J Neurophysiol 124: 715–727, 2020.
First published July 29, 2020; doi:10.1152/jn.00046.2020.

RESEARCH ARTICLE | *Higher Neural Functions and Behavior*

Monkeys and humans implement causal inference to simultaneously localize auditory and visual stimuli

Jeff T. Mohl,[1,2,3] John M. Pearson,[1,2,3,4,5] and Jennifer M. Groh[1,2,3,4]

[1]*Duke Institute for Brain Sciences, Duke University, Durham, North Carolina;* [2]*Center for Cognitive Neuroscience, Duke University, Durham, North Carolina;* [3]*Department of Neurobiology, Duke University, Durham, North Carolina;* [4]*Department of Psychology and Neuroscience, Duke University, Durham, North Carolina; and* [5]*Department of Biostatistics and Bioinformatics, Duke University Medical School, Durham, North Carolina*

Submitted 31 January 2020; accepted in final form 26 July 2020

Mohl JT, Pearson JM, Groh JM. Monkeys and humans implement causal inference to simultaneously localize auditory and visual stimuli. *J Neurophysiol* 124: 715–727, 2020. First published July 29, 2020; doi:10.1152/jn.00046.2020.—The environment is sampled by multiple senses, which are woven together to produce a unified perceptual state. However, optimally unifying such signals requires assigning particular signals to the same or different underlying objects or events. Many prior studies (especially in animals) have assumed fusion of cross-modal information, whereas recent work in humans has begun to probe the appropriateness of this assumption. Here we present results from a novel behavioral task in which both monkeys (*Macaca mulatta*) and humans localized visual and auditory stimuli and reported their perceived sources through saccadic eye movements. When the locations of visual and auditory stimuli were widely separated, subjects made two saccades, while when the two stimuli were presented at the same location they made only a single saccade. Intermediate levels of separation produced mixed response patterns: a single saccade to an intermediate position on some trials or separate saccades to both locations on others. The distribution of responses was well described by a hierarchical causal inference model that accurately predicted both the explicit "same vs. different" source judgments as well as biases in localization of the source(s) under each of these conditions. The results from this task are broadly consistent with prior work in humans across a wide variety of analogous tasks, extending the study of multisensory causal inference to nonhuman primates and to a natural behavioral task with both a categorical assay of the number of perceived sources and a continuous report of the perceived position of the stimuli.

NEW & NOTEWORTHY We developed a novel behavioral paradigm for the study of multisensory causal inference in both humans and monkeys and found that both species make causal judgments in the same Bayes-optimal fashion. To our knowledge, this is the first demonstration of behavioral causal inference in animals, and this cross-species comparison lays the groundwork for future experiments using neuronal recording techniques that are impractical or impossible in human subjects.

behavioral modeling; binding; causal inference; multisensory processing

INTRODUCTION

Perception is inherently multisensory. Often, information from one sensory modality can reduce uncertainty about another, such as reading the lips of a speaker to improve speech comprehension (Sumby and Pollack 1954). However, combining such visual and auditory cues is only beneficial if they originate from the same source in the environment. In the lip-reading example above, the observer must correctly pair the sight of the lip movements with the sound of the person speaking. This is an example of a causal inference (CI) problem: determining which source(s) are most likely to have caused specific sensory observations.

Recent behavioral studies in humans have modeled such multisensory perception hierarchically, involving an assessment of the relative likelihood of two causal scenarios (same source or different sources), which then influences how sensory information is interpreted to perform a given behavioral task (e.g., localization of a particular stimulus source) (Acerbi et al. 2018; Dokka et al. 2015, 2019; Körding et al. 2007; Mahani et al. 2017; Rohe and Noppeney 2015b; Sato et al. 2007; Shams and Beierholm 2010; de Winkel et al. 2017; Wozny et al. 2010). Often, these models are based on an idealized Bayesian observer that optimally accounts for uncertainty when making a causal inference judgment (Körding et al. 2007), though heuristic alternatives have been proposed such as fixed-criterion models, which implement an arbitrary distance threshold (Acerbi et al. 2018), or model-selection models, which choose the most likely possibility and discard the others (Wozny et al. 2010). Regardless of which specific form of model is implemented, causal inference models provide a very accurate description of human behavior across a wide variety of tasks.

In contrast to the richness of these behavioral models, previous multisensory neurophysiological research in animals has generally assumed that the animals fuse the visual and auditory sources (Angelaki et al. 2009; Battaglia et al. 2003; Stein and Stanford 2008; Wallace et al. 2004). This discrepancy of approach has limited the inferences that can be drawn connecting neural and behavioral observations, including at the theoretical level (Cuppini et al. 2017; Fetsch et al. 2013; Lochmann and Deneve 2011; Ma and Rahmati 2013). One challenge has been that, to date, the tasks that have been used

J. Mohl (jeffrey.mohl@duke.edu).

to probe causal inference in humans have involved behavioral reports that are arbitrarily associated to the stimulus at hand, such as using button presses, verbal reports, or cursor movements to indicate the location of a sound and/or visual stimulus (Körding et al. 2007; Rohe and Noppeney 2015b; Wallace et al. 2004; Wei and Körding 2009; Wozny et al. 2010). These scenarios pose challenges for animal training, where it is generally easier to shape naturally occurring responses than induce arbitrary stimulus-response associations de novo.

