI (MA) is shown because the beta effects are very similar to the one shown here if using the other 2 decision strategies (PM and MS). Red asterisk = FWE-corrected *p*-value ≤ 0.05 across all models and all GLM regressors. The T (task) regressor is dummy coded in a way that the auditory task = 0 and visual task = 1, and the AR (aud. reliability) regressor is coded in a way that low AR = 0 and high AR = 1. All regressors are standardised independently using z-score.

The effect of squared disparity captures whether the bias increases nonlinearly with disparity, as predicted by causal inference. A reliability-weighted cue combination is instead 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). Confirming the nonlinear bias suggested by causal inference this GLM revealed a significant effect of squared disparity ($t(14) = -9.28$, mean $\beta = -0.21$, SEM = 0.022). Note that I chose the absolute crossmodal bias as the dependent variable for two reasons. First, this GLM (**Eq. 7**) was implemented mainly to provide an alternative analysis to elucidate why CI outperforms the other models (Fusion and Segregation) in describing the behavioural responses. In particular, by comparing how the GLM effects would have been expected under the three different computational hypotheses we could gain a clear insight about what behavioural characteristics are unique to each model prediction (**Figure 3.4C**). The signature of reliability-dependent crossmodal interaction (T×AR) is present for both the Fusion and CI models, but is absent for the Segregation model. The nonlinear dependency of absolute bias on disparity (negative effect of Disparity^2 on abs(crossmodal bias)) is a unique characteristic of the CI model, but not of others. These insights facilitate an interpretation of the beta coefficients in this GLM applied directly to the model-free absolute bias derived from participants' behavioural reports (**Figure 3.3E**). Crucially, it is important to mention that the regressor of task (T) is included in the GLM mainly because the interaction between T and AR is included to test the effect of reliability-dependent cue weighting (a main effect needs to be added to measure the interaction reliably from a statistical perspective). The task effect turns out to be non-significant not only in the GLMs

based on model-predicted biases (**Figure 3.4C**), but also in the GLM for behavioural biases (**Figure 3.3E**). Indeed, a task effect should have been non-trivial if one unisensory cue strongly dominated over another unisensory cue, resulting in a strong imbalance of the absolute bias measure in different modality-specific tasks. However, this is not seen in the present data, which is probably due to the fact that in pilot experiment the sensory reliabilities were calibrated individually such that the visual reliability roughly lies between the two levels of auditory reliability). Second, using absolute bias as the dependent variable enables methodological consistency throughout the analyses in the present study: neural analysis uses cerebral measures of 'absolutely' bias, which can be extracted as 'distance' from a neural representational dissimilarity matrix (RDM) directly. As can be seen in **section 3.3.6** of this chapter, the bias will be derived as the representational distance (i.e., values in relevant MEG RDM cells) between each of the 64 conditions on one hand (e.g., a condition with auditory rate = 9.1 Hz and visual rate = 16.4 Hz in auditory task), and the congruent condition with the same task-relevant rate on the other hand. In sum, using absolute bias as dependent variable differentiates computational hypotheses and enables a parsimonious analytic approach to neural data within the RSA framework.

As expected, reaction times (RTs) also varied systematically with the experimental manipulations (**Figure S3.2A and S3.2B**). In particular, participants' responses were generally slower (yet increasing nonlinearly) for larger disparity, showing the additional effort required to make judgements facing disparate multisensory evidence. In addition, the RTs in *congruent* multisensory trials were significantly faster than those in unisensory trials with the same rate and task ($t(14) = -2.6$, $p = 0.021$; 2-sided paired-sample t-test), confirming a general facilitation of responses by multisensory congruency. Of note, the RTs in conflicting multisensory

trials were *not* faster than the RTs in unisensory trials. In fact, the RTs in conflicting multisensory conditions increased as the crossmodal disparity increased (**Figure S3.2 E rightmost panel**). These results indicate that the RT facilitation (vs. slowing) by multimodal inputs was dependent on crossmodal congruency (vs. conflict). From the perspective of sensory evidence accumulation, the RT facilitation (relative to unisensory baseline) by congruent audiovisual inputs can be understood as an enhanced drift-rate coefficient in the combined condition (k_{av}), which is computationally equivalent to the square root of the sum of the squares of the drift rates (k_a for auditory and k_v for visual) of the two unisensory evidence accumulation processes ($k_{av}^2 = k_a^2 + k_v^2$, Drugowitsch et al., 2014). However, this optimal multisensory facilitation is achieved if and only if one assumes that the posterior probability of common cause between audiovisual inputs is equal to 1 (see Drugowitsch et al., 2014 for a detailed mathematical proof for optimality). Now consider a Bayesian-like gain weighting between the forcedly combined sensory evidence and the task-relevant unisensory evidence according to the momentary posterior probability of common cause computed at each time step *t*. This allows me to model the combined evidence accumulation by a unified drift-diffusion framework:

$$\dot{x} = \dot{x_{av}} p\left(c = 1 | \vec{X}_{vis,1\ldots t}, \vec{X}_{aud,1\ldots t}\right) + \dot{x}_{task-relevant}[1 - p\left(c = 1 | \vec{X}_{vis,1\ldots t}, \vec{X}_{aud,1\ldots t}\right)]$$

Where $\dot{x}$ is the momentary sensory evidence whose cumulative sum over time guides the final choice behaviour, and the dot symbol above *x* indicates the derivative r.w.t time. $\dot{x_{av}}$ is the momentary sensory evidence that corresponds to a forced fusion scenario given absolute belief about common cause between audiovisual inputs:

$$\dot{x_{av}} = \sqrt{\frac{k_v^2}{k_a^2 + k_v^2}}\dot{x_v} + \sqrt{\frac{k_a^2}{k_a^2 + k_v^2}}\dot{x_a}$$

Where the unisensory evidence $\dot{x_v} = k_v\, r_v + \eta(t)$ and $\dot{x_a} = k_a\, r_a + \eta(t)$, with k represents drift rate, and $\eta(t)$ is a unit variance Gaussian white noise process. And the posterior probability of common cause at time *t* can be derived via a recursive Bayesian updating of the posterior probability at time *t*-1 by a momentary sensory likelihood function $p\left(x_{vis,t}, x_{aud,t} \middle| c = 1\right)$:

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

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. The accumulated evidence $X(T) = \int_0^T \dot{x} dt$ is then fed to a sigmoidal decision function $1/(1 + exp(-bX(T)))$ to derive the posterior probability of reporting a high rate (i.e., above the mean of the four rate levels ranging from 9.1 Hz to 20 Hz), where *b* controls the slope of the sigmoidal curve and is set to 0.1 in the current simulation (**Figure S3.2E leftmost panel**). In the simulation for this model, I set the task-relevant visual rate (in a visual task) to 9.1 Hz, and the task-irrelevant auditory rate varying from a congruent rate (9.1 Hz) to an increasingly incongruent rate (12.7, 13.6 to 20 Hz). If we assume that this Bayesian agent makes a decision about reporting a rate percept at the time when the trajectory of the posterior probability passes a threshold, the RT can be derived as the time-step at which the probability trajectory cuts through the threshold. In comparison to the unisensory visual condition (V-only; shown by the green trajectory), the simulated RT profile (**Figure S3.2 middle**) qualitatively replicates the RT profile as a function of audiovisual disparity in the empirical data (**Figure S3.2 rightmost**). In conclusion,

this simulation using a Bayesian drift-diffusion model provides a computational interpretation of the relationships between unisensory and multisensory RTs, and explains why RT facilitation and slowing can occur concurrently given different combinations of multimodal inputs.

3.3.5 MEG data reveal a spatiotemporal hierarchy of multisensory representations

Next, we investigated when and where the brain represents sensory information in a manner as predicted by each computational model. Specifically, we asked whether neural representations of different candidate computations emerge simultaneously or sequentially and possibly within the same or distinct brain regions. We adopted a multivariate approach using cross-validated representational similarity analysis (RSA, Walther et al., 2016). RSA assesses neural representations by quantifying the statistical association between two dissimilarity matrices: **1.** an MEG representational dissimilarity matrix (RDM) that quantifies the pairwise dissimilarities of brain activity in response to different experimental conditions and **2.** a model RDM that quantifies the hypothesised computational nature of brain representations. The model RDMs were constructed based on the rate estimates predicted by each model fitted to the behavioural responses of each participant (**Figure 3.5A** illustrates the group-averaged model RDMs across participants**)**. Given the condition-wise differences in reaction times (**Figure S3.2**), we carried out separate RSAs by aligning the data to stimulus onset and to trial-by-trial response time.

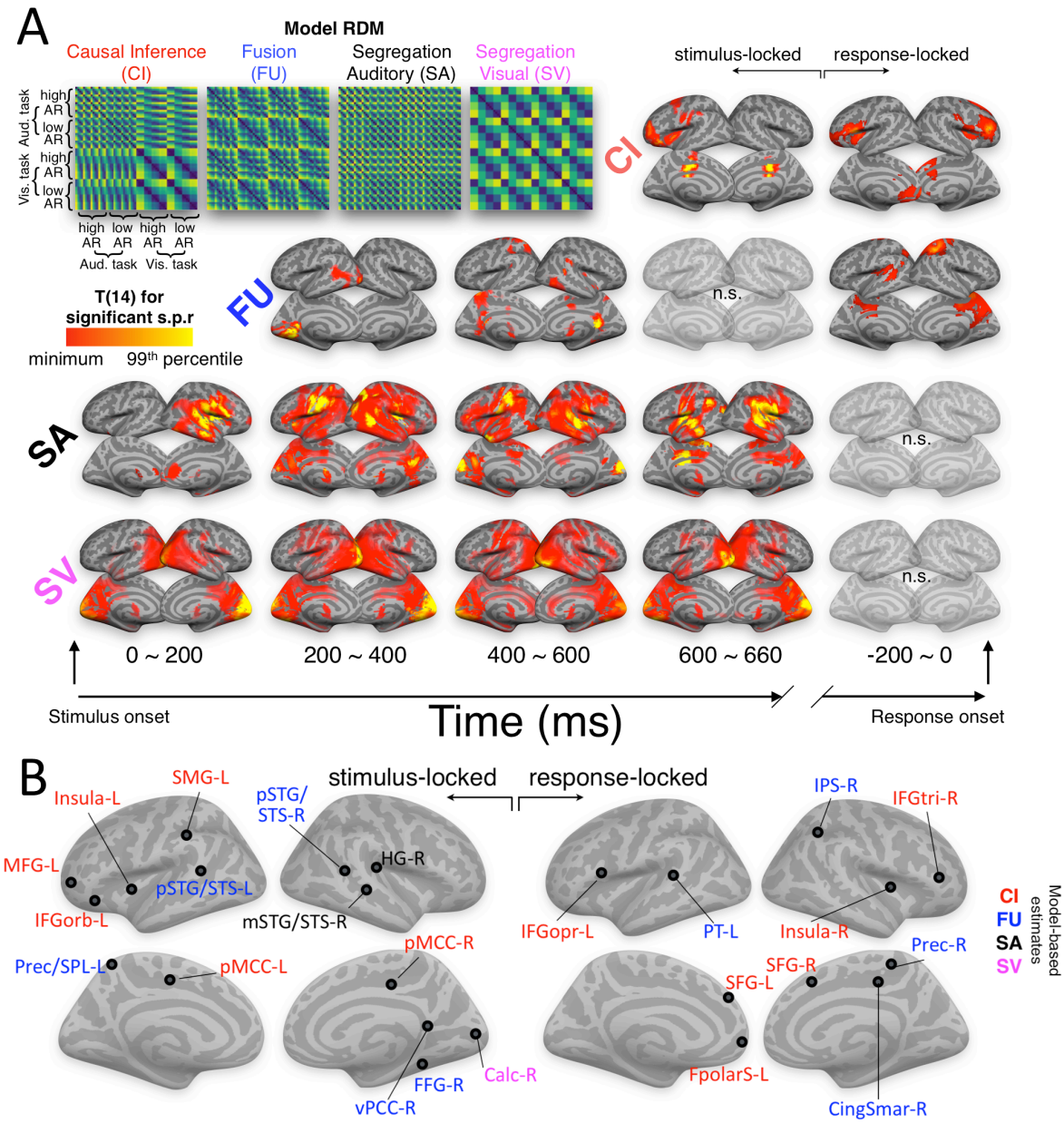


Figure 3.5. Spatiotemporal evolution of multisensory representations. Evolution of sensory representations as revealed by model-based representational similarity analysis (RSA). **(A)** Upper left: group-averaged model representational dissimilarity matrices (RDMs) visualising the ranked distance between model-predictions of condition-wise event rates. The other figures display the cortical surface projection onto FreeSurfer template of the T-maps of group-level significant MEG variance uniquely explained by each model (semi-partial correlation, s.p.r) in stimulus-locked and response-locked data. All effects were significant at $p \leq 0.05$ FWE-corrected (see **Methods**). Colour-coded is the peak T-value of each voxel across time within the respective epoch for visualisation purpose. Light empty surfaces denote non-significant (n.s.) effects. **(B)** Anatomical locations of local and global peaks of RSA effect, colour-coded by the respective significant model (see also **Table 3.1** and **Figure S3.6)**. See also **Figures S2.3** and **S2.4**.

Table 3.1: Selective encoding of candidate models in MEG source activity. The table lists global (*) and local selective-encoding peaks (RSA effects; Montreal Neurological Institute (MNI) coordinates). Anatomical labels based on the Automated Anatomical Labelling (AAL) and Destrieux atlases. s.p.r = group-averaged semi-partial rank correlation between MEG and model RDM (smoothed native-space s.p.r maps normalised to Diffeomorphic Anatomical Registration Through Exponentiated Lie Algebra, i.e., DARTEL template); SEM = standard error of the mean. All effects significant at $p \leq 0.05$ FWE corrected (see **Methods**). L = left hemisphere, R = right hemisphere. See also **Figure S3.3**.

Analysis: Model	Anatomical label	Latency	T(14)	s.p.r (SEM)	MNI coordinates		
Stimulus locked: Segregation Visual (SV)	Calc-R*	140 ms	18.96	0.126 (0.007)	14	-84	8
Stimulus locked: Segregation Auditory (SA)	mSTG/STS-R	180 ms	3.08	0.034 (0.011)	48	-32	-2
	HG-R*	220 ms	4.34	0.026 (0.006)	34	-28	12
Stimulus locked: Fusion (FU)	pSTG/STS-L	220 ms	2.94	0.021(0.007)	-60	-50	8
	Prec/SPL-L	500 ms	3.42	0.02 (0.006)	-10	-56	62
	pSTG/STS-R	500 ms	3.33	0.013 (0.004)	46	-42	16
	vPCC-R*	540 ms	4.29	0.02 (0.005)	10	-46	6
	FFG-R	540 ms	2.96	0.02 (0.007)	34	-46	-20
Stimulus locked: Causal Inference (CI)	pMCC-L*	620 ms	4.26	0.014 (0.003)	-10	-14	44
	pMCC-R	620 ms	4.03	0.018 (0.004)	10	-22	36
	IFGorb-L	620 ms	3.44	0.018 (0.005)	-42	24	-8
	Insula-L	620 ms	3.39	0.024 (0.007)	-38	0	2
	MFG-L	620 ms	3.25	0.017 (0.005)	-42	48	2
	SMG-L	620 ms	2.73	0.012 (0.004)	-56	-32	40
Response locked: FU	IPS-R*	-180 ms	5.14	0.036 (0.007)	34	-52	44
	Prec-R	-180 ms	5.05	0.037 (0.007)	4	-38	54
	CingSmar-R	-180 ms	4.39	0.018 (0.004)	14	-28	36
	PT-L	-180 ms	3.22	0.025 (0.008)	-50	-38	8
Response locked: CI	IFGopr-L*	-180 ms	4.96	0.022 (0.005)	-42	10	8
	FpolarS-L	-180 ms	4.55	0.023 (0.005)	-14	60	-12
	IFGtri-R	-180 ms	4.54	0.018 (0.004)	34	32	12

	Insula-R	-180 ms	3.65	0.015 (0.004)	32	0	16
	SFG-R	-140 ms	3.45	0.019 (0.005)	14	20	48
	SFG-L	-140 ms	3.01	0.01 (0.003)	-4	46	36

Note: HG: Heschl's gyrus; mSTG/STS: middle superior temporal gyrus/sulcus; pSTG/STS: posterior superior temporal gyrus/sulcus; Calc: Calcarine; Prec: Precuneus; SPL: Superior parietal lobule; STGlat: Lateral superior temporal gyrus**;** FFG: Fusiform gyrus; vPCC: Posterior-ventral cingulate gyrus; SMG: Supramarginal gyrus; pMCC: middle-posterior cingulate gyrus and sulcus; IFGorb: Inferior frontal gyrus (pars orbitalis); MFG: Middle frontal gyrus. IPS: Intraparietal sulcus; PT: Planum temporale; CingSmar: Marginal branch of the cingulate sulcus; IFGopr: Inferior frontal gyrus (pars opercularis); FpolarS: Transverse frontopolar sulcus; IFGtri: Inferior frontal gyrus (pars triangularis); SFG: Superior frontal gyrus.

The results of the stimulus-locked RSA revealed a systematic and gradual progression of neural representations from segregated unisensory representations to reliability-weighted fusion, to CI across cortical space and time (**Figure 3.5A**; **Table 3.1**). We quantified the selective representation of each model (e.g., CI) that is not already explained by the representations of other models (fusion and segregation) using semi-partial correlation (s.p.r). The earliest MEG activations significantly reflecting one of the candidate computations (''RSA effects,'' hereafter) were those pertaining to segregated unisensory estimates starting around 100 ms after the stimulus onset. These were localized within the respective sensory cortices (bilateral calcarine cortex starting from ~ 100 ms for segregated visual representations; auditory cortex starting from ~ 140 ms for segregated auditory representations). Subsequently, the MEG activity began to reflect sensory representations formed by reliability-weighted fusion, with significant clusters emerging around 180 ms to 260 ms in the left superior temporal gyrus and later the precuneus and superior parietal lobule, the ventral posterior cingulate, and the posterior superior temporal gyrus. Finally, MEG activity reflecting representations as predicted by CI emerged around 620 ms in dorso- and ventrolateral prefrontal cortices (the left inferior frontal cortex,

in particular), frontopolar and insular cortices, and the middle-posterior cingulate cortex.

The results of the response-locked RSA also suggested a difference between parietal and frontal regions in reflecting fusion and CI. This revealed an RSA effect of fusion within the parietal-temporal lobe (right precuneus and IPS around 220 ms to 140 ms prior to response onset; **Figure 3.5A**; **Table 3.1**) but CI within the frontal lobe (bilateral IFG and frontal pole around 220 ms to 140 ms; bilateral superior frontal gyri around 180 ms to 100 ms). Encoding of segregation models was no longer found in this response-locked analysis.

We carried out control analyses to rule out alternative explanations for the coexisting representations of fused and inference representations. First, the significant RSA effects reported above were contributed by the majority of the participants (**Figure S3.3),** and across these the two effects were uncorrelated (FWE-corrected $p > 0.73$ for stimulus-locked ROIs, and $p > 0.66$ for response-locked ROIs; left-sided maximum-statistics permutation tests), which would not be the case if some participants' MEG activity was explained by fusion and that of others by causal inference. Together with the fact that none of the participants' behaviour was better described by fusion than CI (**Table S3.1)**, this suggests that the coexistence of distinct neural representations is unlikely just a by-product of pooling neural data from different participants whose behaviours were best fit by either model.

To further substantiate the observation that the representations of fusion precede those of causal inference (CI) we explicitly characterised the temporal sequence of the peak RSA effects for fusion and CI. We extracted the latency of the peak T statistics of each RSA effect (**Figure S3.4**) and derived a latency contrast for

each pair of the respective ROIs (latency of fusion minus latency of CI). Across the 30 pairs of stimulus-locked ROIs, the peak RSA effects of fusion emerged significantly earlier than those of CI (two-sided one-sample t-test against zero: $t(29) = -7.19$, $p < 10^{-7}$, 95% bias-corrected and accelerated [BCa] bootstrap confidence interval of the mean [-0.21, -0.12] s; **Figure S3.4C-D**). This result was also confirmed for the response-locked ROIs ($t(23) = -3.24$, $p = 0.004$, 95% BCa bootstrap confidence interval of the mean [-0.023, -0.006] s; **Figure S3.4E-F**).

3.3.6 Distinct markers of sensory integration in posterior and anterior brain regions

We characterised the computational properties of the neural representations revealed by the RSA in further detail. Note that these neural representations are defined both by their location (ROI) and time point of the respective RSA effect (**Table 3.1**). First, we quantified the dependency of the crossmodal bias captured by these representations on the two key features of context-dependent integration: **1.** the interaction between task relevance and sensory reliability, and **2.** the quadratic effect of disparity. In line with the predictions of fused and inference (CI) models, we expected fusion regions to exhibit an interaction between task and reliability, and a positive linear dependency of bias on disparity. By contrast, CI regions should exhibit a nonlinear relationship between bias and disparity. For this analysis, we focused on computation-diagnostic neural representations that potentially contain these key markers.

We implemented a GLM to provide an in-depth understanding of the computational properties of the representations in the ROIs unveiled by the RSA. We

focused on the variance of the brain RDMs reflecting the representation of the various candidate computational models. These computation-diagnostic brain RDMs were derived for each ROI (the closest grid point to the ROI's MNI coordinate deformed to native space) within a cross-validated rank regression framework that prevented over-fitting whilst not favouring *a priori* any particular computational model:

(Eq. 8): MEG RDM = β_0 + β_1 (CI RDM) + β_2 (FU RDM) + β_3 (SA RDM) + β_4 (SV RDM)

where the model RDMs: causal inference = CI, fusion = FU, segregation auditory = SA, segregation visual = SV. We used a leave-one-participant-out cross-validation scheme for estimating the computation-diagnostic RDMs specific to each individual. That is, we derived the betas for predicting in one single GLM the MEG RDMs of all but one participant (independent ranking of data from each participant), and then used these betas to estimate the computation-diagnostic MEG RDM of the left-out participant (using this participant's own model RDMs).

An initial analysis aimed to characterise the dependency of a measure of crossmodal bias derived from the computation-diagnostic MEG RDMs on two key features of flexible multisensory integration: **1.** the interaction of task relevance with sensory reliability, and **2.** the quadratic effect of disparity. The modelling approach mirrored the analysis of behavioural data (**Figure 3.3E**) to predict bias within a rank GLM using, as regressors, task, auditory reliability, the task × reliability interaction, disparity and the squared disparity (**Eq. 7** in this chapter). Here, the cerebral measures of bias and disparity were extracted from the MEG RDMs. The bias was derived as the representational distance (i.e., relevant MEG RDM cells) between each of the 64 conditions on one hand (e.g., a condition with auditory rate = 9.1 Hz and visual rate = 16.4 Hz in auditory task), and the congruent condition with the same task-relevant

rate on the other hand (in this example: 9.1 Hz for both auditory and visual rates; still in auditory task). The disparity for each of the 64 conditions was derived as the representational distance between the two respective congruent conditions with the same task-relevant vs. task-irrelevant rate (in this example: the two congruent conditions with rates of 9.1 and 16.4 Hz respectively; still in auditory task). Permutation-based RFX inference was used for significance testing with correction for multiple comparisons separately for stimulus- and response-locked ROIs (FWE = 0.05; **Figures 2.6A** and **Figure S3.5**).

We also visualised the representational geometries in each ROI. We used inter-individual difference multidimensional scaling (INDSCAL; de Leeuw and Mair, 2011; Ashby et al., 1994; **Figures 2.6C** and **Figure S3.6A-B**) to avoid a known distortion of multidimensional-scaling solutions caused by averaging distances across individuals (Ashby et al., 1994; see **chapter 2**). Here, we used INDSCAL models with different dimensionalities to visualise the influence of different experimental factors upon the neural representations: two-dimensional models for the influence of auditory and visual rates and of their discrepancy, and three-dimensional models for the influence of task and sensory reliability.

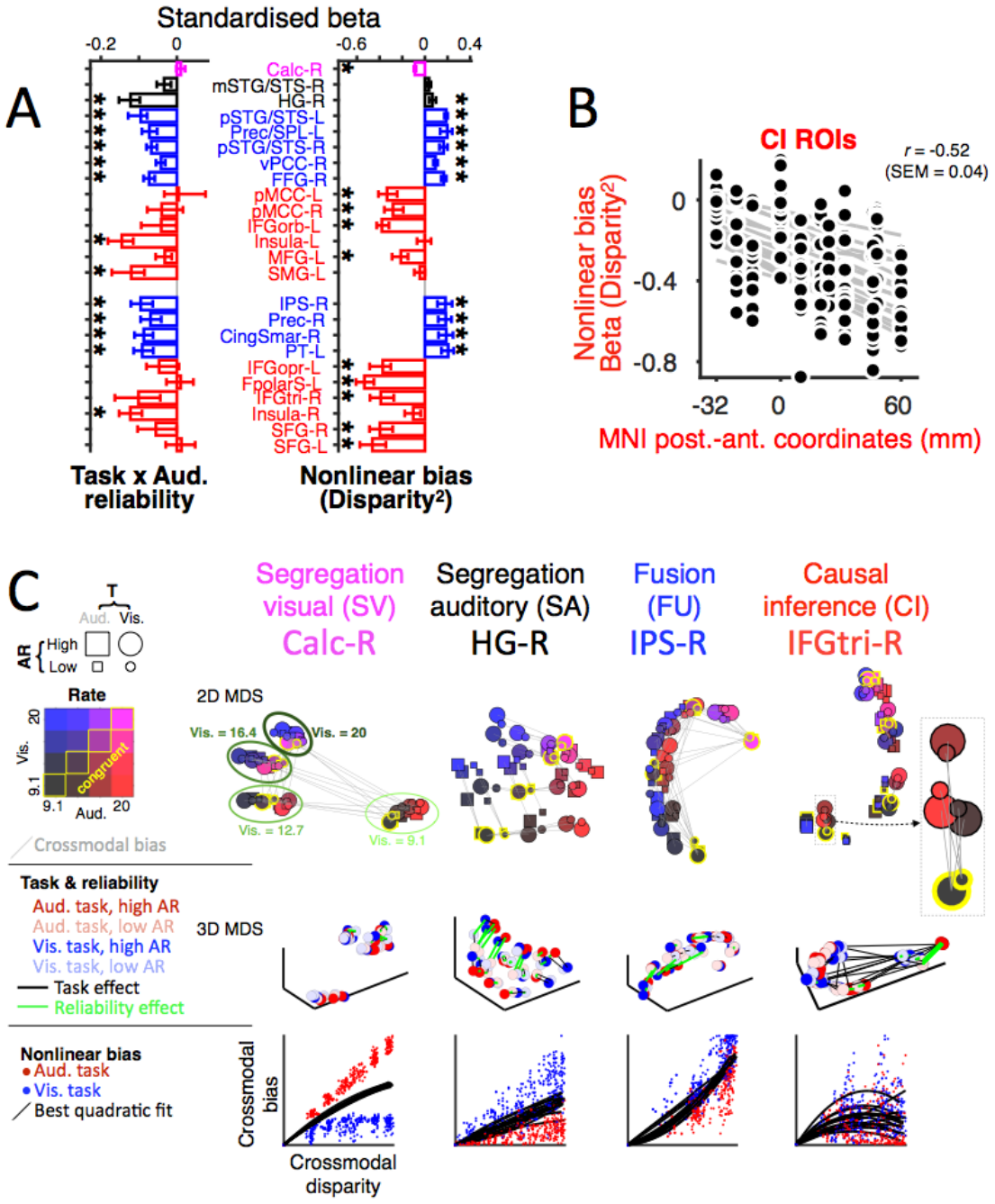


Figure 3.6. Functional and anatomical characterisation of the neural representations of interest (ROI). (**A**) General linear modelling (GLM) of the influence of the task 3 reliability interaction (signature of reliability-weighted fusion) and the quadratic disparity (signature of CI) on crossmodal bias. GLM was applied to the computation- diagnostic RDMs from each ROI (see Methods). Bar height and error bars, mean and 95% bootstrap confidence interval of the standardised GLM coefficients across participants (cf. Figure S3.5 for all ROIs; asterisk, $p \leq 0.05$ FWE-corrected; Methods). (**B**) Negative correlation between anatomical MNI posterior-anterior coordinates and GLM betas for quadratic disparity across CI ROIs. Dots, individual participants for each ROI; grey lines, linear fit for each participant. (**C**) Inter-individual difference multidimensional scaling (MDS) was used to visualize the representational geometry of neural RDMs. Experimental conditions that are closer in

MDS space evoke similar computation-diagnostic MEG responses (Methods). For simplicity, we focus on one region for each model (see Figure S3.6 for all ROIs). The two-dimensional (2D) MDS models emphasise the computational relevance of the two main dimensions: auditory (horizontal) and visual rate (vertical); grey lines indicate the neural distances reflecting crossmodal bias. Shaded green ellipse highlight the 4 primary clusters separating conditions associated with the 4 visual rates in the 2D MDS of Calc-R. Zoomed cluster in 2D MDS of IFGtri-R illustrates disparity-dependent nonlinear bias: representations at higher crossmodal disparity (red circles) being pulled toward the task-relevant rate (yellow contours). The three-dimensional (3D) MDS models emphasize the effects of task relevance and auditory reliability (colour coded). The scatterplots visualize the neurally encoded disparity against crossmodal bias. See also Figures S2.5 and S2.6.

As expected, a significant interaction between task and reliability in shaping bias emerged in all parietal and posterior-temporal fusion ROIs ($t(14) \leq -4.84$, FWE-corrected $p < 0.012$). We also found this interaction in the primary auditory cortex (HG-R, $t(14) = -8.34$, $p < 0.0001$) and some ROIs exhibiting RSA CI effects (SMG and insula; $t(14) \leq -5.45$, $p < 0.0036$; **Figure 3.6A left**). By contrast, the nonlinear dependency of bias on disparity characteristically distinguished fusion and CI ROIs, as only frontal CI regions (e.g., Fpolar and IFGtri) exhibited such well-pronounced dependency ($t(14) \leq -5.74$, $p < 0.0022$ for the 9 significant CI ROIs in **Figure 3.6A right**).

These results suggest a posterior-to-anterior gradient of the influence of multisensory disparity (**Figure 3.6B**). To support this, we correlated the GLM coefficients of the squared disparity with the MNI posterior-anterior coordinates of the CI ROIs. For both stimulus- and response-locked ROIs, this correlation was significant ($t(14) \leq -6.62$, $p < 10^{-5}$), with across-participants mean (SEM) of -0.34 (0.05) and -0.68 (0.06) respectively. In line with an asymmetry between the primary visual and auditory cortex (Rohe & Noppeney, 2016), we also observed a significant

negative quadratic influence of disparity in occipital (Calc-R in **Figure 3.6A,** $t(14) = -17.6$, $p = 10^{-4}$; also see **Figure 3.6C**) but not temporal regions.

