ANNUAL REVIEWS

Annual Review of Neuroscience

Perceptual Inference, Learning, and Attention in a Multisensory World

Uta Noppeney

Donders Institute for Brain, Cognition and Behavior, Radboud University, 6525 AJ Nijmegen, The Netherlands; email: u.noppeney@donders.ru.nl

Annu. Rev. Neurosci. 2021. 44:449–73

First published as a Review in Advance on April 21, 2021

The *Annual Review of Neuroscience* is online at neuro.annualreviews.org

https://doi.org/10.1146/annurev-neuro-100120-085519

Keywords

perceptual inference, learning, attention, multisensory, Bayesian, recalibration

Abstract

Adaptive behavior in a complex, dynamic, and multisensory world poses some of the most fundamental computational challenges for the brain, notably inference, decision-making, learning, binding, and attention. We first discuss how the brain integrates sensory signals from the same source to support perceptual inference and decision-making by weighting them according to their momentary sensory uncertainties. We then show how observers solve the binding or causal inference problem—deciding whether signals come from common causes and should hence be integrated or else be treated independently. Next, we describe the multifarious interplay between multisensory processing and attention. We argue that attentional mechanisms are crucial to compute approximate solutions to the binding problem in naturalistic environments when complex time-varying signals arise from myriad causes. Finally, we review how the brain dynamically adapts multisensory processing to a changing world across multiple timescales.

Contents

INTRODUCTION 450
INTEGRATING SENSORY INFORMATION FOR PERCEPTUAL INFERENCE AND DECISION-MAKING 451
SOLVING THE BINDING PROBLEM: CAUSAL INFERENCE 454
THE MULTIFARIOUS INTERPLAY BETWEEN MULTISENSORY INTEGRATION AND ATTENTION 459
ADAPTING MULTISENSORY PROCESSING TO A DYNAMIC WORLD 462
SUMMARY 466

INTRODUCTION

In busy traffic, our senses are inundated with myriad signals: the noise of a truck passing by at high speed, the smell of fumes, and the sight of flashing traffic lights and other pedestrians. When deciding whether to cross the street, we should estimate the truck's speed by integrating motion information from audition and vision and yet avoid misperceiving the truck as talking and flashing. The effortless ease with which we constantly transform subsets of these signals into seamless percepts masks the complexities of the computations and neural mechanisms involved.

While all of our senses constantly furnish uncertain information about the current state of the world, they are specialized for gathering different types of information. Foveal vision impresses with its spatial precision at daytime but deteriorates in darkness and is blind to sources behind us. Audition informs us about sources outside our field of view or occluded by visual clutter; moreover, it exceeds vision in its temporal precision. These considerations illustrate the extraordinary benefits that we gain from integrating information across the senses, combining their complementary strengths and overcoming their individual frailties. Multisensory integration is a key strategy for the brain to resolve perceptual ambiguities and reduce uncertainties about the world (Ernst & Bülthoff 2004, Fetsch et al. 2013, Noppeney 2020).

Perceptual inference and accurate decision-making require integrating information in a way that is suitably sensitive to the constraints of the physical world, observers' uncertainties, and the brain's limited computational resources. Most notably, multisensory perception relies on solving the binding or causal inference problem: deciding which signals come from a common cause and integrating just those accordingly (Körding et al. 2007a, Sato et al. 2007, Shams & Beierholm 2010). Given a common cause, integration involves weighting noisy signals along with prior knowledge according to their relative precisions (i.e., inverse of uncertainty or variance) (Ernst & Banks 2002, Fetsch et al. 2013). The key to multisensory perception is thus the estimation and representation of different types of ignorance, including prior, sensory, and causal uncertainty. Priors that are hardwired in brain structure, fine-tuned during neurodevelopment, or dynamically learned at multiple timescales based on everyday experiences further inform the interpretation of the noisy sensory inputs. They can pertain to environmental properties (e.g., spatial priors), to the world's causal structure (i.e., causal priors), or, as hyperpriors, to the precision of the sensory signals.

Multisensory integration is closely intertwined with attentional mechanisms. Binding can increase stimulus salience and thereby attract attentional resources. Conversely, attention influences how and the extent to which we combine signals for perceptual inference and decisions (Talsma et al. 2010). In sum, multisensory perception raises some of the most fundamental questions for

neural processing, notably probabilistic computations (Ma & Jazayeri 2014, Pouget et al. 2013), inference, binding, learning, and attention.

In this review, we first discuss how the brain integrates sensory signals that emanate from the same source into more precise estimates. We then consider situations that entail causal uncertainty, because signals can come from common or independent sources. In these situations, the brain needs to arbitrate between sensory integration and segregation. Formally, this goes beyond estimating a simple environmental property and involves inferring a causal model of the world (i.e., structure inference). Next, we describe the complex interactions between multisensory perception and attention. We argue that attentional mechanisms are particularly important in naturalistic environments in which complex time-varying signals can arise from numerous sources. Finally, we review how the brain dynamically adapts multisensory processing to changes in the environment and its sensorium across multiple timescales. We review behavioral, computational, and neural research focusing mainly on humans, though we also refer to the most relevant research involving other species.

Our review is guided by models of normative Bayesian inference. Following von Helmholtz's (1867) notion of unconscious inference, these models posit that observers form a probabilistic generative model of possible sensory signals that is inverted during perceptual inference (Kersten et al. 2004). Bayesian probability theory sets a benchmark of optimal behavior against which observers' performances are compared. Crucially, exact Bayesian inference is computationally intractable for the brain, with its limited computational resources, in all but the simplest laboratory scenarios. In this review, we use Bayesian observer models as departure points to assess the extent to which observers perform statistically optimal computations in simple experimental settings and how they approximate these optimal solutions when facing resource constraints in progressively more complex scenarios (Ma 2012, Shen & Ma 2016). In particular, we explore the idea that attentional mechanisms are recruited in the service of these approximations.

INTEGRATING SENSORY INFORMATION FOR PERCEPTUAL INFERENCE AND DECISION-MAKING

Our senses provide complementary and redundant information about the environment. For instance, color and pitch are complementary, being experienced solely by either vision or audition. By contrast, information about an object's location can be gathered jointly from audition and vision and may in this sense be considered redundant. Critically, because sensory signals are inevitably degraded by various sources of internal and external noise (Faisal et al. 2008), integrating redundant information across the senses is an important way for the brain to reduce perceptual uncertainty.

Within a normative Bayesian framework, this second scenario can be accounted for by a generative model in which a single source emits several sensory signals, each corrupted by independent Gaussian noise. During inference, as formalized in the recognition model, a Bayesian observer should therefore integrate sensory inputs mandatorily into a more reliable (i.e., less variable) percept by weighting them in accordance with their precisions, giving greater emphasis to more reliable signals (Ernst & Bülthoff 2004, Hillis et al. 2002). Under statistical optimality, the reliability of the multisensory estimates should be equal to the sum of the unisensory reliabilities. Hence, a maximal gain in reliability (or reduction in uncertainty) by a factor of two is obtained when the two unisensory signals are equally reliable. In the following, we refer to this recognition model as the forced fusion model because it does not allow sensory signals to be processed independently. Under uninformative priors, it is equivalent to maximum likelihood estimation.

A key question is whether human observers statistically optimally integrate sensory signals from a common source, that is, whether they weight signals in proportion to their relative

reliabilities and obtain a multisensory variance reduction as predicted by the forced fusion model. Suboptimalities arise if the brain cannot access or estimate its own sensory uncertainties. Experimentally, statistical optimality is assessed by comparing observers' sensory weights and multisensory variances to the predictions of the forced fusion model that are made solely based on observers' unisensory variances.

Over the past two decades, a wealth of studies have explored these questions across a variety of sensory combinations and tasks. Overall, their results turned multisensory integration into a poster child for human statistical optimality. Statistical near-optimality was shown for visuohaptic shape (Ernst & Banks 2002), audiovisual rate (Raposo et al. 2012), and audiovisual spatial discrimination (Alais & Burr 2004). The observation of near-optimal integration even in experiments in which sensory uncertainties varied randomly across trials suggests that the brain can access sensory uncertainty near-instantaneously. Likewise, observers integrated audiovisual signals near-optimally in object or phoneme categorization tasks that require the brain to combine signals with weights that adequately reflect the sensory noise and the distributional properties of the task-relevant categories (Bankieris et al. 2017, Bejjanki et al. 2011). Further, many multisensory illusions such as spatial ventriloquism and McGurk or sound-induced flash illusions can be explained by normative integration of signals that are brought into small conflicts (Alais & Burr 2004, Bejjanki et al. 2011, McGurk & MacDonald 1976, Shams et al. 2005). While we focus in this review on integration across the senses, it should be noted that the normative principles of the forced fusion model also apply to cue combination within a single sense [e.g., combining stereo and texture cues for slant judgments (Knill & Saunders 2003)].

Critically, evidence for statistical optimality has never been unequivocal. Most notably, two early influential studies of audiovisual spatial classification disagreed on whether observers integrate signals optimally or overweight visual signals (Alais & Burr 2004, Battaglia et al. 2003). In a large-scale preregistered study, we replicated observers' visual overweighting but observed optimal audiovisual variance reduction consistent with the forced fusion model (Meijer et al. 2019). In additional simulations, this surprising dissociation was explained by the greater sensitivity with which experimental procedures and analysis approaches detect deviations from optimality for sensory weights than they do for multisensory variances. In line with this conjecture, more studies have indicated a tendency for observers to overweight signals from the sensory modality that is usually more informative for the particular task (Burr et al. 2009, Butler et al. 2010, Ernst & Banks 2002, Fetsch et al. 2009, Gepshtein & Banks 2003, Rosas et al. 2005). Deviations from optimal weights may even be more prominent in naïve observers that have not been exposed to the extensive training and stimulus familiarization that are characteristic for this type of psychophysics research (for discussion, see Alais & Burr 2004). One idea is that modest suboptimalities arise because the brain estimates sensory reliability by combining noisy information from current signals with a reliability prior that incorporates observers' prior assumptions about a modality's reliability based on lifelong experience (Battaglia et al. 2003) and/or previous stimuli (see the section titled Adapting Multisensory Processing to a Dynamic World). Putative deviations from statistical optimality may also arise when model assumptions about observers' priors, cost functions, or the structure and sources of noise are not met (for reviews, see Meijer & Noppeney 2020, Rahnev & Denison 2018). For instance, sensory noise may be non-Gaussian or correlated across the senses. Further, decisional noise may corrupt perceptual estimates after sensory integration.

Most importantly, the structure of the standard forced fusion model is limited in that it does not incorporate observers' decisional dynamics. Accounting for observers' decisional dynamics is important, because many decisions in everyday life, such as deciding whether to cross a street, are made under time pressure. If observers maximize not only their response accuracy but also their speed, this will introduce complex speed-accuracy trade-offs, thereby altering what is

considered optimal. Traditionally, computational modelling in multisensory integration did not account for these speed-accuracy trade-offs, instead being dominated by the dichotomy of non-dynamic forced fusion and race models. While forced fusion models made predictions for only response choices and not response times, race models ignored response choices and the signals' time-varying reliabilities. Instead, race models compared multisensory and unisensory response time distributions to assess whether observers accumulate information independently or interactively across the senses (Miller 1982, Otto & Mamassian 2012). Only recently have drift-diffusion models (Forstmann et al. 2016, Gold & Shadlen 2007) been adapted to multisensory decision-making. For multisensory decisions, they integrate sensory evidence at each time point weighted by their momentary reliabilities and accumulate this integrated evidence over time until a decisional threshold. Consistent with the model's predictions, observers were shown to discriminate heading direction by combining visual motion and vestibular acceleration information (i.e., different physical quantities) weighted by their momentary reliabilities that evolved with different time courses (Drugowitsch et al. 2014, 2015). Moreover, observers' putatively suboptimal visuovestibular integration for heading discrimination based on a standard forced fusion model turned out to be optimal when observers' speed-accuracy trade-offs and the varying signal reliabilities were accounted for (Drugowitsch et al. 2014, 2015). Conversely, differences in speed-accuracy trade-offs between unisensory and multisensory conditions may also explain the perhaps surprising finding of multisensory integration benefits that appear to be supraoptimal when compared to standard forced fusion models (Nikbakht et al. 2018, Raposo et al. 2012).