Accordingly, in this study we developed a task that can be deployed in both humans and animals, to provide a direct comparison of multisensory causal inference across species. We leveraged an innate behavior with an intuitive task (localizing the source of a stimulus) and response type (orienting to said source using saccadic eye movements). By requiring subjects to report both auditory and visual targets on each trial by making saccades to the perceived source of each stimulus, we could ascertain whether they perceived the stimuli to be fused versus segregated and where exactly they perceived them to be. This provided reports of both explicit (number of saccades) and implicit (location of saccades) causal inference on each trial.

We find that monkey and human subjects behave similarly and that their behavior reflects similar causal inference strategies. Subjects tended to make one saccade when the visual and auditory stimuli were located in the same spatial position and two saccades when they were widely separated (e.g., greater than 12° apart). The transition between these modes was well described by models that incorporate causal inference, consistent with previous reports concerning human performance in other similar tasks (Acerbi et al. 2018; Körding et al. 2007; Rohe and Noppeney 2015b; Wozny et al. 2010). These results suggest that this single/dual saccade task provides a reliable assay of multisensory causal inference that can be deployed in both humans and animal models and sets the stage for future work bridging between the neural and behavioral domains.

MATERIALS AND METHODS

Ethics Statement

All procedures involving human subjects were approved by the Duke University Institutional Review Board (IRB protocol number 1885), and all participants provided written informed consent. All animal procedures conformed to the guidelines of the National Institutes of Health (NIH Pub. No. 86-23, Revised 1985) and were approved by the Institutional Animal Care and Use Committee of Duke University (Protocol Registry Number A115-15-04).

General Procedures

Human subjects (n = 7, 4 female) were involved in this study. Subjects had apparently normal hearing and normal or corrected-to-normal vision. Informed consent was obtained from all participants before testing, and all subjects received monetary compensation for participation.

Two adult rhesus monkeys (*Macaca mulatta*) participated (*monkey J*, and *monkey Y*, both female). Under general anesthesia and in sterile surgery we implanted a head post holder to restrain the head and a scleral search coil to track eye movements (Judge et al. 1980). After recovery with suitable analgesics and veterinary care, we trained the monkeys in the experimental task.

Behavioral Paradigm

We created a novel multisensory task closely related to tasks commonly used in the literature (Körding et al. 2007; Rohe and Noppeney 2015b; Wallace et al. 2004; Wozny et al. 2010). This paradigm used a dual-report design, where subjects reported both a causal judgment (one or two targets, explicit causal inference) and the target locations (implicit causal inference) on every trial.