In a second analysis, we modelled the representational geometries of computation-diagnostic representations using multi-dimensional scaling (MDS). This projects the neural representations onto a few axes that capture the dimensions along which the candidate representations differ (de Leeuw & Mair, 2011; Ashby et al., 1994). Exemplar ROIs for each computational model are shown in **Figures 2.6C** (for all ROIs see **Figure S3.6**). In early visual cortex (Calc) the representation varied primarily with visual rate, irrespective of task and reliability. By contrast, in early auditory regions (HG) representations were modulated by auditory reliability, by task, and revealed an influence from the visual modality, resulting in a seemingly bimodal representation (2D MDS). In ROIs reflecting fusion (e.g., IPS-R) the representational geometry collapsed the rates across modalities and disparities into a single dimension, but changed with reliability (green lines in 3D MDS). Finally, in ROIs reflecting causal inference (e.g., IFGtri-R) the representational geometry varied with all three factors (disparity, reliability and task; 3D MDS) and hence exhibited the highest computational flexibility of all ROIs.

3.3.7 Complementary evidence for sensory fusion in parietal-temporal regions and causal inference in frontal regions

To further corroborate the parieto-frontal gradient in the neural representations of sensory fusion (FU) and CI, we implemented a bootstrap-based comparison of each model-predicted sensory representation in explaining the local pattern of MEG activity based on group-level exceedance probabilities (P_{exc}).

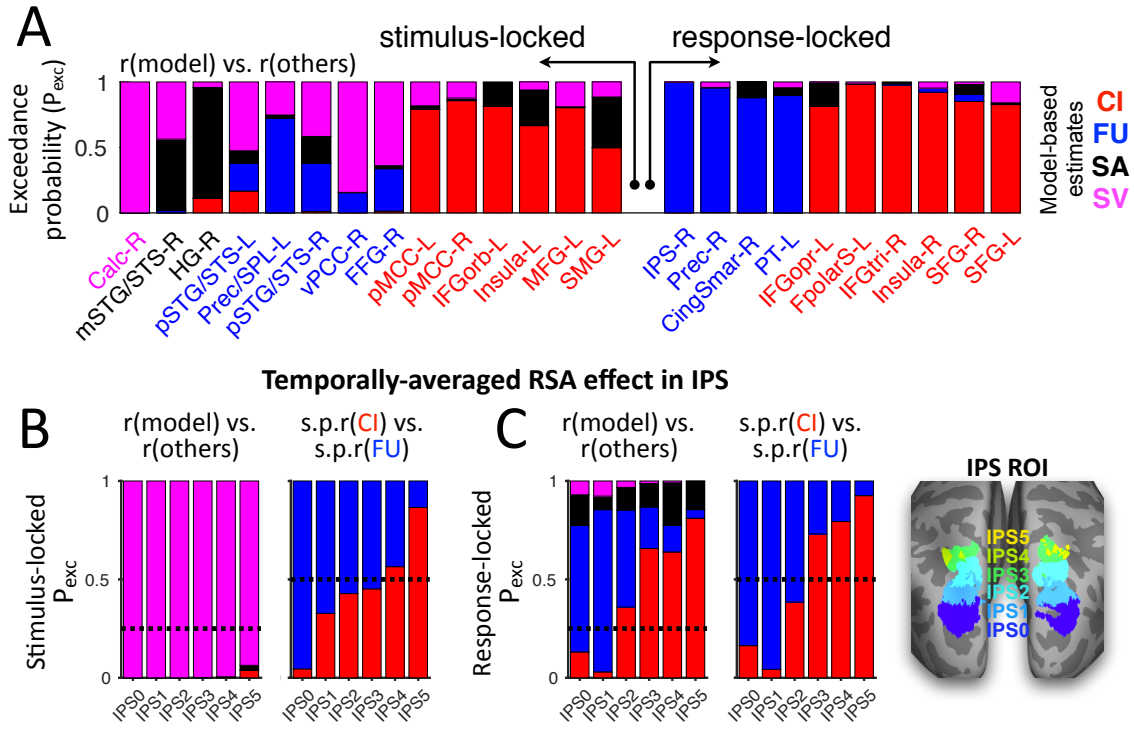


Figure 3.7. Model comparison based on exceedance probabilities. Exceedance probability (P_{exc}) indexes the belief that a region encodes a given model more likely than alternative models across participants. **(A)** P_{exc} for each of the ROIs derived from the RSA (cf. **Figure 3.5**). **(B-C)** Analysis of exceedance probabilities within the inferior parietal lobe for comparison with (Rohe & Noppeney, 2015b), for stimulus-locked (B) and response-locked data (C). **Left panels:** P_{exc} comparing each of the four models with all of the alternative models derived independently for each IPS ROI using time-averaged RSA statistics (encoding coefficient r, Fisher Z scale). **Right panels:** Time-averaged P_{exc} comparing the selective encoding (s.p.r, Fisher Z scale) of causal inference and fusion models. IPS ROIs were based on a probabilistic atlas by Wang et al. (2015).

ROIs with fusion or CI RSA effects exhibited considerably high P_{exc} values for the respective models, in particular in the response-locked data (P_{exc} > 0.88 among all 4 fusion ROIs, with the highest P_{exc}(FU) in IPS-R = 0.997; P_{exc} > 0.81 among all CI ROIs, with the highest P_{exc}(CI) around FpolarS-L = 0.979; **Figure 3.7A**). In the stimulus-locked data, the ROIs with CI RSA effects also exhibited high P_{exc} for CI and the parietal ROIs with fusion RSA effects exhibited high P_{exc} for fusion (Prec/SPL-L, P_{exc}(FU) = 0.72). Some occipital-temporal regions contain a

graded organisation suggesting also a contribution of segregated sensory representations apart from fusion. The superior temporal lobe comprised not only regions representing fusion (bilateral pSTG/STS) but more "bimodal" regions (mSTG/STS) that maintain the representations of both unisensory rates.

A control analysis was performed to rule out the possibility that the ROIs with RSA fusion effects simply reflect a fixed linear integration without weighting each modality by its relative reliability. We compared reliability-weighted fusion against a simpler model of fusion that ignores the trial-by-trial variations in auditory reliability. The average P_{exc} of the full reliability-weighted fusion model was as high as 0.875 (SEM = 0.048) across the fusion ROIs.

These results confirm a parieto-frontal gradient whereby the respective regions predominantly encode the sensory estimates based on fusion and causal inference. This is seemingly at odds with a previous study that described a gradient from fusion to causal inference locally within the intraparietal sulcus, based on a P_{exc} analysis of model encoding in fMRI data (Rohe & Noppeney, 2015b). To make the present data directly comparable to this previous study, we quantified model exceedance probabilities within anatomically defined IPS sub-regions (Wang et al., 2015). This confirmed a gradient along the posterior-anterior axis of the IPS, with the more posterior regions being dominated by fusion and the more anterior regions dominated by causal inference (**Figure 3.7B** and **2.7C**). Hence, while our data also support a graded representation of CI in the parietal lobe, the whole-brain analysis emphasises the involvement of a wider network that chiefly involves the frontal cortex.

3.3.8 Prefrontal cortex drives flexible behaviour in conflicting environments

Last, we asked which of the candidate ROIs is directly predictive of behaviour. The preceding results reveal that multiple regions reflect distinct multisensory computations. Yet, for a given ROI, a significant model-encoding RSA effect by itself does not imply a driving role for the full flexibility of participants' behaviour (Kriegeskorte & Douglas, 2018). Hence, to identify in which candidate ROIs the MEG activity was directly predictive of participants' responses, we implemented an RSA to assess the association between the MEG RDMs and the behavioural RDM: the pairwise absolute distance between the trial-averaged behavioural responses in different conditions.

To investigate in which ROIs the MEG activity directly drives participants' categorisation behaviour, rather than only reflecting the representation of a specific model, we conducted the following analysis. We implemented an RSA measuring the Spearman rank correlation between participant-specific response-locked MEG RDMs and behavioural RDMs, with the latter comprising pairwise absolute distance between the trial-averaged behavioural responses of different conditions. RFX significance testing followed the same permutation approach as for the model-based RSA to ascertain whether the group-average correlation in each ROI was significantly larger than zero (FWE = 0.05 across response-locked ROIs). We first assessed this neuro-behavioural correlation for the entire RDMs pertaining all 64 conditions. To identify the neural underpinning of the disparity-dependent adaptive behaviour, we also quantified the dependency of this neuro-behavioural correlation on disparity. We focused on the ROIs characterised by a significant overall behavioural relevance in the first step, and quantified the RSA effect of behaviour independently for the 24

conditions with small-disparity (absolute disparity = 3.6 Hz; **Figure 3.8**) and another 24 conditions with large-disparity (absolute disparity > 3.6 Hz). We then tested for a significant modulatory influence of the disparity on the RSA effect of behaviour by permuting independently the row and columns of the small and large disparity RDMs, and contrasting their Fisher Z transformed correlation with the respective portions of the behavioural RDM (large minus small; two-sided RFX inference; FWE = 0.05).

We focused on the ROIs identified in the response-aligned data to account for differences in reaction time between conditions. As above, we considered each ROI at its specific latencies of model encoding (see **Table 3.1**). The CI ROIs within the lateral PFC (IFG and SFG) and insula exhibited significant neuro-behavioural correlations ($t(14) \geq 3.01$, FWE-corrected $p \leq 0.036$; **Figure 3.8**). An additional contrast between conditions with large and small crossmodal disparities revealed that activity within the ventrolateral PFC (IFGtri-R) particularly predicted behavioural responses when the two modalities were highly conflicting ($t(14) = 2.94$, $p = 0.043$).

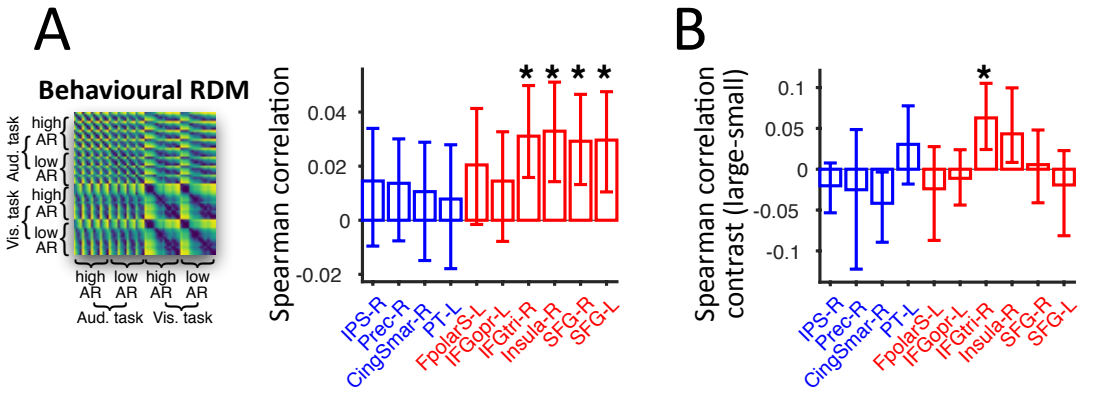


Figure 3.8. Contribution of frontal regions to flexible behaviour. (**A**) RSA was applied to assess the neural representation of behavioural reports (behavioural RDM; left shown = average across participants). Bar height and error bars = mean and 95% bootstrap confidence interval. (**B**) Disparity-modulated RSA effect of behavioural report was assessed by contrasting brain-behaviour correlations between conditions with small and large disparities (two-sided permutation test; small disparity = 3.64 Hz, large disparity > 3.64 Hz). (A-B) Asterisk = $p \leq 0.05$ FWE-corrected (see **Methods**).

3.3.9 Additional control analyses ruling out task-based modulation as the sole determinant for neural effect

Additional control analyses were performed to assert that the neural effect of CI was not solely driven by task-based modulation of neural activity or modality-specific attentional effect. Two analyses, one model based, the other model free, were presented as follows to provide converging evidence with the fingerprint analysis (**Figure 3.6**) showing that the CI ROIs are characterised by nonlinear bias, a feature that is unique to CI whereas the task-modulated segregation does not possess.

First, for the model-based control analysis, an additional control model (SEGtr) was included in the model comparison based on exceedance probability (P_{exc}; **Figure 3.7**). Exceedance probability indexes the belief at group level that a brain region encodes the sensory representation predicted by a given model more likely than those predicted by alternative models. This control model SEGtr describes the *task-relevant segregated* sensory representations whereby participants' perception of the stimulus rates reflects the visual rates for visual reports and auditory rates for auditory reports. Note that the sensory representations for this SEGtr model are in part different from those for SA (segregation auditory) and SV (segregation visual) that were used in the original neural analysis shown in **Figure 3.5**. Specifically, SA represents the auditory rates in both auditory and visual tasks, and SV represents the visual rates in both auditory and visual tasks, and thus these two *veridical* unisensory segregation representations describe the *task-irrelevant* sensory estimates in half of the conditions when pooling over the two tasks for constructing RDMs.

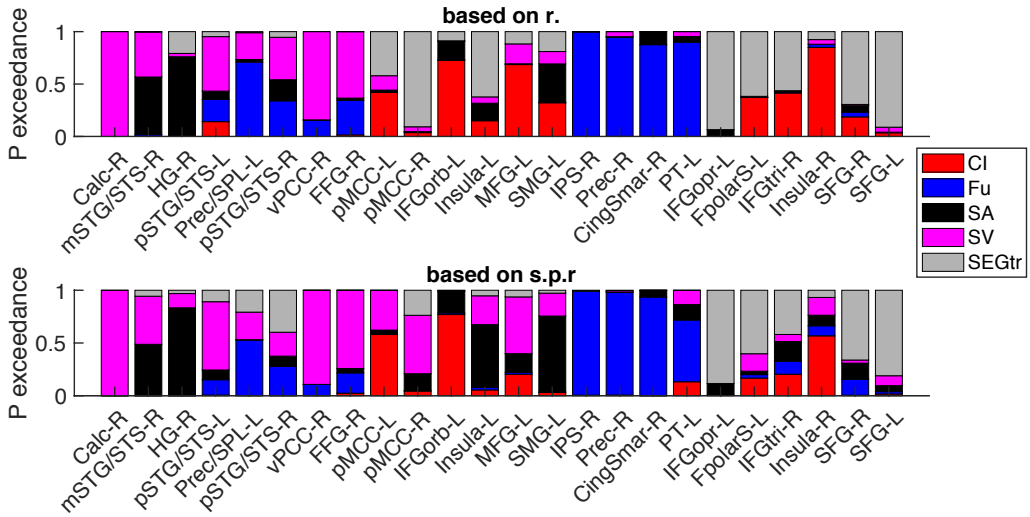


Figure 3.9 (related to Figure 3.7). Model comparison shown as P_{exc} profile after including a control model SEGtr (i.e., task-relevant segregation). Chance level $P_{exc} = 0.2$ due to 5 models in this analysis. The upper panel is directly related to panel A of Figure 3.7 (P_{exc} was derived based on the RSA correlation coefficients) and the lower panel corresponds to the same P_{exc} analysis but using semi-partial correlations (s.p.r; similar to the right panels of B and C of **Figure 3.7**).

Table 3.2: Summary of model comparison results after including the control model SEGtr (related to **Figure 3.9**). Text colours correspond to the colour coding of respective brain ROIs selectively encoding different model RDMs as revealed by the RSA in main analysis (CI = red, FU = blue, SA = black, SV = pink, which are consistent with the colour coding in **Figure 3.7**; none of the SA/SV ROIs were found satisfying any of the criteria in this table).

Criterion	Based on RSA r.	Based on RSA s.p.r
P_{exc} (CI) > 0.2	IFGorb-L MFG-L SMG-L pMCC-L FpolarS-L IFGtri-R Insula-R	IFGorb-L MFG-L pMCC-L IFGtri-R Insula-R
P_{exc} (CI) > P_{exc} (SEGtr)	IFGorb-L MFG-L SMG-L pMCC-L pSTG/STS-L FFG-R Insula-R PT-L Prec-R	IFGorb-L Insula-L MFG-L SMG-L pMCC-L FFG-R Insula-R PT-L Prec-R
P_{exc} (CI) > 0.2 and P_{exc} (CI) > P_{exc} (SEGtr)	IFGorb-L MFG-L SMG-L pMCC-L Insula-R	IFGorb-L MFG-L pMCC-L Insula-R

As shown in **Figure 3.9**, the CI model still explains the variance in neural activity better than the task-relevant segregation model does, for 4 out of 6 stimulus-locked ROIs (IFGorb-L, MFG-L, SMG-L, and pMCC-L, see **Table 3.2**) where significant selective neural representations of CI-predicted sensory estimates were identified using whole-brain RSA (see **Figure 3.5** and **Table 3.1**; and also **Figure 3.7**). **Table 3.2** provides detailed summary statistics of the ROIs for which 1) P_{exc} (CI) > 0.2 (chance level) after including the control model, 2) P_{exc} (CI) > P_{exc} (control model SEGtr), and 3) a conjunction of both P_{exc} (CI) > 0.2 and P_{exc} (CI) > P_{exc} (SEGtr). These results remained largely the same even if obtaining P_{exc} based on the semi-partial correlations reflecting the unique contribution of each computational model to the neural variance. The results of this control analysis thus speaks towards a robust effect of CI in MEG activity emerging around 620 ms (stimulus-locked) in dorso- and ventrolateral prefrontal cortices (note the paucity of SEGtr representation in the left inferior frontal cortex and middle frontal cortex) as well as the middle-posterior cingulate cortex and anterior part of the inferior parietal cortex (SMG-L).

Interestingly, some frontal regions (SFG for example) identified as representing CI predicted sensory estimates near the time of response selection also encode the task-relevant representations predicted by SEGtr, in line with past studies supporting a role of dorso-lateral PFC in modality-specific attentional modulation (Mayer, Ryman, Hanlon, Dodd, & Ling, 2010; Talsma, Senkowski, Soto-Faraco, & Woldorff, 2016). However, there is a clear evidence for above-chance P_{exc} supporting the representation of CI rather than SEGtr in multiple inferior and anterior PFC regions including FpolarS-L, IFGtri-R, and Insula-R.

Second, a model-free control analysis was conducted to complement the above exceedance-probability approach. This analysis was inspired by the model-free approach used in a previous study (Rohe & Noppeney, 2016). Here, I quantified the modulatory effect of crossmodal disparity on the neural weights of two orthogonal variables, i.e., the task-relevant vs. task-irrelevant stimulus rates. These neural weights were computed as the RSA correlation coefficients between the cross-validated brain RDM and the RDMs of task-relevant versus task-irrelevant stimulus rates (aggregated across the 2 modality-specific tasks), respectively. To assess the influence of crossmodal disparity on neural weighting between task-relevant vs. task-irrelevant rates, these neural weights were obtained independently for different levels of crossmodal disparity, similarly to the method used by Rohe and Noppeney (2016).

If we assume that in a brain region the task-relevant segregation thoroughly accounts for the neural data, then the neural weight for the task-relevant rate should dominate over that for the task-irrelevant rate in a consistent manner irrespectively of the disparity factor. However, the brain regions that genuinely encode CI should exhibit a significant modulatory effect of disparity upon how the neural weight is assigned to the task-relevant vs. task-irrelevant rate. That is, the brain should increase the weight of task-relevant rate (and/or decreases the weight of task-irrelevant rate) for larger crossmodal disparities. Hence, I formalised this logic with a GLM predicting the neural weights (Fisher Z transformed RSA correlation coefficients) using 3 regressors: Task relevance (task-irrelevant minus task-relevant), Disparity (larger minus smaller), and their interaction. The CI effect (nonlinear weighting of disparate crossmodal information) should be reflected by a significantly negative coefficient for the interaction term in this GLM. To clarify how this control analysis was conducted, I decompose the procedure into three discrete steps:

Step 1: construct task-relevant (target) vs. task-irrelevant (distractor) rate RDM

To construct stimulus-rate RDMs, I used the rate reports in multisensory congruent conditions, just the same as in the main analyses. These RDMs were of size 64 conditions by 64 conditions (aggregating the two tasks). For illustration purposes, the two group-average RDMs are shown here for the task-relevant (target) vs. task-irrelevant (distractor) rate RDMs. Since these RDMs were based upon raw behavioural responses, they have a model-free nature.

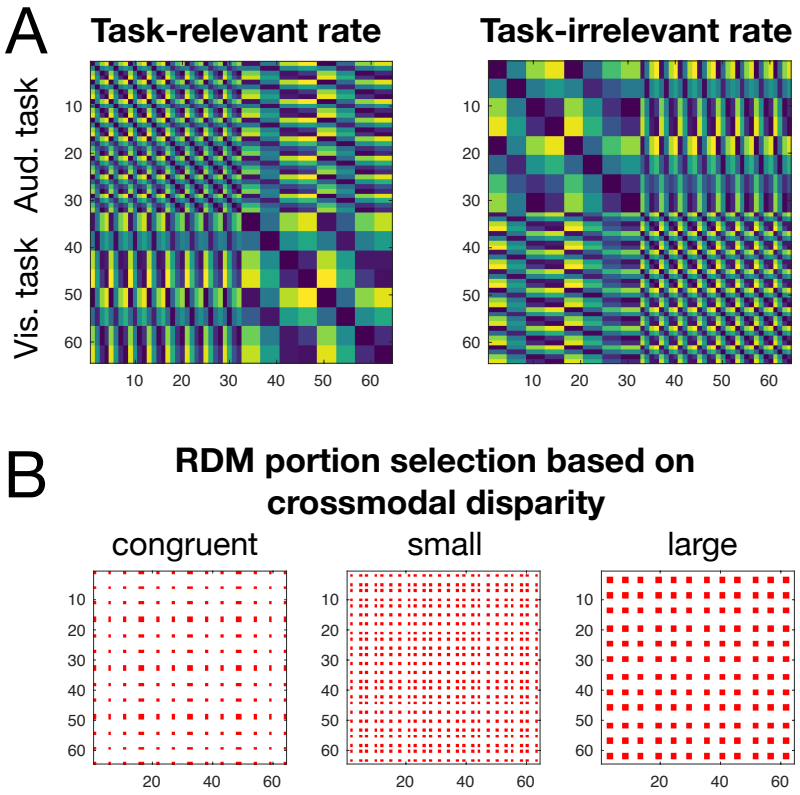


Figure 3.10. RDM construction and portion selection for model-free control analysis. (**A**) Model-free RDMs of task-relevant rate (target) and task-irrelevant rate (distractor). (**B**) The portion of RDM (displayed in red) that correspond to the respective experimental conditions with different levels of crossmodal disparity.

Step 2: select RDM portion based on AV disparity, and do RSA within each disparity

In this step, I obtained two Fisher-Z-transformed RSA correlation coefficients, r(brain RDM, target RDM) and r(brain RDM, distractor RDM), independently for each of the 3 disparity levels, i.e., congruent: disparity = 0, small disparity = 3.64 Hz, and large disparity >= 7.27 Hz.

Step 3: statistical inference

I used a GLM to predict those RSA coefficients. The regressors were Task relevance (distractor minus target), Disparity (larger vs. smaller), and their interaction. Statistical inference was based on maximum-statistics permutation tests (multiple comparisons across all 3 regressors and all brain regions; FWE = 0.05; 2-sided). For analysis consistency with the main analysis presented in this chapter, for each participant the mean RSA coefficient was taken across the ROIs of the same model 'class' (see **Table 3.1**).

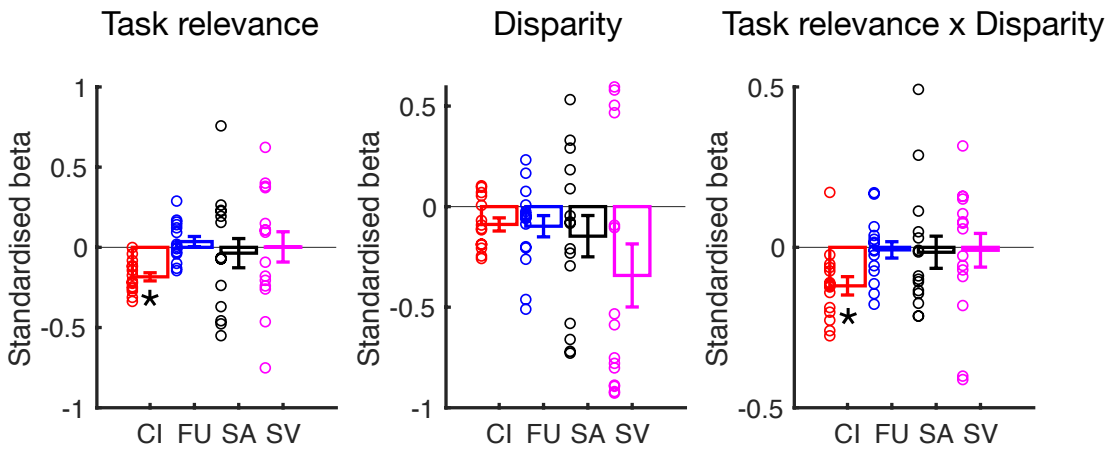


Figure 3.11. Beta coefficients of a GLM predicting the RSA effects of rate representation. A negative effect of task relevance (distractor minus target) indicates the neural weight of task relevant rate is significantly larger than the neural weight of task-irrelevant rate. A negative interaction between task relevance and disparity (larger vs. smaller) indicates a downweighing of task-irrelevant rate (and/or increasing neural weight of task-relevant rate) as the disparity increases. Significant effects are marked by * (assessed through maximum-statistics permutation across all

effects and brain regions; FEW = 0.05). Regressors are standardised. Bar colours correspond to the colour coding of respective brain ROIs selectively encoding different models as revealed by the RSA in main analysis (CI = red, FU = blue, SA = black, SV = pink, which are consistent with the colour coding in **Figure 3.7**).

As shown in the leftmost panel of **Figure 3.11**, a significant main effect of 'Task relevance' indicates that the CI regions reflect more of the target rate than the distractor rate according to task contexts. Crucially, however, task-based modulation could not explain the negative interaction between Task relevance and Disparity in shaping the MEG activity in these CI regions (rightmost panel of **Figure 3.11**; for completeness, the results for individual ROIs are shown in **Figure 3.12**). Such a negative interaction effect only occurs in the CI regions (identified in model-based RSA) but not in other regions, indicating an additional confirmation for the nonlinear bias as a signature of CI in neural activity.

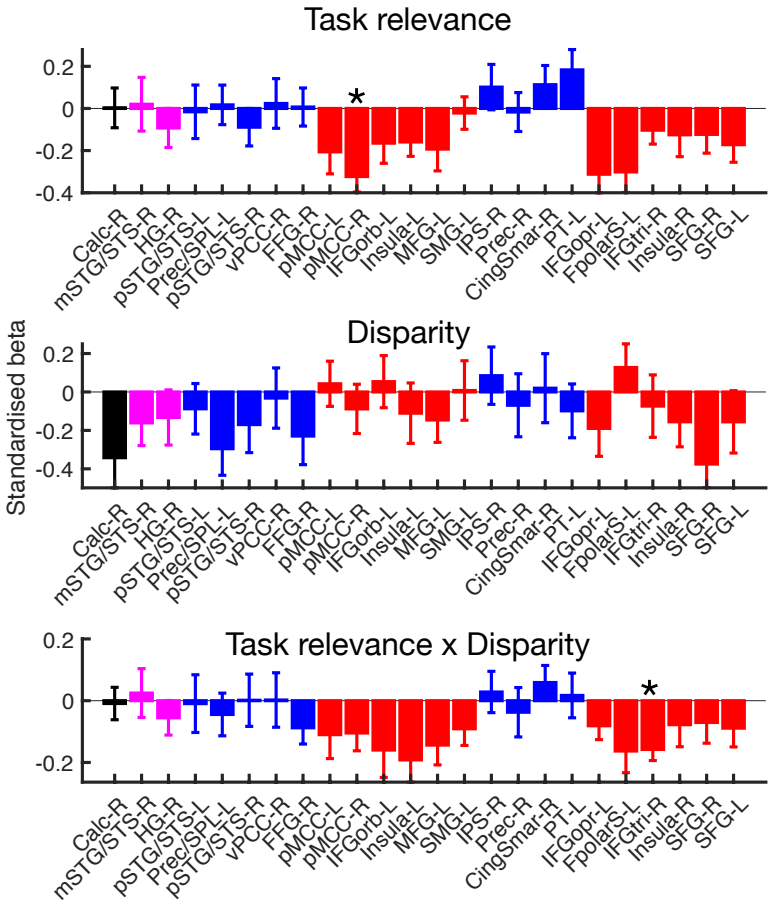


Figure 3.12. Beta coefficients of a GLM predicting the RSA effects of rate representation. Same GLM as **Figure 3.11** but implemented for each single ROI independently. Bar colours correspond to the colour coding of respective brain ROIs selectively encoding different models as revealed by the RSA in main analysis (CI = red, FU = blue, SA = black, SV = pink, which are consistent with the colour coding in **Figure 3.7**).

In summary, the above two complementary control analyses provide extra converging evidence asserting the nonlinear dependency of crossmodal bias on disparity in the CI regions identified in the main RSA results (presented earlier in this chapter; see **section 3.3.5**). The model-free analysis here did not rely on any fitting of theoretical models of multisensory integration and provided standard statistical inference that was not readily accessible using P_{exc} statistics. P_{exc} serves as a useful but relatively "crude" instrumental tool for quantitative analysis because it alone provides no direct statistical inference. Specifically, even if a model is associated with a larger P_{exc} compared to all other models, we do not know from an inference point of view whether the P_{exc} for the winning model is significantly higher than the P_{exc} for the other competing models. The model-free approach is capable of overcoming this issue and hence provides presumably more convincing (yet converging) conclusion.

Last but not least, future studies using an alternative cueing paradigm might help to further disentangle these different explanatory mechanisms. For example, such a paradigm could withhold the cueing about which modality to report until an unpredictable cueing signal (e.g., 'report the auditory rate you just heard' or 'report the visual rate you just saw') at a post-stimulus latency that is sufficiently long to help dissociating the cueing from information processing. This design presumably requires holding in working memory the information about the rates in both modalities, and could be extremely interesting if combined with additional working memory

manipulations to measure mnemonic hidden states in neural activities (Wolff, Jochim, Akyürek, & Stokes, 2017). Alternatively, a more conventional paradigm requiring 'dual reports' for both modalities (Körding et al., 2007) can be used to reduce the chance of 'cue-switching' between modalities or of gambling to just attend to one modality for a reduction of cognitive load (because dual-report performance can be pondered to assess the extent to which participants are doing so).