At the neural level, multisensory interactions are known to be pervasive in neocortex, starting at the primary cortical level and increasing progressively along the cortical hierarchy (Atilgan et al. 2018; Bizley et al. 2007; Fiebelkorn et al. 2010; Gau et al. 2020; Ghazanfar & Schroeder 2006; Kayser et al. 2007, 2008; Lakatos et al. 2007; Martuzzi et al. 2007; Metzger et al. 2020; Noesselt et al. 2007; Schroeder & Foxe 2002; Werner & Noppeney 2010a,b). This ubiquity of multisensory interactions raises the question of which of those interactions reflect the reliability weighting evident in observers' behavior. Noninvasive whole-brain imaging has suggested that reliability-weighted integration arises in a variety of mid-level sensory and association cortices depending on the information to be integrated. For instance, human functional magnetic resonance imaging (fMRI) revealed reliability-weighted integration of visuohaptic shape information in parietal and fusiform cortices (Helbig et al. 2012) and of audiovisual spatial information in parietal cortices (Rohe & Noppeney 2018). For audiovisual speech comprehension, visual reliability modulated the functional connectivity between the superior temporal sulcus with visual and auditory areas (Nath & Beauchamp 2011).

Focusing on the dorsal medial superior temporal area (MSTd), a series of elegant neurophysiological studies in nonhuman primates investigated how neural populations and single neurons integrate visual and vestibular motion for heading discrimination. Behaviorally, macaques integrated visual and vestibular motion signals near-optimally with a modest vestibular overweighting, as observed in humans (Fetsch et al. 2012). Likewise, in MSTd, neural populations and single neurons with congruent visuotactile heading preferences integrated visual and vestibular signals weighted by their relative reliabilities (Fetsch et al. 2012), leading to a greater sensitivity (Gu et al. 2008). Further, electrical microstimulation and chemical inactivation confirmed the causal relevance of the neural computations in MSTd for heading discrimination (Gu et al. 2012).

A more recent study compared the roles of MSTd and lateral intraparietal area (LIP) in visuovestibular heading discrimination (Hou et al. 2019). Intriguingly, MSTd integrated visual and vestibular motion information within brief time windows, while LIP accumulated the momentary evidence proportional to the visual speed and the absolute value of vestibular acceleration over time. Thus, the choice-related activity in LIP ramped up with different time courses for visual,

vestibular, and visuovestibular heading discrimination. Additional computational analyses showed that LIP activity was consistent with the activity of a neural network, more specifically its integration layer, which implements near-optimal multisensory decision-making via linear invariant probabilistic population coding.

Within this theoretical framework, neural population activity is thought to encode probability distributions over inputs, with the amplitude of the neuronal responses being proportional to the reliability of its inputs (Beck et al. 2008, Ma et al. 2006). As a result, neurons can implement reliability-weighted integration by linearly combining their inputs across sensory modalities and time with fixed neural weights. In other words, the synaptic weights do not depend on the sensory reliability for implementing reliability-weighted integration. Probabilistic population coding thus provides an elegant and simple solution not only for the static forced fusion case but also for dynamic integration processes as observed in LIP, in which sensory evidence is integrated across the senses and time until a decisional threshold (Hou et al. 2019).

Alternatively, reliability-weighted integration may be mediated by mechanisms of divisive normalization (Ohshiro et al. 2011), a canonical neural computation previously implicated in nonlinear stimulus interactions in primary visual cortex, motion integration, and attentional modulation (Carandini & Heeger 2011). Within models of divisive normalization, multisensory neurons initially combine their unisensory inputs linearly. Reliability weighting arises because, after a nonlinear transformation, the output of each multisensory neuron is normalized by the net activity of a pool of neurons. While divisive normalization does not guarantee statistical optimality, it can explain not only reliability weighting but also a whole basket of neural response profiles that are frequently observed in multisensory processing such as inverse effectiveness (Stanford et al. 2005) and multisensory response suppression (Ohshiro et al. 2017).

SOLVING THE BINDING PROBLEM: CAUSAL INFERENCE

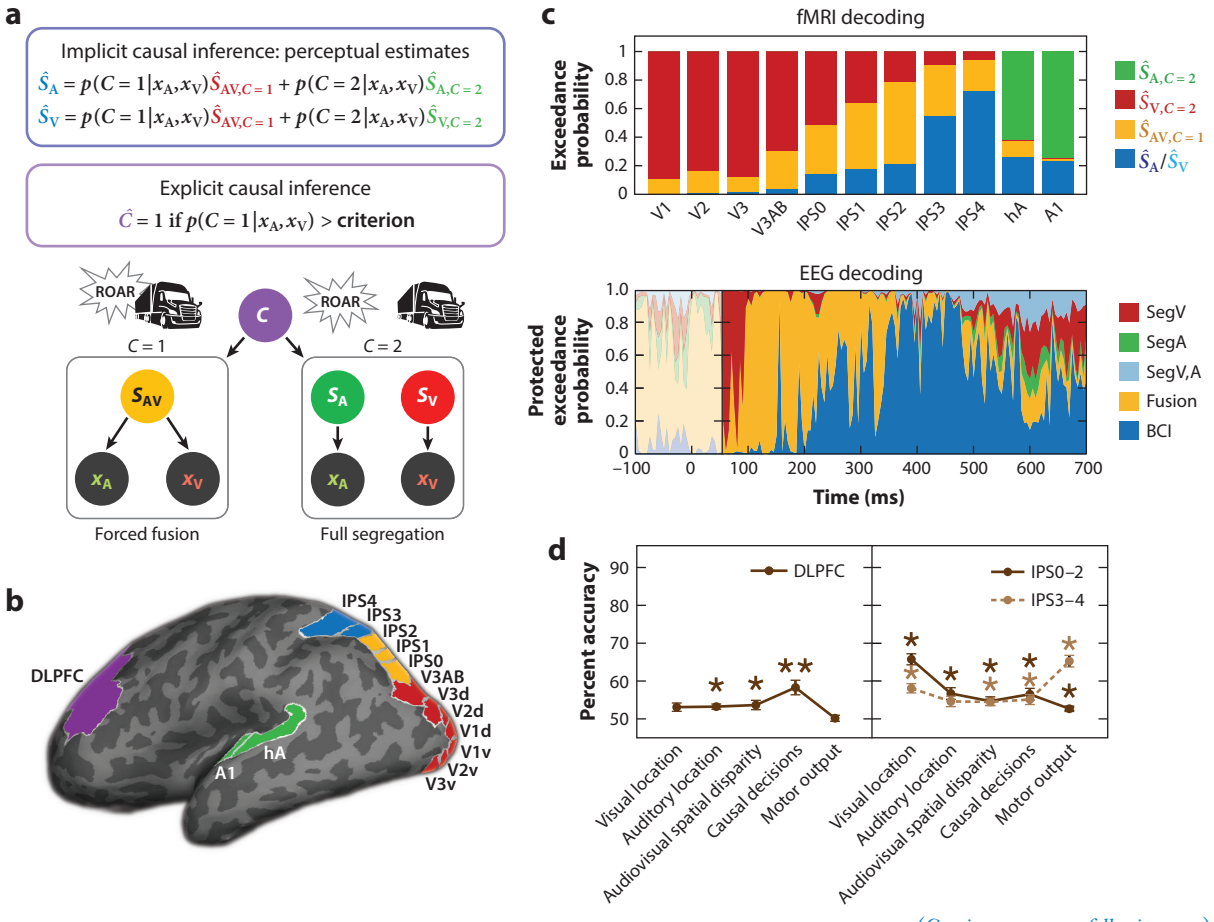
In complex environments, observers need to solve the binding or causal inference problem—determining whether signals come from common sources and should hence be integrated or else be treated independently. Ample evidence has shown that perceptual illusions, metamers, and, more generally, crossmodal biases break down at large intersensory conflicts when common sources are unlikely (Bertelson & Radeau 1981, Hillis et al. 2002, Lewald & Guski 2003, Magnotti & Beauchamp 2017, Magnotti et al. 2013, Rohe & Noppeney 2015b, Wallace et al. 2004). Likewise, multisensory benefits such as gains in precision are smaller than predicted by the forced fusion model when auditory and visual signals are temporally uncorrelated (Locke & Landy 2017, Parise et al. 2012).

Bayesian causal inference accounts for this binding problem by explicitly modelling the potential causal structures (here, common versus independent sources) that could have generated the sensory signals (Körding et al. 2007a, Sato et al. 2007, Shams & Beierholm 2010) (**Figure 1**). During inference, signals from common sources are integrated weighted by their relative reliabilities (i.e., forced fusion). Signals from independent sources are processed independently (i.e., segregation). Critically, observers do not know and need to infer the signals' causal structure by combining evidence from temporal, spatial, or higher statistical (e.g., phonetic, semantic) correspondences with a causal prior that quantifies observers' prior tendency to bind signals. To account for observers' causal uncertainty, a final perceptual estimate is computed by averaging the estimates of the environmental property (e.g., spatial location) under the potential causal structures weighted by their respective posterior probabilities [i.e., model averaging; for other decision functions, see Wozny et al. (2010)].

To study Bayesian causal inference, observers are typically presented with, for example, audiovisual signals that vary randomly in their conflict sizes across trials. In explicit causal inference

tasks, they report whether signals come from common or independent sources. In implicit causal inference tasks, they report their auditory and/or visual percept rather than their audiovisual percept as in the forced fusion experiments discussed in the previous section. In these tasks, observers' causal inference implicitly influences their perceptual estimates. Hierarchical Bayesian causal inference provides a principled explanation for the inverted U-shaped function that describes how observers' explicit common-cause judgments decline with positive (e.g., auditory leading) and negative (e.g., visual leading) temporal, spatial, or other conflicts (Bertelson & Radeau 1981, Lewald & Guski 2003, Magnotti & Beauchamp 2017, Magnotti et al. 2013, Rohe & Noppeney 2015b, van Wassenhove et al. 2007, Wallace et al. 2004). Likewise, the influence of causal inference on observers' perceptual estimates can explain that crossmodal biases (e.g., spatial, rate, numerosity, heading direction) depend nonlinearly on conflict size (Acerbi et al. 2018, de Winkel et al. 2018, Mohl et al. 2020, Rohe & Noppeney 2015b). Further, a difference in causal prior or binding tendency has been shown to provide a computational explanation for the finding that integration between senses is less tolerant to cue conflicts than integration within a sense (Hillis et al. 2002, Hospedales & Vijayakumar 2009).