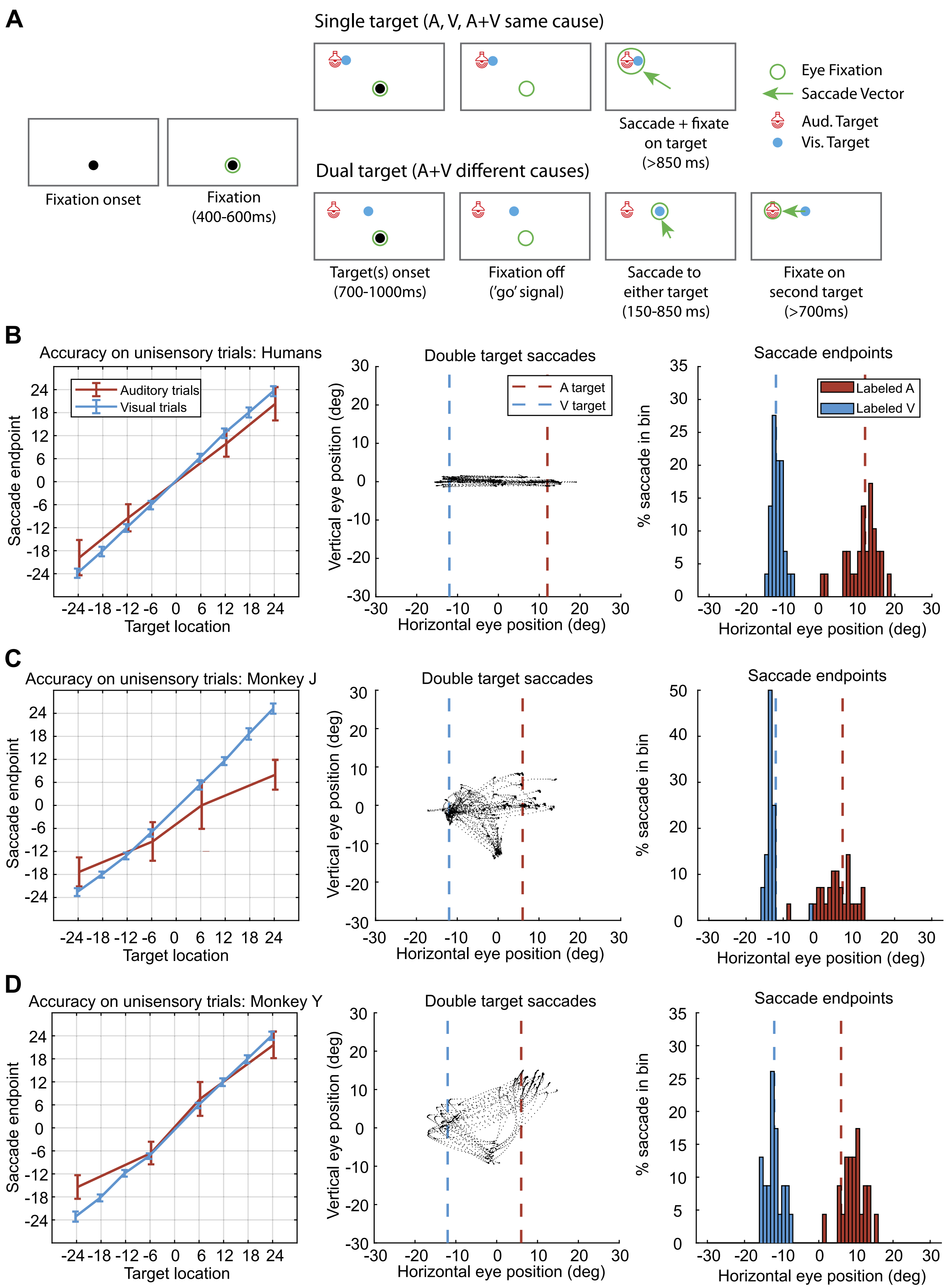
Subjects were seated in an anechoic chamber at a distance of 1.25 m from a row of speakers and LEDs located on the horizontal plane. Eye movements were monitored either via magnetic eye coil (monkeys; Riverbend) or video eye tracker (humans; SR Research Eyelink 1000). Eye tracking was calibrated against a set of visual targets spanning the response range at the start of each experimental session before beginning data collection.

While fixating at a central point (0° horizontal, variable vertical offset of −12 to +12° for monkey subjects varied from day to day), subjects were presented with either a light (green LED), sound (white noise), or both at one of 8 visual (± 6–24° in 6° increments) or 4 auditory (±6 and ±24° for monkeys; ±12 and ±24° for humans) locations. Targets were paired such that each combination of ipsilateral pairs was used (8 pairs per side, for 16 pairs), plus 4 contralateral pairs (±12° visual paired with either contralateral auditory location) for a total of 20 dual conditions. After a brief delay (600–900 ms) the fixation light was extinguished, and subjects reported percepts by making saccades to the perceived stimulus location and then maintaining fixation at that target location. On conditions with multiple targets, subjects were instructed (humans) or trained (monkeys) to make sequential saccades to each target in any order and then to hold fixation at the second target until the end of the trial. The timing of the task was such that subjects needed to make both saccades in rapid succession and so could not adopt a strategy of waiting until the trial ended (or not) before making a decision about the second saccade (see Fig. 1*A*).

For monkeys, this dual saccade behavior was incentivized by providing a juice reward on well-separated dual trials (>12° separation) only if both stimuli were correctly localized (within ±5° for visual targets, ±8° for auditory targets). These target windows were also used for coincident target trials, and monkeys received a reward only if they maintained fixation within the window until the end of a trial (i.e., did not make a second saccade out of the window). This was done to ensure that monkeys would correctly perform both the single-target, single-saccade trials as well as the dual-target, dual-saccade trials when there was little ambiguity about the number of targets. Ambiguous trials (6–12° separation) were rewarded randomly (50%) provided that the monkey made a saccade to at least one of these target windows. Importantly, this criterion was only imposed for the purposes of obtaining juice rewards, and a separate criterion that did not depend on reported location accuracy or number of saccades was used to exclude trials during preprocessing (described in the following subsection).