3.4 Chapter discussion

We characterised the computational strategies in the service of flexible multisensory integration, and identified their time-resolved underpinnings in source-localised MEG activity. Our behavioural data suggest that when faced with trial-varying reliability and discrepancy of crossmodal information, humans arbitrate between integration and segregation in a manner captured by Bayesian causal inference. At the neural level, we unveil that the distinct computations required for flexible multisensory perception (segregation, fusion and causal inference) coexist, but each dominates at different times and in distinct regions. Consistent with previous reports, these results show that the initially segregated unisensory signals are fused in temporal and parietal lobes. However, this fused information gives way to more flexible representations formed under multisensory causal inference in the frontal cortex, which drive behaviour in volatile environments with discrepant multimodal cues.

3.4.1 Temporal hierarchy of multisensory computations

Previous studies have posed reliability-weighted fusion and causal inference as rival accounts of multisensory perception, and sought to obtain empirical evidence that favours one over the other (Acerbi et al., 2018; Körding et al., 2007; Magnotti & Beauchamp, 2017; Parise et al., 2012; Roach et al., 2006; Rohe & Noppeney, 2015a; Wozny et al., 2010). In the spirit of this endeavour, we found that human behaviour during a rate categorisation task was consistently captured by a causal inference model. This fared best in quantitative model comparison and was the only model being able to describe the reduction in crossmodal bias that emerged when sensory cues were most disparate. Critically, the neural data challenge the dichotomy between fusion and causal inference as separate accounts of multisensory perception. Rather, they support the notion that perception is a hierarchical process relying on the explicit representations of distinct multisensory computations orchestrated over several brain regions (Rohe & Noppeney, 2015b; Rohe et al., 2018; Kayser & Shams, 2015). Our results suggest that representations as predicted by each model coexist, and unveil the functional hierarchy of the underlying computations in distinct regions and over different timescales.

We observed a systematic temporal sequence whereby the neural underpinnings of sensory segregation and reliability-weighted fusion were succeeded by those of causal inference. This cascade suggests a specific computational scheme for causal inference at the systems level: causal inference effectively relies on a weighted combination of sensory estimates predicted by fusion and segregated signals from *a posteriori* estimates, with the relative contribution depending on the inferred level of crossmodal disparity. Thus, one may interpret fusion as a component that explicitly feeds into the computations for inference (Körding et al., 2007; Rohe &

Noppeney, 2015a, 2015b; Wozny et al., 2010). Consistent with the early emergence of fusion, previous evidence suggests that multisensory integration starts at the level of early sensory cortices (Kayser & Logothetis, 2007; Lakatos et al., 2007; Lewis & Noppeney, 2010), and neuroimaging data support physiological correlates of reliability-weighted fusion starting around 120 ms post-stimulus onset (Aller & Noppeney, 2018; Boyle et al., 2017). Along this line, behavioural studies also suggested that fusion may be a rather automatic process. For example, crossmodal biases tend to be stronger when participants respond faster or after acquiring only little sensory evidence (Noppeney et al., 2010). By contrast, causal inference requires additional processing time as it capitalises on evaluating the degree of sensory discrepancy, maintaining beliefs over latent causes, and possibly exploring distinct decision strategies. Indeed, the conscious segregation of multimodal cues that usually tend to be fused requires additional time and effort (Gau & Noppeney, 2016). In line with these reports, and based on a principled assessment of candidate representations in spatiotemporally resolved brain activity, we demonstrate a systematic emergence of fused sensory representations within parietal-temporal regions before those predicted by causal inference in frontal regions.

3.4.2 Graded context-dependent computations along parietal-frontal pathway

Behaviour adapts to contextual modulations, such as trial-varying reliability of individual senses or changes in the congruency of sensory information (Angelaki et al., 2009; Ma & Pouget, 2008). Previous modelling studies have formalised two hallmarks of context-dependent integration: **1.** giving more credence to the more reliable modality (Alais & Burr, 2004; Ernst & Bülthoff, 2004), and **2.** refraining

from distraction by irrelevant information from an apparently distinct causal origin (Körding et al., 2007; Roach et al., 2006). The observers in our study exhibited both hallmarks as shown by their perceptual biases. Resolving the sensory conflict and inferring the sensory causal structure require time, and this was also visible in the nonlinear dependency of RTs on disparity. Importantly, the observers' brain activities directly reflected both hallmarks of flexible multisensory computation. Our complementary analysis approaches to the MEG data, relying on either selective model encoding or group-level model exceedance probabilities, corroborated a parieto-frontal gradient reflecting a progression from the neural representations adapting to sensory reliability to representations adapting to disparity. Whilst these results resonate with a previous study suggesting the coexistence of distinct context-dependent multisensory computations (Rohe & Noppeney, 2015b, 2016), our findings elaborate this perspective of computational gradient along a wider parietal-frontal network that supports causal inference in more anterior regions. This view also fits with a general understanding of the functional differences between parietal and frontal regions, whereby both accumulate sensory information but parietal cortex potentially encodes more of the sensory aspect whereas frontal regions underlie the context-dependent transformation of evidence into choice (Erlich, Brunton, Duan, Hanks, & Brody, 2015; Hanks et al., 2015).

Across the brain, we observed multiple and functionally heterogeneous multisensory representations on which flexible inference can capitalise (Bizley, Jones, & Town, 2016). First, we found a weak but noticeable auditory bias in V1, confirming multisensory interactions in early sensory cortex (Kayser & Logothetis, 2007; Lakatos et al., 2007). Second, within the temporal lobe we revealed the coexistence of bimodal (in the absence of integration) and truly fused representations,

in line with a topographical organisation of uni- and multisensory representations (Beauchamp et al., 2004a). Third, we found that the neural representation of fusion was stronger in parietal than temporal regions (Rohe & Noppeney, 2016), which may pertain to the abstract nature of the rate stimuli (Chafee, 2013; Raposo, Kaufman, & Churchland, 2014). Yet, these regions did not exhibit the characteristic down-weighting of task-irrelevant information at larger disparities indicative of inference. Rather, parietal activity appeared to scale with the task-irrelevant information, suggesting that these regions may automatically project evidence to one dimension (Ganguli et al., 2008; Luyckx, Nili, Spitzer, & Summerfield, 2018; Suzuki & Gottlieb, 2013).

By contrast, we found that the frontal cortex has a privileged role in reflecting both hallmarks of context-dependent integration, but features subtle representational variations across regions. Noteworthy, the nonlinear influence of disparity increased systematically along the posterior-to-anterior axis. Whilst representations within more posterior regions (e.g., the insula) appear to emphasise the reliability-based weighting, representations in anterior regions mostly depended on disparity (e.g., frontopolar cortex and inferior frontal gyrus). Such heightened sensitivity of more anterior regions to the causal structure resonates with the general hierarchical organisation of the PFC, whereby anterior regions are involved in representing abstract rules (Badre & D'Esposito, 2009). By combining the localisation of specific candidate representations in MEG activity with an analysis probing their choice-predictability, our results further demonstrate the behavioural relevance of frontal representations. Here the more prominent role of the dorsomedial and ventrolateral, rather than the anterior PFC, in driving behaviour fits with the role of these regions in

inferring probable causes of observed contingencies for retrieving goal-relevant behavioural strategies (Donoso et al., 2014).

3.4.3 The frontal lobe in multisensory and domain-general inference

A key aspect of behavioural flexibility is the ability to perceive the environments through multiple senses and to exploit the most appropriate modalities for the task at hand. Given the role of the frontal cortex in subserving general reasoning and adaptive behaviour (such as inferring the reliability of different decision strategies), it is perhaps not surprising that we found the key driver of flexible multisensory behaviour in the frontal lobe (Donoso et al., 2014; Koechlin & Summerfield, 2007; Tomov et al., 2018). Yet, previous studies promoted divergent views on the neural basis of multisensory perception. In fact, many studies have emphasised the role of superior temporal and parietal cortices in sensory integration (Beauchamp et al., 2004b; Calvert, 2001). In part, this may have arisen from the specific search for bimodal representations in some, and for sensory fusion in other studies, or the aim to identify the earliest convergence of multisensory signals (Kayser & Logothetis, 2007). However, neuroanatomical evidence has long implied the prefrontal cortex as a convergence zone for multisensory information (Jones & Powell, 1970), with the ventrolateral prefrontal cortex receiving projections from auditory and visual cortices and association areas implementing sensory fusion (Romanski, 2012). In the context of multisensory integration, the PFC has been highlighted as a domain-general structure responsible for evaluating sensory discordance (Adam & Noppeney, 2010; Calvert, 2001; Hein et al., 2007; Noppeney et al., 2010), and has been implied in forming beliefs over inferential states, based on expectations or prior experience, in face of sensory uncertainty (Gau & Noppeney, 2016; Kayser & Kayser, 2018; Noppeney et al., 2010).

Yet, the specific computations underlying frontal multisensory representations, and their role in driving behaviour, remained elusive. Our results reconcile the previous literature and suggest that the frontal cortex implements a flexible strategy capitalising on distinct computational solutions (fusion vs. segregation) when processing volatile multisensory information. Thereby perception effectively amplifies the behavioural significance either of segregated representations established early on in a trial and maintained within sensory-specific regions, or of fused representations formed within temporal and parietal association regions later on. These results help to unify previous studies on general adaptive behaviour with those elucidating the function of frontal regions in multisensory integration. In fact, prefrontal cortex seems to support a domain-general mechanism for selecting among multiple candidate strategies for solving a problem at hand, when handling both multisensory and other information. As such, the role of the PFC is unlikely about merging sensory information alone, but arbitrating between competing strategies of how the most appropriate sensory representation should be formed to guide behaviour.

Chapter appendix I

This appendix provides supplemental materials containing 2 tables and 7 figures.

Table S3.1: Comparison of candidate models and their explanatory power for behavioural data (related to Figure 3.3). CI = causal inference (with decision strategies: MA = model averaging, PM = probability matching, MS = model selection, see texts). R^2 = generalised coefficient of determination (Nagelkerke, 1991); here, across-participants mean (SEM). BIC = Bayesian Information Criterion, AICc = Akaike Information Criterion (corrected for sample size). Alternative models are compared to the best-fitting model CI (MA). Smaller relative BIC/AICc indicates better explanation of data among alternative models. Expected posterior p = probability that a given model could have generated the data for a randomly selected participant; Exceedance probability (P_{exc}) = probability that a given model is more likely than any other models; protected P_{exc} further corrected for chance level (Rigoux et al., 2014). Note: Across-participants minimum BIC and AICc difference between fusion and the optimal model CI (MA) = 108 and 113, respectively.

	CI (MA)	CI (PM)	CI (MS)	Likelihood	Fusion	Segregation
Parameters (n)	10	10	10	9	9	9
$\boldsymbol{R^2}$	0.869 (0.0078)	0.865 (0.0072)	0.856 (0.0086)	0.854 (0.01)	0.700 (0.18)	0.851 (0.22)
Relative BIC (sum)	0	128	393	345	5179	469
Expected posterior *p* (BIC evidence)	0.4	0.102	0.103	0.174	0.0476	0.172
$\boldsymbol{P_{exc}}$ **(BIC evidence)**	0.85	0.0122	0.0126	0.063	0.0017	0.061
Protected $\boldsymbol{P_{exc}}$ **(BIC evidence)**	0.65	0.0571	0.0574	0.093	0.049	0.092
Relative AICc (sum)	0	128	393	424	5257	547
Expected posterior *p* (AICc evidence)	0.578	0.146	0.113	0.059	0.048	0.055
$\boldsymbol{P_{exc}}$ **(AICc evidence)**	0.99	0.0063	0.0025	0.0004	0.0002	0.0003
Protected $\boldsymbol{P_{exc}}$ **(AICc evidence)**	0.985	0.0073	0.0036	0.0014	0.0013	0.0014

Table S3.2: Optimal model parameters (related to Figure 3.3). Model abbreviations are the same as in Table S1. Numbers indicate the mean (SEM) and [median] across participants (N = 15). The first 7 parameters (sensory noise parameters, σ denoted, and the mean of central prior μp) have unit of Hz. The models were parameterised using Gaussian sensory likelihoods for noisy visual and auditory representations (with variance σ_v^2 and σ_a^2 around true physical rate), a Gaussian central prior (mean μp and variance σ_p^2) capturing response bias towards intermediate rate choices, and a common-source prior probability (p_c, ranging from 0 to 1) quantifying the a priori integration tendency. *Note*: p_c for the Fusion, Segregation and Likelihood models were fixed at 1, 0 and 0.5, respectively, and were thus not free parameters. σ(a1) and σ(a4), the standard deviation of auditory noise corresponding to the lowest (a1 = 9.09 Hz) and highest (a4 = 20 Hz) rate level, respectively. σ(v1) and σ(v4) for visual noise; *k*a and *k*v, power coefficients for auditory and visual unisensory psychophysical function, respectively (see "Sensory noise function").

Parameter	**CI (MA)**	CI (PM)	CI (MS)	Likelihood	Fusion	Segregation
σp	6.94 (0.89) [6.13]	6.01 (0.44) [5.77]	6.26 (0.51) [5.81]	6.41 (0.72) [6.4]	8.51 (1.30) [5.5]	5.72 (0.28) [5.66]
μp	19.6 (1.35) [17.6]	19.2 (1.30) [17.7]	19.3 (1.73) [17.4]	18.4 (0.58) [18.02]	34.5 (5.55) [19.3]	17.3 (0.31) [17.4]
σ(a1 high)	2.34 (0.21) [2.24]	2.29 (0.2) [2.22]	2.38 (0.21) [2.31]	2.32 (0.18) [2.27]	3.28 (0.26) [2.98]	2.38 (0.21) [2.39]
σ(a1 low)	4.28 (0.56) [3.49]	4.15 (0.52) [3.25]	4.45 (0.58) [3.62]	4.15 (0.51) [3.45]	4.46 (0.51) [3.72]	4.46 (0.59) [3.63]
σ(v1)	3.67 (0.61) [2.84]	3.57 (0.57) [2.76]	3.81 (0.62) [2.98]	3.79 (0.62) [3.01]	4.45 (0.78) [3.50]	3.80 (0.61) [2.99]
σ(a4 high)	10.9 (3.15) [8.42]	10.47 (3.2) [5.73]	10.88 (3.11) [6.46]	10.15 (2.77) [3.93]	7.60 (2.17) [4.24]	11.2 (3.05) [6.49]
σ(v4)	8.28 (1.45) [7.03]	8.23 (1.36) [6.24]	10.60 (3.02) [7.04]	7.21 (1.07) [5.54]	8.78 (3.13) [4.34]	8.43 (1.24) [7.07]
***k*a**	-0.51 (2.64) [0.65]	-3.04 (2.46) [-7.43]	-1.03 (2.48) [-4.50]	0.14 (2.07) [0.28]	5.79 (2.63) [9.44]	-2.36 (2.4) [-0.73]
***k*v**	-1.37 (2.51) [0.84]	-0.77 (2.42) [-1.73]	1.95 (2.55) [2.10]	-2.42 (2.59) [-2.51]	3.73 (3.04) [8.11]	-3.99 (2.37) [-3.49]
***p*c**	0.379 (0.077) [0.35]	0.372 (0.087) [0.37]	0.44 (0.075) [0.48]	0.5	1	0

Figure S3.1. Frequency response compensation using customised inverse filter. The inverse filtering methods eliminate the spectral distortion of the sound stimuli induced by the frequency response of the tubephone system (Giordano et al., 2018), resulting in flat frequency response up to approximately 5000 Hz. L = Left-ear channel, R = Right-ear channel.

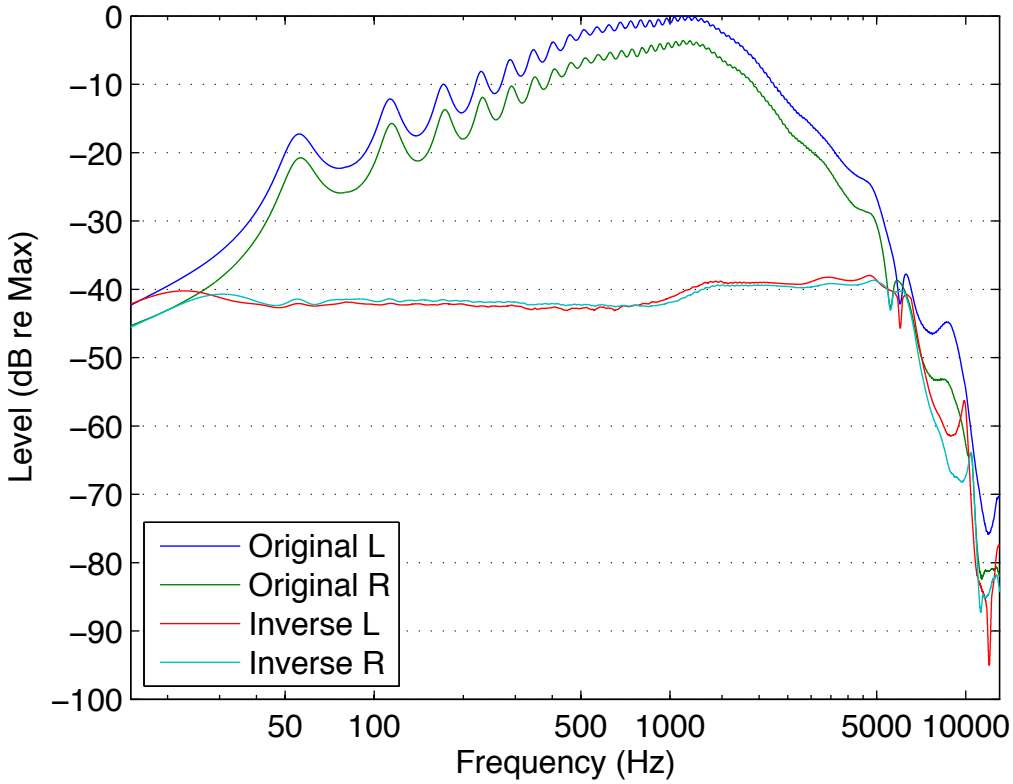


Figure S3.2. Analysis of reaction times (RTs) and response bias (related to Figure 3.3)**. (A)** RT as a function of task (T), auditory reliability (AR) and disparity. The horizontal dashed line marks the grand mean RT across all multisensory trials from all participants (N = 15). To avoid overlapping error bars, the ± SEM for high AR condition is shown as semi-transparent and using dashed lines. Unisensory RTs are displayed in blue/green/pink. **(B)** GLM predicting RT based on task, AR, task × AR, abs(Disparity) and squared disparity (i.e., $Disparity^2$). Red asterisk * = significant ($p \leq 0.05$ corrected across the five GLM effects); Small circles = participant-specific betas. **(C)** Cumulative probability of RT in multisensory trials (N = 1408) for each individual participant (light grey curves) and across participants (thick black curve). From this we derived a cut-off time for the stimulus-locked analysis window as the RT corresponding to the 5^{th} percentile of the distribution (red dashed line, RT = 0.69 s). Individual participants' 5^{th} percentiles were in the range of [0.54, 0.84] s. Semi-transparent green band marks the latencies investigated in the stimulus-locked RSA [-0.1, +0.7] s relative to stimulus onset. Circles (in **A**) and bar height (in **B**) = mean and error bars (bands) = ± 1 SEM across participants. **(D)** Extended GLM (for Figure 3.3E): abs(crossmodal bias) ~ intercept + task + AR + task × AR + abs(Disparity) + $Disparity^2$ + task × abs(Disparity) + task × $Disparity^2$ + task × AR × abs(Disparity) + task × AR × $Disparity^2$. T = task, AR = auditory reliability. Red asterisk * = significant ($p \leq 0.05$ FWE-corrected across the 11 GLM effects); Error bars = ± 1 SEM; circles = participant-specific betas. The (non)-significance of the first 5 effects (T, AR, T × AR, abs(Disparity) and $Disparity^2$) matches that observed in the reduced GLM with the 5 main regressors in Figure 3.3E. (**E**) Leftmost and middle panels: Simulated predictions of a Bayesian drift diffusion model. The leftmost panel shows 5 trajectories (mean of 1,000 simulated diffusion traces) for 5 different conditions with different combinations of audiovisual rate inputs. The green curve is the prediction of a drift diffusion process given only unisensory input. The posterior probability is the predicted probability of reporting a high rate (i.e., > 14.5 Hz). The middle panel shows the RTs derived as the threshold-cutting latencies. Colour coding corresponds to the probability trajectories. The rightmost panel shows the RT data in the rate categorisation task in chapter 3.

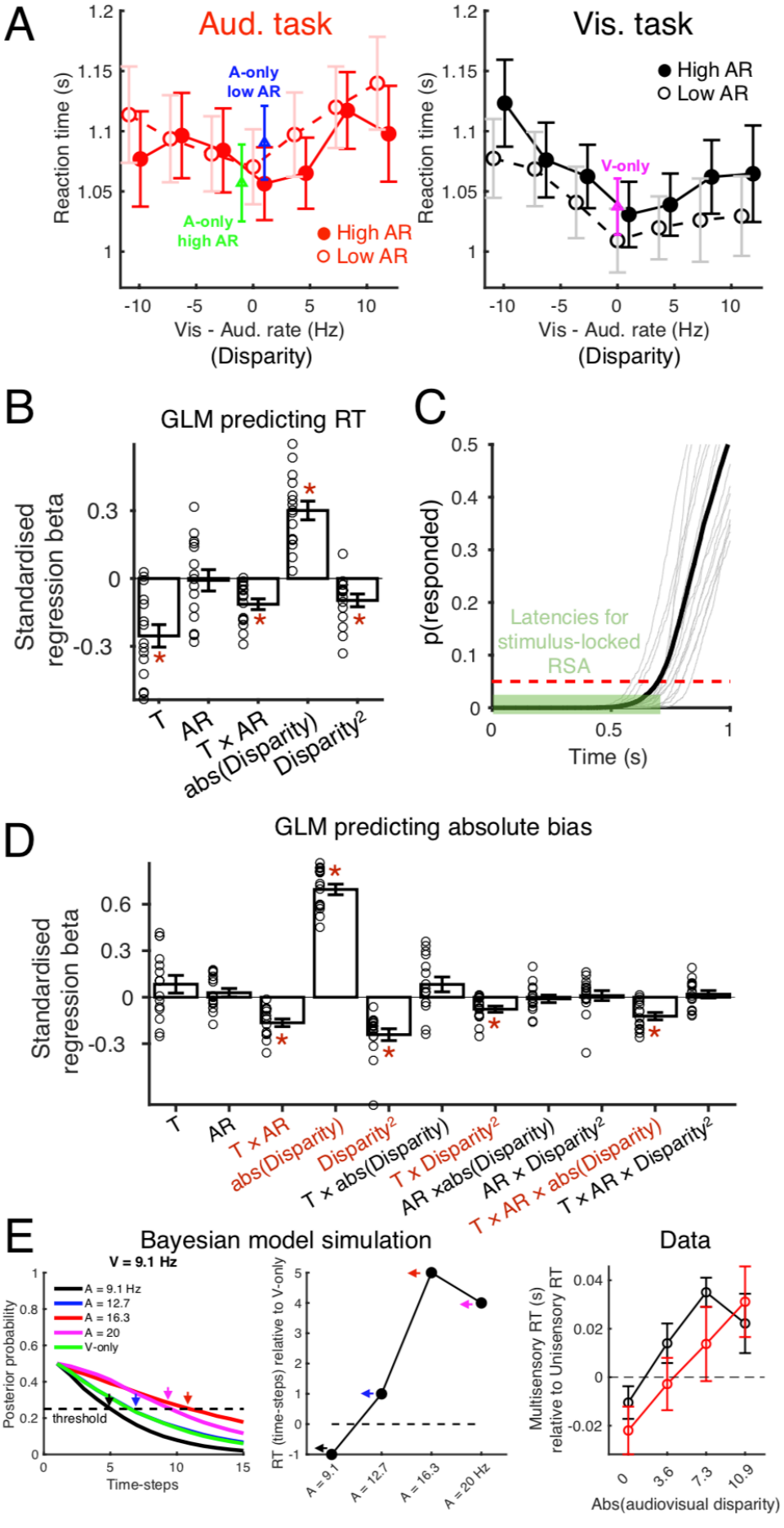

A
Aud. task
Vis. task
Reaction time (s)
A-only low AR
A-only high AR
High AR
Low AR
V-only
Vis - Aud. rate (Hz)
(Disparity)
B
GLM predicting RT
Standardised regression beta
T
AR
T x AR
abs(Disparity)
Disparity²
C
p(responded)
Latencies for stimulus-locked RSA
Time (s)
D
GLM predicting absolute bias
T x Disparity²
T x abs(Disparity)
AR xabs(Disparity)
AR x Disparity²
T x AR x abs(Disparity)
T x AR x Disparity²
E
Bayesian model simulation
V = 9.1 Hz
A = 9.1 Hz
A = 12.7
A = 16.3
A = 20
V-only
Posterior probability
threshold
Time-steps
RT (time-steps) relative to V-only
Data
Multisensory RT (s) relative to Unisensory RT
Abs(audiovisual disparity)

Figure S3.3. Selective encoding of sensory representations predicted by different models (related to Figure 3.5 and Table 3.1). Selective brain encoding of fusion (FU) and causal inference (CI) quantified by semi-partial correlation (s.p.r) between model and MEG RDMs. Semi-partial correlations for each individual participant (N = 15) and the group average, denoted by filled and empty circles, respectively. Error bars = ± 1 SEM.

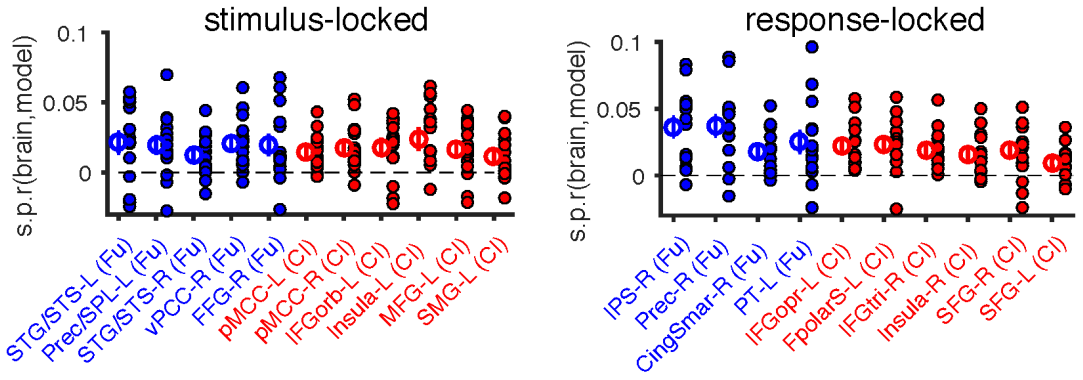


Figure S3.4. Temporal sequence of significant RSA model encoding in MEG activity (related to Figure 3.5). **(A)** and **(B)** CI = causal inference, FU = Fusion, SA = Segregation auditory, SV = Segregation visual model-based estimates. Colour code differentiates the four model-based estimates and the respective brain regions that selectively encode each model. Black horizontal bars mark the peak latencies at which the selective encoding of each model-based representation was significant (FWE--corrected $p \leq 0.05$). See **Table 1** for the latencies and spatial locations of the respective global and local peaks. All time courses represent the mean (SEM) across participants. **(C)** and **FWE-:** time courses of RSA effects for Fusion and CI (T statistics of s.p.r). Thick: grand mean time courses across ROIs; thin: ROI-specific effects. Circles: ROI-specific peak T(s.p.r) quantifying the most robust selective encoding of each model at group level. **(D)** and **(F)**: peak latency contrasts between pairs of ROIs. From the time courses of RSA effects we extracted the peak latencies of Fusion and CI effects and derived a latency contrast for each pair of the respective ROIs (latency of Fusion – latency of CI).