Most intriguingly, Bayesian causal inference makes qualitatively characteristic predictions for the trial-to-trial relationship between observers' implicit (e.g., spatial localization) and explicit



(*Caption appears on following page*)

Figure 1 (*Figure appears on preceding page*)

(*a*) The generative model of Bayesian causal inference explicitly models the potential causal structures that could have generated the sensory signals. For the audiovisual localization example, it models whether the sight of the truck and the truck's looming motor noise are generated by common ($C = 1$) or independent ($C = 2$) sources (Körding et al. 2007a). In the case of a common source, the audiovisual location (S_{AV}) is drawn from the prior spatial distribution. In the case of independent sources, the auditory (S_A) and visual (S_V) locations are drawn independently from this prior spatial distribution. Auditory (x_A) and visual (x_V) inputs are generated by adding independent sensory noise. Hierarchical Bayesian causal inference makes predictions for two related inference tasks: observers' explicit causal inferences, i.e., decisions about whether signals come from common or independent causes, and their perceptual estimates that are implicitly informed by their causal inference. Panel *a* adapted with permission from Noppeney (2020). (*b*) Bayesian causal inference is performed by encoding several spatial estimates along the cortical hierarchy. The regions of interest are shown on the surface of an inflated brain. For illustrational purposes, the regions are colored in correspondence with the estimates with which they are most closely associated from the BCI (panel *a*) based on the neuroimaging results shown in panels *c* and *d*. (*c*) In an auditory/visual spatial localization task, observers were presented with brief audiovisual signals at variable levels of spatial disparity; fMRI (*top row*) and EEG (*bottom row*) decoding reveal the spatiotemporal evolution of Bayesian causal inference in spatial perception. Early activity (<100 ms) in auditory and visual areas is associated with segregation [i.e., separate auditory (*green*) or visual (*red*) estimates], later activity (100–250 ms) in posterior parietal areas with forced fusion, and finally, activity in anterior parietal areas (350–450 ms) with Bayesian causal inference. Anterior parietal activity forms spatial estimates that arbitrate between sensory integration and segregation depending on the signals' causal structure. The exceedance probabilities for the different spatial estimates (fMRI) or models (EEG) are indexed by the length of the colored areas of each bar (note that the *y* axis indicates the cumulative exceedance probabilities). Panel *c* adapted from Rohe & Noppeney (2015a) for fMRI and Aller & Noppeney (2019) for EEG. (*d*) Explicit causal inference in spatial perception: In a common source judgment task, observers were presented with audiovisual signals that were spatially congruent or incongruent. We adjusted the spatial disparity individually for each participant to threshold performance, allowing us to dissociate their causal decisions from the signals' physical spatial disparity. Further, we counterbalanced the mapping from observers' causal decisions to their motor choices over multiple runs. The line plots show the fMRI decoding accuracy for visual location, auditory location, audiovisual spatial disparity, causal decisions, and motor output for DLPFC, IPS0–2, and IPS3–4 ($* = p < 0.05$, $** = p < 0.005$). DLPFC was the only region that selectively encoded observers' causal decisions. By contrast, the parietal cortices jointly encoded visual location, auditory location, audiovisual spatial disparity (i.e., congruent versus incongruent), causal decisions, and motor output. Panel *d* adapted from Mihalik & Noppeney (2020). Abbreviations: A1, primary auditory cortices; BCI, Bayesian causal inference model; DLPFC, dorsolateral prefrontal cortex; EEG, electroencephalogram; fMRI, functional magnetic resonance imaging; Fusion, forced fusion model; hA, higher auditory cortices; IPS, intraparietal sulcus; SegA, full segregation auditory model; SegV, full segregation visual model; SegV,A, full segregation model, in which the full segregation auditory or visual estimates are reported depending on whether the auditory or visual modality is task relevant; V, primary and higher order visual cortices.

causal inference. Because the two inferences are affected by the same sensory noise on each trial, observers' perceptual estimates from one sensory modality should be biased toward those from the other sensory modality when common sources are perceived and repulsed for independent-source judgments (Körding et al. 2007a). Indeed, one previous behavioral study corroborated this characteristic relationship between attractive/repulsive perceptual biases and observers' causal judgments in a dual task paradigm (Wallace et al. 2004). However, a more recent study failed to replicate this qualitative behavioral profile (Rohe & Noppeney 2015b).

Further, when the Bayesian causal inference model was fitted jointly across multiple sensory reliability levels, observers' crossmodal biases and response variability differed markedly from the model's predictions (Rohe & Noppeney 2015b). To elucidate potential sources of suboptimalities more formally, a recent study of visuovestibular heading discrimination compared Bayesian causal inference with a variety of other approximate strategies (Acerbi et al. 2018). The study showed that observers arbitrate between sensory integration and segregation based on a fixed conflict size without taking into account their sensory uncertainty. This is the first model-based indication that even in these simple scenarios observers do not perform consistent with normative Bayesian causal inference but resort to approximate, that is, suboptimal, strategies.

Within experiments studying Bayesian causal inference, we can also ask whether the forced fusion principles hold on the subset of trials that were congruent or included a small conflict. Results showed that observers' integration performance deviated substantially from the forced fusion principles. For instance, on audiovisual trials with a small conflict, observers reported

different auditory and visual percepts giving a stronger weight to the task-relevant sensory modality. Even for collocated audiovisual signals, observers' auditory and visual percepts were associated with different variances (Rohe & Noppeney 2018). This failure to fuse synchronous and collocated audiovisual signals into one unified percept may reflect observers' remaining causal uncertainty, because collocated signals occurred randomly interspersed with spatially disparate signals. Moreover, observers may have lowered their causal prior or binding tendency because they were instructed to attend to and selectively report their percept in one sensory modality (see the next section for further discussion of the role of attention in multisensory processing).

At the neural level, cross-sensory correspondences and conflicts have been widely recognized as key determinants of multisensory interactions since the seminal neurophysiological work by Stein and colleagues (for a review, see Stein & Stanford 2008). In the superior colliculus, multisensory response enhancement for spatiotemporally coincident signals turns into suppression for spatiotemporally disparate signals (Meredith & Stein 1983, Meredith et al. 1987). Within neocortex, the computation of multisensory correspondences appears to rely on a widespread system of regions. Temporal synchrony and correlations have been associated with primary sensory and superior temporal sulci based on a wealth of neuroimaging and neurophysiology research (Lee & Noppeney 2011a, 2014; Lewis & Noppeney 2010; Miller & D'Esposito 2005; Noesselt et al. 2007; Powers et al. 2012). Notably, time-varying visual signals enhanced the encoding of temporally correlated auditory signals in primary auditory cortices, even when anesthesia largely obliterated top-down influences. Initial causal inference based on temporal correlations and coincidence alone may thus occur in sensory cortices during anesthesia (Atilgan et al. 2018). Relatively simple processing units similar to the Hassenstein-Reichard motion detector (Parise & Ernst 2016) or mechanisms of phase resetting (Kayser et al. 2008, Lakatos et al. 2007, Mercier et al. 2013, Zumer et al. 2021) may potentially support the computation of multisensory coincidence, correlations, and response enhancement for nearly synchronous stimuli in early sensory areas.

By contrast, the computation of audiovisual spatial disparity that involves transformations across different reference frames (i.e., eye versus head centered) and representational formats (i.e., topographic versus hemifield code) (Maier & Groh 2009) relies predominantly on planum temporal and parietal cortices (Mihalik & Noppeney 2020), key players in auditory and visual spatial processing (Ortiz-Rios et al. 2017, Rauschecker & Tian 2000, Schlack et al. 2005). Finally, phonetic correspondences have been associated with superior temporal cortices and semantic correspondences with temporal/fusiform cortices (Calvert et al. 2000, Hein et al. 2007, Lee & Noppeney 2011b, Noppeney et al. 2008). Given that cross-sensory correspondences are computed within a widely distributed neural system at variable poststimulus times, a key challenge is to understand how the brain dynamically combines them via bottom-up, top-down, and lateral connections for causal and perceptual inference. In particular, dorsolateral prefrontal cortex may be important for accumulating evidence from diverse correspondences about the world's causal structure and in turn controlling sensory information flow and integration via recurrent loops across the cortical hierarchy (Gau & Noppeney 2016, Noppeney et al. 2010).

To characterize how the brain uses cross-sensory correspondences for spatial perception consistent with Bayesian causal inference, we presented observers with brief, simple audiovisual signals in synchrony but at variable spatial disparities (Aller & Noppeney 2019; Rohe & Noppeney 2015a, 2016). Consistent with Bayesian causal inference, observers perceived the sound location shifted toward the visual location (and vice versa) at small spatial disparities, but these crossmodal spatial biases were attenuated at large spatial disparities. Multivariate fMRI/electroencephalogram (EEG) analyses showed that the brain accomplishes Bayesian causal inference by dynamically encoding multiple spatial estimates across the cortical hierarchy (**Figure 1**). Early activity (50–100 ms) in

primary sensory cortices encoded the unisensory spatial estimates (i.e., segregation). Later activity (100–200 ms) in posterior parietal cortices merged signals into audiovisual spatial estimates (i.e., fusion). Even later activity (350–450 ms) in anterior parietal cortices integrated signals at small spatial disparities but segregated them at large spatial disparities. They thus formed spatial estimates that took into account the signal's causal structure as predicted by Bayesian causal inference.

Next, we asked where and how the brain infers the signal's causal structure based on audiovisual synchrony and varying spatial correspondences. In an explicit causal inference task, observers decided in each trial whether signals came from common or independent sources (Mihalik & Noppeney 2020). To dissociate observers' causal decisions from the signals' physical spatial disparity, we adjusted audiovisual spatial disparity individually for each participant to threshold performance. Our results showed that the dorsolateral prefrontal cortex encoded mainly observers' causal decisions irrespective of physical spatial disparity or observers' motor responses. By contrast, a circuitry encompassing the frontal eye fields and the parietal cortices encoded auditory/visual locations, physical spatial disparity, observers' causal decisions, and their motor responses. Collectively, these results suggest that parietal cortices may form spatial representations informed by causal decisions encoded in dorsolateral prefrontal cortices.

Recent human magnetoencephalography (MEG)/EEG studies moved beyond spatial perception to investigate how the brain performs Bayesian causal inference for estimating the number of events (Rohe et al. 2019) or the temporal rate (Cao et al. 2019) of auditory/visual pulse trains. These studies replicated the dynamic encoding of segregation, fusion, and Bayesian causal inference estimates across the cortical hierarchy. Yet, not surprisingly, the timing and regions differed from those observed for spatial tasks. For instance, a recent MEG study associated auditory and visual rate estimates informed by Bayesian causal inference with late activity in anterior frontal areas at about 600–660 ms after stimulus onset and 200–0 ms before the observer's motor response. This late timing and anterior frontal dominance most likely arose because observers needed to estimate the auditory/visual rate over 550-ms pulse trains and map those estimates onto arbitrary categories and motor choices (Cao et al. 2019).

Likewise, an EEG study of the sound-induced flash illusion associated Bayesian causal inference estimates of the number of flashes/beeps with neural activity at about 550–650 ms after stimulus onset (Rohe et al. 2019), which corresponds to 350–450 ms after stimulus offset, as reported above for spatial tasks. This study also investigated the relationship between prestimulus alpha oscillations, previously implicated in inhibitory and attentional mechanisms, and observers' trial-specific causal prior (i.e., binding tendency). The study showed that lower prestimulus alpha power and specific alpha phases were associated with stronger audiovisual binding. However, while observers' causal prior dynamically adapted to the sensory input statistics as expected for a Bayesian learner (see the section titled Adapting Multisensory Processing to a Dynamic World), this was not the case for prestimulus alpha power (Rohe et al. 2019). Thus, spontaneous fluctuations in prestimulus alpha power dynamically set the functional neural system into states that facilitate or inhibit interactions between sensory modalities independent of the sensory input statistics.

In sum, accumulating evidence has shown that the brain dynamically computes Bayesian causal inference across the cortical hierarchy. Only late neural activity in association cortices encoded perceptual estimates that were susceptible to influences from other sensory modalities depending on the signals' causal structure. However, because these studies directed observers' attention to either the auditory or visual modalities prior to stimulus onset, they conflated Bayesian causal inference with modality-specific attention. Cast in attentional terms, Bayesian causal inference has enabled the brain to filter out signals from the unattended modality when these signals were likely to arise from independent sources, a perspective explored further in the next section.