Trial Filtration and Saccade Detection

All trials were included provided that the subject held fixation through the go cue and then made at least one saccade, without enforcing any restrictions on saccade accuracy. All saccades that occurred between the go cue and the end of the trial were considered valid reports. Trials occasionally ended early due to loss of eye tracking signal midtrial, or if the subject failed to enter at least one of the target windows centered on each target. This could result in labeling trials as single-saccade (i.e., unity reports), when in reality the subject just did not have adequate time to make two saccades. We therefore excluded multistimulus trials that ended less than 600 ms after the go cue (shorter than the minimum duration for a successfully completed single-saccade trial). This was done to minimize the


A
Single target (A, V, A+V same cause)
Eye Fixation
Saccade Vector
Aud. Target
Vis. Target
Saccade + fixate on target (>850 ms)
Fixation onset
Fixation (400-600ms)
Dual target (A+V different causes)
Target(s) onset (700-1000ms)
Fixation off ('go' signal)
Saccade to either target (150-850 ms)
Fixate on second target (>700ms)
B
Accuracy on unisensory trials: Humans
Auditory trials
Visual trials
Saccade endpoint
Target location
Double target saccades
A target
V target
Vertical eye position (deg)
Horizontal eye position (deg)
Saccade endpoints
Labeled A
Labeled V
% saccade in bin
C
Accuracy on unisensory trials: Monkey J
D
Accuracy on unisensory trials: Monkey Y

number of trials that ended before the subject's full response had been recorded.

Saccades were defined as any eye movement exceeding 50° per second and followed by at least 30 ms of very little eye movement (maximum velocity <25°/s). Saccades of less than 3° were considered corrective and were not included as responses in subsequent analyses.

Occasionally (7% of trials), subjects made three saccades, which satisfied the above criteria on a given trial, either saccading back and forth between target locations or returning eye position back to the central location in anticipation of the upcoming trial. In these cases, the trial was considered a "two-cause" judgment for purposes of the unity judgment response, and the least accurate saccade was discarded for the localization response value. Subjects very rarely (0.8% of trials) made more than three saccades, and these trials were treated in the same way (discarding the least accurate saccades).

Behavioral Modeling

We implemented a class of causal inference models that is common in human behavioral multisensory research (see Shams and Beierholm 2010 for review). These models arbitrate between two sensory processing strategies. The first strategy treats sensory stimuli as completely independent, amounting to unisensory estimation of the parameter of interest (in this case, location of the source) for each stimulus. The second implements the established maximum-likelihood form of cue integration, which has been shown to provide excellent descriptions of human behavior in conditions where the disparity between multisensory cues is small or the cues are mandatorily fused (Alais and Burr 2004; Ernst and Banks 2002; Knill 2007). Different models of causal inference then combine these two estimates according to specific rules, resulting in predictions that can be compared with behavior in our task.

Below we briefly describe the important components of our models and refer interested readers to recent work from Acerbi and colleagues for a much more thorough treatment of this class of models (Acerbi et al. 2018). We begin by describing the cases for location estimation under given causal assumptions (one or two cases) and then describe how these estimates are combined according to different causal inference strategies to produce both judgments about number of targets (unity judgment task) and location of stimulus source(s) (localization task). It is important to note that, while each "task" can be fit independently by different models, the subjects themselves performed the tasks jointly and simultaneously.

Fused and segregated sensory localization. For all stimuli, we assume that stimuli (S_A, S_V, S_{AV}) give rise to internal representations (x_A, x_V) on which inferences are based. That is, the model takes the point of view of the observer, for whom the internal representations, x, are data and the external sources, S, must be inferred. We furthermore assume a generative model, in which the latter are noisy versions of the former: $p(x_A|S_A) = N(S_A, \sigma_A^2)$, $p(x_V|S_V) = N(S_V, \sigma_V^2)$, with S denoting the actual location of the source of the respective stimulus and σ reflecting the modality specific sensory standard deviation (a free parameter). From these internal representations, estimates about stimulus locations for a given causal structure ($c = 1$, common cause, *Eq. 1*; $c = 2$, independent causes, *Eq. 2*) can be computed via Bayes' rule:

$$p(S_{AV}|x_A, x_V, c = 1) = \frac{p(x_A|S_{AV})p(x_V|S_{AV})p(S_{AV}|c = 1)}{p(x_A, x_V)} \quad (1)$$

$$p(S_A, S_V|x_A, x_V, c = 2) = \frac{p(x_A|S_A)p(x_V|S_V)p(S_A, S_V|c = 2)}{p(x_A, x_V)} \quad (2)$$

where for the $c = 1$ case the source S is assumed to be the same for both the auditory and visual stimuli.

Location prior. The subject is assumed to have some prior over possible stimulus locations. A common choice in this type of model is to assume that the subjects have an independent, identical prior over both sensory stimuli,

$$p(S_{AV}|c = 1) = N(S_{AV}|\mu_p, \sigma_p^2) \quad (3)$$

while for the two causes,

$$p(S_A, S_V|c = 2) = N(S_A|\mu_p, \sigma_p^2)N(S_V|\mu_p, \sigma_p^2) \quad (4)$$

where μ_p is the mean of the prior (here taken as 0, the location straight ahead of the subject) and σ_p is the prior standard deviation. The selection of such a prior enforces a belief that subjects have learned that targets are more likely to appear at low eccentricity but have not learned the exact locations and frequencies of each specific target or target pair. Fitting such a prior allows us to incorporate some intersubject variability, and induces a compressive bias which is compatible with many psychophysical results (Odegaard et al. 2015), without enforcing strong assumptions about the subject's perfectly learning the actual target distribution. Alternative priors for the form of this bias could be chosen here to reflect either a stronger assumption (for instance, making the prior distribution exactly match distribution the of target locations) or a weaker assumption (such as a flat prior over all potential locations) (Acerbi et al. 2018).