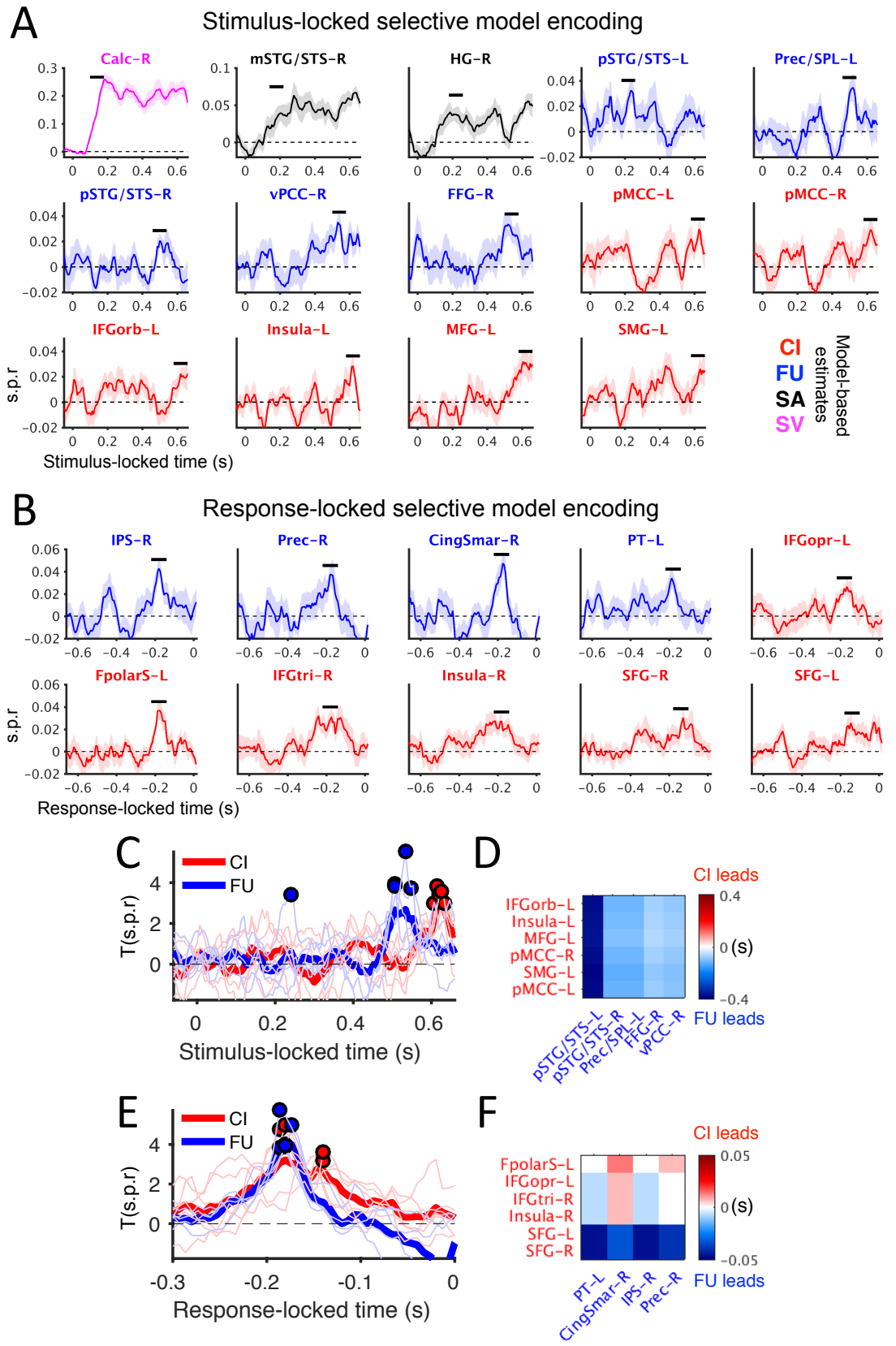

A
Stimulus-locked selective model encoding
Calc-R
mSTG/STS-R
HG-R
pSTG/STS-L
Prec/SPL-L
pSTG/STS-R
vPCC-R
FFG-R
pMCC-L
pMCC-R
IFGorb-L
Insula-L
MFG-L
SMG-L
CI
FU
SA
SV
Model-based estimates
s.p.r
Stimulus-locked time (s)
B
Response-locked selective model encoding
IPS-R
Prec-R
CingSmar-R
PT-L
IFGopr-L
FpolarS-L
IFGtri-R
Insula-R
SFG-R
SFG-L
s.p.r
Response-locked time (s)
C
CI
FU
T(s.p.r)
Stimulus-locked time (s)
D
CI leads
IFGorb-L
Insula-L
MFG-L
pMCC-R
SMG-L
pMCC-L
pSTG/STS-L
pSTG/STS-R
Prec/SPL-L
FFG-R
vPCC-R
FU leads
E
CI
FU
T(s.p.r)
Response-locked time (s)
F
CI leads
FpolarS-L
IFGopr-L
IFGtri-R
Insula-R
SFG-L
SFG-R
PT-L
CingSmar-R
IPS-R
Prec-R
FU leads

Figure S3.5. Full GLM results for all brain ROIs (related to Figure 3.6). Asterisk * = significant ($p \leq 0.05$ FWE-corrected across the five GLM effects and ROIs). Bar height and error bar = mean and 95% bootstrap confidence interval of the standardised GLM coefficients across participants, respectively.

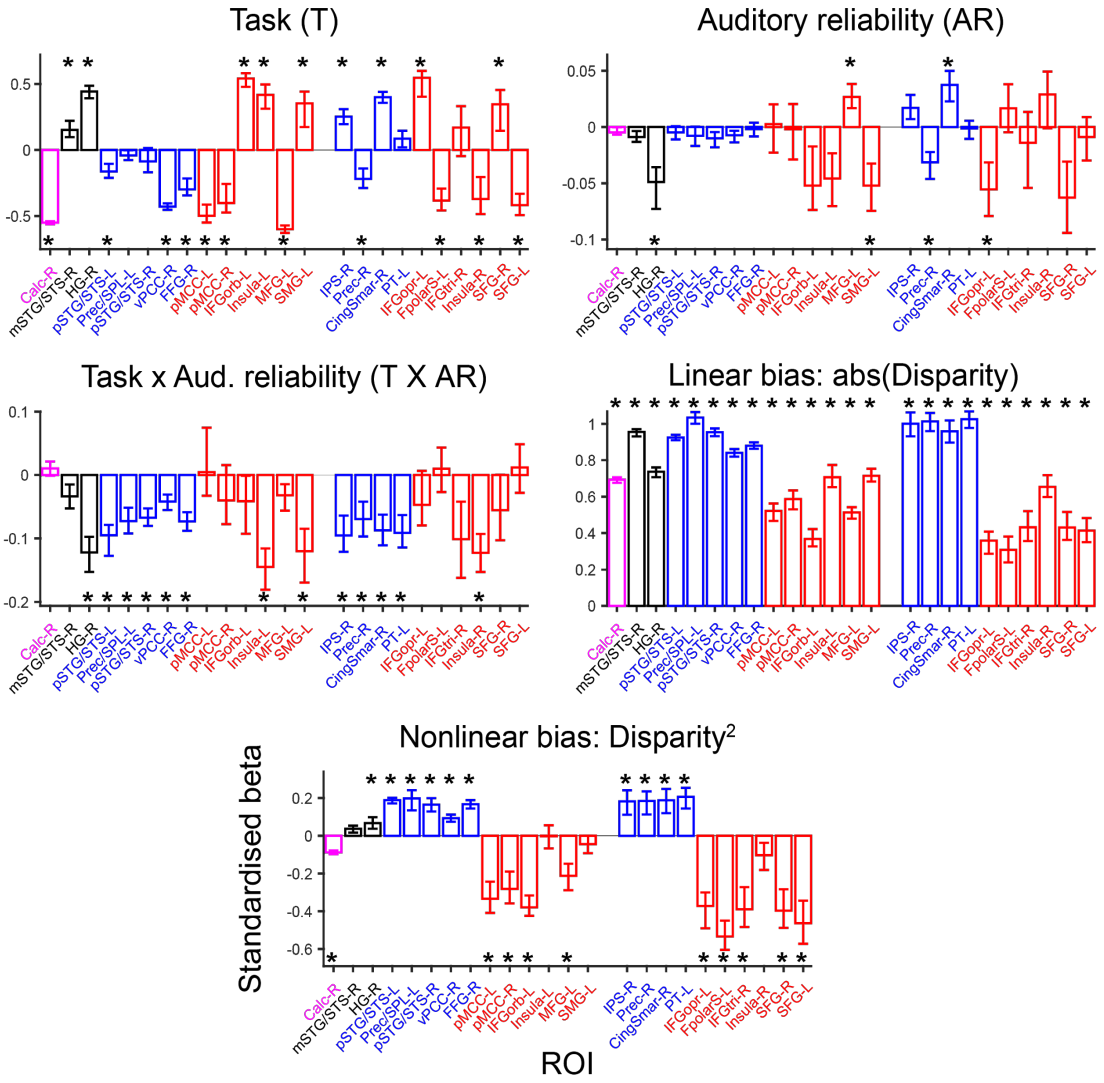


Figure S3.6. Analysis of computation-diagnostic representations in all ROIs (related to Figure 3.6). **(A)** and **(B)** Inter-individual difference multidimensional scaling (MDS) of the computation-diagnostic RDM in each of the selected ROIs (A = 2D MDS; B = 3D MDS). **(C)** Scatterplot of the measures of disparity and bias (ranks) recovered from the computation-diagnostic neural RDMs.

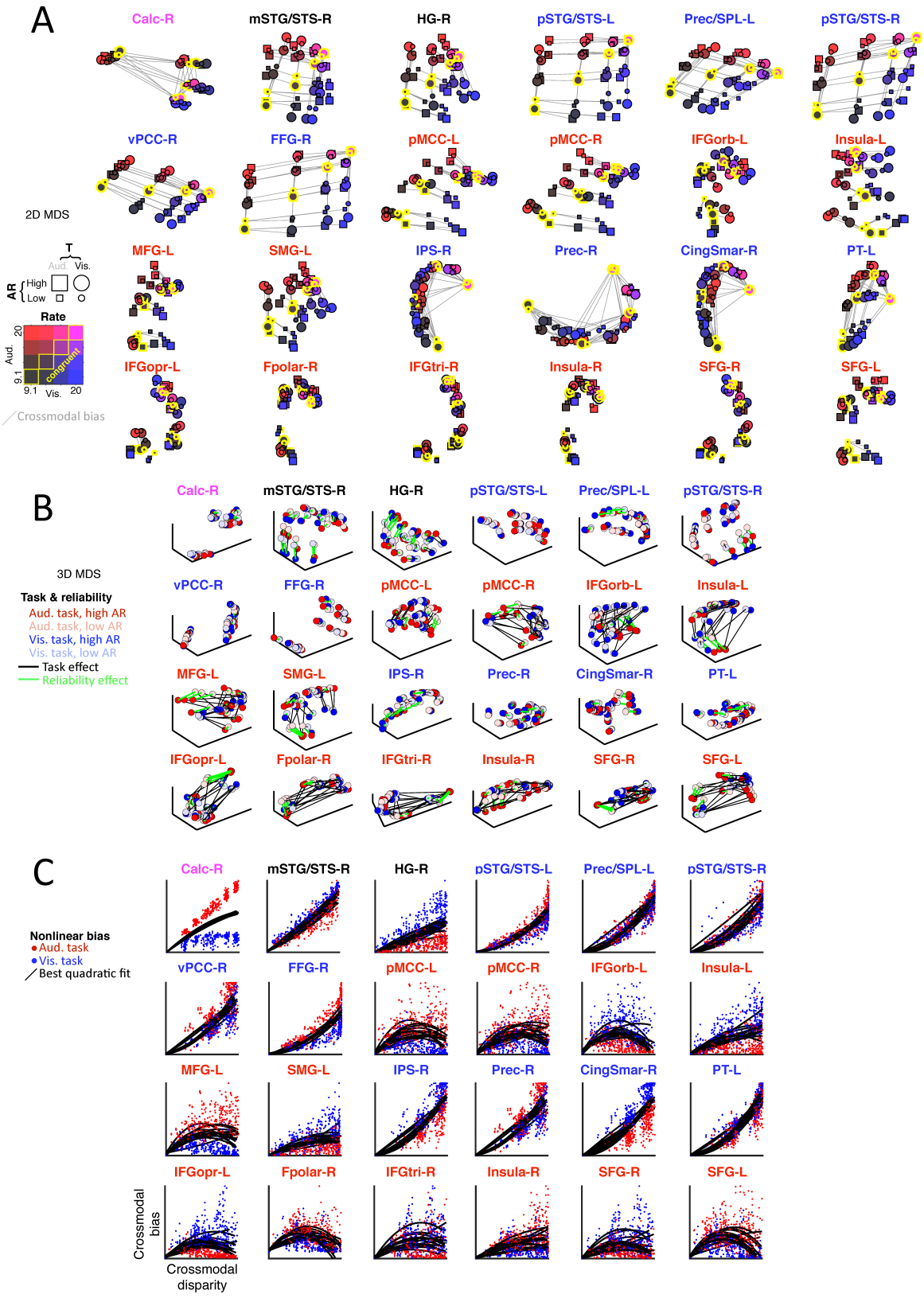
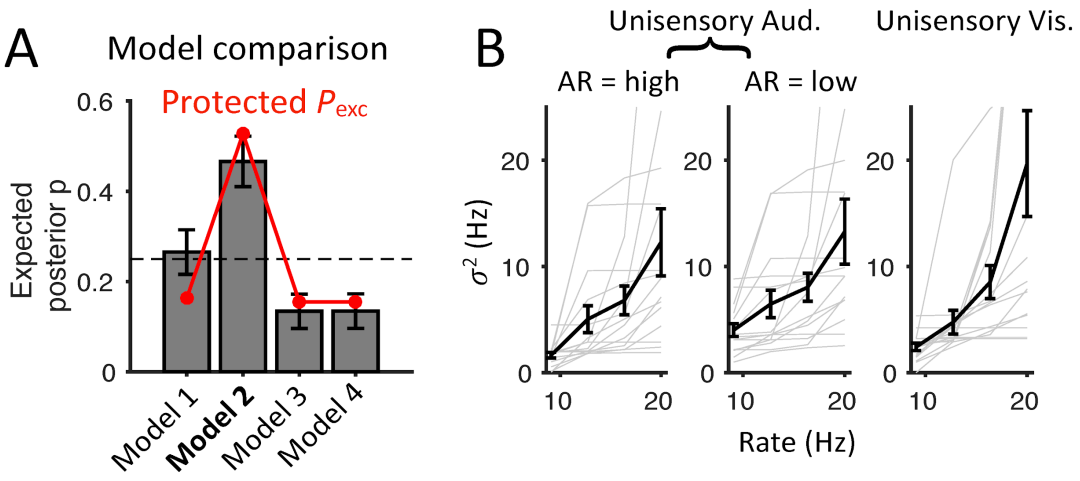


Figure S3.7. Sensory noise function (related to Methods)**. (A)** Random-effects model comparison of 4 candidate noise models [parameters]. (**Model 1**): power noise is both modality-specific and reliability-specific [σ(a1 high), σ(a1 low), σ(v1), σ(a4 high), σ(v4), ka-low, ka-high, kv] (Note: e.g., σ(a1 high) = SD of auditory noise at the lowest rate for high auditory reliability; kv = power coefficient for visual noise); (**Model 2**): power noise is modality-specific but reliability-independent [σ(a1 high), σ(a1 low), σ(v1), σ(a4 high), σ(v4), ka, kv]; (**Model 3**): power noise is modality- and reliability-independent (noise function just being shifted up and down by an offset) [σ(a1 high), σ(a1 low), σ(v1), σ(a4 high), kav]; (**Model 4**): constant noise across rates [σ(a1 high), σ(a1 low), σ(v1)]. **(B)** Estimates of sensory noise (Model 2) in unisensory auditory and visual conditions as a function of rate levels. Thick black lines and error bars: mean and ± SEM across participants (N = 15), respectively; thin grey lines: individual participants. AR = auditory reliability.



Chapter appendix II

Temporal extent of RSA searchlight

A choice of 80 ms was made based on the need to capture the information represented in both spatial and temporal dimensions of the MEG (note that the signals were never averaged within the time window for any of the presented analyses). To estimate reliably the MEG representational dissimilarity patterns, a sufficient temporal extent of the searchlight is preferable.

We implemented a control analysis to show that the RSA results remain largely unchanged when varying the temporal extent of the searchlight (**Figure S3.8** below; Note the increased jaggedness of these time courses for shorter RSA searchlights, which is an expected consequence of the decreased reliability of the neural distances estimates when fewer time points are considered).

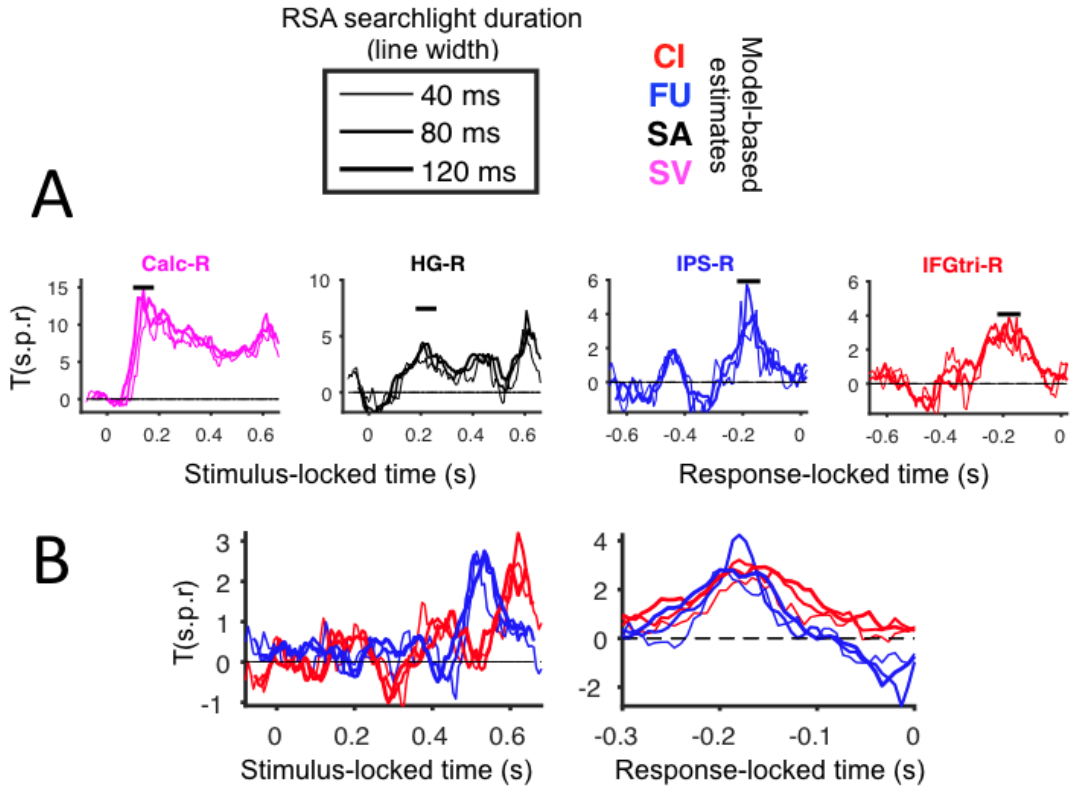


Figure S3.8. Temporal extent of RSA window. (**A**) Influence of the temporal extent of the RSA searchlight on model-encoding time courses. T(s.p.r) denotes the group-level T statistics of semi-partial correlation between model and MEG RDM (related to Figure S3.4A-B). (**B**) Average T(s.p.r) time course across ROIs selectively encoding fusion or CI model estimates (related to Figure S3.4C and E).

Cross-generalisation of RSA neural states across time

Analysing the temporal generalisation of model representations is used to complement an analysis of neural sensory representations. Specifically, we measured the temporal generalisation of either the neural RDMs themselves (independent of any model) or of specific model representations. These are different concepts that we quantified in the RSA domain as different aspects of the cross-temporal variance of the RDMs. That is, we calculated the variance shared between the RDMs in a specific ROI at time *t*1 and time *t*2 ("Neural RDM generalisation analysis"), or the variance shared about a specific candidate model ("Model representation generalisation analysis").

The "neural RDM generalisation analysis" (shown in **Figure S3.9)** reveals a predominant correlation between the RDMs for neighbouring latencies. This could arise from strong autocorrelations in the MEG data and/or the RDMs themselves. Critically, the structure of these time × time variance maps does not support the presence of oscillatory patterns. The "model representation generalisation analysis" indeed reveals some meaningful generalisation patterns. For example, the representation of visual rate (SV) in Calc-R appears to be strongly redundant (high common SV variance) across stimulus-locked latencies after around 200 ms. Critically, in all of the ROIs we observe a strong resemblance between common variances closer to the diagonal and model encoding time courses (**Figure S3.9**, bottom panel). This is not surprising because the diagonal of the 2D time × time

common variance maps (excluded here from computations) equals the squared correlation between the model and the RDM at a specific time point. Yet, it remains possible that these off-diagonal elements may be influenced by statistical artefacts driven by the time-varying model encoding itself. As such, we envision that a substantial amount of development work will be required to reach a clear (and statistical) interpretation of such time-generalisation measures in the future.

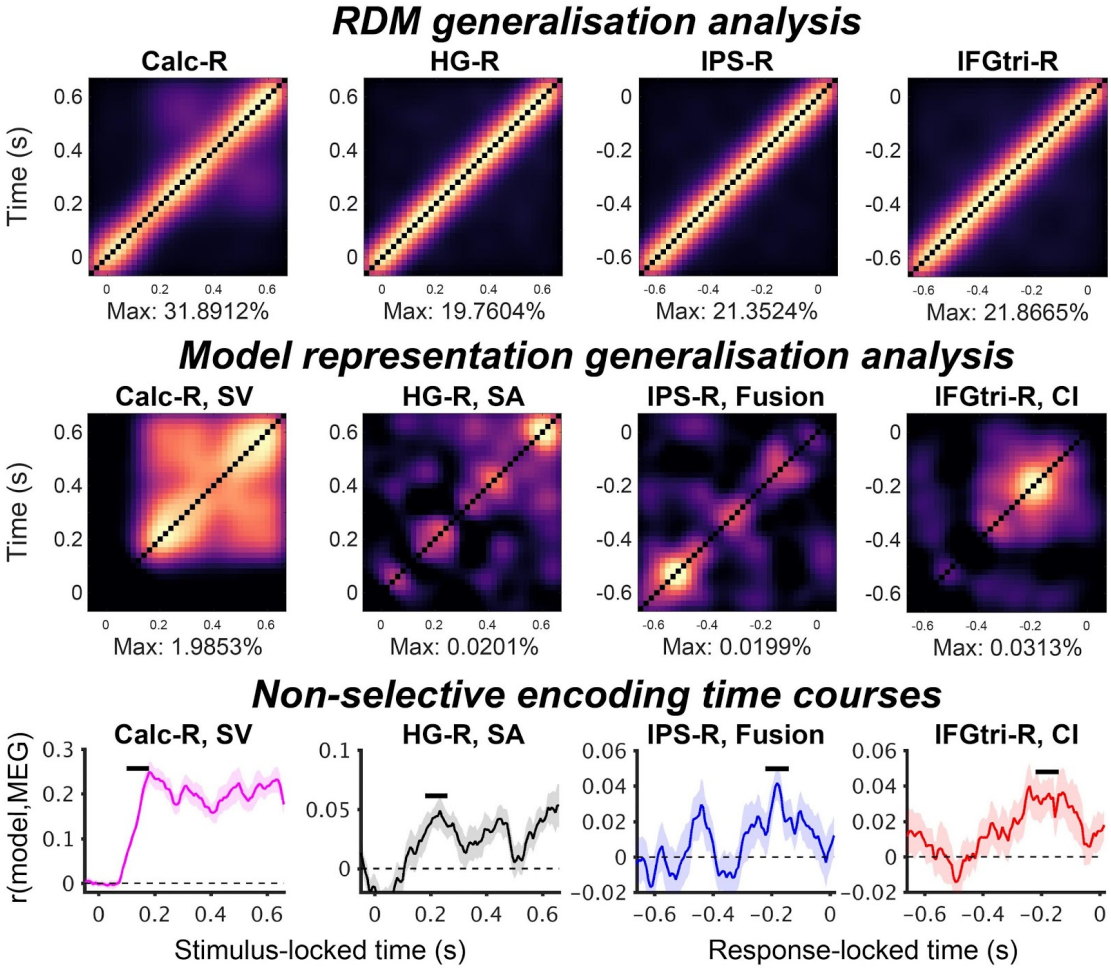


Figure S3.9. Temporal generalisation of neural RDMs and of model representations (RDM and model variance shared between pairs of different-time searchlights, respectively). Colormaps in the time × time variance maps range from 0 to the 95th percentile. The maximum variance is reported below each map. The last row reports the non-selective encoding time course for each ROI to facilitate their comparison with the structure of the maps measuring the generalisation of model representations (from **Figure S3.4**). For both analyses, we initially computed variance

components independently for each participant. The "time × time" maps of participant-specific variance were then converted to Fisher's Z scale for averaging.

Crossmodal bias versus error rate

A traditional measure of accuracy (error rate) was not focused in the analysis of behavioural data. However, we argue that the crossmodal bias essentially serves as a continuous (parametric) quantification of performance error. In fact, the bias is a more sensible measure than percent correct in capturing the nature of the crossmodal interaction as it is directly and parametrically related to the stimulus dimension.

To support this, we used the error rate as an alternative measure of behaviour and found largely the same results as obtained using crossmodal bias (**Figure S3.10**). However, and as expected, error rate proves ineffective for capturing the nonlinear quadratic effect of disparity: although the mean effect of squared disparity is still negative (as expected), the effect is less consistent across participants (**Figure S3.10**, right panel; in comparison with **Figure 3.3E** in the paper). This is because for larger disparities the bias can capture well the responses that are closer to the task-relevant target rate despite not necessarily matching the target rate. By contrast, error rate cannot distinguish between these subtleties given its discrete nature. For this reason, we refrained from analysing performance accuracy and focused on bias, which has been in fact more widely used in previous studies using similar paradigms (e.g., Angelaki et al., 2009; Rohe & Noppeney, 2015a, 2015b, 2016).

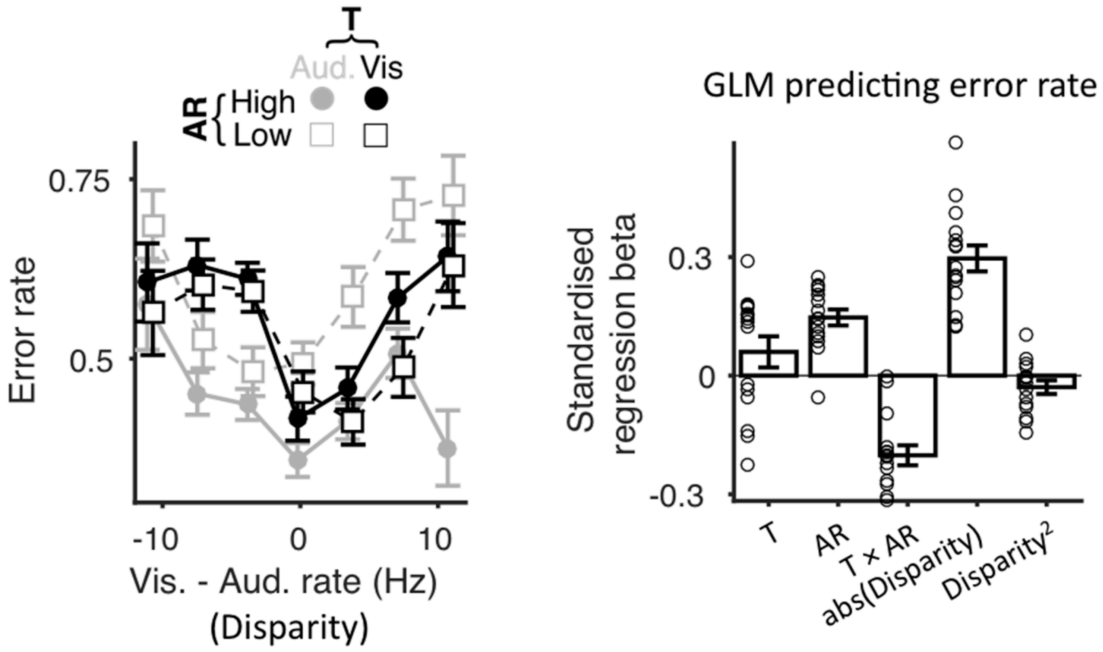


Figure S3.10. Using Error rate as behavioural measure, the negative effect for squared disparity is still observed but becomes less consistent across participants.

Multivariate noise normalisation for MEG distance estimation

It has been well-known that for multivariate neuroimaging data modelling, it is important to take into account signal variations (some voxels are noisier than other voxels) as well as noise correlations (neighbouring voxels are more likely to be characterized by correlated noise structures). For MEG signals, the noise covariance structures should involve the temporally connected samples. An estimate of the structure of the noise (Σ) was obtained from the residual structure of an GLM estimation. Then, the response estimate coefficients for condition $\boldsymbol{b_k}$ can be 'pre-whitened' as follows:

$$\boldsymbol{b}_k^w = \boldsymbol{b}_k \Sigma^{-1/2}$$

Σ was regularized by shrinking it towards the diagonal matrix (using optimal shrinkage factor, i.e., the parameter that minimises the expected squared loss of the covariance estimator (see Ledoit & Wolf, 2004; Walther et al., 2016).

Consequently, this multivariate noise normalisation procedure renders the noise component of the voxel response patterns approximately independent and identically distributed, both spatially and temporally across the voxels and across time points within each RSA searchlight.

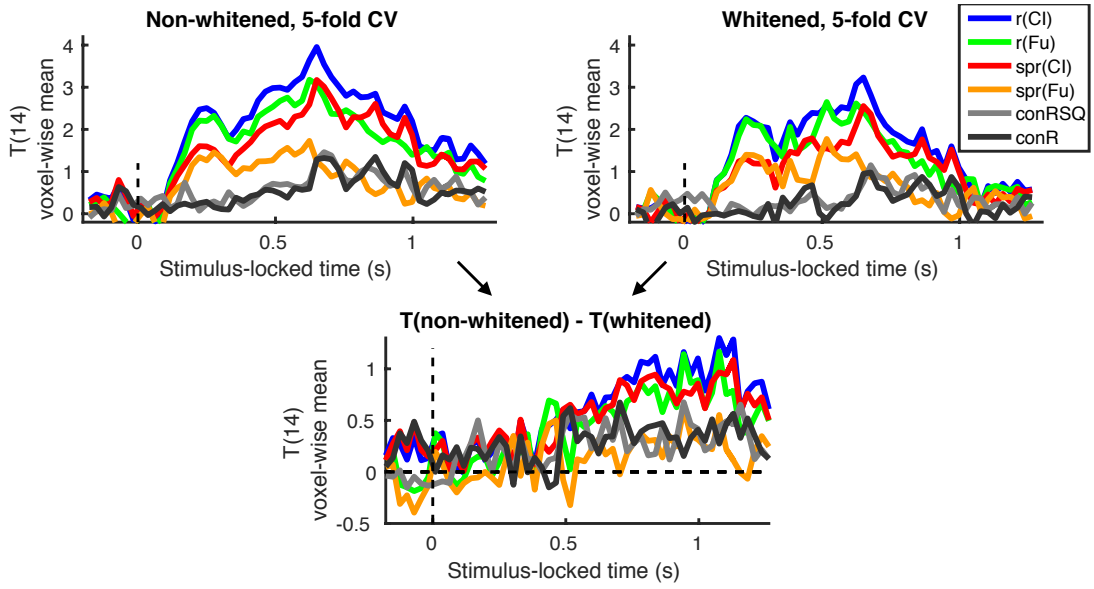


Figure S3.11. The effect of spatiotemporal pre-whitening (multivariate noise normalisation) on the T(14) time courses of RSA model encoding.

Control analysis using real-time eye-tracking data

In spite of the fact that the task performed by the participants was of non-spatial nature, i.e., temporal rate judgement, a control analysis was carried out here to rule out any possibility that the model encoding results were driven by

electrooculographic activity (Dijkstra, Mostert, de Lange, Bosch, & van Gerven, 2018; Mostert et al., 2018). To this end, time-varying RDMs of eye-movement activity were computed using the condition-wise cross-validated Euclidean distance of horizontal and vertical positions of the eye gaze collected using an EyeLink 1000 Eye tracker (SR Research) in real-time along the MEG recording. This was done in the same fashion as in the MEG RSA, except for the pre-whitening since there were only 2 channels of eye data. Then, T(s.p.r) statistics across participants for each computational model (CI, Fusion, Segregation Auditory, and Segregation Visual) was derived. As shown in **Figure S3.12**, the encoding timecourses are flat and noisy, with low T values throughout the whole analysis window locked to the stimulus onset.