THE MULTIFARIOUS INTERPLAY BETWEEN MULTISENSORY INTEGRATION AND ATTENTION

The interplay between multisensory interactions and attention is complex and multifaceted (Talsma et al. 2010). Both can express similar neural and behavioral signatures such as increases in neural responses, precision of sensory representations, and perceptual sensitivity (Ernst & Bülthoff 2004, Maunsell 2015). In this section, we first focus on how top-down attention can influence multisensory processing. We then turn to how multisensory interactions may increase stimulus salience and thereby attract attentional resources (i.e., bottom-up attention).

Ample evidence has shown that the brain can eliminate, suppress, or amplify the influence of sensory information on perceptual inference and decisions—a phenomenon commonly ascribed to various forms of attention, acting on sensory modalities, features, or space. Yet, this descriptive formulation is agnostic as to why and how these attentional influences come about. Moving beyond the use of attention as a descriptive term, we distinguish between attention as normative statistical inference and attention as a process to implement approximate inference in response to resource constraints (Dayan et al. 2000, Dayan & Solomon 2010). In the former, task or context informs the brain that some signals are more relevant than others, making it computationally sensible to suppress or even discard the less relevant signals by adjusting priors or cost functions. Changes in priors alter observers' posterior distribution and thereby their perceptual estimates and uncertainties. By contrast, cost functions influence solely the readout of the perceptual estimate from the posterior distribution, leaving observers' uncertainties unchanged. However, when observers recurrently sample from the environment, cost functions can also alter their uncertainties by influencing the sampling of information from the environment. Critically, adjusting priors or cost functions to incorporate information from the task or context may eliminate the influence of less relevant information, consistent with normative principles. By contrast, in other situations such as in the face of myriad signals, discarding information may be required because of the brain's limited computational resources, but it is not optimal for task performance. Selective attention then forms an algorithmic realization of approximate, that is, suboptimal, inference. Taking Bayesian causal inference as a departure point, we next discuss how attention as normative inference and as a process to compute approximate solutions in response to resource constraints shapes multisensory processing at the behavioral and neural levels, although we note that it is impossible to cover the wide spectrum of neural mechanisms previously implicated in attention (e.g., for a review, see Maunsell 2015). We illustrate these two distinct aspects of attention focusing on attention to sensory modalities, types of information (e.g., spatial versus phonetic), and spatial locations (e.g., left versus right hemifield).

Almost all research to date has studied Bayesian causal inference closely intertwined with modality-specific attention by making observers report their percept solely in one sensory modality throughout an entire run. This seems a natural experimental choice, because in everyday life, observers also optimize sensory processing during stimulus presentation for the particular information they need to make an inference about. Yet, this experimental choice introduces intimate links between Bayesian causal inference and modality-specific attention, because the relevance of signals in the unreported sensory modality depends on the causal structure of the two signals. Signals in the unreported modality are informative when they happen to come from the same source as those in the reported modality, but they are otherwise uninformative and interfering. With increasing certainty that signals come from independent sources, a normative observer should therefore progressively discard information from the unreported modality, which results in complex nonlinearities in the decision processes (Dayan & Solomon 2010, Yu et al. 2009). Consistent with this conjecture, observers' categorization of a sound source was influenced more strongly by

incongruent visual information at shorter response times than at longer response times, when more evidence had been accumulated about the signals' causal structure (Jones et al. 2019, Noppeney et al. 2010). These results indicate that modality-specific attention arises from dynamic interactions with causal inference leading to a progressive elimination of signals from the unreported sensory modality in separate-source situations.

Paradigms in which observers attend (prior/during stimulus presentation) and report (poststimulus) their percept in the same sensory modality are also condemned to assess the joint influence of modality-specific attention that occurs prior to, during, and after stimulus processing. Only recently have studies started to dissociate these aspects in pre/post-cueing paradigms. They demonstrated that modality-specific attention influences observers' causal priors, reliability weighting, and task-relevant readout. For instance, pre-cueing observers to attend to one rather than both sensory modalities attenuated overall sensory noise and decreased observers' causal prior, that is, their prior binding tendency, thereby enabling a faster elimination of influences from the irrelevant sensory modality (Badde et al. 2020). In EEG, this may be reflected in altered early multisensory interactions at about 50 ms poststimulus (Talsma et al. 2006). Other psychophysics and fMRI studies pre-cued human observers to attend to the visual (or auditory) modality and post-cued them 400 ms after stimulus presentation to report the location of either the attended or unattended signal (A. Ferrari & U. Noppeney, unpublished manuscript). Prestimulus attention influenced multisensory processing by enhancing the precision and hence the influence of spatial representations of the attended sensory modality in the integration processes in mid-level visual and posterior parietal cortices. By contrast, poststimulus report controlled the readout of the task-relevant estimates possibly from heterogeneous neural populations in anterior parietal cortices (Hou et al. 2019). Because the pre-cues in this study were 50% valid (i.e., uninformative), this study also shows that human observers can optimize statistical inference voluntarily according to task instructions even when knowing that following instructions does not improve their performance. However, anecdotal comparison across studies suggests that the impact of modality-specific attention depends also on cue validity (i.e., 50% versus 100%), as expected under normative inference (Rohe & Noppeney 2015a; A. Ferrari & U. Noppeney, unpublished manuscript).

Tasks that direct observers' attention to various (e.g., spatial, phonetic) types of information are another way to mold causal and perceptual inference. The most decisive evidence comes from a dual task design in which observers reported both the sound location and the perceived phoneme for audiovisual viseme/phoneme stimuli that varied in their spatial disparity and phonetic correspondences. The study revealed a double dissociation with the McGurk illusion (i.e., phoneme task), dependent on phonetic but not spatial correspondences, and the ventriloquist illusion (i.e., spatial task), sensitive to spatial disparity but not to the availability of phonetic correspondences (Bertelson et al. 1994, Bishop & Miller 2011). While recent work has unraveled small influences of phonetic or semantic correspondences on spatial ventriloquism (Delong & Noppeney 2021, Kanaya & Yokosawa 2011), the task-dependent susceptibilities to various cross-sensory conflicts remain uncontested (for reviews, see Chen & Spence 2017, Chen & Vroomen 2013). Put simply, causal inference or binding relies more strongly on the congruency of the information that is relevant for the particular perceptual task (i.e., spatial congruency for localization and phonetic congruency for phoneme categorization). As discussed above, these task dependencies may result from changes in observers' priors or cost functions. For instance, observers may accumulate more information about the task-relevant correspondences so that those are more precisely estimated and hence receive greater weights in causal inference (Rohe & Noppeney 2015b). However, from a normative perspective one may argue that observers should combine spatial and phonetic correspondences to arbitrate between sensory integration and segregation in speaker localization tasks. The discarding of precious correspondence information may thus reflect the brain's limited

computational resources that make the concurrent computation of reliable phonetic and spatial correspondences impossible. Consistent with this conjecture, the McGurk illusion falters under concurrent demanding auditory or visual detection tasks (Alsius et al. 2005, 2014), again suggesting that the extraction of fine-grained phonetic information breaks down in the face of resource constraints.

Spatial attention is arguably the most important form of attention for multisensory processing, starting from simple lab experiments with only two sensory signals to complex naturalistic situations with myriad signals. Even in simple lab experiments with only two signals, spatial attention may mold perceptual inference via multiple mechanisms. Consistent with normative principles, prior knowledge about where something happens may, for instance, increase observers' precision about their estimates inside the spatial spotlight by eliminating noise from elsewhere, which in turn impacts causal inference and the relative weighting of signals inside and outside the attentional focus (Rohe & Noppeney 2015b). Alternatively, observers may pursue heuristics and simply integrate signals that co-occur within an attentional spotlight.

The multiplicity of mechanisms by which spatial attention can influence multisensory processing with even opposite effects on observers' perceptual outcome may explain why early influential studies did not reveal any effects of spatial attention on audiovisual integration in spatial ventriloquist paradigms (Bertelson et al. 2000). Since then, however, psychophysics and neuroimaging have indicated a profound impact of spatial attention on multisensory processing. Most notably, EEG and fMRI have shown effects of spatial attention on interactions of spatially disparate, yet synchronous, brief audiovisual inputs in planum temporale from 220 ms onward, that is, at latencies associated with integrating audiovisual signals into spatial representations (Bonath et al. 2007, Busse et al. 2005, Donohue et al. 2011). Likewise, the McGurk illusion that is considered to be relatively immune to spatial correspondences occurs more frequently when spatially focused attention is directed to the location of the sound (Tiippana et al. 2011), suggesting that spatial attention may enhance the integration of spatially disparate phoneme/viseme signals.

Solving the binding or causal inference problem consistent with normative principles becomes progressively more challenging with an increasing number of signals. Even for relatively simple lab experiments with only two auditory and two visual signals, a normative observer needs to compute the posterior probability over more than ten possible source combinations ranging from one common source to four independent sources. Further, in these situations, temporal, spatial, and other correspondences play dual roles—they guide binding within and between sensory modalities. A recent study assessed the influence of spatial attention on multisensory processing in a more complex sound-induced flash illusion paradigm; in this study, observers were presented with two flash-beep sequences bilaterally and reported the number of flashes they perceived, for example, in their left hemifield (Kumpik et al. 2014). While the sound-induced flash illusion is typically insensitive to spatial disparity, under spatial attention the reported number of flashes in the attended hemifield depended more strongly on the number of flashes/beeps in the attended rather than the unattended hemifield (Bizley et al. 2012). These results suggest that spatial attention suppresses the influence of signals in the unattended hemifield and possibly enhances the integration of collocated signals within an attentional spotlight. Likewise, when observers were presented with two speaker videos bilaterally and the two corresponding auditory speech streams in the center, the auditory speech stream of the attended video was associated with greater intertrial coherence and neural encoding/decoding of the acoustic envelope in auditory cortices (for MEG, see Park et al. 2018; for EEG, see Crosse et al. 2015, O'Sullivan et al. 2015; for electrocorticography, see Zion Golumbic et al. 2013; and for fMRI, see Fairhall & Macaluso 2009). These attentional benefits exceeded even those observed for pure auditory stimulation, highlighting the importance of

synergistic interactions between multisensory integration and spatial attention in these so-called cocktail party scenarios.

Scenarios in our natural environment are even richer, including myriad sources that may emit signals in one or more sensory modalities. In these complex multisensory scenes, solving the binding problem exactly is almost certainly intractable for the brain with its limited computational resources. The brain cannot compute the posterior probability distribution over all possible source combinations. Further, in many situations, such as when searching for a friend in a busy restaurant, observers do not even know the source location a priori in order to allocate spatial attention accordingly. How does the brain solve the causal inference problem in these complex, multisource environments to form a seemingly coherent percept? One idea is that the brain computes approximate solutions to the binding problem via attentional mechanisms that sequentially select and assess subsets of sensory signals within an attentional spotlight for causal inference and binding (see Treisman & Gelade 1980). By recurrently shifting the attentional spotlight, the brain then gathers progressively more information about the multisource environment, thereby forming an approximate posterior distribution. Consistent with such a serial strategy, observers' response times to visual targets in audiovisual search scenarios have been shown to increase linearly with the number of visual distractors (Alsius & Soto-Faraco 2011, Fujisaki et al. 2006).