Unity judgment. Above, we have focused on a generative model in which the number of causes, c, is known. But the task given to the observer is to make inferences about c from internal representations, x. The choice of causal inference strategy determines how the observer model decides between the $c = 1$ and $c = 2$ cases when presented with sensory stimuli. In general, this choice can follow Bayesian principles, non-Bayesian heuristics (i.e., fixed criterion), or strategies that do not actually implement causal inference at all (i.e., forced fusion). There is considerable behavioral work exploring the relative merits of both Bayesian and heuristic forms of causal inference in humans, which is outside the scope of this paper (Acerbi et al. 2018; Odegaard and Shams 2016; Rohe and Noppeney 2015b; Wozny et al. 2010). Instead, we present a Bayesian form of causal inference and contrast this with a maximally flexible null model that does not perform causal inference at all.

The Bayesian causal inference strategy will compute the posterior probability of the $c = 1$ and $c = 2$ cause cases, given sensory information, as follows,

$$p(c|x_A, x_V) = \frac{p(x_V, x_A|c)p(c)}{p(x_A, x_V)} \quad (5)$$

where $p(c)$ reflects the prior probability of a common cause $p(c = 1) = 1 - p(c = 2) = p_{common}$, which is left as a free parameter

Fig. 1. Subjects accurately localize multiple stimuli in a novel behavioral paradigm. *A*: each trial begins with fixation at a central target location. After a variable stimulus presentation interval, subjects respond by making saccades to the sensory target(s). For single target trials [*top*: either unisensory trials or trials with coincident auditory (A; Aud.) and visual (V, Vis.) stimuli], subjects make a single saccade to the perceived location. For multiple target trials (*bottom*), subjects make two saccades in rapid succession to each target in any order. These two trial types and all target combinations were interleaved throughout the session. *B*, *left*: human subjects were able to localize both visual (blue) and auditory (red) stimuli at all stimulus locations when presented alone on unisensory trials. Note that auditory responses have higher standard deviation, indicating lower perceptual accuracy (error bars show SD). Example eye traces (*middle*) and extracted saccade end points (*right*) for a representative dual-stimulus condition with targets well separated in space (24° separation). *Right*: saccade end points are color coded according to whether they were labeled as auditory or visual responses on the given trial (see MATERIALS AND METHODS). *C* and *D*: same as in *B* but for two monkey subjects. Target separation for *middle* and *right* is 18° (rather than 24° for humans) due to a difference in targets used between species (see MATERIALS AND METHODS).

(reflecting an unknown innate bias that may vary subject to subject). Because there are only two possibilities for causal state in this paradigm, this can be written as

$$p(c=1 \mid x_V, x_A) = \frac{p(x_V, x_A \mid c=1)p_{common}}{p(x_V, x_A \mid c=1)p_{common} + p(x_V, x_A \mid c=2)(1-p_{common})} \quad (6)$$

The sensory likelihoods depend on the choice of prior in the previous section, according to

$$p(x_A, x_V \mid c=1) = \int p(x_A \mid S_{AV})p(x_V \mid S_{AV})p(S_{AV} \mid c=1)dS_{AV} \quad (7)$$

$$p(x_A, x_V \mid c=2) = \int p(x_A \mid S_A)p(S_A \mid c=2)dS_A \times \int p(x_V \mid S_V)p(S_V \mid c=2)dS_V \quad (8)$$

For the simple normal prior these can be solved analytically, but for other forms of prior numerical integration is required. To ensure fairness during the model comparison steps, all likelihoods are computed using the same method (numerical integration).

To transform this posterior probability distribution into an experimentally observed binary response (1 or 2 saccades), we must specify a decision rule. For the Bayesian model, we assume subjects report whichever causal scenario has the highest posterior probability. That is,

$$\Pr(\text{choose unity} \mid x_A, x_V) = \frac{\lambda}{2} + (1-\lambda)[\![\Pr(c=1 \mid x_A, x_V) > 0.5]\!] \quad (9)$$

where λ represents the lapse rate (the subject randomly makes a response) and $[\![\cdot]\!]$ is the Iverson bracket, which is 1 when the statement inside is true and 0 otherwise.

The null model for the unity judgment task is that the subject randomly responds with either one or two saccades at some arbitrary ratio. Such a response pattern would indicate that the subject had learned or decided that a mixture of single and dual saccades was required but did not use any information about the distance between the targets to make this decision. We implement this by fitting a fixed rate of single-saccade responses, equivalent to p_{common}, that is common across all combinations of x_A, x_V.