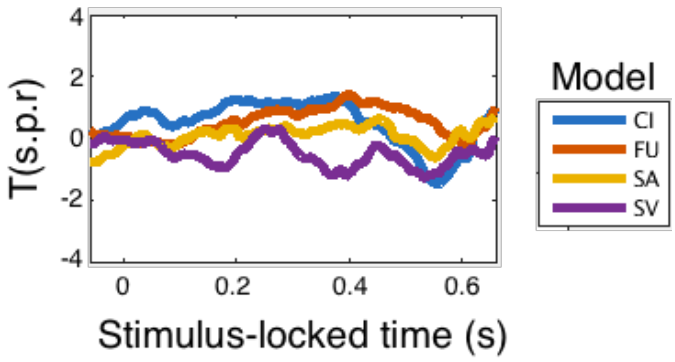


Figure S3.12. Control analysis of eye-tracking data. T(14) statistics of s.p.r selective-encoding timecourses for the sensory representations associated with different candidate computational models derived using *x* and *y* eye-gaze positions (i.e., brain RDM is replaced with eye-gaze RDM constructed from bivariate eye-movement data).

Dissociating model-predicted sensory representation from motor activities

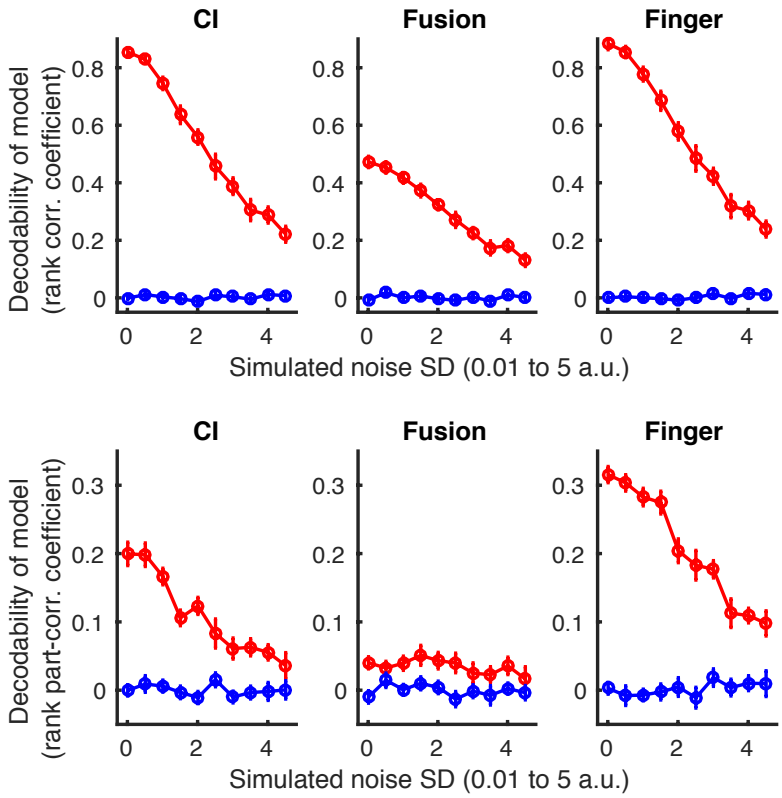


Figure S3.14. Simulation result of the ERF-correction procedure controlling for motor-related activities in MEG signals. Red: RSA effect of model encoding using raw neural activity containing pure motor-related signals, blue: RSA effect of model encoding of model using ERF-corrected activity. Horizontal axis shows different levels of condition-irrelevant basline noise in the neural activity. Upper panels: non-selective RSA effect (rank correlation, r), lower panels: selective RSA effect (semi-partial rank correlation, s.p.r). Error bars represent SEM across 15 subjects. CI: causal inference model, Fusion: reliability-weighted fusion model, Finger: condition-specific mean motor response.

In order to elaborate the ERF-correction approach that has been used to dissociate motor activity from condition-specific sensory representation, I provide a simulation here. Twenty-two blocks of artificial MEG responses were simulated as a function of finger code by contaminating the true block-specific behavioural responses (i.e., button indices: 1, 2, 3, 4) with trial-by-trial random Gaussian noise. The resulting neural pattern would thus contain no condition-specific pattern about information representation except for the finger representational geometry. The ERF

correction was then applied to these simulated MEG data by subtracting from each trial (condition) the mean across those conditions associated with the same finger code, in other words, by mean-centring the data within fingers. I then cross-validated the simulated MEG pattern using exactly the same 5-folds approach as in the main analysis. I didn't whiten the data because the current simulations were based on a hypothetical single vertex at a single time latency. I finally decoded three model RDMs from the simulated neural pattern using Spearman semi-partial rank correlations. The three models were CI, Fusion and a Finger model (condition-wise Euclidean distances between 22 block-specific multivariate observations). **Figure S3.13** shows that without ERF correction the finger model and CI model are saliently represented in the simulated MEG particularly when the simulated noise level is low (red curves). This observation is due to inter-correlation between these 2 models (i.e., the CI model explained choices well). However, the ERF correction removes the neural pattern that is specific to fingers, thus the association between finger geometry and neural pattern becomes close to zero despite slight fluctuation due to random noise in the simulations. CI model is also not decodable after ERF correction because the remaining neural pattern is just Gaussian noise.

The results show that the ERF-correction approach dissociates the neural encoding of causal inference from motor plans, by taking the advantage of the condition-rich design, which makes it possible for a finger to be used in many distinct experimental conditions.

Neural representation of posterior belief about common cause

In order to investigate when and where the neural activities reflected the belief over causal structure, the CI predicted posterior belief $p(c = 1|r_v, r_a)$ (see **Eq. 8** in

chapter 1) was mapped to the source-localised MEG pattern activities. I took the log-odds (i.e., log posterior ratio) for the whole distribution (20,000 samples), and then take the mean of log-odds, i.e., $\log(p(c = 1|r_v, r_a)/p(c = 2|r_v, r_a))$ across these samples for each condition. The posterior-belief RDM was constructed based on point estimates of the condition-specific mean Log odds.

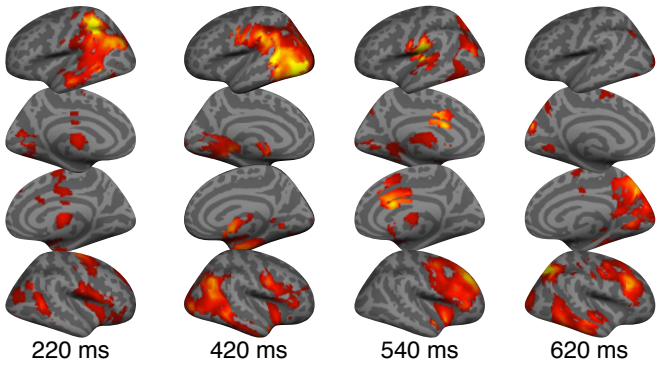


Figure S3.14. Neural representation of CI model-predicted posterior belief of common cause. The activations represent the significant effects that were significant at p <= 0.05 FWE-corrected. The maps show peak T-value of each voxel across time within the respective epoch for visualisation purpose. The time points represent the centre of RSA windows locked to stimulus onset (t = 0).

Neural representation of the log-odds appears to progress from sensory and parietal regions at earlier latencies (220 to 420 ms post-stimulus onset) towards more anterior regions such as the frontal cortex at later latencies (540 to 620 ms). On one hand, these later representations in the frontal cortex appear to overlap with the anatomical locations and time windows of the neural representations of the CI representations revealed in the main analysis. This is interesting because the average correlation between CI and posterior-belief RDMs was only 0.22 across participants. On the other hand, the posterior-belief representation seems to start at early latencies in sensory areas (occipital and temporal regions around 220 ms; **Figure S3.14**). It can be speculated that this could due to either a genuine selective representation of the

belief of common cause or a reflection of the variance explained by the physical disparity of audiovisual rates, which is correlated with the model-predicted posterior belief. Future work could consider orthogonalising experimentally the physical and subjective disparity, e.g., to collect a sufficiently large number of trials (for robust estimation of behavioural and neural RDM) for audiovisual stimuli with a set of fixed physical disparities, and assess the neural encoding of subjective belief of common cause (e.g., via explicit report or rating) *within* each physical disparity.

Chapter 4. Behavioural readout of temporal unfolding in multisensory inference

Chapter abstract

Recent human neuroimaging studies (**chapter 3** MEG study; also see Aller & Noppeney, 2019) have shed light on a systematic temporal sequence whereby the neural representations of sensory segregation and reliability-weighted fusion were succeeded by those of CI. A remaining challenge concerns the link between these neural representations decoded from macroscopic brain activities and their behavioural relevance. In this chapter, using paradigms such as speeded-judgement under deadline-induced time pressure (Behavioural Experiment I) as well as sensory masking (Behavioural Experiment II), I explicitly quantify how the unfolding of neural representations can be read out to drive behaviour. Collectively, slower vs. faster behavioural choices reflect more of CI vs. sensory fusion, respectively, indicating that the mapping from neural representation to behavioural choice corroborates the temporal hierarchy revealed in the neural data.

4.1 Chapter introduction

Human multisensory inference has recently been characterised as involving sensory fusion, segregation, or a flexible weighting between fusion and segregation by virtue of sensory 'causal inference' (CI; Körding et al., 2007; Rohe & Noppeney, 2015b). One theoretical work suggests that multisensory inference is a monolithic

process whereby different types of neurons maintain sensory fusion and segregation concurrently in a network model (Zhang et al., 2019). An alternative view, however, is that the computations underlying fusion and CI are structured in time, so that different processes dominate at different post-stimulus latencies. This view postulates that the brain rapidly derives a fused sensory estimate for computational expediency and, later and if required, filtering out irrelevant signals based on the inferred sensory cause(s) for bias minimisation. There has been emerging and converging neural evidence for a temporal structure of this transition between the separate processes (Aller & Noppeney, 2019; Cao et al., 2019; Rohe, Ehlis, & Noppeney, 2019). Leveraging multivariate analytical approaches to time-resolved imaging data, these recent studies have been able to reveal a detailed picture of how the human brain forms the neural representations reflecting sensory fusion and CI in a sequential manner in time and across different brain regions (see **chapter 3**).

Critically however, an open question remains to be addressed as to when and how the brain reads out these candidate neural representations to guide flexible behaviour. A common conclusion drawn across these past imaging studies has been that CI almost completely dominates over sensory fusion in describing the behavioural outputs whilst the mind keeps an explicit neural code for the dual processes (Aller & Noppeney, 2019; Cao et al., 2019; Rohe et al., 2019; Rohe & Noppeney, 2015b). This naturally brings us to the question—in what circumstances could the early fused sensory estimates be used to drive behaviour? One possibility is that if rapid decision-making is explicitly requested under greater time pressure, a higher probability of sensory fusion could be observed in the behavioural readout. This prediction had not been tested systematically, at least in part due to the

assumption about a stable tendency towards multisensory integration modelled as long-term prior (Odegaard & Shams, 2016; but see Körding et al., 2007).

Using a multisensory spatial judgement task under deadline-induced speed pressure, we show that a Bayesian CI model describes the behavioural responses pooled over all trials, but the tendency of fusing multimodal signals progressively increases as response speed increases, with faster choice approximating to a linear fused estimate. The findings presented in this chapter suggest a systematic behavioural readout of the structured temporal dynamics of cue combination during a single decision, enriching the view of the multiplexed timescales of multisensory plasticity across trials (Rohe et al., 2019; Wozny & Shams, 2011), days (Ernst, 2007; Seitz, Kim, & Shams, 2006), and the lifespan (Murray, Lewkowicz, Amedi, & Wallace, 2016).

4.2 Behavioural Experiment I

4.2.1 Methods

4.2.1.1 Participant details and sensory stimuli

Twenty-two human volunteers (M = 23.6, SD = 3.5 years of age; nine males) participated in this study. All participants had normal or corrected-to-normal vision, reported normal hearing, and gave informed consent prior to the study. The study was conducted in accordance with the Declaration of Helsinki and had ethical approval from the University of Oxford (MS-IDREC-C1-2015-236). The experiment was conducted in a sound-attenuated booth. Visual stimuli were transient spatial signal presented on a 23'' display (Dell UltraSharp U2312HMt; 60 Hz refresh rate; dark grey background). This visual spatial signal consisted of a two-dimensional Gaussian

cloud of 25 white dots (0.25° diameter for each dot; see **Figure 4.1A** below). The standard deviation (SD) of the Gaussian cloud along the vertical dimension was fixed at 1.5°, whilst that along the horizontal dimension was varied across trials to manipulate the reliability of the visual signal about its centre location. Thus, there were 2 different levels of visual reliability controlled by trial-varying horizontal SD of the visual cloud (high reliability: 1.5°, low reliability: 8°). Acoustical stimuli were delivered binaurally via Sennheiser HD650 headphones at an RMS level of 75 dB SPL (sampled at 48 kHz). The spatial sound stimuli were generated by convolving a white-noise burst with a set of head-related transfer functions (HRTFs; the CIPIC database, Algazi, Duda, Thompson, & Avendano, 2001). The sound onset and offset were cosine ramped (3 ms). Stimulus presentation was controlled via Matlab (Mathworks, Inc., Natick, MA) using routines from the Psychtoolbox 3.0.12 (Brainard, 1997). Participants were seated in front of the computer screen (viewing distance of 35 cm) and with their chin supported by a chinrest.

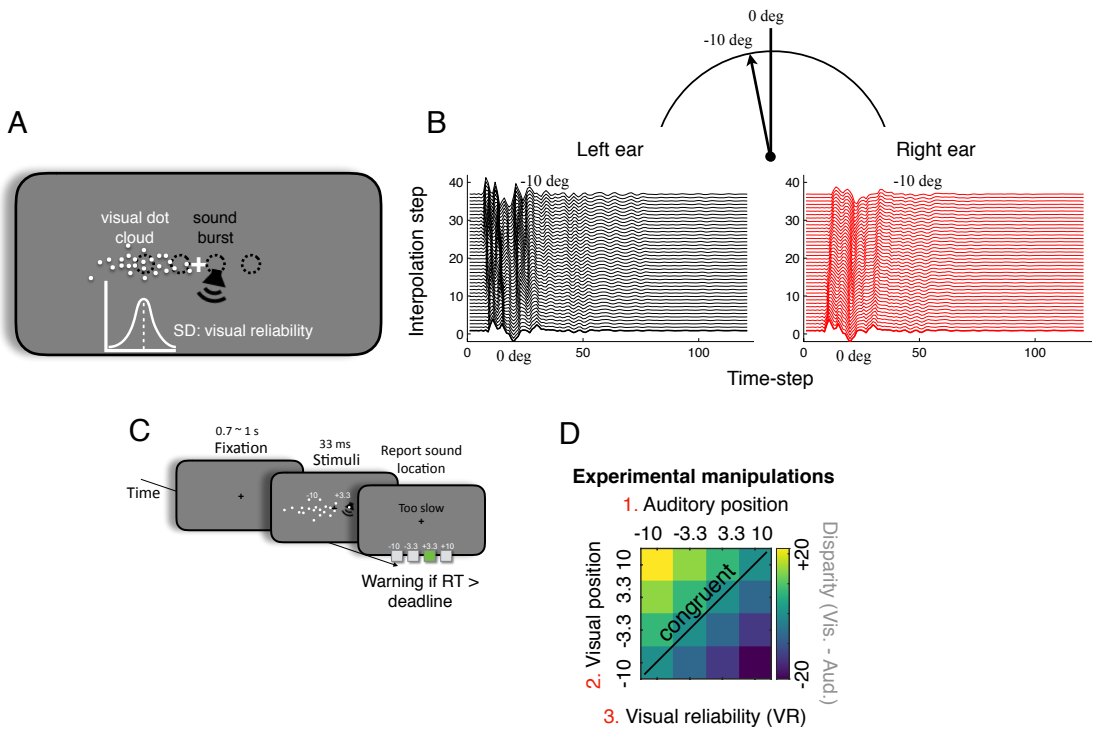


Figure 4.1. Experimental stimuli and design. (A) The visual signal consisted of 25 white dots on a dark background, whose spatial positions were draw online on each trial from a two-dimensional Gaussian distribution centred on one of the four hidden source locations, and with variable horizontal standard deviation controlling for the reliability of visual signal about its central source location. **(B)** Illustration of the head-related transfer functions (HRTFs) generated by interpolating in the time domain between 2 extreme HRTFs measured using sound sources placed at -10 and 0 degrees relative to the midline. **(C)** Sequence of speeded auditory localization task. Participants were required to report the location of a transient auditory white-noise burst as accurately and as quickly as possible. Only auditory signal (speaker symbol) needs to be reported (correct button is marked in green for illustration purpose but not shown to participants). Warning feedback was only shown for RT exceeding a deadline (see Methods). (**D**) Experimental manipulation features a 4 (visual location)-by-4 (auditory location)-by-2 (visual reliability) design, resulting in 32 different multisensory conditions.

4.2.1.2 Task and experimental manipulations

In an auditory-localisation task, participants made speeded judgements concerning the horizontal location of a transient auditory signal (noise burst played from virtual sources) in the presence of a visual spatial signal by pressing one of four

buttons (**Figures 4.1A and 4.1C**). The task comprised thirty-two conditions, defined by the factorial combination of four locations for the visual and auditory (−10°, −3.3°, 3.3°, 10° relative to the midline; varying along the azimuth), and two levels of visual reliability (high vs. low, as determined by the horizontal standard deviation of the Gaussian cloud, 1.5° vs. 8°, respectively). The audiovisual spatial signals were onset-synchronized and were 33.3 ms in duration. The participants were instructed to maintain fixation on a small cross (white; 0.4°) in the centre of the screen. On each trial, the audiovisual spatial signals were presented following a random fixation delay, sampled from a truncated exponential distribution (range: 0.7–1.2 s; mean: 0.82 s). No feedback on the correctness of participants' spatial judgement was provided. Within each block, each of the 32 experimental conditions was repeated 4 times, and all conditions were presented in a random order.

To encourage speeded responses, an adaptive staircase procedure (3-down/1-up rule) was implemented to adjust the RT deadline on a trial-by-trial basis, maintaining the probability of exceeding the deadline on each trial at around 0.206 (García-Pérez, 2011). The RT deadline was initialised at 800 ms, decreased by 40 ms every time participant successfully responded within the deadline on 3 consecutive trials, and increased by 38 ms for the next trial if exceeding the deadline (note that these trial-by-trial step sizes for RT increase and decrease follow an optimal ratio suggested by theoretical work, García-Pérez, 2011). Participants were shown a warning message "Too slow" if they exceeded a deadline, but only when the deadline was longer than 350 ms post-stimulus onset (this is similar to the principle of the "carry-on boundary," García-Pérez, 1998) for the purpose of preventing participants from sacrificing accuracy for extremely fast responses. Each participant completed 14–17 blocks (≥ 1792 trials in total; about 60 repeated trials for each of the 32

conditions) on a single day. Before starting this main task, all of the participants performed an initial training session with unisensory stimuli (localisation for auditory-only or visual-only stimuli).

4.2.2 Results

4.2.2.1 Fast choices under time pressure approximate to linear fusion

We tested the hypothesis that the temporal sequence of neural representations underlying fusion and CI can be read out in behaviour as a function of reaction times (RTs). A key behavioural pattern that distinguishes fusion from the CI process is the relationship between crossmodal bias (i.e., deviation of participants' response from the actual auditory location) and crossmodal disparity (distance between visual and auditory locations). Specifically, we expected that faster responses would be characterised more by a linearly growing bias as a function of disparity as predicted by the reliability-weighted fusion, whereas slower choices would be characterised more by a nonlinear bias as predicted by CI.

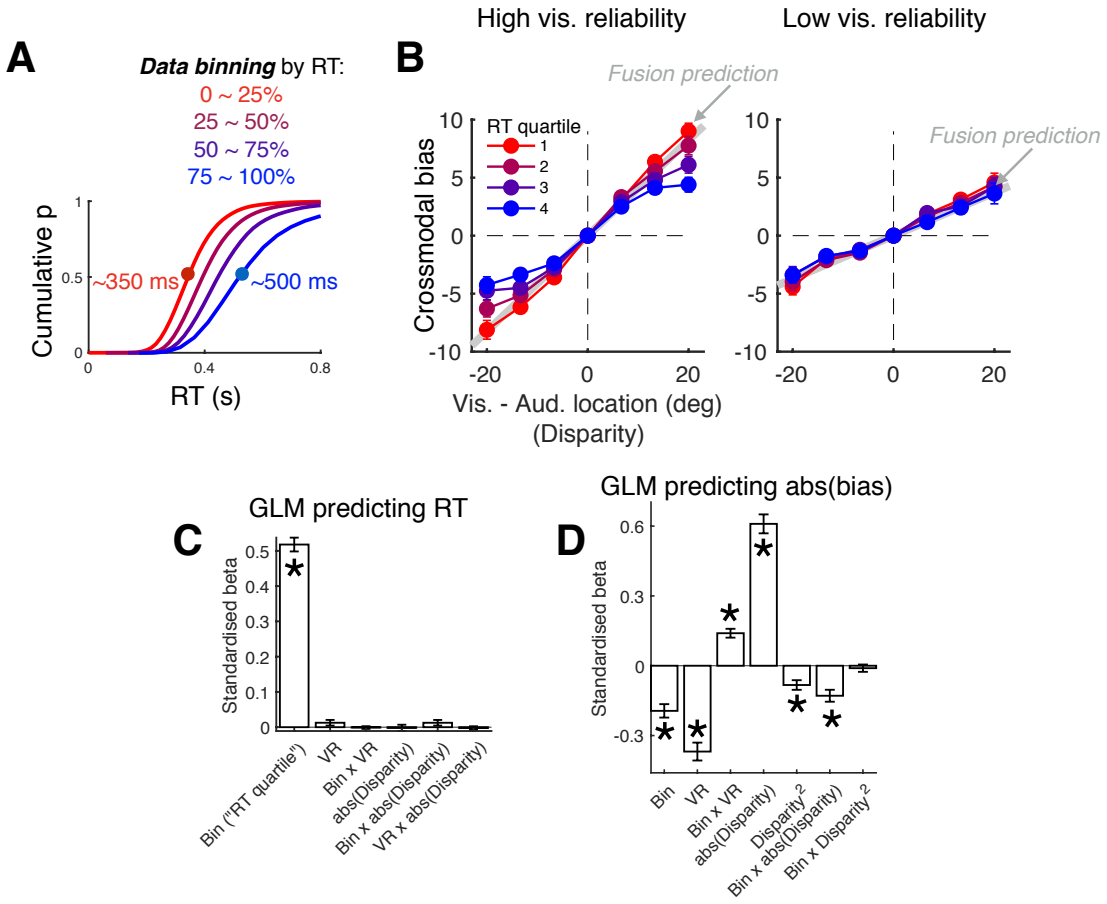


Figure 4.2. Behavioural analysis. (**A**) Cumulative distribution functions of RT for quartile data bins, aggregated for all participants and conditions. Dots mark mid-point RTs. (**B**) Crossmodal bias over auditory localisation as a function of disparity, reliability and RT bin. Units = degree. Circles and error bars = mean and ± SEM across participants (N = 22). Grey diagonal lines = group-mean prediction of a reliability-weighted fusion model fitted to the fastest RT bin. (**C**) Standardised regression coefficients quantifying the influence of RT bin, visual reliability (VR; low minus high), absolute disparity, and interactions among these on trial-by-trial RT. (**D**) Standardised regression coefficients quantifying the influence of RT bin, VR, and the linear and quadratic effects of the absolute disparity, and interactions among these on the absolute bias. Asterisk = significant effect (2-sided permutation tests; FWE = 0.05), error bars = ± 1 SEM.

Hence, we sorted individual observer's responses according to their RTs and assigned each trial to one of 4 bins (quartile splits; **Figure 4.2A and 4.2B**), independently for each experimental condition and for each block. To assess the extent to which different conditions were assigned trials with vastly different RTs, we used a general linear model (GLM) to predict how the actual RT depended on the RT

bin, visual reliability, crossmodal disparity, and the interactions between these (all effects were assessed using maximum-statistics permutation controlling for multiple comparisons, family-wise error FWE = 0.05). This GLM revealed the RT bin as the only significant effect on RT values ($t(21) = 26.5$, FWE-corrected $p = 0.0001$, mean $\beta = 0.52$, SEM = 0.0196), but not other effects ($abs[t(21)] \leq 1.56$, $p \geq 0.137$, mean $abs(\beta) \leq 0.013$, SEM ≈ 0.0066; **Figure 4.2C**).

Next, we used a GLM to predict how the magnitude (i.e., absolute value) of the crossmodal bias depended on RT bin, visual reliability, their interaction, as well as crossmodal disparity (**Figure 4.2D**). We included an effect of squared crossmodal disparity in this analysis to capture whether the bias scales nonlinearly with disparity, as predicted by the CI process. This nonlinearity was indeed significant for the overall responses ($t(21) = -3.99$, $p = 0.0015$, mean $\beta = -0.083$, SEM = 0.021). The bias was dependent on visual reliability, indicating reliability-weighted influence of the task-irrelevant visual cue ($t(21) = -9.59$, $p < 0.001$, mean $\beta = -0.37$, SEM = 0.039). Importantly, this GLM revealed a significant effect of RT bin ($t(21) = -6.72$, $p < 0.001$, mean $\beta = -0.194$, SEM = 0.029), and a significant interaction between RT bin and linear dependency of the bias on disparity ($t(21) = -5.03$, $p < 0.001$, 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 the prediction of a linear fusion principle for faster RTs (**Figure 4.2B**; the fusion model was fit to the fastest RT bin for this illustration). The mean beta coefficient for the interaction between RT bin and squared disparity was negative (mean $\beta = -0.011$), indicating that a tendency of increasing nonlinearity for slower RT, but this interaction was not significant ($t(21) = -0.66$, $p > 0.1$, SEM = 0.015). A GLM applied to an alternative median RT split of the data produced the same results (see **Figure 4.3A** and **4.3B**).

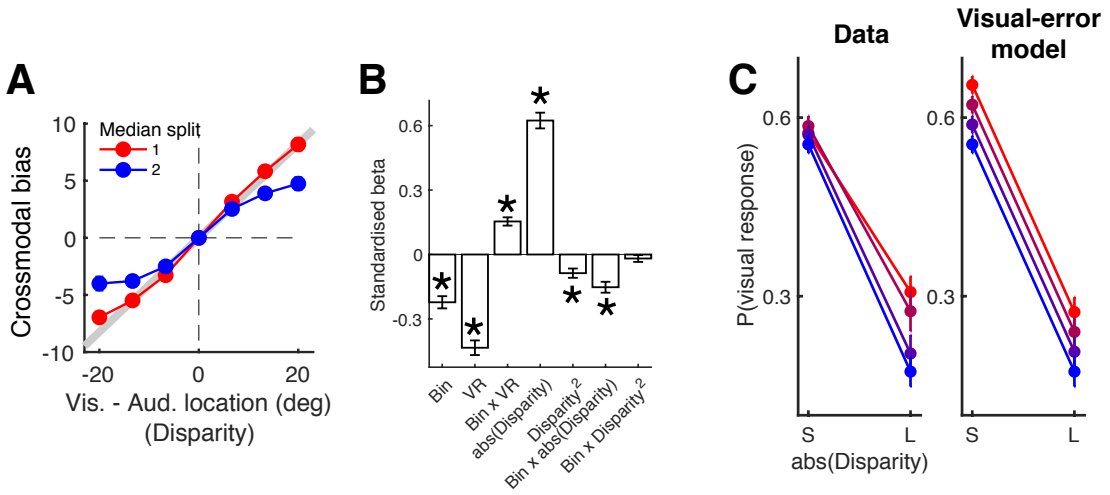


Figure 4.3. Control analyses for behaviour in Experiment I. (A) Crossmodal bias over auditory localisation for median RT split (Figure 4.2B). **(B)** GLM predicting absolute bias (Figure 4.2D). (**C**) Probability of reporting task-irrelevant visual cue as a function of absolute disparity and RT level (left panel), and qualitative predictions of an alternative visual-error model (right panel; see text). S = small disparity, L = large disparity, error bars = ± 1 SEM.

Moreover, these results persisted after controlling for the difference in the probability of responding to (task-irrelevant) visual location across RT bins for each incongruent condition (**Figure S4.1**). In addition, comparable RTs in unisensory practice trials (visual-only or auditory-only measured separately) can be treated as evidence rejecting any salient imbalance between unisensory processing of the spatial cues as the main account for the observed time-dependency of the bias (**Figure S4.2**). We also ruled out the possibility that the time-dependency of the bias was solely due to a more frequent tendency to erroneously report the visual cue under time pressure, using a logistic regression. We compared the actual data to the prediction of a 'visual-error model' for the probability of reporting the visual location (**Figure 4.3C**). This visual-error model describes the behaviour of an agent who presses more often the button corresponding to the visual location under deadline-induced time pressure

(e.g., analogous to the Colavita visual dominance effect; Spence, Parise, & Chen, 2012). To simulate the behaviour of this visual-error model, I considered the actual responses of the participants in the 4th RT bin (slowest; blue lines in **Figure 4.3C**) as a baseline and artificially increased the probability of choosing the visual location by the same amount across all conditions for the data in faster RT bins. From the 3rd to the 1st RT bin (under quartile split), the visual response probability was increased linearly by 0.033, 0.067, and 0.1, respectively. This visual-error model can mimic the effect of an increasing average crossmodal bias for faster RT bin, but it predicts no interaction between disparity and RT because this model disregards different conditions (i.e., parallel lines for different RT bins; **Figure 4.3C right**). As shown in **Figure 4.3C left**, however, the abs(Disparity) × RT bin interaction was indeed significantly predicting the probability of reporting visual location in the real data (logistic regression: $t(21) = 5.95$, $p < 0.001$, mean $\beta = 0.12$, SEM = 0.02), regardless of whether the small disparity condition included the congruent conditions or not (see **Figure S4.3B:** $t(21) = 6.51$, $p < 0.0001$, mean $\beta = 0.105$, SEM = 0.016). Collectively, these control analyses ruled out the alternative explanation that the higher crossmodal bias (more fusion) in faster decisions was solely induced by mistakenly reporting a task-irrelevant cue under time pressure.