Yet, the slope of this linear increase is shallower and almost flat when the spatially uninformative auditory signals (e.g., pulse trains, speech) that evolve in synchrony with the visual target features (e.g., luminance, articulatory facial movements) are salient by virtue of their signal strength and temporal structure (Stacey et al. 2014, Van der Burg et al. 2010). Most notably, brief, salient, spatially uninformative beeps can make a temporally correlated visual target pop out among distractors (pip and pop), which was associated in EEG with early (50 ms poststimulus) audiovisual interactions and later changes in the N2pc component, reflecting spatial bottom-up attraction of attentional resources (Van der Burg et al. 2008, 2011). Because the audiovisual benefit was greatest when the auditory signal was synchronous or slightly lagging behind the visual target, it is unlikely to reflect simple alertness effects. Instead, the brain appears to compute temporal correspondences between a salient beep and multiple concurrent visual signals at least partly via preattentive parallel processes. Likewise, viewing facial articulatory movements enables observers to search for and detect corresponding speech signals among auditory distractor signals rather independently of the number of auditory distractor signals (Alsius & Soto-Faraco 2011). These results dovetail nicely with other studies showing that a single visual signal amplifies the encoding of a temporally correlated auditory signal in primary auditory cortices even under anesthesia (Atilgan et al. 2018). Collectively, they raise the intriguing possibility that the brain may compute audiovisual saliency maps that not only benefit from the complementary spatial precision of vision (Itti & Koch 2001, Li 2002) and temporal precision of audition (Kayser et al. 2005) but also incorporate an initial tentative solution to the binding or causal inference problem based on temporal correspondences. These spatiotemporal salience maps go beyond linear combinations of auditory and visual salience maps to attract observers' attention to audiovisual events in a complex, dynamic world.

ADAPTING MULTISENSORY PROCESSING TO A DYNAMIC WORLD

Adapting dynamically to changes in the environment and the sensorium is a fundamental challenge facing the brain throughout the life span. Changes in sensory statistics evolve across multiple timescales ranging from milliseconds to years. The brain should adapt faster in a volatile world, when signals in the distant past are no longer relevant for the future, than in a stable world, when variations over time are more likely to reflect random fluctuations. While the previous section described how priors incorporate information from a task or context, this section reviews how

priors dynamically adapt to changes in the statistical structure of the environment across multiple timescales, focusing on changes in the signals' reliabilities and in the world's causal structure and its properties (e.g., objects' locations). Finally, we turn to how the brain calibrates the senses to keep sensory estimates internally consistent and accurate with respect to the outside world.

In our natural environment, sensory reliabilities typically evolve slowly. For instance, observers receive progressively more reliable spatial information from the looming noise of an approaching truck. A normative Bayesian learner should capitalize on this slow temporal dynamic and estimate the reliabilities of the sensory signals by combining information from present and past sensory inputs—the latter being formally incorporated in hyperpriors about sensory reliability. Consistent with these predictions, a recent study showed that observers integrate audiovisual signals weighted by reliabilities estimated from sensory inputs from up to four seconds in the past (Beierholm et al. 2020). Computational modelling showed that an optimal Bayesian as well as an exponential learner could capture this behavior equally well. Critically, while both learners rely more strongly on recent inputs, only the Bayesian learner adapts its learning rate dynamically based on its uncertainty about its reliability estimates. Collectively, these results extend current models (e.g., forced fusion, Bayesian causal inference) in which sensory reliabilities are estimated instantaneously and independently for each stimulus. They also raise the possibility that the brain resorts to approximate strategies of exponential discounting to adapt to the changing reliabilities of the sensory inputs.

To guide multisensory inference, the brain needs to learn not only about signals' reliabilities but also about the statistical structure over signals from various sensory modalities. Cross-sensory priors influence the integration of sensory signals and enable the prediction of unobserved cue values in one modality (e.g., predator's size) from observed cue values in another modality (e.g., acoustic scale of its vocalization). Importantly, observers' prior binding tendency (i.e., causal prior) needs to adapt dynamically to changes in the world's causal structure. For instance, observers should progressively reduce their causal prior when they listen to one person's speech and view another person's facial movements. Indeed, observers are less likely to integrate signals into McGurk or sound-induced flash illusions after a series of incongruent and/or asynchronous signals (Gau & Noppeney 2016; Nahorna et al. 2012, 2015; Rohe et al. 2019). At the neural level, fMRI research has shown that the dorsolateral prefrontal cortex combines prior causal expectations (based on previous stimuli) with sensory correspondence cues (from current stimuli) to flexibly control whether signals are integrated or segregated (Gau & Noppeney 2016). The brain thus controls information flow in multisensory inference—as in classical Stroop paradigms (Kerns et al. 2004)—via control mechanisms that dynamically adapt to the changing causal structure of the environment (Van Wanrooij et al. 2010). At a longer timescale, learning prior cross-sensory distributions enables observers to integrate previously unrelated features such as visuospatial and novel echolocation cues (Negen et al. 2018), luminance and stiffness (Ernst 2007), or visual and vestibular self-motion cues for yaw and roll axes (Kaliuzhna et al. 2015). Yet, the brain's ability to integrate novel cues in a statistically optimal manner may be limited to closely related features such as rotational self-motion cues from yaw and roll axes (Kaliuzhna et al. 2015). The integration of visuospatial and echolocation cues has already fallen short of the normative predictions (Negen et al. 2018), and the integration of genuinely novel pairs of cues (e.g., pitch, color) may be even more challenging or impossible.

Observers' dynamic beliefs about the world's causal structure should influence not only perceptual inference but also learning. Observers should learn and update (e.g., spatial) priors jointly across sensory modalities when common sources are likely but separately otherwise. Surprisingly, when visual and tactile signals occurred solely as unisensory events and with different probabilities over left and right hemifields (Mengotti et al. 2018), observers' reaction times and event-related potentials for these unisensory events depended jointly on the spatial probabilities of both sensory

modalities, with a stronger influence given to the sensory modality of the stimulus (Eimer et al. 2004, Spence & Driver 1996). These crossmodal influences arose even when cues informed observers about the sensory modality of the upcoming stimulus with 100% validity prior to stimulus presentation (Mengotti et al. 2018). Thus, contrary to normative principles, observers were not able to learn independent spatial priors for each sensory modality and allocate their attentional resources accordingly (for a review of crossmodal attention, see Spence 2014). Neuroimaging research has only started to disentangle this complex interplay between (spatial) expectation and attention in the human brain (for a review, see Summerfield & Egner 2009). For instance, a recent fMRI study manipulated spatial attention and expectations selectively in audition and assessed their effects on neural responses to auditory and visual stimuli. The study showed that attentional resources were controlled interactively across the senses via frontoparietal cortices, while spatial expectations were encoded in auditory and parietal cortices independently for each sense (Zuanazzi & Noppeney 2019).

So far, we have assumed that causal inference needs to dissociate solely whether sensory estimates differ because of noise or because signals come from independent sources. Yet, sensory estimates can also disagree because of modality-specific biases. Notably, physical growth, ageing, or entering a room with reverberant acoustics can introduce biases by profoundly altering the sensory cues that guide the brain's construction of spatial representations. To maintain internal consistency between the senses (unsupervised calibration) and external accuracy with respect to the outside world (supervised calibration), the brain constantly needs to recalibrate the senses. Experimentally, cross-sensory recalibration can be invoked by introducing a sensory conflict along one dimension (e.g., spatial) while enforcing audiovisual binding via correspondence cues along another dimension (e.g., time). For instance, exposure to synchronous, yet spatially misaligned, audiovisual signals induces a bias in observers' perceived sound location toward the previously presented visual stimulus even when presented alone, a phenomenon termed the ventriloquist aftereffect (Bertelson et al. 2006, Woods & Recanzone 2004). Thus, multisensory recalibration leads to attractive biases in contrast to the repulsive effects typically observed in unisensory adaptation (e.g., for contrast, see Bao & Engel 2012; for tilt, see Schwartz et al. 2007).

Despite extensive experimental evidence for recalibration, the computational principles of unsupervised recalibration remain controversial and may also differ across tasks and contexts. Most notably, unlike the repulsive effects in unisensory adaptation (Stocker & Simoncelli 2005), the attractive biases in cross-sensory recalibration can in principle result from changes in priors and/or likelihoods. For instance, a recent study suggested that exposure to McGurk stimuli (e.g., an auditory "pa" paired with a visual "ga") changes observers' priors over sensory features of a particular phoneme category (Olasagasti & Giraud 2020). By contrast, cross-sensory recalibration of space or heading motion has been attributed to changes in observers' likelihoods (Wozny & Shams 2011a). Here, the idea is that shifts in the means of the likelihood functions (i.e., as incorporated by bias terms) enable the brain to keep sensory representations in coregistration. Yet, how exactly the brain updates sensory likelihoods is a matter of debate. Some theoretical accounts intimately link the computations of recalibration with multisensory perceptual inference and suggest that likelihoods are updated by the difference between sensory inputs and observers' forced fusion (Burge et al. 2010, Gharamani et al. 1997) or Bayesian causal inference estimate (Sato et al. 2007). Sensory modalities would then recalibrate according to their reliabilities, adapting more when they are less reliable and, when Bayesian causal inference is put into the equation, when conflicting signals are perceived as coming from one source. Other accounts argue that perceptual inference and recalibration pursue distinct goals—the former reducing sensory uncertainty, the latter increasing perceptual accuracy by attenuating modality-specific biases. Sensory modalities should therefore recalibrate irrespective of their sensory reliabilities with fixed weights that reflect

observers' beliefs about modality-specific biases (Ernst & di Luca 2011; Zaidel et al. 2011, 2013). Further, because the updates are computed based on the difference between individual sensory estimates, recalibration should be greater when conflicting signals are perceived as coming from different sources. Results to date are too limited and inconsistent to arbitrate between these accounts. While recalibration depended on sensory reliabilities (Burge et al. 2010) and/or observers' inferred causal structures (Wozny & Shams 2011a,b) in some studies, it was largely independent from them in others (Di Luca et al. 2009; Zaidel et al. 2011, 2013).

Possibly, cross-sensory recalibration may also rely on different computational principles, neural mechanisms, or even circuitries depending on stimulus statistics (e.g., spatial versus phonetic), adaptation duration, and task context (for unisensory adaptation, see Bao & Engel 2012, Schwartz et al. 2007). In support of multiple mechanisms, recalibration arises across several timescales from milliseconds (Bosen et al. 2017, 2018; Wozny & Shams 2011b) to minutes (Bertelson et al. 2006, Woods & Recanzone 2004) and even days (Zwiers et al. 2003), with the length of spatial recalibration influencing its frequency selectivity (Bruns & Röder 2015, Woods & Recanzone 2004) and the reference frames of the underlying spatial representations (Kopčo et al. 2009). Intriguingly, a recent elegant psychophysics study has shown that the effects of long-term audiovisual spatial recalibration are cancelled transiently by short-term recalibration into the opposite direction, yet they then reappear (Watson et al. 2019). This recalibration rebound has previously been observed in unisensory adaptation (Bao & Engel 2012) and sensorimotor learning (Smith et al. 2006). It suggests that spatial recalibration can evolve independently at two distinct timescales so that long-term recalibration effects reappear when short-term recalibration effects taper off more rapidly. Recalibration across multiple timescales enables the brain to adapt flexibly to brief (e.g., reverberant acoustics) and more prolonged (e.g., physical growth) changes in sensory statistics (Bosen et al. 2017, 2018; Watson et al. 2019).

An open question is how observers infer whether intersensory discrepancies result from noise, signals coming from independent sources, or a variety of perturbations that arise at multiple timescales. While an optimal Bayesian learner should solve this credit assignment problem by constantly updating their estimates about the signals' causes, properties, and various perturbations as well as their uncertainties about those estimates over time [e.g., Kalman filter (Kording et al. 2007b)], it is likely that observers need to compute approximate solutions. Consistent with this conjecture, a double exponential model can capture human (spatial) recalibration at two distinct timescales (Bosen et al. 2018, Watson et al. 2019). At the neural level, audiovisual spatial recalibration has been shown to affect neural processing throughout the dorsal auditory processing stream from primary auditory to dorsolateral prefrontal cortices (Park & Kayser 2019), with early auditory and parietal activity involved in adaptive coding of continuous auditory space and later frontoparietal activity reflecting observers' decisional uncertainty involved in mapping those recalibrated spatial estimates onto decisional choices (Aller et al. 2021, Zierul et al. 2017). It is unknown whether recalibration across different timescales relies on common, partially overlapping, or different neural systems.