Localization with causal inference. For the localization component of the task, subjects must arbitrate between the two possible causal conclusions (*Eqs. 3* and *4*). We assume this involves some reweighting of the two estimates that is dependent on the sensory percept,

$$p(\hat{S}_A, \hat{S}_V \mid x_A, x_V) = w_1(x_A, x_V)p(S_A = S_V \mid x_A, x_V, c=1) + [1 - w_1(x_A, x_V)]p(S_A, S_V \mid x_A, x_V, c=2) \quad (10)$$

where w_1 is a decision weight applied to the $c = 1$ condition.

Typically, Bayesian models of causal inference use a model averaging (MA) strategy to set these weights. This refers to reweighting the two possibilities according to the posterior probability, such that $w_1(x_A, x_V) = \Pr(c = 1 \mid x_A, x_V)$ (Acerbi et al. 2018; Cao et al. 2019; Körding et al. 2007). Alternatively, subjects could adopt a model selection (MS) strategy. This means they determine which causal structure has the highest posterior probability (*Eq. 6*) and then adopt that strategy for the localization component. In this case the weight is equivalent to *Eq. 9*, such that the weight applied to the $c = 1$ condition is 1 when that is the most likely causal scenario and 0 otherwise.

We compare these models with a null probabilistic fusion (PF) strategy, which is analogous to the null strategy considered in the previous section for the unity judgment case. In this model, subjects randomly select from either the fused or segregated location estimates according to some fixed probability (defined by the free parameter p_{common}), which does not depend on the sensory input, such that $w_1(x_A, x_V) = \Pr(c = 1) = p_{\text{common}}$. This model includes as special cases the strategies of *1*) always fusing the stimuli (always indicating locations intermediate of the two targets), *2*) always segregating the stimuli (never indicating intermediate locations), or *3*) some random mixture of the two. The critical difference between this model and the CI models discussed above is that the mixture strategy *3*) does not depend on target separation. Rather, subjects are assumed to randomly switch between fusion and segregation at a fixed rate set by p_{common}. Importantly, this model differs from the above models only in terms of the predicted location of saccade responses, and not in number of saccades expected, because these two reports (location and number of saccades) are made independent for the purposes of model fitting (see next section).

Comparing with behavioral data. The above response estimates are dependent on internal variables, x_A and x_V, which are not accessible to the experimenter. To obtain distributions that can be compared with data, *Eq. 10* must be marginalized across the internal variables:

$$p(\hat{S}_A, \hat{S}_V \mid S_A, S_V) = \int\int p(\hat{S}_A, \hat{S}_V \mid x_A, x_V)p(x_A \mid S_A)p(x_V \mid S_V)dx_A dx_V \quad (11)$$

We compute this distribution using numerical integration for each of the 20 combinations of visual and auditory targets and then use the resulting distributions to calculate likelihood of the observed response distributions (saccade end points) for the purposes of parameter fitting.

A technical concern in the above is that two-source trials appear to contain two data points (the locations of the auditory and visual sources), while single-source trials contain only one. However, we can equally well consider a joint distribution of auditory and visual estimated source locations $(\hat{S}_A, \hat{S}_V)$ defined on a 2D plane in 1×1° bins. In this formulation, whereas two-source trials each provide a single (2D) data point consisting of the auditory and visual saccade end points, single-source trials are modeled as likewise providing a single 2D data point with both numbers constrained to be equal. That is, plotted in a 2D plane, the two-cause model places probability mass over the entire plane, while the single-cause model places it only along the diagonal. This allows us to fit a model that equally weights both single- and dual-saccade trials, as each trial type contributes only a single data point for purposes of model fitting.

For models that assume some form of causal inference, the unity judgment lapse rate discussed above, λ, will also affect the reported locations (because the subject may make only a single saccade, even though the targets are well separated). To account for this, we included a chance for a single saccade to either the visual or auditory locations rather than the fused location. It was assumed to be equally likely to make an auditory or visual saccade, and the total probability of such saccades was $\lambda/2$.

Model Fitting

Models were fit using a maximum likelihood approach to determine the set of parameters that best explains the provided data. This was accomplished using MATLAB's fminsearch function to minimize the negative log likelihood of the data under each of the above models. The search was initialized using the best starting parameters from a uniform grid search across 3,125 initial parameter settings (corresponding to 4 evenly spaced initialization points for each of the 5 starting parameters), conducted before fitting.

Model Comparison

We used a Bayesian random-effects model comparison to determine which of the above-described models provided the best fits for the observed behavioral data (Rigoux et al. 2014). We performed this model comparison based on the Bayesian Information Criterion (BIC): BIC = $-2\text{LL} + k \times \ln(n)$ where LL is the log-likelihood of the data under the model, k is the number of free parameters, and n is the number of data points. We then determined each models posterior frequency and protected exceedance probability using the variational Bayesian analysis toolbox (Daunizeau et al. 2014).