4.2.2 Faster choices reflect higher tendency to fuse multisensory signals

Next, we sought to understand the strategies participants used to combine audiovisual cues for the spatial localisation task by using computational modelling. As shown in **Figures 4.2D** and **4.3B**, the average growth rate of crossmodal bias

diminishes for discrepant audiovisual information that is unlikely to originate from a common cause (i.e., reflecting a nonlinear dependency of the bias on disparity) as captured by the significantly negative effect of squared disparity. We thus began by describing this nonlinear bias in overall behavioural responses using a Bayesian CI model. The 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 sound location. The model was parameterised with the variances of the two visual sensory likelihoods associated with different reliabilities ($\sigma^2_{\text{vis1,}}\, \sigma^2_{\text{vis2}}$), the variance of the auditory sensory likelihood (σ^2_{aud}), a probabilistic belief that the cues had a common cause (p_c, ranging from 0 to 1; "integration tendency," hereafter), and the variance of a spatial centrality prior (σ^2_p) capturing potential bias to respond preferentially towards the central locations. A CI model with model-averaging strategy (see **chapter 1**: **Eq. 10**) outperformed the CI models with alternative decision strategies (BIC score pooled over participants for 'probability-matching' and 'model-selection' relative to model-averaging was 62.02 and 410.7, respectively; see **chapter 1**: **Eq. 11**). This best CI model with model-averaging strategy also outperformed a fusion model (with fixed $p_c = 1$; relative BIC: 1847.5) and a segregation model (with fixed $p_c = 0$; relative BIC: 9114.8), consistent with previous reports (Cao et al., 2019; Rohe & Noppeney, 2015a). Hence, only the model-averaging CI model with a free p_c is able to capture the nonlinear bias in participants' overall behavioural responses (mean $R^2 = 84.1\%$, SEM = 1.04%).

Further, the key question for our purposes is how the changes in crossmodal bias as a function of RT observed in behaviour can be reflected in the parameters of this CI model, which represent distinct component processes of CI. Specifically, the integration tendency (p_c) describes a top-down belief about the probability of

audiovisual inputs sharing a common cause, whereas the variance of sensory likelihoods describes noise at the sensory processing stage and the centrality bias describes response prior, respectively. To identify the origin of the changing crossmodal bias as a function of RT, we compared the goodness-of-fit of seven CI models, each allowing only one parameter to vary freely or fixing all parameters across the two RT bins (i.e., based on median split; see **Figure 4.3A**). In addition to the five basic model parameters, a motor noise parameter was introduced (practically achieved by adding Gaussian noise with variance of σ^2_{motor} to the task-relevant spatial estimates $\hat{S}_{\text{aud}}$) in order to assess whether the RT-dependent crossmodal bias was due to change in variability of action selection. The best fitting CI strategy for overall responses (model-averaging) was used for these seven models, but the model comparison results were similar if using other strategies or if adopting the best decision strategy for each participant (12 participants were best described by model-averaging, 6 by probability-matching, and 4 by model-selection).

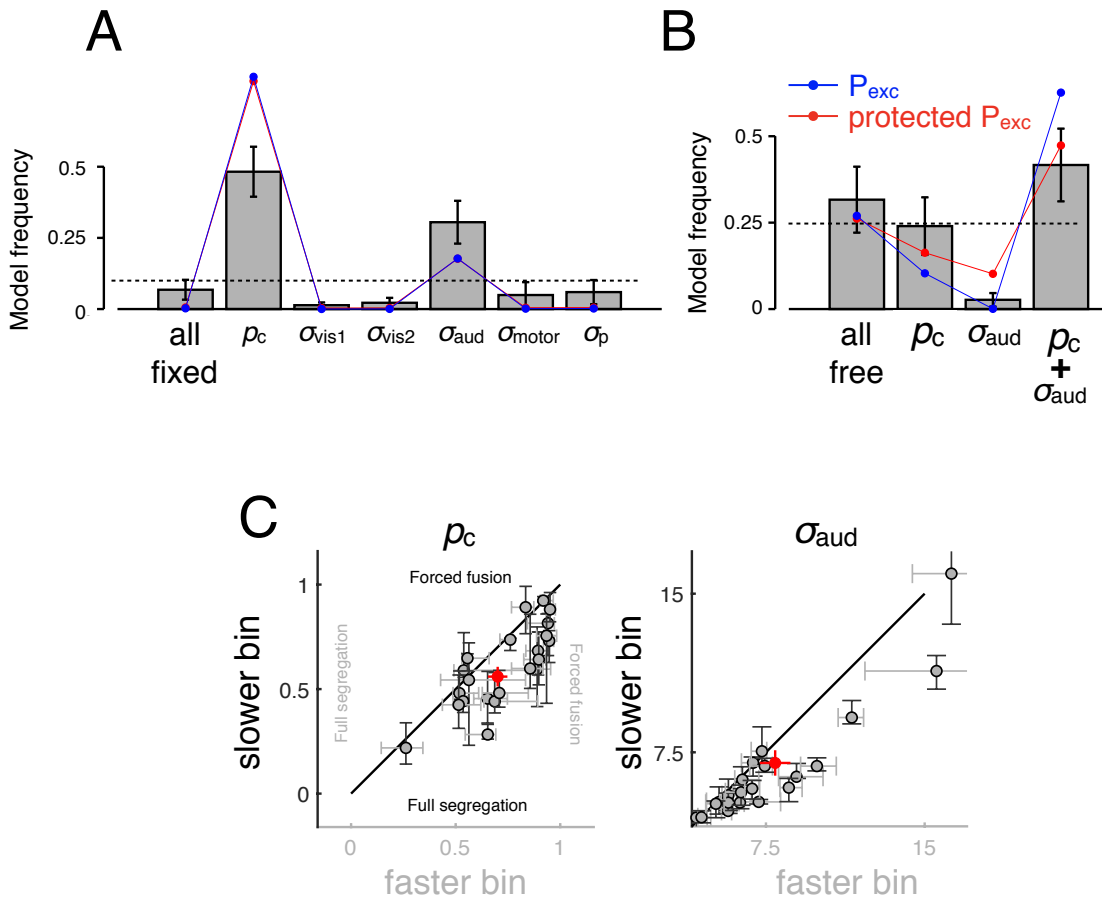


Figure 4.4. Computational modelling. (**A**) Random-effect model comparison for 7 CI models, each allowing only one parameter to vary freely (or fixing all parameters, as shown by the left most bar) across the two RT bins. The blue and red dots indicate the exceedance probability (P_{exc}) or protected P_{exc} (controlling for chance level) of a model against other models at group level. (**B**) Random-effect model comparison for 4 CI models, either containing both free integration tendency and free auditory noise, only one of these, or more free parameters (see Table S1 for details). **(C)** Probability of common cause (p_c, integration tendency) and the standard deviation of auditory noise estimated from the full model with 11 free parameters. Grey markers indicate medians; error bars represent interquartile range (25% to 75% percentile) of 10 best estimates (sorted by Log-Likelihood) among 2,000 Monte Carlo model fits with different initial parameter guesses. Red markers represent the grand mean of the medians (error bars: ± 1 SEM).

A Bayesian random-effects model comparison showed that introducing a time-varying integration tendency (p_c) or a time-varying auditory variance σ^2_{aud} significantly improved the model fit relative to a baseline model in which all model parameters were fixed, whereas other models featuring time-varying visual sensory

likelihoods, motor noise or spatial prior reduced the model fit due to increased model complexity (**Figure 4.4A**). The model with time-varying p_c had the highest exceedance probability ($P_{exc} > 0.8$; the probability that a model is more likely than other models at the group level; Stephan et al., 2009) and the highest expected posterior model probability (or model frequency, bars in **Figure 4.4A**; the probability that a given model generated the data for a randomly selected participant) among all model variants. The prominent role of this time-varying p_c in explaining the observed changing crossmodal bias as a function of RT was further emphasised when assessing a model with both time-varying p_c and σ^2_{aud} (**Figure 4.4B**; $P_{exc} = 0.63$). Introducing time-varying p_c on top of a time-varying σ^2_{aud} model led to a gain in model fit.

A final analysis assessed a model with more free parameters than the above models. The estimated integration tendency was stronger (higher p_c) in faster responses than in slower responses (two-tailed paired-sample Wilcoxon signed-rank test: $Z = 2.71$, Bonferrori-corrected $p = 0.033$, $r = 0.41$), and the estimated auditory noise was higher in faster responses than in slower responses ($Z = 3.29$, Bonferrori-corrected $p = 0.0049$, $r = 0.49$) (**Figure 4.4C**). The other parameters, by contrast, were not significantly different for the 2 RT bins (see **Table S1** and **Figure S4.4**).

4.3 Behavioural Experiment II

This experiment relies on a sensory masking method, and aims to test the predictions of a temporal sequence of multisensory representations. This mechanism of sequential computation has been evidenced in **chapter 3** (MEG study) and Behavioural Experiment I in the current chapter (decision-making under deadline-induced time pressure). Such a temporal sequence predicts that the neural

representations of a fusion estimate are more pronounced at earlier post-stimulus time than those of the Bayesian CI estimate whose computation requires maintaining posterior belief over causal structures and cue disparity across senses. The data in Behavioural Experiment I support a more pronounced behavioural readout of the rapid fusion representation in faster choices. I use Experiment II to corroborate the mechanistic account of sequential computation by ruling out alternative factors such as neural gain modulation induced by decision urgency or motor preparation signals (see discussion in this chapter). To this end, a visual mask is used to disrupt the unisensory visual representation after a delay relative to stimulus onset, long enough to permit rapid fusion (150 ms) but short enough to dampen CI in spatial domain (200 ms; see Aller & Noppeney, 2019). Within this paradigm, a delayed "go" signal also dissociates action selection from sensory processing and inferential stages.

4.3.1 Methods

Twelve healthy adults participated in Experiment II. On each trial, after a period of fixation (1 ~ 1.5 s) participants were presented with a 16.6-ms (1 visual frame) audiovisual spatial stimulus (**Figure 4.5**). The visual stimulus was a 25-dots Gaussian cloud of a single reliability level (horizontal SD = 4.75°, i.e., the mean value across the two SD levels used in Experiment I; vertical SD was fixed at 1.5°). The auditory stimulus was a 16.6-ms white-noise burst spatialised through convolution with HRTFs. There were 3 different types of conditions (differing in signal modality), randomly interleaved across trials. In the auditory-only condition, only the auditory burst was presented; in the visual-only condition, only the visual Gaussian cloud was presented; in the audiovisual condition, a compound signal was presented by synchronising the onsets of the visual and auditory stimulus. The frequency of encountering visual-only and auditory-only trials was relatively rare compared with

that of encountering an audiovisual trial (unisensory vs. multisensory: 12.4% vs. 75.2%; **Figure 4.5B**).

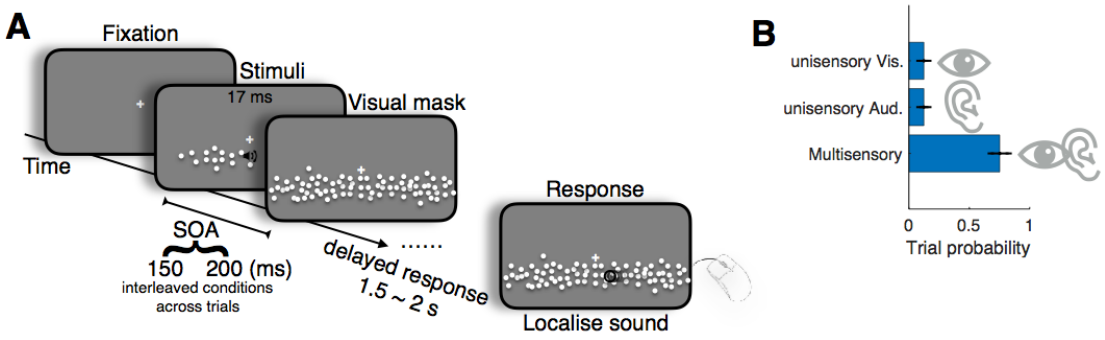


Figure 4.5. Task and design for Experiment 2. (**A**) Sequence of a single trial for the masking paradigm. Participants were required to move a computer mouse to control for horizontal position of a cursor on the screen and click to indicate judgement of the position of a transient auditory noise burst. The cursor did not appear until a variable delay of 1.5 ~ 2 s after stimulus onset. (**B**) Probability of different conditions.

Analogous to Experiment 1, there were multiple candidate hidden locations for the stimulus to appear (9 participants: −10°, −5°, 0°, 5°, 10° relative to the midline; 3 participants: −10°, −3.3°, 3.3°, 10°), and the locations of the signals in different modalities were independently manipulated to generate combinations of audiovisual cues with differing spatial congruency. Participants were required to only report the location of the auditory burst as long as they heard it (i.e., in the auditory-only and the audiovisual conditions), and they were required to report the visual location only in the visual-only conditions (because there was no sound to report; the number of visual-only trials was low: ~12%). The inclusion of infrequent unisensory trials made it possible to validate the effect of the visual mask (see below) and also served as a baseline measure capturing any potential tendency of responding to central locations as evidenced in past studies (Körding et al., 2007; Rohe & Noppeney, 2015a).

Regardless of the signal modality, a visual mask was presented 150 ms or 200

ms after the stimulus onset, and remained on the screen hereafter until the next trial. The visual mask consisted of an array of horizontal Gaussian clouds (each one was generated online independently; horizontal SD = 4.75°; vertical SD = 1.5°; centre distance = 2.5°) subtending the full width of the screen and superimposing on the visual stimulus cloud. The 2 levels of stimulus-to-mask onset asynchrony (SOA) constituted an orthogonal factor to other design factors (visual location, auditory location, and modality). Each condition was repeated around 10 times. Participants moved a mouse cursor to localise the auditory burst. Continuous response was used for the purpose of obtaining a sensible measure for true spatial localisation without changing the nature of the localisation task to a categorisation task (see Rohe & Noppeney, 2015a for relevant discussions). The cursor occurred only after a delay, which was 1.5 ~ 2 s relative to the stimulus onset. This was to dissociate the motor response from stimulus processing.

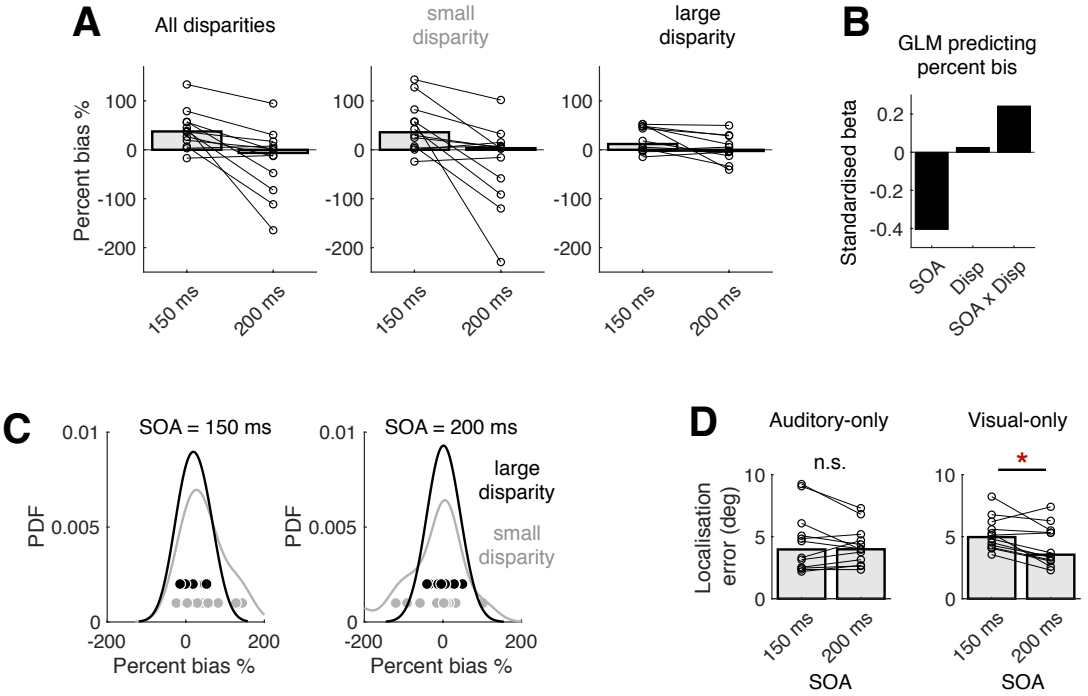


Figure 4.6. Data analyses for Behavioural Experiment II. (**A**) Percent bias, computed as (auditory response – auditory signal location)/(visual signal location –

auditory signal location) and converted to percentage scale. Conditions with congruent audiovisual signal locations are excluded to avoid infinite values. (**B**) a GLM predicting percent bias using stimulus-to-mask onset asynchrony (SOA; 150 ms vs. 200 ms), audiovisual disparity ("Disp": small vs. large), and their interaction. This GLM is used for qualitative demonstration for the mean positive trend for the SOA × Disp interaction over participants. (**C**) Distribution of percent bias over participants as a function of disparity and SOA showing the negative (repulsive) bias was mainly found in the small disparity condition at the longer SOA. (**D**) Localisation error in the unisensory conditions measured as the mean of the trial-by-trial absolute deviation of participants' response from the signal location.

4.3.2 Analyses and Results

In Experiment II, I sought to provide a complementary evaluation for the hypothesis that multisensory fusion is a relatively rapid process compared to the more elaborated CI process. A sensory mask was applied at different latencies during the course of the multisensory inferential process to interrupt the unisensory short-term memory trace. According to the computational principle of a CI process, computing and evaluating the potential disparity between unisensory cues is a crucial component mechanism, which per se relies on maintaining at least partial access to the individual sensory representations encoded in primary sensory cortices along with working memory. If the assumption about an early fusion process around 120~150 ms for spatial perception is correct, one would expect the masking effect at the shorter SOA (150 ms; presumably long enough for rapid fusion to happen) dampening CI to a greater extent (i.e., promoting fusion) than that at the longer SOA. To provide a systematic evaluation for this prediction, percent crossmodal bias was derived as a function of these experimental manipulations. Percent bias represents the relative crossmodal bias on the perceived auditory location 'in the direction' of the visual signal, providing a sensible 'signed' measure for CI when continuous localisation responses are collected (Körding et al., 2007; Rohe & Noppeney, 2015a; Wallace et

al., 2004). The percent bias was computed as participants' auditory localisation response in the audiovisual conditions minus the auditory signal location, then divided by the deviation of the visual signal location to the auditory signal location (i.e., visual signal location minus auditory signal location). To control for a potential response tendency towards the midline (Körding et al., 2007; Rohe & Noppeney, 2015a), the auditory and visual signal locations here were measured as the subjective location estimates from unisensory conditions for each participant and for each SOA.

The percent biases for the shorter SOA was significantly larger than those at the longer SOA (two-sided paired-sample test: $t(11) = 3.48$, $p = 0.0051$; **Figure 4.6A**). And this result held for an analysis separately for each audiovisual disparity level (small: $t(11) = 2.97$, $p = 0.013$; large: $t(11) = 2.27$, $p = 0.044$; **Figure 4.6B**). Overall, the percent biases at the shorter SOA per se were significantly larger than 0 ($t(11) = 3.54$, $p = 0.0046$; **Figure 4.6A**) whereas those at the longer SOA were not significantly different from zero across participants ($t(11) = -1.17$, $p = 0.27$). The SOA-induced difference in percent bias for the smaller disparity was slightly larger than that for the larger disparity, as suggested by a marginally significant contrast between the disparities for the SOA-induced difference in percent bias ($t(11) = 2.07$, $p = 0.062$), and qualitatively shown as a positive effect of the interaction between SOA and disparity upon the percent bias in a GLM of the percent bias over participants (**Figure 4.6B** is for illustration purpose only for this positive beta). As shown in **Figure 4.6A** (middle), this interaction between SOA and disparity was primarily due to the negative (repulsive) percent bias for some participants, which could be interpreted as originating from a well-known repulsion bias (particularly seen in continuous localisation response; Rohe & Noppeney, 2015a; Wallace et al., 2004; see **Figure 4.6C**) for small audiovisual disparity, but could also be attributable

to small sample size of participants and trials in the current experiment. Lastly, the mean absolute localisation error relative to the physical signal location was compared across different SOAs for the unisensory conditions (interleaved randomly among audiovisual trials; **Figure 4.6D**). The visual mask at the shorter SOA induced larger visual localisation error ($t(11) = 2.99$, $p = 0.012$) but did not affect the auditory localisation error ($t(11) = 0.95$, $p = 0.36$). The visual mask appeared to be disrupting effectively unisensory visual spatial representation as most participants reported uncertainty about the visibility of the visual cloud and about their visual localisation judgement.

4.4 Chapter discussion

I characterised the behavioural readout of a recently found sequential neural representation evolving from sensory fusion to CI in the human brain (Aller & Noppeney, 2019; Cao et al., 2019; Rohe et al., 2019). Our behavioural and modelling analyses reveal that human participants respond to multimodal signals differently under variable response speeds, with faster decisions increasingly dominated by a higher probability of fusing multimodal signals linearly weighted by their relative reliability whilst slower choices reflecting more of the Bayesian CI process. This greater fusion in faster decisions is associated with a larger crossmodal bias in the auditory spatial judgement induced by concurrent visual spatial signal. A quantitative modelling approach identifies dynamic integration tendency as constituting a prominent source of this time-dependent crossmodal bias during a single decision.

Revealing the direct behavioural relevance of the neural sequence underlying multisensory inference unfolding has crucial implications. First, establishing the

direct link between sensory representations in neural activity and whether and when these representations are used to guide behaviour facilitates the understanding of how component processes involved in flexible multisensory inference are implemented in the brain in a biologically plausible manner (Kriegeskorte & Douglas, 2018; Panzeri et al., 2017). Our finding of a systematic temporal dynamics in integration tendency whereby the process of reliability-weighted fusion was succeeded by that of CI suggests a cascade of behavioural readouts of the information formed under different inferential processes. This cascade suggests a specific computational scheme for CI at the systems level: CI effectively relies on a weighted combination of sensory estimates predicted by fusion and segregated signals from a posteriori estimates, with the relative contribution depending on the inferred degree of crossmodal disparity. Second, these results help unify previous studies on multisensory inference. The timing concerning early emergence of fusion in the fastest decisions appears to fit with past human imaging data that support neural correlates of reliability-weighted fusion around 120-150 ms post-stimulus onset (Aller & Noppeney, 2019; Boyle et al., 2017), taking into account additional motor delay. Along this line, previous behavioural studies also suggested that sensory fusion may be an automatic process, i.e., crossmodal bias tends to be stronger when human observers respond faster or after acquiring only little sensory evidence (De Winkel et al., 2017; Noppeney et al., 2010). Instead, CI relies on extra processing time as it capitalises on maintaining beliefs over more than one latent causal structures of the world. Indeed, a conscious segregation of multimodal signals that usually tend to be fused requires additional time and effort (Gau & Noppeney, 2016). A dual processing system with multiplexed timescales of distinct inferential strategies also reconciles the empirical neural data revealing sequential computations and theoretical framework suggesting concurrent

coding of integration and segregation (by virtue of congruent and opposite neurons; Zhang et al., 2019). The sequential process provides a possible mechanism whereby congruent neurons code for a default fusion computation (Gu, Angelaki, & DeAngelis, 2008) while higher-order brain structures (e.g., frontal cortex; Gu, Cheng, Yang, DeAngelis, & Angelaki, 2016) dynamically weight the connectivity to congruent and opposite neurons in other regions (e.g., temporal-parietal; Chen, DeAngelis, & Angelaki, 2013; Zhang et al., 2019) to perform flexible CI.

For the speeded judgement data of Experiment I, our model comparison reveals the higher tendency of projecting common-cause belief (p_c) over multisensory inputs as constituting an important source of the early fusion in behaviour. We also observed that the noise estimate of auditory sensory likelihood appears to be higher in faster estimates. This result may resonate with the general benefit of maintaining a higher p_c for fusing noisy inputs across multiple sensory channels in order to achieve better precision. Alternatively, this could be due to the adaptive deadline modulating an "urgency signal" and its associated neural gain modulation, which changes the gain of sensory inputs to decision circuits (Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012; Murphy, Boonstra, & Nieuwenhuis, 2016), or due to cortical feedback from higher-order cortices to sensory regions (Noppeney et al., 2010). These possible accounts may inspire future research avenues. Of note, however, the explanatory power of a time-dependent auditory-noise parameter is soaked up by that of dynamic p_c when they are both allowed to be free parameters. This corroborates an important role of the dynamic integration tendency in accounting for early fusion.

Efficient CI requires the brain to maintain access to the representations of individual cues beyond an integrated estimate. This may rely on maintaining the

sensory representations in working memory over the course of a trial to support perceptual decisions even for transient stimulus inputs of less than 100 ms (Mostert, Kok, & de Lange, 2015). At the neural level, a set of distributed neural representations and their functional interplay could provide a plausible neural underpinning for the maintenance of unisensory representations (Cao et al., 2019; Rohe & Noppeney, 2015b, 2016). In the brains of humans and other animals, multiple and functionally heterogeneous multisensory representations have been identified, which flexible inference can potentially capitalise on (Bizley, Jones, & Town, 2016). At the behavioural level, the results from the sensory masking paradigm (Experiment II) corroborate this view in that the mask-induced disruption of the unisensory representation might interfere with the evaluation of intersensory disparity, thus interfering with a crucial component process for CI. The large repulsion bias observed in a few participants appeared surprising. However, in line with previous studies, it was observed predominantly for small-disparity conditions (Körding et al., 2007; Wallace et al., 2004). The delay of the visual mask (200 ms compared to 150 ms) may have increased the chance of the visual short-term memory trace being fed into the process of disparity evaluation (Ratcliff & Rouder, 2000; Smith & Ratcliff, 2009). Because the visual mask was always present across all conditions, such an increase in the persistence of the unisensory visual representation in working memory may enhance the sense of cue conflict between modalities and the tendency of separating cues. The negative repulsion bias may at least in part be attributable to this tendency of 'compensating' for a crossmodal bias by pushing the response away from the visual signal. This longer SOA (200 ms) also seems to fit with the earliest CI latency observed in a recent study on the timecourse of spatial CI (Aller & Noppeney, 2019).

In the data of Experiment II, the visual mask effectively reduced the visibility of the visual stimulus as evidenced in the large localisation error in the masked visual-only conditions (nearly comparable to the auditory localisation error). A previous study has shown that a robust audiovisual spatial biasing effect is observed even when participants judge the flash distractor as barely invisible (Delong et al., 2018). The magnitude of percent bias does not increase as a function of the actual accuracy of visual localisation measured on trials classified by participants as 'invisible' (see Rohe, 2014). This previous evidence suggests that sensory fusion may indeed capitalise on the earliest (on the scale of faster than 50-ms post-stimulus onset) multisensory interactions mediated via colliculo-pulvinar circuits (Cappe, Rouiller, & Barone, 2009; Hackett et al., 2007) or via direct pathways connecting primary sensory cortices (Cappe & Barone, 2005; Falchier et al., 2002).

Chapter appendix

This appendix provides supplemental materials containing 1 table and 4 figures.

Table S1 (Related to Figure 4.4; Experiment I). Parameter estimates and fit quality of the full model and the p_c-only or σ_{aud}-only causal inference models fitted to individual observers.

Model parameter		**Model (number of free parameters)**									
		Full (11)		p_c-only (7)		σ_{aud}-only (7)		σ_{aud} and p_c (8)			
p_c	fast	.73 (.37)	$Z = 2.71$ $p = .033$	.81 (.34)	$Z = 3.84$ $p < .001$			.82 (.29)	$Z = 2.87$ $p = .0082$		
	slow	.69 (.43)	$r = 0.41$	.58 (.41)	$r = 0.58$	0.65 (0.41)		.72 (.49)	$r = 0.43$		
σ_{vis1}	fast	3.27 (2.49)	$Z = -1.87$ $p = .31$								
	slow	3.83 (2.91)	$r = 0.28$	3.42 (1.77)		3.46 (1.91)		3.86 (2.27)			
σ_{vis2}	fast	10.7 (9.39)	$Z = -0.08$ $p > .1$								
	slow	8.87 (5.51)	$r = 0.012$	9.70 (2.91)		10.1 (5.54)		11.5 (6.83)			
σ_{aud}	fast	6.32 (3.79)	$Z = 3.29$ $p = .0049$			7.16 (3.51)	$Z = 4.11$ $p < .0001$	9.31 (4.91)	$Z = 3.98$ $p < .001$		
	slow	5.58 (2.05)	$r = 0.49$	6.21 (2.67)		5.76 (2.38)	$r = 0.62$	7.69 (3.53)	$r = 0.6$		
σ_{motor}	fast	1.81 (1.29)	$Z = 1.64$ $p > .1$								
	slow	1.41 (1.26)	$r = 0.24$	1.32 (1.33)		1.22 (1.06)		1.52 (1.67)			
σ_p		8.51 (47.8)		9.85 (48.9)		10.8 (43.1)		12.6 (57.3)			
Goodness-of-fit											
R^2		82.8% (1.12%)		81.5% (1.06%)		81.4% (1.02%)		82.8% (1.02%)			
sum(ΔBIC)		83		208		210		0			

Note: All parameter estimates are reported in medians with interquartile range (IQR) of the optimal parameter estimate (with highest Log-Likelihood among all 2,000 fits for each individual observer) across observers (N = 22). The fit quality is assessed by the goodness-of-fit quantities (R^2 and Bayesian Information Criterion or BIC). The R^2 terms describe the fraction of variance in data that is explained by the model. R^2 [mean (±sem)] is the generalized coefficient of determination (The BIC (the smaller the value, the better the fit) is computed as $-2LL + k\ln(n)$, where LL denotes the Log-Likelihood of the fitted model, k the number of free parameters and n the total number of data points (trials), ln is natural logarithm. ΔBIC is the BIC difference relative to the full model. Summary of ΔBIC over observers are reported here. Two-tailed paired-sample Wilcoxon signed-rank tests for the parameters related to fast vs. slow RT bins are summarised to the right of each pair of model estimates. Effect size $r = abs(Z)/\sqrt{44}$. The spatial prior was modelled as a time-invariant parameter in our model fits for two reasons. First, an analysis on the congruent trials (A = V) revealed that the response speed had no effect on the linear slope measure quantifying the relationship between the reported and physical signal locations ($t_{(21)} <$

1.88, $p > 0.14$, Cohen's $dz < 0.4$). Second, Bayesian model comparisons showed that a model featuring only time-varying spatial prior explained participants' behaviour decisively worse than models featuring other time-varying parameters.