Recalibration can also arise to resolve temporal or high-order statistical (e.g., phonetic) cross-sensory conflicts. Temporal recalibration is particularly important to compensate for stimulus-dependent variability in cross-sensory timings (Fujisaki et al. 2004). Most prominently, audiovisual timings may vary because light and sound differ in their traveling speed, sensory transduction, or transmission processes. Similar to spatial recalibration, temporal recalibration can arise concurrently at fast and long timescales (Van der Burg et al. 2015). Observers can even simultaneously recalibrate audiovisual timing in opposite directions for different stimuli, which may enable the brain to account for differences in audiovisual timing for multiple simultaneous stimuli that differ in luminance or distance from the observer (Roseboom & Arnold 2011). Initial EEG and

drift-diffusion modelling have shown that evidence accumulation in synchrony judgements changes in accordance with observers' recalibrated timing (Simon et al. 2018). Likewise, audiovisual phonetic conflicts recalibrate observers' auditory phoneme percept (Bertelson et al. 2003, Olasagasti & Giraud 2020) and corresponding neural representations in temporal cortices (Kilian-Hütten et al. 2011). In comparison to the sustained effects of long-term spatial recalibration, the effects of phonetic recalibration taper off more quickly, possibly because phonetic recalibration serves different computational goals and deals with faster temporal statistics. For instance, phonetic recalibration is thought to enable observers to adapt rapidly to various speakers.

Collectively, this body of research documents the brain's ability to adapt effectively to changes in the environment and its sensorium across multiple timescales by adjusting priors, hyperpriors, and likelihoods. However, we are still lacking data and computational models to help us understand how the brain computes approximate solutions to the causal inference and credit assignment problem in complex environments in which intersensory discrepancies arise dynamically because of sensory biases or signals coming from different sources.

SUMMARY

We have used multisensory processing as a microcosm to review how the brain tackles some of the most fundamental challenges for neural processing, namely, inference- and decision-making, binding, attention, and learning. Over the past two decades, mounting evidence has shown that the brain integrates signals near-optimally weighted according to their momentary uncertainties for perceptual inference and decisions. In situations that entail causal uncertainty, the brain arbitrates between sensory integration and segregation approximately consistent with the principles of Bayesian causal inference. At the neural level, the brain accomplishes this feat by dynamically encoding multiple perceptual estimates that segregate, integrate, and flexibly combine information depending on the world's causal structure and observers' perceptual goals along the cortical hierarchy. Crucially, solving the causal inference problem exactly is intractable for the brain, with its limited resources, in all but the simplest laboratory scenarios. We have discussed how the brain computes approximate solutions in progressively complex multisource environments and argued that attentional mechanisms may be recruited in the service of these approximations. Finally, we have described how the brain adapts dynamically to changes in the sensory statistics arising from changes in the environment and sensorium across multiple timescales.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I am grateful to my collaborators in the various studies mentioned, in particular Tim Rohe, Ulrik Beierholm, Mate Aller, Agoston Mihalik, David Meijer, Ambra Ferrari, Arianna Zuanazzi, Samuel Jones, Remi Gau, Patrycja Delong, and Johanna Zumer.

LITERATURE CITED

Acerbi L, Dokka K, Angelaki DE, Ma WJ. 2018. Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLOS Comput. Biol.* 14:e1006110

Alais D, Burr D. 2004. Ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14:257–62

Aller M, Mihalik A, Noppeney U. 2021. Audiovisual adaptation is expressed in spatial and decisional codes. bioRxiv 2021.02.15.431309. **https://doi.org/10.1101/2021.02.15.431309**

Aller M, Noppeney U. 2019. To integrate or not to integrate: temporal dynamics of hierarchical Bayesian causal inference. *PLOS Biol.* 17:e3000210

Alsius A, Möttönen R, Sams ME, Soto-Faraco S, Tiippana K. 2014. Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Front. Psychol.* 5:727

Alsius A, Navarra J, Campbell R, Soto-Faraco S. 2005. Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* 15:839–43

Alsius A, Soto-Faraco S. 2011. Searching for audiovisual correspondence in multiple speaker scenarios. *Exp. Brain Res.* 213:175–83

Atilgan H, Town SM, Wood KC, Jones GP, Maddox RK, et al. 2018. Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. *Neuron* 97:640–55.e4

Badde S, Navarro KT, Landy MS. 2020. Modality-specific attention attenuates visual-tactile integration and recalibration effects by reducing prior expectations of a common source for vision and touch. *Cognition* 197:104170

Bankieris KR, Bejjanki VR, Aslin RN. 2017. Sensory cue-combination in the context of newly learned categories. *Sci. Rep.* 7:10890

Bao M, Engel SA. 2012. Distinct mechanism for long-term contrast adaptation. *PNAS* 109:5898–903

Battaglia PW, Jacobs RA, Aslin RN. 2003. Bayesian integration of visual and auditory signals for spatial localization. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* 20:1391–97

Beck JM, Ma WJ, Kiani R, Hanks T, Churchland AK, et al. 2008. Probabilistic population codes for Bayesian decision making. *Neuron* 60:1142–52

Beierholm U, Rohe T, Ferrari A, Stegle O, Noppeney U. 2020. Using the past to estimate sensory uncertainty. *eLife* 9:e54172

Bejjanki VR, Clayards M, Knill DC, Aslin RN. 2011. Cue integration in categorical tasks: insights from audio-visual speech perception. *PLOS ONE* 6:e19812

Bertelson P, Frissen I, Vroomen J, de Gelder B. 2006. The aftereffects of ventriloquism: patterns of spatial generalization. *Percept. Psychophys.* 68:428–36

Bertelson P, Radeau M. 1981. Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Percept. Psychophys.* 29:578–84

Bertelson P, Vroomen J, de Gelder B. 2003. Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychol. Sci.* 14:592–97

Bertelson P, Vroomen J, de Gelder B, Driver J. 2000. The ventriloquist effect does not depend on the direction of deliberate visual attention. *Percept. Psychophys.* 62:321–32

Bertelson P, Vroomen J, Wiegeraad G, de Gelder B. 1994. Exploring the relation between McGurk interference and ventriloquism. In *Proceedings of the Third International Congress on Spoken Language Processing (ICSLP 94), Yokohama, Japan, September 18–22, 1994*, pp. 559–62. Baixas, Fr.: Int. Speech Commun. Assoc.

Bishop CW, Miller LM. 2011. Speech cues contribute to audiovisual spatial integration. *PLOS ONE* 6:e24016

Bizley JK, Nodal FR, Bajo VM, Nelken I, King AJ. 2007. Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cereb. Cortex* 17:2172–89

Bizley JK, Shinn-Cunningham BG, Lee AK. 2012. Nothing is irrelevant in a noisy world: Sensory illusions reveal obligatory within- and across-modality integration. *J. Neurosci.* 32:13402–10

Bonath B, Noesselt T, Martinez A, Mishra J, Schwiecker K, et al. 2007. Neural basis of the ventriloquist illusion. *Curr. Biol.* 17:1697–703

Bosen AK, Fleming JT, Allen PD, O'Neill WE, Paige GD. 2017. Accumulation and decay of visual capture and the ventriloquism aftereffect caused by brief audio-visual disparities. *Exp. Brain Res.* 235:585–95

Bosen AK, Fleming JT, Allen PD, O'Neill WE, Paige GD. 2018. Multiple time scales of the ventriloquism aftereffect. *PLOS ONE* 13:e0200930

Bruns P, Röder B. 2015. Sensory recalibration integrates information from the immediate and the cumulative past. *Sci. Rep.* 5:12739

Burge J, Girshick AR, Banks MS. 2010. Visual-haptic adaptation is determined by relative reliability. *J. Neurosci.* 30:7714–21

Burr D, Banks MS, Morrone MC. 2009. Auditory dominance over vision in the perception of interval duration. *Exp. Brain Res.* 198:49–57

Busse L, Roberts KC, Crist RE, Weissman DH, Woldorff MG. 2005. The spread of attention across modalities and space in a multisensory object. PNAS 102:18751–56

Butler JS, Smith ST, Campos JL, Bülthoff HH. 2010. Bayesian integration of visual and vestibular signals for heading. *J. Vis.* 10:23

Calvert GA, Campbell R, Brammer MJ. 2000. Evidence from functional magnetic resonance imaging of cross-modal binding in the human heteromodal cortex. *Curr. Biol.* 10(11):649–57

Cao Y, Summerfield C, Park H, Giordano BL, Kayser C. 2019. Causal inference in the multisensory brain. *Neuron* 102:1076–87.e8

Carandini M, Heeger DJ. 2011. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.* 13:51–62

Chen L, Vroomen J. 2013. Intersensory binding across space and time: a tutorial review. *Atten. Percept. Psychophys.* 75:790–811

Chen Y-C, Spence C. 2017. Assessing the role of the 'unity assumption' on multisensory integration: a review. *Front. Psychol.* 8:445

Crosse MJ, Butler JS, Lalor EC. 2015. Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J. Neurosci.* 35:14195–204

Dayan P, Kakade S, Montague PR. 2000. Learning and selective attention. *Nat. Neurosci.* 3:1218–23

Dayan P, Solomon JA. 2010. Selective Bayes: attentional load and crowding. *Vis. Res.* 50:2248–60

de Winkel KN, Katliar M, Diers D, Bulthoff HH. 2018. Causal inference in the perception of verticality. *Sci. Rep.* 8(1):5483

Delong P, Noppeney U. 2021. Semantic and spatial congruency mould audiovisual integration depending on perceptual awareness. *Sci. Rep*. In press

Di Luca M, Machulla TK, Ernst MO. 2009. Recalibration of multisensory simultaneity: Cross-modal transfer coincides with a change in perceptual latency. *J. Vis.* 9:7

Donohue SE, Roberts KC, Grent-'t-Jong T, Woldorff MG. 2011. The cross-modal spread of attention reveals differential constraints for the temporal and spatial linking of visual and auditory stimulus events. *J. Neurosci.* 31:7982–90

Drugowitsch J, DeAngelis GC, Angelaki DE, Pouget A. 2015. Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making. *eLife* 4:e06678

Drugowitsch J, DeAngelis GC, Klier EM, Angelaki DE, Pouget A. 2014. Optimal multisensory decision-making in a reaction-time task. *eLife* 3:e03005

Eimer M, van Velzen J, Driver J. 2004. ERP evidence for cross-modal audiovisual effects of endogenous spatial attention within hemifields. *J. Cogn. Neurosci.* 16:272–88

Ernst MO. 2007. Learning to integrate arbitrary signals from vision and touch. *J. Vis.* 7:7

Ernst MO, Banks MS. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415:429–33

Ernst MO, Bülthoff HH. 2004. Merging the senses into a robust percept. *Trends Cogn. Sci.* 8:162–69

Ernst MO, di Luca M. 2011. Multisensory perception: from integration to remapping. In *Sensory Cue Integration*, ed. J Trommershäuser, KP Kording, MS Landy, pp. 225–50. Oxford, UK: Oxford Univ. Press

Fairhall SL, Macaluso E. 2009. Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. *Eur. J. Neurosci.* 29:1247–57

Faisal AA, Selen LP, Wolpert DM. 2008. Noise in the nervous system. *Nat. Rev. Neurosci.* 9:292–303

Fetsch CR, DeAngelis GC, Angelaki DE. 2013. Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nat. Rev. Neurosci.* 14:429–42

Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE. 2012. Neural correlates of reliability-based cue weighting during multisensory integration. *Nat. Neurosci.* 15:146–54

Fetsch CR, Turner AH, DeAngelis GC, Angelaki DE. 2009. Dynamic reweighting of visual and vestibular cues during self-motion perception. *J. Neurosci.* 29:15601–12

Fiebelkorn IC, Foxe JJ, Molholm S. 2010. Dual mechanisms for the cross-sensory spread of attention: How much do learned associations matter? *Cereb. Cortex* 20:109–20