Data and Code Availability

Data and custom MATLAB code generated as part of this work can be found on GitHub at https://doi.org/10.5281/zenodo.3900181.

RESULTS

Behavioral Paradigm

Subjects (7 humans and 2 monkeys) were seated in a dark, anechoic chamber facing a row of colocated speakers and LEDs. Trials were randomly interleaved and consisted of either unisensory (single auditory or visual stimulus) or multisensory (both modalities, presented at the same time and for the same duration) stimuli. We used a delayed saccade paradigm, where subjects were required to maintain fixation at a central location for a variable interval during stimulus presentation, before making a saccade to indicate a response. This delay period ensures that each trial has a comparable amount of sensory processing time (avoiding biases toward more rapidly detected cues) and also avoids responses that may be confounded by insufficient motor planning time (Ottes et al. 1984, 1985). Multisensory trials had varying amounts of spatial separation between the auditory and visual targets, ranging from 0° (coincident) to 36°. On every trial, subjects were required to report the location of both the auditory and visual stimulus. For unisensory or multisensory-coincident trials (Fig. 1*A*, *top*), subjects made a single saccade to the perceived location of the stimulus source and then held fixation at that point. For multisensory-separate trials, subjects made two saccades in rapid succession, one to each of the perceived sources. It is important to note that data were not excluded if subjects failed to make two saccades in this condition, as it is expected that subjects will often perceive the two stimuli as coincident if the spatial separation is small (6–12°). This dual-report design allowed characterization of both explicit (one versus two saccades) and implicit (location of fused percept or independent percepts) causal inference on each trial.

First, we sought to determine whether both monkeys and humans perform the above task in a qualitatively similar manner. Subjects were able to localize both the auditory (red) and visual (blue) stimuli when presented on unisensory trials (Fig. 1, *B–D*, *left*). Visual localization was much more accurate and less biased (across subjects SD = 1.08 ± 0.07°, mean absolute error = 0.49 ± 0.19°) than auditory localization (SD = 4.07 ± 0.36°, error = 4.80 ± 1.61°), consistent with the higher sensory reliability of visual information for spatial localization tasks (Alais and Burr 2004; Witten and Knudsen 2005). On trials where the targets were well separated, subjects accurately localized both targets by making two saccades in rapid succession (Fig. 1, *B–D*, *middle*). The distribution of end points from these saccades were extracted and formed the basis for comparison with model predicted distributions (Fig. 1, *B–D*, *right*). These results demonstrate that both human and monkey subjects grasped the primary goals of the task and accurately reported both visual and auditory targets on a single trial.

Audiovisual Causal Inference

Causal inference, as it relates to multisensory localization, has two primary characteristics that can be compared with the data collected in our task. Explicit causal inference is the most straightforward and amounts to simply determining which of two (or more) causal scenarios is most likely to have generated the perceived sensory inputs (Fig. 2*A*) (Chen and Spence 2017; Wallace et al. 2004). In our task, this judgment is reported via the number of saccades made on a given trial. A subject performing explicit causal inference would therefore be expected to report perceiving separate sources (by making two saccades) more often when the targets are well separated and to