Figure S4.1. Control analysis of crossmodal bias with equalised probability of choosing visual location across RT bins (related to Figure 4.2; Experiment I). (A) An illustrative example of data equalisation applied to the original data (left panel) and the resulting multinomial distribution of response probabilities (right panel) for the 16 combinations of auditory vs. visual signal locations, for the condition of high visual reliability in a single participant. The data equalisation was performed for both visual reliability conditions and for all participants. Specifically, the observed probabilities of reporting visual location across different RT bins were equalised for each incongruent condition, and the response probabilities of the other 3 'non-visual' locations within each RT level were scaled to ensure probabilities sum to 1. This provides a rigorous test on the time-varying crossmodal bias by removing any RT-induced difference in the probabilities of reporting the visual location while keeping unchanged the multinomial response pattern associated with the other 3 non-visual choices. Imagine a case where 200 trials of responses to a certain condition are collected and classified into relatively slow and fast bins according to the median RT. The condition specifies a reliable visual signal at -10° and an auditory signal at 3.3°. In those slower trials, the response counts associated with the four choices (-10°, -3.3°, 3.3°, 10°), are 10, 20, 40, and 30, respectively. Further suppose that the faster trials are simply characterized by a higher chance of reporting the visual signal by mistake (e.g., 10 trials) than slower trials, and the multinomial response pattern in the remaining 90 trials is identical to that in the slower trials, the resulting response probabilities across the 100 faster trials would be 0.19, 0.18, 0.36 and 0.27, associated with the four choices. Applying the above-mentioned probability-equalisation procedure to the data would result in identical response probabilities across the two RT bins (0.145, 0.19, 0.285, 0.38). In other words, our control analysis ensured that any remaining time-varying crossmodal bias would be genuinely attributable to the changes over time in the underlying response pattern rather than in the response tendency elicited by the visual signal on incongruent trials. (**B**) Absolute visual bias after controlling for the difference over time in response tendency elicited by the visual signal. (**C**) GLMs predicting absolute crossmodal bias for 4 RT bins (left) and 2 RT bins (right).

Control analysis equalising probabilities of visual response across RT bins

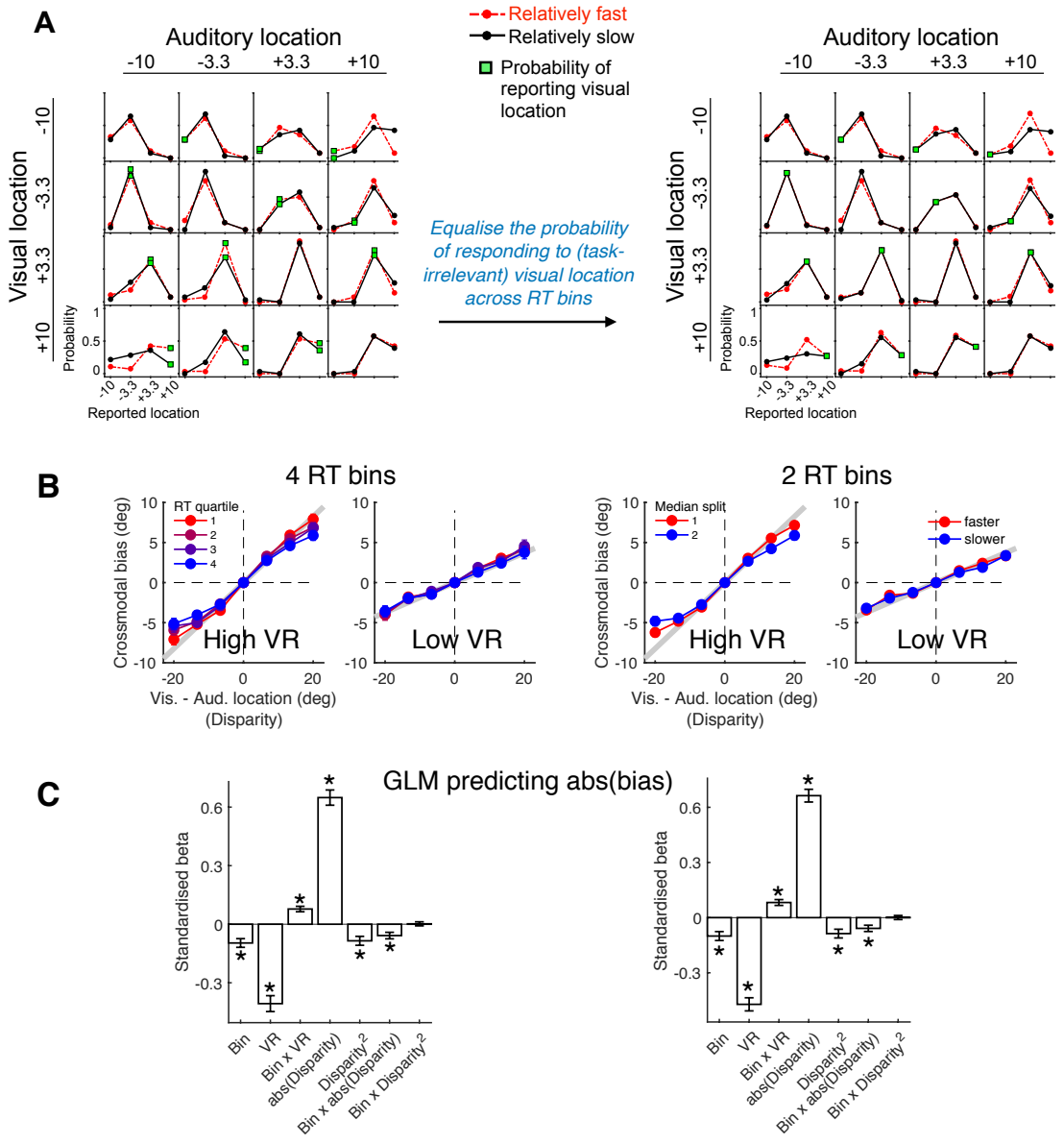


Figure S4.2. Reaction time (RT) and model-predicted unisensory noise (standard deviation of Gaussian sensory likelihood) in unisensory conditions for Experiment I. Vis. = Visual-only condition, H = High visual reliability, L = Low visual reliability, Aud. = Auditory-only condition. These unisensory conditions were tested in practice trials before the main experiment.

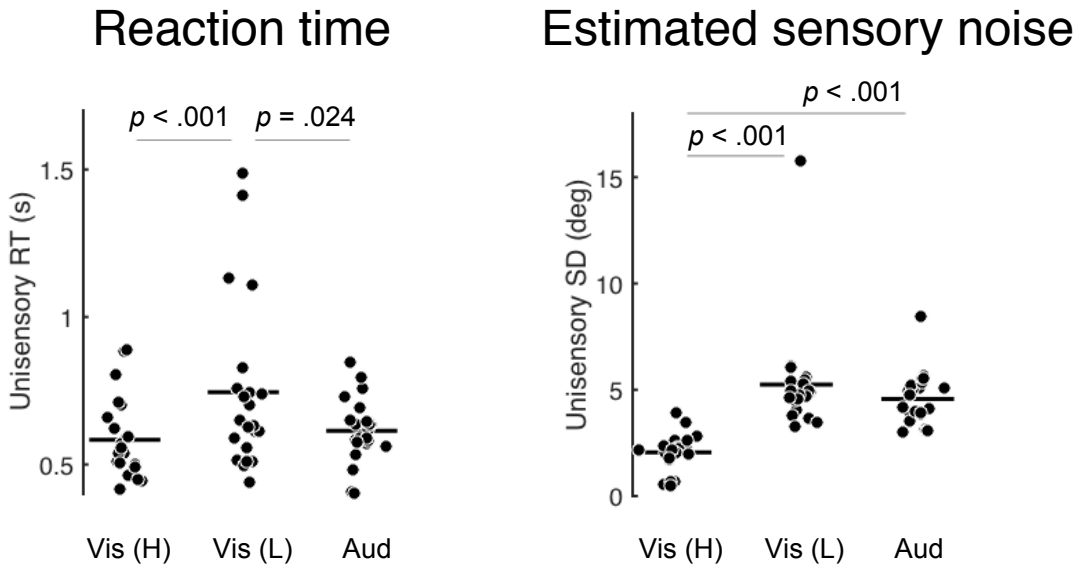


Figure S4.3. Probability of reporting the task-irrelevant visual cue as a function of the absolute disparity and RT level (left panels), and qualitative predictions of an alternative visual-error model (right panels). (**A**) and (**B**) demonstrate that whether or not the congruent conditions are included in the small disparity level does not change the result that the interaction between abs(Disparity) and RT bin is significant in real data but not in the predictions of the visual-error model.

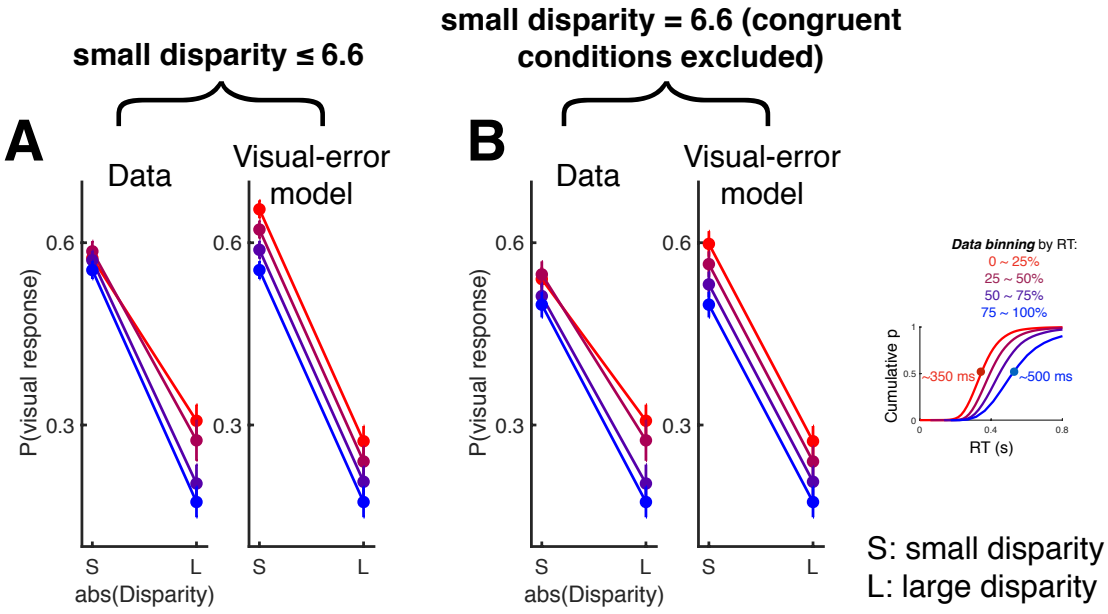
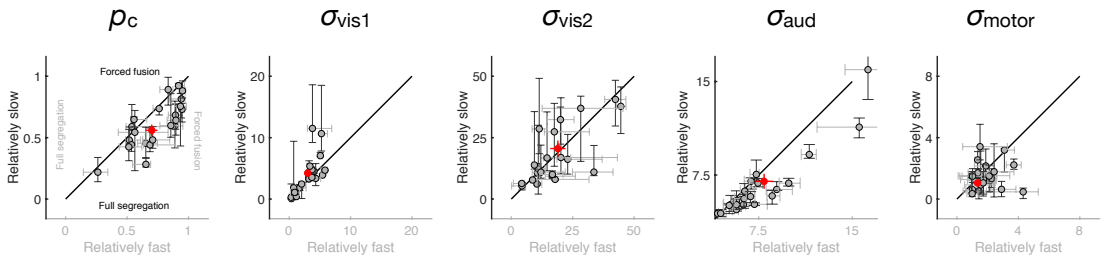


Figure S4.4. (Related to Figure 4.4 and Table S1). Parameter estimates of CI model with free p_c and σ^2_{aud}. Grey markers indicate medians; error bars represent interquartile range (25% to 75% percentile) of 10 best estimates (sorted by Log-Likelihood) among 2,000 Monte Carlo model fits with different initial parameter guesses. Red markers represent the grand mean of the medians (error bars: ± 1 SEM).



Chapter 5. Conclusions and general discussion

This thesis has used computational modelling approaches to investigate the neural underpinnings of flexible multisensory decision-making in the context of sensory integration and causal structure inference. I have drawn upon the Bayesian models discussed in **chapter 1** when characterising the computational strategies used by human participants for solving multisensory inference problems. In **chapter 3**, I used time- and source-resolved magnetoencephalography (MEG) to test predictions of a principle of sequential neural computations. Combined with the multivariate techniques discussed in **chapter 2**, this permitted a systematic characterisation of the spatiotemporal unfolding of relevant sensory computations across the human cerebral cortex. In **chapter 4**, I quantified the behavioural relevance of this principle of sequential computation and demonstrated when and how the neural representations formed under different causality assumptions are read out to drive behaviour.

In this final chapter, I discuss some key insights that can be drawn from these neuroimaging and psychophysical studies, and how they might be interpreted in the context of a broader literature on multisensory integration and adaptive behaviour.

5.1 Temporal sequence for multisensory inference

5.1.1 Temporal unfolding of multisensory causal structure inference

The neural data in **chapter 3** support a mechanistic account of flexible multisensory inference, i.e., it is a hierarchical process relying on the explicit representations of distinct multisensory computations orchestrated over different timescales and across distributed brain regions (Bizley, Jones, & Town, 2016; Kayser & Shams, 2015). Crucially, the human brain arranges the relevant component processes in a systematic temporal sequence whereby the neural underpinnings of sensory segregation and reliability-weighted fusion are succeeded by those of causal inference (CI).

Such a temporally-structured computational dynamic appears to serve as a general principle underlying multisensory inference. Firstly, this temporal sequence has been revealed in this thesis using not only rate stimuli (essentially in the temporal domain) but also spatial stimuli, not only categorisation task with discrete choices (**chapter 3** and Behavioural Experiment I in **chapter 4**), but also localisation task with continuous responses (Behavioural Experiment II in **chapter 4**). Secondly, the finding of this temporal sequence has replicated in independent studies using alternative imaging modality such as the electroencephalogram (EEG) and alternative multivariate analysis techniques such as a linear support-vector-machine decoder (Aller & Noppeney, 2019; also see Rohe et al., 2019 using numerosity stimuli). Interestingly, however, the specificity of when the sensory predictions by CI dominate neural activity appears to differ across the various recent studies (350 ms in Aller & Noppeney, 2019; later than 600 ms in Cao et al., 2019 and Rohe et al., 2019). This difference about precise CI timing is unlikely solely because of the difference in

neuroimaging modality (MEG vs. EEG, which in principle differ in their sensitivity to the orientation of the current flows of neuronal sources and their sensitivity to the inhomogeneity of tissue conductivity). The two studies using magnitude-related stimuli (rate and numerosity) revealed strikingly similar timecourses of CI, i.e., culminating in the neural activity around 620 to 660 ms post-stimulus (MEG: Cao et al., 2019; EEG: Rohe et al., 2019). It is more plausible to interpret the CI timing difference as due to the difference in stimulus attribute. The computational unfolding for spatial inference seems to be faster in general than that for the rate or numerosity inference. This makes sense in that the decision about rate and numerosity presumably requires integrating discrete pieces of sensory evidence (pulsatile fluctuations) over time (see **Figure S3.2** in **chapter 3** for long RTs around 1 s). Another mutually nonexclusive possibility is that the neural coding scheme for these distinct types of sensory cues might be different: whereas rate and numerosity judgements require converting a peripheral temporal spiking pattern to a magnitude code, location cues may simply be coded as spatial representations in the periphery and be read out by downstream stages in a faster fashion.

Despite these stimulus-related timing variabilities, the temporal unfolding from fusion to CI during the course of a single decision seems to be a generalised mechanism of flexible multisensory inference. It is interesting to notice that the latencies of early readout of fusion in behaviour across the two psychophysical experiments (**chapter 4)** coincides with the latencies of the neural signature reflecting fusion (Aller & Noppeney, 2019, in the case of spatial processing). Future research could further investigate the generalisability of this temporal sequence for audiovisual signals in other higher-order cognitive domains (e.g., voice and face; Campanella & Belin, 2007) or in other combinations of sensory modalities (Wozny, Beierholm, &

Shams, 2008). One prediction is that the brain may use the same temporal principle to organise computations but emphasise some specialised neuroanatomical structures (e.g., posterior STS for audiovisual speech and voice-face integration).

5.1.2 The role of neural selective representation of sensory fusion

A wealth of behavioural evidence suggests that when faced with trial-varying reliability and discrepancy of crossmodal information, humans select actions flexibly following the principle of Bayesian CI. For behaviours in response to multisensory inputs characterised by a wide range of crossmodal disparities, CI has nearly completely dominated over sensory fusion (Cao et al., 2019; Körding et al., 2007; Odegaard et al., 2015; Odegaard & Shams, 2016a; Odegaard et al., 2016b; Rohe & Noppeney, 2015a, 2015b; Wozny et al., 2010). Why does the brain still represent sensory information in a manner as predicted by each computational model when some of these neural representations (e.g., fusion) doesn't *seem* to guide action selection explicitly? In what follows, I discuss two non-mutually exclusive explanations.

One plausible explanation is that fusion may allow rapid inference through frugal computations, for example implemented in feedforward circuits (Ma et al., 2006; Ohshiro et al., 2011, 2017), and serves as a good rule of thumb for many past or future circumstances giving rise to correlated multimodal signals in the natural world (Parise & Ernst, 2016). Fusion is typically overshadowed by CI in situations where the speed of decision-making is not a requirement. Another framing of this view is that rapid fusion is putatively coded as a nearly 'default' strategy in certain brain structures for computational expediency. Fast decisions under time pressure

promote a readout of the sensory estimates formed under this default process, as evidenced in **chapter 4**.

Another possibility is that fusion is a component process that explicitly feeds into the neural computations for CI. This view can draw upon evidence from the MEG activity recorded near the time of response onset (**chapter 3**; **Figure 3.5A**). Shortly prior to the trial-by-trial response onset, the neural representations of fusion and CI co-exist but with that of fusion preceding that of CI on the scale of 40 to 50 ms. To illustrate this finding, I computed the cross-correlation between the timecourses of neural representations of fusion and CI and extracted the peak lag time on the cross-correlation function. The grand mean response-locked RSA timecourses show that fusion precedes CI, and CI seems temporally more sustained than fusion prior to the response onset. This result replicates across different RSA statistics (**Figure 5.1**). One interpretation is that this neural representation of sensory information in a manner as formed by fusion is a 'residual' effect of feeding component computation into the CI process near the time of forming a final decision. Alternatively, this effect might be attributable to a subset of conditions with small disparities. However, a non-significant contrast between the small vs. large disparities for the RSA fusion effect tends to rule out this account (analysis not shown here). A detailed investigation for this re-emergence of fusion in the brain activity near response time remains for future research to discern. Analytical tools such as directed functional connectivity (Giordano et al., 2017) could help shed light on whether it is better explained by local encoding or by modulatory influence from other brain regions, considering the rich reciprocal connections between prefrontal and posterior parietal cortices (Buschman & Miller, 2007; Rohe & Noppeney, 2018).

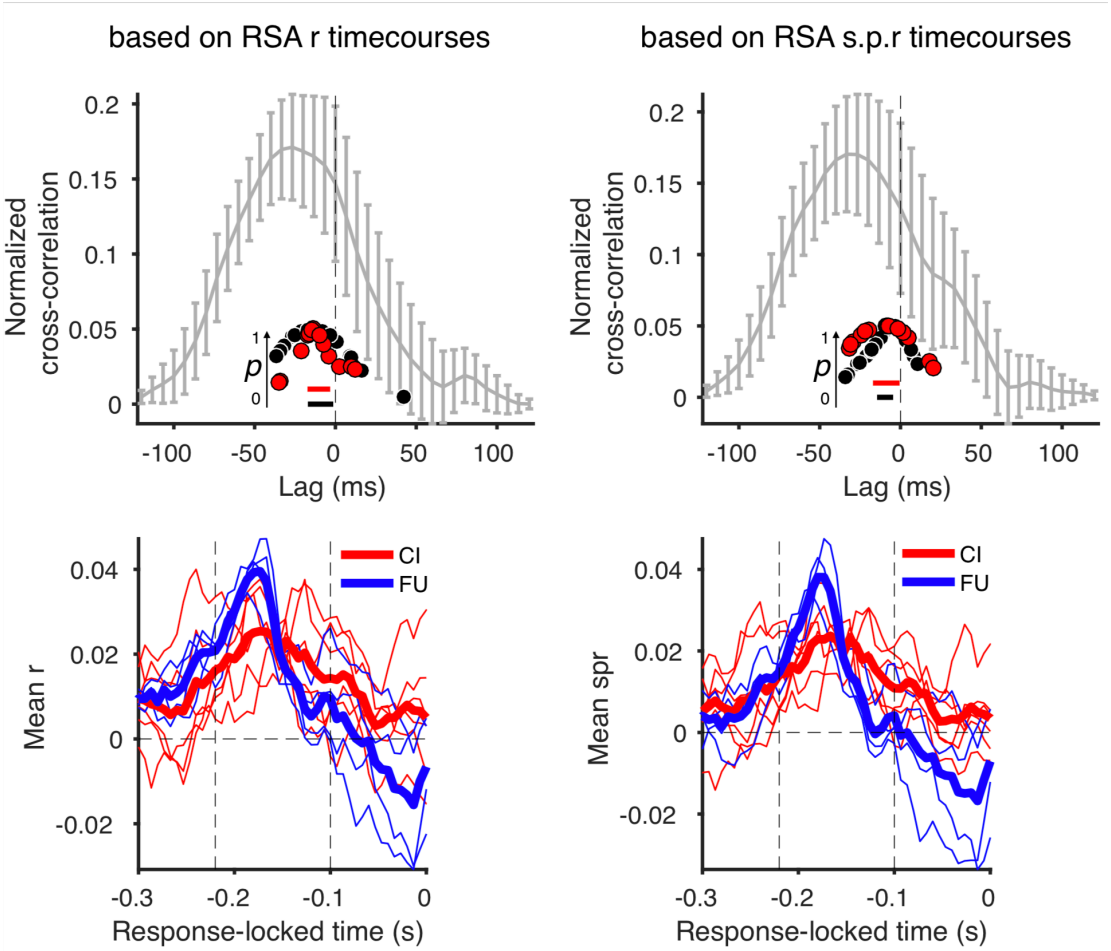


Figure 5.1. Fusion leads CI by ~ 50 ms near response time. (**A**) Normalised cross-correlation (xcorr) as a function of the time lag. The grey curve is the mean xcorr across parietal-to-frontal ROI pairs and participants, the error bars represent SEM over participants. Black circle: each parietal-to-frontal pair (across-participant mean); red circle: each participant (across-pair mean). Horizontal bars: 95% bias-corrected and accelerated [BCa] bootstrap confidence interval of the mean. The height of each circle is the probability estimated using *ksdensity* and normalised to peak = 1. **(B)** Each thin trajectory is the participants-averaged timecourse of a certain brain region. Thick curve depicts the grand mean across all participants and pairs. FU = fusion, CI = causal inference. The grand mean timecourse for the RSA model encoding shows that fusion precedes CI, and CI seems more temporally sustained than fusion. This result holds across different RSA statistics, r and s.p.r.

5.2 Parietal-frontal network for multisensory inference

5.2.1 Posterior parietal cortex represents information in low-dimensional space

The neural activity localised to the posterior parietal cortex appeared to scale with the task-irrelevant crossmodal information (**chapter 3**), suggesting that this brain structure may automatically project incoming evidence to a low-dimensional manifold (Ganguli et al., 2008; Luyckx et al., 2019; Suzuki & Gottlieb, 2013; for a review on this neural coding scheme, see Summerfield, Luyckx, & Sheahan, 2019). It is as if the posterior parietal cortex represents pulsation rate (Cao et al., 2019), numerosity (Nieder, 2016), and space (Rohe & Noppeney, 2015b, 2016) using a domain-general coding scheme. For example, in the MEG data (**chapter 3**; **Figure 3.6C**), the representational geometry of neural activities over this brain structure collapsed the rates across modalities and disparities into a single dimension (leading to large crossmodal biases). The posterior parietal regions also seemed to project concurrent audiovisual evidence about space onto a fused dimension disregarding inter-sensory discordance (Rohe & Noppeney, 2016). This theory of low-dimensional neural coding may shed light on the mechanism that the posterior parietal cortex uses to fuse multisensory evidence, as seen in those space- or magnitude-related stimuli tested in these various recent neuroimaging studies of multisensory CI.

Future work might expand on this view and start to investigate how the neural weighting based on reliability (as seen in **chapter 3**) might play a role in re-scaling the low-dimensional manifold such as changing the relational positions of items being coded neurally. Of note, however, it is highly plausible to argue that the neural (Rohe & Noppeney, 2016, 2018) and behavioural weights (Meijer et al., 2019) assigned to

different sensory modalities might depend not only on the sensory reliabilities as predicted the MLE fusion model (Ernst & Banks, 2002) but also on task-relevance of a given modality. **Chapter 3** demonstrated a result fitting with view. The neural representations of sensory information in the IPS were slightly influenced by a task-relevance factor (i.e., to report the pulsation rate of signal in a certain sensory modality; see **Figure 3.6C**), which seems to relax the constraint of MLE fusion. Despite it was not a motivation for this thesis, this task-dependent modulation of sensory weighting during fusion can indeed be captured using our multivariate analyses.

5.2.2 Choice predictability of neural activity in frontal vs. parietal cortex

There has been a much-debated topic about the relative contributions of parietal and frontal regions to accumulating evidence and determining a categorical decision (Brody & Hanks, 2016). Several studies have suggested that parietal regions reflect evidence accumulation, whilst frontal regions not only reflect accumulation but also the transformation of evidence into a categorical choice or the short-term memory for such (Erlich et al., 2015; Goard, Pho, Woodson, & Sur, 2016; Hanks et al., 2015). Also, a study using a similar rate task as used here found that parietal inactivation differentially affected behavioural sensitivity to visual and auditory stimuli (Raposo et al., 2014), suggesting that parietal cortex may be functionally closer to sensory regions than to the actual decision process (Goard et al., 2016). The MEG results in this thesis are in line with this work, in particular the gradual transition from mandatory fusion to more flexible behaviour mediated by frontal regions, as well as the choice-relevance of frontal regions. This view also fits with a general understanding of the functional differences between parietal and frontal

regions, whereby both accumulate sensory information, but parietal cortex potentially encodes more of the sensory aspect whereas frontal regions underlie the context-dependent transformation of evidence into choice (Erlich et al., 2015; Hanks et al., 2015).

5.2.3 Where is CI represented in the brain?

I use this section to emphasise a key point that draws upon the neural data in **chapter 3**. That is, the neural representation of sensory information formed under CI follows a *gradient* in the cerebral cortex. Our complementary analysis approaches to the MEG data, relying on either selective model encoding or group-level model exceedance probabilities, corroborated the parieto-frontal gradient reflecting a progression from the neural representations adapting to sensory reliability to representations adapting to disparity. Whilst these results resonate with a previous study suggesting the coexistence of distinct context-dependent multisensory computations in the IPS (Rohe & Noppeney, 2015b), they elaborate this perspective of computational gradient along a wider parietal-frontal network that supports CI in more anterior regions.

On one hand, one candidate brain structure (anterior parietal cortex) that has been implicated as reflecting the sensory information predicted by CI was also identified in the MEG data in this thesis (i.e., the supramarginal gyrus). On the other hand, RSA effect of CI was much stronger along a cortical axis pointing towards more anterior regions. The prefrontal cortex (PFC) appeared to have a privileged role in reflecting both hallmarks of context-dependent integration, but featured subtle representational variations across regions. The nonlinear influence of disparity

increased systematically along the posterior-to-anterior axis. Whilst representations within more posterior regions (e.g., the insula) appear to emphasise the reliability-based weighting, representations in anterior regions mostly depend on disparity (e.g., frontopolar cortex and inferior frontal gyrus). Such heightened sensitivity of more anterior regions to the structure of sensory causes resonates with the general hierarchical organisation of the PFC, whereby anterior frontal regions are recruited for abstract rule representation (Badre & D'Esposito, 2009).

One interpretation drawn upon this representational gradient for multisensory CI in the brain is that such an inference indeed relies on a distributed brain network (Bizley, Jones, & Town, 2016; Kayser & Shams, 2015). It could be an interesting research direction for future work to further interrogate which of the brain structures within this distributed network (e.g., dorsomedial and ventrolateral PFC, anterior parietal cortex, etc.) are causally driving CI and which merely reflect. Besides, we are still far from an understanding of CI at the circuit level (Bizley, Jones, & Town, 2016) and of where to find a mechanistic link between network models on single units (Zhang et al., 2019) and the neural representational principle as revealed in macroscopic human brain signals. I believe evidence from causal manipulations is likely going to provide useful insights, and the development of process models should be valuable contribution to this field, but this remains for future research to discern.

Finally, I think there is a great opportunity to draw upon the convergent evidence about the dynamic computational principle as an inspiration for future research in multisensory inference and learning. This thesis has shown that structures of sensory cause can be inferred by humans to build internal models of the world and to select actions in a flexible manner. However, a related question is: why does the

mind infer structures the way it does? The answer could at least in part be that structural inference facilitates causal knowledge generalisation and supports flexible behaviour in novel contexts. Future studies might build on this work by further investigating how the posterior belief about sensory cause can be factorised into bottom-up and top-down components (**chapter 3** Appendix II) and how learning about the causal structures of the environment dynamically updates the representation of posterior belief over the causal structure in both behaviour and the brain.

References

Acerbi, L., Dokka, K., Angelaki, D. E., & Ma, W. J. (2018). Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLoS Computational Biology*, *14*(7), e1006110.

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian Adaptive Direct Search. In *Proc. Advances in Neural Information Processing Systems 30 (NIPS '17)*, Long Beach, USA.

Adam, R., & Noppeney, U. (2010). Prior auditory information shapes visual category-selectivity in ventral occipito-temporal cortex. *NeuroImage*, *52*(4), 1592-1602.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*(3), 257-262.

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio & Acoustics (Cat. No. 01TH8575)* (pp. 99-102).

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

Angelaki, D. E., Gu, Y., & DeAngelis, G. C. (2009). Multisensory integration: Psychophysics, neurophysiology, and computation. *Current Opinion in Neurobiology, 19*(4), 452-458.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage, 38*(1), 95-113.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science, 5*(3), 144-151.

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K., & Bizley, J. K. (2018). Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. *Neuron*, *97*(3), 640-655.

Badre, D., & D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Review Neuroscience*, *10*(9), 659-669.

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron, 66*(2), 315-326.

Baillet, S., Mosher, J. C., & Leahy, R. M. (2001). Electromagnetic brain mapping. *IEEE Signal Processing Magazine, 18*(6), 14-30.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A, 20*(7), 1391-1397.

Beauchamp, M. S., Argall, B. D., Bodurka, J., Duyn, J. H., & Martin, A. (2004a). Unraveling multisensory integration: Patchy organization within human STS multisensory cortex. *Nature Neuroscience*, *7*(11), 1190-1192.