Forstmann BU, Ratcliff R, Wagenmakers EJ. 2016. Sequential sampling models in cognitive neuroscience: advantages, applications, and extensions. *Annu. Rev. Psychol.* 67:641–66

Fujisaki W, Koene A, Arnold D, Johnston A, Nishida S. 2006. Visual search for a target changing in synchrony with an auditory signal. *Proc. Biol. Sci.* 273:865–74

Fujisaki W, Shimojo S, Kashino M, Nishida S. 2004. Recalibration of audiovisual simultaneity. *Nat. Neurosci.* 7:773–78
Gau R, Bazin PL, Trampel R, Turner R, Noppeney U. 2020. Resolving multisensory and attentional influences across cortical depth in sensory cortices. *eLife* 9:e46856
Gau R, Noppeney U. 2016. How prior expectations shape multisensory perception. *NeuroImage* 124:876–86
Gepshtein S, Banks MS. 2003. Viewing geometry determines how vision and haptics combine in size perception. *Curr. Biol.* 13:483–88
Gharamani Z, Wolpert DM, Jordan MI. 1997. Computational models of sensorimotor integration. In *Advances in Psychology*, Vol. 119: *Self-Organization, Computational Maps, and Motor Control*, ed. P. Morasso, V Sanguineti, pp. 117–47. Amsterdam: Elsevier
Ghazanfar A, Schroeder C. 2006. Is neocortex essentially multisensory? *Trends Cogn. Sci.* 10:278–85
Gold JI, Shadlen MN. 2007. The neural basis of decision making. *Annu. Rev. Neurosci.* 30:535–74
Gu Y, Angelaki DE, DeAngelis GC. 2008. Neural correlates of multisensory cue integration in macaque MSTd. *Nat. Neurosci.* 11:1201–10
Gu Y, DeAngelis GC, Angelaki DE. 2012. Causal links between dorsal medial superior temporal area neurons and multisensory heading perception. *J. Neurosci.* 32:2299–313
Hein G, Doehrmann O, Müller NG, Kaiser J, Muckli L, Naumer MJ. 2007. Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *J. Neurosci.* 27:7881–87
Helbig HB, Ernst MO, Ricciardi E, Pietrini P, Thielscher A, et al. 2012. The neural mechanisms of reliability weighted integration of shape information from vision and touch. *NeuroImage* 60:1063–72
Hillis JM, Ernst MO, Banks MS, Landy MS. 2002. Combining sensory information: mandatory fusion within, but not between, senses. *Science* 298:1627–30
Hospedales T, Vijayakumar S. 2009. Multisensory oddity detection as Bayesian inference. *PLOS ONE* 4:e4205
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* 104:1010–21.e10
Itti L, Koch C. 2001. Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2:194–203
Jones SA, Beierholm U, Meijer D, Noppeney U. 2019. Older adults sacrifice response speed to preserve multisensory integration performance. *Neurobiol. Aging* 84:148–57
Kaliuzhna M, Prsa M, Gale S, Lee SJ, Blanke O. 2015. Learning to integrate contradictory multisensory self-motion cue pairings. *J. Vis.* 15:10
Kanaya S, Yokosawa K. 2011. Perceptual congruency of audio-visual speech affects ventriloquism with bilateral visual stimuli. *Psychon. Bull. Rev.* 18:123–28
Kayser C, Petkov CI, Augath M, Logothetis NK. 2007. Functional imaging reveals visual modulation of specific fields in auditory cortex. *J. Neurosci.* 27:1824–35
Kayser C, Petkov CI, Lippert M, Logothetis NK. 2005. Mechanisms for allocating auditory attention: an auditory saliency map. *Curr. Biol.* 15:1943–47
Kayser C, Petkov CI, Logothetis NK. 2008. Visual modulation of neurons in auditory cortex. *Cereb. Cortex* 18:1560–74
Kerns JG, Cohen JD, MacDonald AW 3rd, Cho RY, Stenger VA, Carter CS. 2004. Anterior cingulate conflict monitoring and adjustments in control. *Science* 303:1023–26
Kersten D, Mamassian P, Yuille A. 2004. Object perception as Bayesian inference. *Annu. Rev. Psychol.* 55:271–304
Kilian-Hütten N, Valente G, Vroomen J, Formisano E. 2011. Auditory cortex encodes the perceptual interpretation of ambiguous sound. *J. Neurosci.* 31:1715–20
Knill DC, Saunders JA. 2003. Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vis. Res.* 43:2539–58
Kopčo N, Groh JM, Lin IF, Shinn-Cunningham BG. 2009. Reference frame of the ventriloquism aftereffect. *J. Neurosci.* 29:13809–14
Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L. 2007a. Causal inference in multisensory perception. *PLOS ONE* 2:e943
Kording KP, Tenenbaum JB, Shadmehr R. 2007b. The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nat. Neurosci.* 10:779–86

Kumpik DP, Roberts HE, King AJ, Bizley JK. 2014. Visual sensitivity is a stronger determinant of illusory processes than auditory cue parameters in the sound-induced flash illusion. *J. Vis.* 14:12

Lakatos P, Chen C-M, O'Connell MN, Mills A, Schroeder CE. 2007. Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron* 53:279–92

Lee H, Noppeney U. 2011a. Long-term music training tunes how the brain temporally binds signals from multiple senses. *PNAS* 108:E1441–50

Lee H, Noppeney U. 2011b. Physical and perceptual factors shape the neural mechanisms that integrate audiovisual signals in speech comprehension. *J. Neurosci.* 31:11338–50

Lee H, Noppeney U. 2014. Temporal prediction errors in visual and auditory cortices. *Curr. Biol.* 24:R309–10

Lewald J, Guski R. 2003. Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Brain Res. Cogn. Brain Res.* 16:468–78

Lewis R, Noppeney U. 2010. Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. *J. Neurosci.* 30:12329–39

Li Z. 2002. A saliency map in primary visual cortex. *Trends Cogn. Sci.* 6:9–16

Locke SM, Landy MS. 2017. Temporal causal inference with stochastic audiovisual sequences. *PLOS ONE* 12:e0183776

Ma WJ. 2012. Organizing probabilistic models of perception. *Trends Cogn. Sci.* 16:511–18

Ma WJ, Beck JM, Latham PE, Pouget A. 2006. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* 9:1432–38

Ma WJ, Jazayeri M. 2014. Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* 37:205–20

Magnotti JF, Beauchamp MS. 2017. A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLOS Comput. Biol.* 13:e1005229

Magnotti JF, Ma WJ, Beauchamp MS. 2013. Causal inference of asynchronous audiovisual speech. *Front. Psychol.* 4:798

Maier JX, Groh JM. 2009. Multisensory guidance of orienting behavior. *Hear. Res.* 258:106–12

Martuzzi R, Murray MM, Michel CM, Thiran JP, Maeder PP, et al. 2007. Multisensory interactions within human primary cortices revealed by BOLD dynamics. *Cereb. Cortex* 17:1672–79

Maunsell JHR. 2015. Neuronal mechanisms of visual attention. *Annu. Rev. Vis. Sci.* 1:373–91

McGurk H, MacDonald J. 1976. Hearing lips and seeing voices. *Nature* 264:691–811

Meijer D, Noppeney U. 2020. Computational models of multisensory integration. In *Multisensory Perception: From Laboratory to Clinic*, ed. K Sathian, VS Ramachandran, pp. 113–33. London: Academic Press

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

Mengotti P, Boers F, Dombert PL, Fink GR, Vossel S. 2018. Integrating modality-specific expectancies for the deployment of spatial attention. *Sci. Rep.* 8:1210

Mercier MR, Foxe JJ, Fiebelkorn IC, Butler JS, Schwartz TH, Molholm S. 2013. Auditory-driven phase reset in visual cortex: Human electrocorticography reveals mechanisms of early multisensory integration. *Neuroimage* 79:19–29

Meredith MA, Nemitz JW, Stein BE. 1987. Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *J. Neurosci.* 7:3215–29

Meredith MA, Stein B. 1983. Interactions among converging sensory inputs in the superior colliculus. *Science* 221:389–91

Metzger BA, Magnotti JF, Wang Z, Nesbitt E, Karas PJ, et al. 2020. Responses to visual speech in human posterior superior temporal gyrus examined with iEEG deconvolution. *J. Neurosci.* 40:6938–48

Mihalik A, Noppeney U. 2020. Causal inference in audiovisual perception. *J. Neurosci.* 40:6600–12

Miller J. 1982. Divided attention: evidence for coactivation with redundant signals. *Cogn. Psychol.* 14:247–79

Miller LM, D'Esposito M. 2005. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *J. Neurosci.* 25:5884–93

Mohl JT, Pearson JM, Groh JM. 2020. Monkeys and humans implement causal inference to simultaneously localize auditory and visual stimuli. *J. Neurophysiol.* 124:715–27

Nahorna O, Berthommier F, Schwartz J-L. 2012. Binding and unbinding the auditory and visual streams in the McGurk effect. *J. Acoust. Soc. Am.* 132:1061–77

Nahorna O, Berthommier F, Schwartz J-L. 2015. Audio-visual speech scene analysis: characterization of the dynamics of unbinding and rebinding the McGurk effect. *J. Acoust. Soc. Am.* 137:362–77
Nath AR, Beauchamp MS. 2011. Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *J. Neurosci.* 31:1704–14
Negen J, Wen L, Thaler L, Nardini M. 2018. Bayes-like integration of a new sensory skill with vision. *Sci. Rep.* 8:16880
Nikbakht N, Tafreshiha A, Zoccolan D, Diamond ME. 2018. Supralinear and supramodal integration of visual and tactile signals in rats: psychophysics and neuronal mechanisms. *Neuron* 97:626–39.e8
Noesselt T, Rieger JW, Schoenfeld MA, Kanowski M, Hinrichs H, et al. 2007. Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *J. Neurosci.* 27:11431–41
Noppeney U. 2020. Multisensory perception: behaviour, computations, neural mechanisms. In *The Cognitive Neurosciences*, ed. D Poeppel, GR Mangun, M Gazzaniga, pp. 141–51. Cambridge, MA: MIT Press
Noppeney U, Josephs O, Hocking J, Price CJ, Friston KJ. 2008. The effect of prior visual information on recognition of speech and sounds. *Cereb. Cortex* 18:598–609
Noppeney U, Ostwald D, Werner S. 2010. Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *J. Neurosci.* 30:7434–46
Ohshiro T, Angelaki DE, DeAngelis GC. 2011. A normalization model of multisensory integration. *Nat. Neurosci.* 14:775–82
Ohshiro T, Angelaki DE, DeAngelis GC. 2017. A neural signature of divisive normalization at the level of multisensory integration in primate cortex. *Neuron* 95:399–411.e8
Olasagasti I, Giraud AL. 2020. Integrating prediction errors at two time scales permits rapid recalibration of speech sound categories. *eLife* 9:e44516
Ortiz-Rios M, Azevedo FAC, Kuśmierek P, Balla DZ, Munk MH, et al. 2017. Widespread and opponent fMRI signals represent sound location in macaque auditory cortex. *Neuron* 93:971–83.e4
O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, et al. 2015. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25:1697–706
Otto TU, Mamassian P. 2012. Noise and correlations in parallel perceptual decision making. *Curr. Biol.* 22:1391–96
Parise CV, Ernst MO. 2016. Correlation detection as a general mechanism for multisensory integration. *Nat. Commun.* 7:11543
Parise CV, Spence C, Ernst MO. 2012. When correlation implies causation in multisensory integration. *Curr. Biol.* 22:46–49
Park H, Ince RAA, Schyns PG, Thut G, Gross J. 2018. Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLOS Biol.* 16:e2006558
Park H, Kayser C. 2019. Shared neural underpinnings of multisensory integration and trial-by-trial perceptual recalibration in humans. *eLife* 8:e47001
Pouget A, Beck JM, Ma WJ, Latham PE. 2013. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* 16:1170–78
Powers AR 3rd, Hevey MA, Wallace MT. 2012. Neural correlates of multisensory perceptual learning. *J. Neurosci.* 32:6263–74
Rahnev D, Denison RN. 2018. Suboptimality in perceptual decision making. *Behav. Brain Sci.* 41:e223
Raposo D, Sheppard JP, Schrater PR, Churchland AK. 2012. Multisensory decision-making in rats and humans. *J. Neurosci.* 32:3726–35
Rauschecker JP, Tian B. 2000. Mechanisms and streams for processing of "what" and "where" in auditory cortex. *PNAS* 97:11800–6
Rohe T, Ehlis A-C, Noppeney U. 2019. The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. *Nat. Commun.* 10:1907
Rohe T, Noppeney U. 2015a. Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLOS Biol.* 13:e1002073
Rohe T, Noppeney U. 2015b. Sensory reliability shapes perceptual inference via two mechanisms. *J. Vis.* 15:22