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004b). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*(5), 809-823.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*(2), 490-509.

Berényi, A., Somogyvári, Z., Nagy, A. J., Roux, L., Long, J. D., Fujisawa, S., ... & Buzsáki, G. (2013). Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals. *Journal of Neurophysiology, 111*(5), 1132-1149.

Bertelson, P., & de Gelder, B. (2004). The psychology of multimodal perception. In C. Spence, & J. Driver (Eds.), *Crossmodal Space and Crossmodal Attention* (pp. 151-177). Oxford/New York: Oxford University Press.

Bertelson, P., Vroomen, J., De Gelder, B., & Driver, J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics, 62*(2), 321-332.

Bizley, J. K., Jones, G. P., & Town, S. M. (2016). Where are multisensory signals combined for perceptual decision-making? *Current Opinion in Neurobiology*, *40*, 31-37.

Bizley, J. K., & King, A. J. (2009). Visual influences on ferret auditory cortex. *Hearing Research, 258*(1-2), 55-63.

Bizley, J. K., Maddox, R. K., & Lee, A. K. (2016). Defining auditory-visual objects: Behavioral tests and physiological mechanisms. *Trends in Neurosciences*, *39*(2), 74-85.

Bizley, J. K., Nodal, F. R., Bajo, V. M., Nelken, I., & King, A. J. (2007). Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cerebral Cortex, 17*(9), 2172-2189.

Blangero, A., & Kelly, S. P. (2017). Neural signature of value-based sensorimotor prioritization in humans. *Journal of Neuroscience, 37*(44), 10725-10737.

Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron, 62*(5), 733-743.

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, *391*(6669), 756.

Boyle, S. C., Kayser, S. J., & Kayser, C. (2017). Neural correlates of multisensory reliability and perceptual weights emerge at early latencies during audio-visual integration. *European Journal of Neuroscience*, *46*(10), 2565-2577.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

Brody, C. D., & Hanks, T. D. (2016). Neural underpinnings of the evidence accumulator. *Current Opinion in Neurobiology, 37*, 149-157.

Buschman, T. J., & Miller, E. K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science*, *315*(5820), 1860-1862.

Butler, J. S., Campos, J. L., & Bülthoff, H. H. (2015). Optimal visual-vestibular integration under conditions of conflicting intersensory motion profiles. *Experimental Brain Research, 233*(2), 587-597.

Caclin, A., Soto-Faraco, S., Kingstone, A., & Spence, C. (2002). Tactile "capture" of audition. *Perception & Psychophysics*, *64*(4), 616-630.

Calvert, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, *11*(12), 1110-1123.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*(12), 535-543.

Cao, Y., Giordano, B. L., Avanzini, F., & McAdams, S. (2016). The dominance of haptics over audition in controlling wrist velocity during striking movements. *Experimental Brain Research, 234*(4), 1145-1158.

Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal inference in the multisensory brain. *Neuron, 102*(5), 1076-1087.

Cappe, C., & Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience, 22*(11), 2886-2902.

Cappe, C., Rouiller, E. M., & Barone, P. (2009). Multisensory anatomical pathways. *Hearing Research, 258*(1-2), 28-36.

Carlin, J. D., & Kriegeskorte, N. (2017). Adjudicating between face-coding models with individual-face fMRI responses. *PLoS Computational Biology, 13*(7), e1005604.

Cavanaugh, J. E. (1997). Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, *33*(2), 201-208.

Chafee, M. V. (2013). A scalar neural code for categories in parietal cortex: Representing cognitive variables as “more” or “less”. *Neuron*, *77*(1), 7-9.

Chan, S. C., Niv, Y., & Norman, K. A. (2016). A probability distribution over latent causes, in the orbitofrontal cortex. *Journal of Neuroscience, 36*(30), 7817-7828.

Chandrasekaran, C., Lemus, L., & Ghazanfar, A. A. (2013). Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection. *Proceedings of the National Academy of Sciences, 110*(48), E4668-E4677.

Chen, A., DeAngelis, G. C., & Angelaki, D. E. (2013). Functional specializations of the ventral intraparietal area for multisensory heading discrimination. *Journal of Neuroscience*, *33*(8), 3567-3581.

Chen, Y. C., & Spence, C. (2011). Crossmodal semantic priming by naturalistic sounds and spoken words enhances visual sensitivity. *Journal of Experimental Psychology: Human Perception & Performance, 37(*5), 1554-1568.

Chen, Y. C., & Spence, C. (2017). Assessing the role of the ‘unity assumption’ on multisensory integration: A review. *Frontiers in Psychology, 8*, 445.

Dahl, C. D., Logothetis, N. K., & Kayser, C. (2009). Spatial organization of multisensory responses in temporal association cortex. *Journal of Neuroscience, 29*(38), 11924-11932.

Darlington, T. R., Beck, J. M., & Lisberger, S. G. (2018). Neural implementation of Bayesian inference in a sensorimotor behavior. *Nature Neuroscience, 21*(10), 1442-1451.

Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Computational Biology*, *10*(1), e1003441.

da Silva, F. L. (2013). EEG and MEG: relevance to neuroscience. *Neuron, 80*(5), 1112-1128.

De Corte, B. J., Della Valle, R. R., & Matell, M. S. (2018). Recalibrating timing behavior via expected covariance between temporal cues. *eLife, 7*, e38790.

Dekel, R., & Sagi, D. (2019). Seeing, fast and slow: The effects of processing time on perceptual bias. *bioRxiv*, 556944.

De Leeuw, J., & Mair, P. (2011). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software, 31*(3), 1-30.

Deneve, S., Latham, P. E., & Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature Neuroscience, 4*(8), 826-831.

De Winkel, K. N., Katliar, M., & Bülthoff, H. H. (2015). Forced fusion in multisensory heading estimation. *PLoS One, 10*(5), e0127104.

De Winkel, K. N., Katliar, M., & Bülthoff, H. H. (2017). Causal inference in multisensory heading estimation. *PloS One*, *12*(1), e0169676.

Diederich, A., & Busemeyer, J. R. (2006). Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift-rate-change, or two-stage-processing hypothesis. *Perception & Psychophysics, 68*(2), 194-207.

Diaconis, P., & Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics, 14*(1), 1-26.

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology, 13*(4), e1005508.

Diedrichsen, J., Provost, S., & Zareamoghaddam, H. (2016). On the distribution of cross-validated Mahalanobis distances. *arXiv preprint arXiv*:1607.01371.

Dijkstra, N., Mostert, P., de Lange, F. P., Bosch, S., & van Gerven, M. A. (2018). Differential temporal dynamics during visual imagery and perception. *eLife, 7*, e33904.

Donoso, M., Collins, A. G., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science, 344*(6191), 1481-1486.

Drugowitsch, J., DeAngelis, G. C., Angelaki, D. E., & Pouget, A. (2015). Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making. *eLife, 4*, e06678.

Drugowitsch, J., DeAngelis, G. C., Klier, E. M., Angelaki, D. E., & Pouget, A. (2014). Optimal multisensory decision-making in a reaction-time task. *eLife, 3*, e03005.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, *32*(11), 3612-3628

Drugowitsch, J., Wyart, V., Devauchelle, A. D., & Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron*, *92*(6), 1398-1411.

Durgin, F. H., Evans, L., Dunphy, N., Klostermann, S., & Simmons, K. (2007). Rubber hands feel the touch of light. *Psychological Science*, *18*(2), 152-157.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Ehrsson, H. H., Spence, C., & Passingham, R. E. (2004). That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. *Science*, *305*(5685), 875-877.

Erlich, J. C., Brunton, B. W., Duan, C. A., Hanks, T. D., & Brody, C. D. (2015). Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat. *eLife, 4*, e05457.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, 7(5):7, 1-14.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Science, 8*(4), 162-169.

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience, 22*(13), 5749-5759.

Fang, W., Li, J., Qi, G., Li, S., Sigman, M., & Wang, L. (2019). Statistical inference of body representation in the macaque brain. *Proceedings of the National Academy of Sciences, 116*(40), 20151-20157.

Feng, S., Holmes, P., Rorie, A., & Newsome, W. T. (2009). Can monkeys choose optimally when faced with noisy stimuli and unequal rewards? *PLoS Computational Biology, 5*(2), e1000284.

Fetsch, C. R., DeAngelis, G. C., & Angelaki, D. E. (2013). Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nature Reviews Neuroscience, 14*(6), 429-442.

Fetsch, C. R., Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (2011). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience, 15*(1), 146-154.

Fetsch, C. R., Turner, A. H., DeAngelis, G. C., & Angelaki, D. E. (2009). Dynamic reweighting of visual and vestibular cues during self-motion perception. *Journal of Neuroscience, 29*(49), 15601-15612.

Fischl, B. (2012). FreeSurfer. *NeuroImage, 62*(2), 774-781.

Foree, D. D., & LoLordo, V. M. (1973). Attention in the pigeon: Differential effects of food-getting versus shock-avoidance procedures. *Journal of Comparative & Physiological Psychology*, *85*(3), 551-558.

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex, 22*(3), 509-526.

Ganguli, S., Bisley, J. W., Roitman, J. D., Shadlen, M. N., Goldberg, M. E., & Miller, K. D. (2008). One-dimensional dynamics of attention and decision making in LIP. *Neuron*, *58*(1), 15-25.

García-Pérez, M. A. (2011). A cautionary note on the use of the adaptive up–down method. *The Journal of the Acoustical Society of America*, *130*(4), 2098-2107.

García-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Research*, *38*(12), 1861-1881.

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. *NeuroImage*, *124*(Pt A), 876-886.

Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *Journal of Neuroscience, 25*(20), 5004-5012.

Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current Biology, 13*(6), 483-488.

Gershman, S. J. (2017). Context-dependent learning and causal structure. *Psychonomic Bulletin & Review*, *24*(2), 557-565.

Gershman, S. J., & Beck, J. M. (2017). Complex probabilistic inference: From cognition to neural computation. In A. Moustafa (Ed.), *Computational models of brain and behavior*. Hoboken, NJ: Wiley-Blackwell.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43-50.

Gibson, J. J. (1933). Adaptation, after-effect and contrast in the perception of curved lines. *Journal of Experimental Psychology, 16*(1), 1-31.

Giordano, B. L., Ince, R. A., Gross, J., Schyns, P. G., Panzeri, S., & Kayser, C. (2017). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. *eLife, 6*, e24763.

Giordano, B. L., Whiting, C., Kriegeskorte, N., Kotz, S. A., Belin, P., & Gross, J. (2018). From categories to dimensions: Spatiotemporal dynamics of the cerebral representations of emotion in voice. *BioRxiv*, 265843. doi:10.1101/265843.

Goard, M. J., Pho, G. N., Woodson, J., & Sur, M. (2016). Distinct roles of visual, parietal, and frontal motor cortices in memory-guided sensorimotor decisions. *eLife, 5*, e13764.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535-574.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*(1), 3-32.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334-384.

Gu, Y., Angelaki, D. E., & DeAngelis, G. C. (2008). Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience, 11*(10), 1201-1210.

Gu, Y., DeAngelis, G. C., & Angelaki, D. E. (2012). Causal links between dorsal medial superior temporal area neurons and multisensory heading perception. *Journal of Neuroscience, 32*(7), 2299-2313.

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage, 173*, 434-447.

Hackett, T. A., De La Mothe, L. A., Ulbert, I., Karmos, G., Smiley, J., & Schroeder, C. E. (2007). Multisensory convergence in auditory cortex, II. Thalamocortical

connections of the caudal superior temporal plane. *Journal of Comparative Neurology, 502*(6), 924-952.

Hadjipapas, A., Hillebrand, A., Holliday, I. E., Singh, K. D., & Barnes, G. R. (2005). Assessing interactions of linear and nonlinear neuronal sources using MEG beamformers: a proof of concept. *Clinical Neurophysiology, 116*(6), 1300-1313.

Hanks, T. D., Kopec, C. D., Brunton, B. W., Duan, C. A., Erlich, J. C., & Brody, C. D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature, 520*(7546), 220-223.

Hanks, T. D., Mazurek, M. E., Kiani, R., Hopp, E., & Shadlen, M. N. (2011). Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *Journal of Neuroscience, 31*(17), 6339-6352.

Hansen, P., Kringelbach, M., & Salmelin, R. (2010). *MEG: An introduction to methods*. Oxford New-York: Oxford University Press.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425-2430.

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *eLife, 7*, e32816.

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *Journal of Neuroscience*, *27*(30), 7881-7887.

Helbig, H. B., Ernst, M. O., Ricciardi, E., Pietrini, P., Thielscher, A., Mayer, K. M., ... & Noppeney, U. (2012). The neural mechanisms of reliability weighted integration of shape information from vision and touch. *NeuroImage, 60*(2), 1063-1072.

Hillebrand, A., & Barnes, G. R. (2002). A quantitative assessment of the sensitivity of whole-head MEG to activity in the adult human cortex. *NeuroImage, 16*(3), 638-650.

Hillebrand, A., & Barnes, G. R. (2005). Beamformer analysis of MEG data. *International Review of Neurobiology, 68*, 149-171.

Hipp, J. F., & Siegel, M. (2013). Dissociating neuronal gamma-band activity from cranial and ocular muscle activity in EEG. *Frontiers in Human Neuroscience, 7*, 338.

Hou, H., Zheng, Q., Zhao, Y., Pouget, A., & Gu, Y. (2019). Neural correlates of optimal multisensory decision making under time-varying reliabilities with an invariant linear probabilistic population code. *Neuron* (in press).

Hunt, L. (2011). *Modelling human decision under risk and uncertainty* (Doctoral dissertation). Retrieved from: https://ora.ox.ac.uk/objects/uuid:244ce799-7397-4698-8dac-c8ca5d0b3e28/download_file?safe_filename=THESIS01&file_format=application%2Fpdf&type_of_work=Thesis

Jackson, C. V. (1953). Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology*, *5*(2), 52-65.

Jones, E. G., & Powell, T. P. S. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain, 93*, 793-820.

Jousmäki, V., & Hari, R. (1998). Parchment-skin illusion: Sound-biased touch. *Current Biology*, *8*(6), R190-R191.

Kayser, S. J., & Kayser, C. (2018). Trial by trial dependencies in multisensory perception and their correlates in dynamic brain activity. *Scientific Reports, 8*, 37-42.

Kayser, C., & Logothetis, N. K. (2007). Do early sensory cortices integrate cross-modal information? *Brain Structure & Function*, *212*(2), 121-132.

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex, 18*(7), 1560-1574.

Kayser, C., Petkov, C. I., Remedios, R., & Logothetis, N. K. (2012). Multisensory influences on auditory processing: Perspectives from fMRI and electrophysiology. In M.M. Murray, & M.T. Wallace (Eds.), *The neural bases of multisensory processes*. Boca Raton (FL): CRC Press/Taylor & Francis.

Kayser, C., & Shams, L. (2015). Multisensory causal inference in the brain. *PLoS Biology, 13*(2), e1002075.

Kerr, J. N., & Denk, W. (2008). Imaging in vivo: Watching the brain in action. *Nature Reviews Neuroscience, 9*(3), 195-205.

Kiani, R., Hanks, T. D., & Shadlen, M. N. (2008). Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. *Journal of Neuroscience, 28*(12), 3017-3029.

King, A. J. (2013). Multisensory circuits. In J. Rubenstein, & P. Rakic (Eds.), *Neural circuit development and function in the healthy and diseased brain: Comprehensive developmental neuroscience*. San Diego, USA: Academic Press.

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M. C., ... & Song, J. H. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage, 46*(3), 786-802.

Koechlin, E., & Hyafil, A. (2007). Anterior prefrontal function and the limits of human decision-making. *Science, 318*(5850), 594-598.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Science*, *11*(6), 229-235.

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.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience, 21*, 1148-1160.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, 4.

Lakatos, P., Chen, C. M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron*, *53*(2), 279-292.

Ledoit, O., & Wolf, M. (2004). Honey, I shrunk the sample covariance matrix. *The Journal of Portfolio Management, 30*(4), 110-119.

Levi, A. J., Yates, J. L., Huk, A. C., & Katz, L. N. (2018). Strategic and dynamic temporal weighting for perceptual decisions in humans and macaques. *eNeuro*, *5*.

Lewis, R., & Noppeney, U. (2010). Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. *Journal of Neuroscience*, *30*(37), 12329-12339.

Ludwig, C. J., Gilchrist, I. D., McSorley, E., & Baddeley, R. J. (2005). The temporal impulse response underlying saccadic decisions. *Journal of Neuroscience, 25*(43), 9907-9912.

Luyckx, F., Nili, H., Spitzer, B., & Summerfield, C. (2019). Neural structure mapping in human probabilistic reward learning. *eLife*, *8*, e42816.

Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in Cognitive Sciences, 16*(10), 511-518.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9*(11), 1432-1438.

Ma, W. J., & Pouget, A. (2008). Linking neurons to behavior in multisensory perception: A computational review. *Brain Research, 1242*, 4-12.

Ma, W. J., & Rahmati, M. (2013). Towards a neural implementation of causal inference in cue combination. *Multisensory Research, 26*(1-2), 159-176.

Macaluso, E., & Driver, J. (2005). Multisensory spatial interactions: A window onto functional integration in the human brain. *Trends in Neuroscience*, *28*(5), 264-271.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms.* Cambridge, UK: Cambridge University Press.

Magnotti, J. F., & Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLoS Computational Biology*, *13*(2), e1005229.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177-190.

Mayer, A. R., Ryman, S. G., Hanlon, F. M., Dodd, A. B., & Ling, J. M. (2016). Look hear! The prefrontal cortex is stratified by modality of sensory input during multisensory cognitive control. *Cerebral Cortex, 27*(5), 2831-2840.

Mazurek, M. E., Roitman, J. D., Ditterich, J., & Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. *Cerebral Cortex, 13*(11), 1257-1269.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748.

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.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, *17*(1), 154-163.

Morgan, M. L., DeAngelis, G. C., & Angelaki, D. E. (2008). Multisensory integration in macaque visual cortex depends on cue reliability. *Neuron, 59*(4), 662-673.

Mostert, P., Albers, A. M., Brinkman, L., Todorova, L., Kok, P., & de Lange, F. P. (2018). Eye movement-related confounds in neural decoding of visual working memory representations. *eNeuro, 5*(4).

Mostert, P., Kok, P., & De Lange, F. P. (2015). Dissociating sensory from decision processes in human perceptual decision making. *Scientific Reports*, *5*, 18253.

Murphy, P. R., Boonstra, E., & Nieuwenhuis, S. (2016). Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nature Communications*, *7*, 13526.

Murray, M. M., Lewkowicz, D. J., Amedi, A., & Wallace, M. T. (2016). Multisensory processes: A balancing act across the lifespan. *Trends in Neurosciences*, *39*(8), 567-579.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691-692.

Nieder, A. (2016). The neuronal code for number. *Nature Reviews Neuroscience, 17*(6), 366-382.

Nieder, A., & Miller, E. K. (2003). Coding of cognitive magnitude: Compressed scaling of numerical information in the primate prefrontal cortex. *Neuron*, *37*(1), 149-157.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H. J., & Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *Journal of Neuroscience*, *27*(42), 11431-11441.

Noppeney, U. (2012). Characterization of multisensory integration with fMRI. In M.M. Murray, & M.T. Wallace (Eds.), *The neural bases of multisensory processes*. Boca Raton (FL): CRC Press/Taylor & Francis.

Noppeney, U., Ostwald, D., & Werner, S. (2010). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *Journal of Neuroscience*, *30*(21), 7434-7446.

O'Reilly, J. X., Jbabdi, S., Rushworth, M. F., & Behrens, T. E. (2013). Brain systems for probabilistic and dynamic prediction: computational specificity and integration. *PLoS Biology, 11*(9), e1001662.

Odegaard, B., & Shams, L. (2016). The brain's tendency to bind audiovisual signals is stable but not general. *Psychological Science*, *27*(4), 583-591.

Odegaard, B., Wozny, D. R., & Shams, L. (2015). Biases in visual, auditory, and audiovisual perception of space. *PLoS Computational Biology*, *11*(12), e1004649.

Odegaard, B., Wozny, D. R., & Shams, L. (2016). The effects of selective and divided attention on sensory precision and integration. *Neuroscience Letters*, *614*, 24-28.

Ohshiro, T., Angelaki, D. E., & DeAngelis, G. C. (2011). A normalization model of multisensory integration. *Nature Neuroscience*, *14*(6), 775-782.

Ohshiro, T., Angelaki, D. E., & DeAngelis, G. C. (2017). A neural signature of divisive normalization at the level of multisensory integration in primate cortex. *Neuron*, *95*(2), 399-411.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence & Neuroscience*, *2011*, 156869.

Ota, N., Gahr, M., & Soma, M. (2018). Couples showing off: Audience promotes both male and female multimodal courtship display in a songbird. *Science Advances*, *4*(10), eaat4779.

Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E., & Fellin, T. (2017). Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron*, *93*(3), 491-507.

Panzeri, S., Macke, J. H., Gross, J., & Kayser, C. (2015). Neural population coding: combining insights from microscopic and mass signals. *Trends in Cognitive Sciences, 19*(3), 162-172.

Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nature Communications*, *7*, 11543.

Parise, C. V., & Ernst, M. O. (2017). Noise, multisensory integration, and previous response in perceptual disambiguation. *PLoS Computational Biology, 13*(7), e1005546.

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. *Current Biology, 22*(1), 46-49.

Partan, S., & Marler, P. (1999). Communication goes multimodal. *Science*, *283*(5406), 1272-1273.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442.

Perrodin, C., Kayser, C., Logothetis, N. K., & Petkov, C. I. (2015). Natural asynchronies in audiovisual communication signals regulate neuronal multisensory interactions in voice-sensitive cortex. *Proceedings of the National Academy of Sciences, 112*(1), 273-278.

Pick, H. L., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics*, *6*(4), 203-205.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review*, *83*(2), 157-171.

Prsa, M., Gale, S., & Blanke, O. (2012). Self-motion leads to mandatory cue fusion across sensory modalities. *Journal of Neurophysiology, 108*(8), 2282-2291.

Qu, T., Xiao, Z., Gong, M., Huang, Y., Li, X., & Wu, X. (2009). Distance-dependent head-related transfer functions measured with high spatial resolution using a spark gap. *IEEE Transactions on Audio, Speech, & Language Processing, 17*, 1124-1132.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, *9*(7), 545-556.

Raposo, D., Kaufman, M. T., & Churchland, A. K. (2014). A category-free neural population supports evolving demands during decision-making. *Nature Neuroscience*, *17*(12), 1784-1792.

Raposo, D., Sheppard, J. P., Schrater, P. R., & Churchland, A. K. (2012). Multisensory decision-making in rats and humans. *Journal of Neuroscience, 32*(11), 3726-3735.

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human perception & Performance, 26*(1), 127-140.

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology, 89*(2), 1078-1093.

Remington, E. D., Narain, D., Hosseini, E. A., & Jazayeri, M. (2018). Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *Neuron, 98*(5), 1005-1019.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *NeuroImage, 84*, 971-985.

Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio-visual integration. *Philosophical Transactions of the Royal Society B: Biological Sciences, 273*, 2159-2168.

Rock, I., & Harris, C. S. (1967). Vision and touch. *Scientific American*, *216*(5), 96-107.

Rock, I., & Victor, J. (1964). Vision and touch: An experimentally created conflict between the two senses. *Science, 143*(3606), 594-596.

Rohe, T. (2014). *Causal inference in multisensory perception and the brain* (Doctoral dissertation). Retrieved from: https://pure.mpg.de/rest/items/item_2160943/component/file_2562288/content

Rohe, T., & Noppeney, U. (2015a). Sensory reliability shapes perceptual inference via two mechanisms. *Journal of Vision, 15*, 1-16.

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

Rohe, T., & Noppeney, U. (2016). Distinct computational principles govern multisensory integration in primary sensory and association cortices. *Current Biology, 26*, 509-514.

Rohe, T., & Noppeney, U. (2018). Reliability-weighted integration of audiovisual signals can be modulated by top-down attention. *eNeuro, 5*(1).

Rohe, T., Ehlis, A., and Noppeney, U. (2019). The neural dynamics of hierarchical Bayesian inference in multisensory perception. *Nature Communications, 10*, 1907.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of Neuroscience, 22*(21), 9475-9489.

Romanski, L. M. (2012). Convergence of auditory, visual, and somatosensory information in ventral prefrontal cortex. In M.M. Murray, & M.T. Wallace (Eds.), *The neural bases of multisensory processes* (Chapter 33). Boca Raton (FL): CRC Press/Taylor & Francis.

Romanski, L. M. (2007). Representation and integration of auditory and visual stimuli in the primate ventral lateral prefrontal cortex. *Cerebral Cortex, 17*(suppl_1), i61-i69.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences, 116*(23), 11537-11546.

Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10-and 12-month-old infants. *Psychological Science*, *16*(12), 995-1001.

Schroeder, C. E., & Foxe, J. (2005). Multisensory contributions to low-level 'unisensory' processing. *Current Opinion in Neurobiology*, *15*(4), 454-458.

Schumacher, S., de Perera, T. B., Thenert, J., & von der Emde, G. (2016). Cross-modal object recognition and dynamic weighting of sensory inputs in a fish. *Proceedings of the National Academy of Sciences, 113*(27), 7638-7643.

Seibold, D. R., & McPhee, R. D. (1979). Commonality analysis: A method for decomposing explained variance in multiple regression analyses. *Human Communication Research*, *5*, 355-365.

Seilheimer, R. L., Rosenberg, A., & Angelaki, D. E. (2014). Models and processes of multisensory cue combination. *Current Opinion in Neurobiology, 25*, 38-46.

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology*, *16*(14), 1422-1427.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385*(6614), 308.

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron, 80*(3), 791-806.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature*, *408*(6814), 788.

Sheppard, J. P., Raposo, D., & Churchland, A. K. (2013). Dynamic weighting of multisensory stimuli shapes decision-making in rats and humans. *Journal of Vision, 13*(6):4, 1-19.

Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review, 116*(2), 283-317.

Sohn, H., Narain, D., Meirhaeghe, N., & Jazayeri, M. (2019). Bayesian computation through cortical latent dynamics. *Neuron, 103*(5), 934-947.

Stanford, T. R., Shankar, S., Massoglia, D. P., Costello, M. G., & Salinas, E. (2010). Perceptual decision making in less than 30 milliseconds. *Nature Neuroscience, 13*(3), 379-385.

Stein, B. E. (Ed.). (2012). *The new handbook of multisensory processing*. MIT Press.

Sugihara, T., Diltz, M. D., Averbeck, B. B., & Romanski, L. M. (2006). Integration of auditory and visual communication information in the primate ventrolateral prefrontal cortex. *Journal of Neuroscience, 26*(43), 11138-11147.

Summerfield, C., Luyckx, F., & Sheahan, H. (2019). Structure learning and the posterior parietal cortex. *Progress in Neurobiology*, 101717.

Suzuki, M., & Gottlieb, J. (2013). Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nature Neuroscience*, *16*(1), 98-104.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences, 14*(9), 400-410.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.

Thomas, G. J. (1941). Experimental study of the influence of vision on sound localization. *Journal of Experimental Psychology, 28*, 167-177.

Thornton, M. A., Weaverdyck, M. E., Mildner, J. N., & Tamir, D. I. (2019). People represent their own mental states more distinctly than those of others. *Nature Communications, 10*(1), 2117.

Tomov, M. S., Dorfman, H. M., & Gershman, S. J. (2018). Neural computations underlying causal structure learning. *Journal of Neuroscience*, *38*(32), 7143-7157.

Trommershauser, J., Körding, K., & Landy, M. S. (Eds.). (2011). *Sensory cue integration*. Oxford University Press.

Tsunada, J., Cohen, Y., & Gold, J. I. (2019). Post-decision processing in primate prefrontal cortex influences subsequent choices on an auditory decision-making task. *eLife, 8*, e46770.

Uchida, N., Kepecs, A., & Mainen, Z. F. (2006). Seeing at a glance, smelling in a whiff: Rapid forms of perceptual decision making. *Nature Reviews Neuroscience, 7*(6), 485-491.

Uetake, K., & Kudo, Y. (1994). Visual dominance over hearing in feed acquisition procedure of cattle. *Applied Animal Behaviour Science*, *42*(1), 1-9.

van Beers, R. J., Wolpert, D. M., & Haggard, P. (2002). When feeling is more important than seeing in sensorimotor adaptation. *Current Biology, 12*(10), 834-837.

Van Veen, B. D., Van Drongelen, W., Yuchtman, M., & Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. *IEEE Transactions on Biomedical Engineering, 44*(9), 867-880.

Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A., & Körding, K. P. (2012). Differential representations of prior and likelihood uncertainty in the human brain. *Current Biology, 22*(18), 1641-1648.

Vroomen, J., Bertelson, P., & De Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics*, *63*(4), 651-659.

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research, 158*(2), 252-258.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage, 137*, 188-200.

Wang, L., Mruczek, R. E., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. *Cerebral Cortex, 25*(10), 3911-3931.

Warren, D. H., & Cleaves, W. T. (1971). Visual-proprioceptive interaction under large amounts of conflict. *Journal of Experimental Psychology, 90*(2), 206-214.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*(3), 638- 667.

Welch, R. B., DutionHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics, 39*(4), 294-300.

Whiteley, L., & Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision, 8*(3):2, 1-15.

Witkin, H. A., Wapner, S., & Leventhal, T. (1952). Sound localization with conflicting visual and auditory cues. *Journal of Experimental Psychology, 43*(1), 58-67.

Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience, 20*(6), 864-871.

Wolters, C. H., Anwander, A., Tricoche, X., Weinstein, D., Koch, M. A., & Macleod, R. S. (2006). Influence of tissue conductivity anisotropy on EEG/MEG field and return current computation in a realistic head model: A simulation and visualization study using high-resolution finite element modeling. *NeuroImage, 30*(3), 813-826.

Wozny, D. R., Beierholm, U. R., & Shams, L. (2010). Probability matching as a computational strategy used in perception. *PLoS Computational Biology, 6*, e1000871.

Wozny, D. R., & Shams, L. (2011). Recalibration of auditory space following milliseconds of cross-modal discrepancy. *Journal of Neuroscience*, *31*(12), 4607-4612.

Yassa, M. A., & Stark, C. E. (2009). A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. *NeuroImage, 44*(2), 319-327.

Yates, J. L., Park, I. M., Katz, L. N., Pillow, J. W., & Huk, A. C. (2017). Functional dissection of signal and noise in MT and LIP during decision-making. *Nature Neuroscience, 20*(9), 1285-1292.

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.