Rohe T, Noppeney U. 2016. Distinct computational principles govern multisensory integration in primary sensory and association cortices. *Curr. Biol.* 26:509–14
Rohe T, Noppeney U. 2018. Reliability-weighted integration of audiovisual signals can be modulated by top-down attention. *eNeuro* 5:ENEURO.0315-17.2018
Rosas P, Wagemans J, Ernst MO, Wichmann FA. 2005. Texture and haptic cues in slant discrimination: reliability-based cue weighting without statistically optimal cue combination. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* 22:801–9
Roseboom W, Arnold DH. 2011. Twice upon a time: multiple concurrent temporal recalibrations of audiovisual speech. *Psychol. Sci.* 22:872–77
Sato Y, Toyoizumi T, Aihara K. 2007. Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. *Neural Comput.* 19:3335–55
Schlack A, Sterbing-D'Angelo SJ, Hartung K, Hoffmann KP, Bremmer F. 2005. Multisensory space representations in the macaque ventral intraparietal area. *J. Neurosci.* 25:4616–25
Schroeder CE, Foxe JJ. 2002. The timing and laminar profile of converging inputs to multisensory areas of the macaque neocortex. *Brain Res. Cogn. Brain Res.* 14:187–98
Schwartz O, Hsu A, Dayan P. 2007. Space and time in visual context. *Nat. Rev. Neurosci.* 8:522–35
Shams L, Beierholm UR. 2010. Causal inference in perception. *Trends Cogn. Sci.* 14:425–32
Shams L, Ma WJ, Beierholm U. 2005. Sound-induced flash illusion as an optimal percept. *Neuroreport* 16:1923–27
Shen S, Ma WJ. 2016. A detailed comparison of optimality and simplicity in perceptual decision making. *Psychol. Rev.* 123:452–80
Simon DM, Nidiffer AR, Wallace MT. 2018. Single trial plasticity in evidence accumulation underlies rapid recalibration to asynchronous audiovisual speech. *Sci. Rep.* 8:12499
Smith MA, Ghazizadeh A, Shadmehr R. 2006. Interacting adaptive processes with different timescales underlie short-term motor learning. *PLOS Biol.* 4:e179
Spence C. 2014. Orienting attention: a crossmodal perspective. In *The Oxford Handbook of Attention*, ed. AC Nobre, S Kastner, pp. 446–73. Oxford, UK: Oxford Univ. Press
Spence C, Driver J. 1996. Audiovisual links in endogenous covert spatial attention. *J. Exp. Psychol. Hum. Percept. Perform.* 22:1005–30
Stacey PC, Murphy T, Sumner CJ, Kitterick PT, Roberts KL. 2014. Searching for a talking face: the effect of degrading the auditory signal. *J. Exp. Psychol. Hum. Percept. Perform.* 40:2106–11
Stanford TR, Quessy S, Stein BE. 2005. Evaluating the operations underlying multisensory integration in the cat superior colliculus. *J. Neurosci.* 25:6499–508
Stein BE, Stanford TR. 2008. Multisensory integration: current issues from the perspective of the single neuron. *Nat. Rev. Neurosci.* 9:255–66
Stocker A, Simoncelli E. 2005. Sensory adaptation within a Bayesian framework for perception. In *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, ed. Y Weiss, B Schölkopf, J Platt, pp. 1291–98. San Diego, CA: NeurIPS
Summerfield C, Egner T. 2009. Expectation (and attention) in visual cognition. *Trends Cogn. Sci.* 13(9):403–9
Talsma D, Doty TJ, Woldorff MG. 2006. Selective attention and audiovisual integration: Is attending to both modalities a prerequisite for early integration? *Cereb. Cortex* 17:679–90
Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG. 2010. The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* 14:400–10
Tiippana K, Puharinen H, Möttönen R, Sams M. 2011. Sound location can influence audiovisual speech perception when spatial attention is manipulated. *Seeing Perceiving* 24:67–90
Treisman AM, Gelade G. 1980. A feature-integration theory of attention. *Cogn. Psychol.* 12:97–136
Van der Burg E, Alais D, Cass J. 2015. Audiovisual temporal recalibration occurs independently at two different time scales. *Sci. Rep.* 5:14526
Van der Burg E, Cass J, Olivers CNL, Theeuwes J, Alais D. 2010. Efficient visual search from synchronized auditory signals requires transient audiovisual events. *PLOS ONE* 5:e10664
Van der Burg E, Olivers CNL, Bronkhorst AW, Theeuwes J. 2008. Pip and pop: nonspatial auditory signals improve spatial visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 34:1053–65

Van der Burg E, Talsma D, Olivers CNL, Hickey C, Theeuwes J. 2011. Early multisensory interactions affect the competition among multiple visual objects. *NeuroImage* 55:1208–18
Van Wanrooij MM, Bremen P, Van Opstal AJ. 2010. Acquired prior knowledge modulates audiovisual integration. *Eur. J. Neurosci.* 31:1763–71
van Wassenhove V, Grant KW, Poeppel D. 2007. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45:598–607
von Helmholtz H. 1867. *Handbuch der Physiologischen Optik*. Leipzig: Leopold Voss
Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA. 2004. Unifying multisensory signals across time and space. *Exp. Brain Res.* 158:252–58
Watson DM, Akeroyd MA, Roach NW, Webb BS. 2019. Distinct mechanisms govern recalibration to audiovisual discrepancies in remote and recent history. *Sci. Rep.* 9:8513
Werner S, Noppeney U. 2010a. Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. *J. Neurosci.* 30:2662–75
Werner S, Noppeney U. 2010b. Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cereb. Cortex* 20:1829–42
Woods TM, Recanzone GH. 2004. Visually induced plasticity of auditory spatial perception in macaques. *Curr. Biol.* 14:1559–64
Wozny DR, Beierholm UR, Shams L. 2010. Probability matching as a computational strategy used in perception. *PLOS Comput. Biol.* 6:e1000871
Wozny DR, Shams L. 2011a. Computational characterization of visually induced auditory spatial adaptation. *Front. Integr. Neurosci.* 5:75
Wozny DR, Shams L. 2011b. Recalibration of auditory space following milliseconds of cross-modal discrepancy. *J. Neurosci.* 31:4607–12
Yu AJ, Dayan P, Cohen JD. 2009. Dynamics of attentional selection under conflict: toward a rational Bayesian account. *J. Exp. Psychol. Hum. Percept. Perform.* 35:700–17
Zaidel A, Ma WJ, Angelaki DE. 2013. Supervised calibration relies on the multisensory percept. *Neuron* 80:1544–57
Zaidel A, Turner AH, Angelaki DE. 2011. Multisensory calibration is independent of cue reliability. *J. Neurosci.* 31:13949–62
Zierul B, Röder B, Tempelmann C, Bruns P, Noesselt T. 2017. The role of auditory cortex in the spatial ventriloquism aftereffect. *Neuroimage* 162:257–68
Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, et al. 2013. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron* 77:980–91
Zuanazzi A, Noppeney U. 2019. Distinct neural mechanisms of spatial attention and expectation guide perceptual inference in a multisensory world. *J. Neurosci.* 39:2301–12
Zumer JM, White TP, Noppeney U. 2021. The neural mechanisms of audiotactile binding depend on asynchrony. *Eur. J. Neurosci.* 52:4709–31
Zwiers MP, Van Opstal AJ, Paige GD. 2003. Plasticity in human sound localization induced by compressed spatial vision. *Nat. Neurosci.* 6:175–81

Annual Review of Neuroscience

Volume 44, 2021

Contents

Sources of Individual Differences in Pain
Jeffrey S. Mogil 1

Current Status of and Perspectives on the Application of Marmosets in Neurobiology
Hideyuki Okano 27

Astrocytes and Behavior
Paulo Kofuji and Alfonso Araque 49

A Common Space Approach to Comparative Neuroscience
Rogier B. Mars, Saad Jbabdi, and Matthew F.S. Rushworth 69

Parkinson's Disease Genetics and Pathophysiology
Gabriel E. Vázquez-Vélez and Huda Y. Zoghbi 87

Neural and Molecular Mechanisms of Biological Embedding of Social Interactions
Ian M. Traniello and Gene E. Robinson 109

Smartphones and the Neuroscience of Mental Health
Claire M. Gillan and Robb B. Rutledge 129

Integrated Patterning Programs During *Drosophila* Development Generate the Diversity of Neurons and Control Their Mature Properties
Anthony M. Rossi, Shadi Jafari, and Claude Desplan 153

Consolidating the Circuit Model for Addiction
Christian Lüscher and Patricia H. Janak 173

Ensheathment and Myelination of Axons: Evolution of Glial Functions
Klaus-Armin Nave and Hauke B. Werner 197

Neocortical Layer 1: An Elegant Solution to Top-Down and Bottom-Up Integration
Benjamin Schuman, Shlomo Dellal, Alvar Prönneke, Robert Machold, and Bernardo Rudy 221

Human Representation Learning
Angela Radulescu, Yeon Soon Shin, and Yael Niv 253

Dense Circuit Reconstruction to Understand Neuronal Computation: Focus on Zebrafish
Rainer W. Friedrich and Adrian A. Wanner 275

The Role of the Medial Prefrontal Cortex in Moderating Neural Representations of Self and Other in Primates
Masaki Isoda 295

Inferring Macroscale Brain Dynamics via Fusion of Simultaneous EEG-fMRI
Marios G. Philiastides, Tao Tu, and Paul Sajda 315

Ion Channel Degeneracy, Variability, and Covariation in Neuron and Circuit Resilience
Jean-Marc Goaillard and Eve Marder 335

Oxytocin, Neural Plasticity, and Social Behavior
Robert C. Froemke and Larry J. Young 359

Physiology and Pathophysiology of Mechanically Activated PIEZO Channels
Ruhma Syeda 383

The Geometry of Information Coding in Correlated Neural Populations
Rava Azeredo da Silveira and Fred Rieke 403

The Cortical Motor Areas and the Emergence of Motor Skills: A Neuroanatomical Perspective
Peter L. Strick, Richard P. Dum, and Jean-Alban Rathelot 425

Perceptual Inference, Learning, and Attention in a Multisensory World
Uta Noppeney 449

Adaptive Prediction for Social Contexts: The Cerebellar Contribution to Typical and Atypical Social Behaviors
Catherine J. Stoodley and Peter T. Tsai 475

Neurophysiology of Human Perceptual Decision-Making
Redmond G. O'Connell and Simon P. Kelly 495

How Cortical Circuits Implement Cortical Computations: Mouse Visual Cortex as a Model
Cristopher M. Niell and Massimo Scanziani 517

Spatial Transcriptomics: Molecular Maps of the Mammalian Brain
Cantin Ortiz, Marie Carlén, and Konstantinos Meletis 547