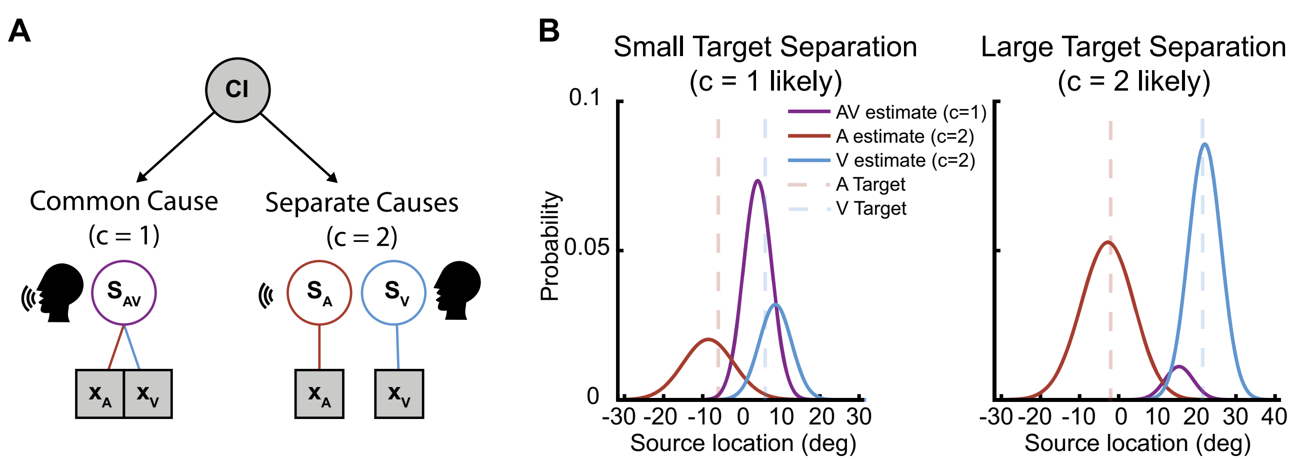


Fig. 2. Causal inference (CI) in audiovisual localization. *A*: the generative model producing sensory percepts (x_A, x_V) is assumed to have two potential causal structures: either sharing a common source S_{AV} (though perturbed by different amounts of sensory noise, *left branch*) or having independent sources S_A and S_V (*right branch*). Causal inference is accomplished by inverting this generative model and determining which branch is most likely given the observations x_A and x_V. *B*: the generated probability distributions are illustrated for two hypothetical example conditions. *Left*: the auditory source (red dashed line; A) is located just to the left of center while the visual source (blue dashed lined; V) is located just to the right of center. Because the targets are relatively close together, it is expected that many of the reports will match the common cause estimate (purple curve; AV) rather than the segregated estimates (red and blue solid curves). The mean of the segregated estimates is slightly shifted from the actual target location, a consequent of causal inference (see main text). *Right*: the targets are further apart in space, resulting in a higher probability of segregated reports and less bias. The distributions are normalized such that the total area under the solid curves sum to one.

report them as sharing a common source (by making a single saccade) when the targets are close together or coincident.

The second characteristic is demonstrated when subjects localize stimuli in a scenario where the causal structure is uncertain. Each of the potential causal structures should result in different source estimates (Körding et al. 2007; Mahani et al. 2017; de Winkel et al. 2017). This means location reports will be implicitly shaped by the explicit judgment, as subjects will more often report intermediate locations (Fig. 2*B*, purple curves) when they judge the stimuli to share a source (leaving aside the possibility for lapses or mistakes; see MATERIALS AND METHODS), and not when they perceive them as separate (Fig. 2*B*, red and blue curves). This will shape the distribution of expected responses (slight offset between peak of solid curves and dashed line) because stimuli that randomly appear to be closer together due to sensory noise are more likely to be fused. Crucially, this shift should depend on the separation between targets; subjects should report more intermediate locations when the targets are close together and more segregated estimates when they are well separated.

We operationalized the above descriptions of causal inference by quantitatively modeling responses, adapting generative models common in the literature (Acerbi et al. 2018; Körding et al. 2007; Wozny et al. 2010). We built a set of several observer models, which reflect different assumptions about the observer's causal inference strategy, and compared them to a null model involving a default strategy instead. Our overarching goal was to ascertain whether and how subjects compared information across modalities to infer the number of causes (referred to as the unity judgment component), and whether they then used that information to inform their localization of those causes (referred to as the localization component).

On each trial, we assumed that some set of sensory sources, S_A and S_V, produced noisy internal measurements x_A and x_V. These noise distributions were assumed to be Gaussian, centered on the source, with sensory variance encoded by the free parameters σ_A^2 and σ_V^2, respectively. We assumed that these sensory variances were used for both the generative model of responses and the observer's internal estimate of causality; that is, both the localization and unity judgment steps. Both auditory and visual stimuli were assumed to share the same, normally distributed prior, centered at 0 with variance σ_{prior}^2.

We considered two separate response strategies for performing the unity judgment component of the task based on these noisy internal measurements: Do subjects use the visual and auditory information on each trial or do they use a default heuristic that amounts to guessing in a probabilistic fashion? The first possibility is captured under an idealized Bayesian observer (Bay; see MATERIALS AND METHODS). For the Bay model, the observer reports a common cause when the posterior probability for that one cause is greater than 0.5, $\Pr(c = 1|x_A, x_V) > 0.5$. The prior probability of common cause, $p_{\text{common}} = \Pr(c = 1)$ is a free parameter, reflecting an unknown bias toward either single or dual saccades.

The default, heuristic possibility is captured under a non-CI probabilistic fusion model (PF). This model does not take the disparity between the visual and auditory stimuli into consideration at all. Rather, participants might always fuse, always segregate, or make a probabilistic choice between the two (the latter indicating that the subject has learned that a mix of behaviors is required). The PF model is equivalent to forced fusion or forced segregation models commonly compared with causal inference models (Cao et al. 2019; Körding et al. 2007; Wozny et al. 2010) but is more general, with the possibility of a response pattern intermediate to these extremes. For the PF model the observer reports a single cause on some fixed percentage of the trials, defined by p_{common}.

Estimating the source location, $p(S|x_A, x_V)$, depends on the assumed causal structure. If the measurements x_A and x_V are assumed to originate from the same source, the localization estimate will be a weighted average reflecting the relative reliability of each of the cues (S_{AV}; Fig. 2*B*, purple), consistent with well-established maximum likelihood models of multisensory integration (Alais and Burr 2004; Ernst and Banks 2002). When the sources are assumed to be independent, the resulting estimates are independent and rely only on the sensory information and the prior (S_A, S_V; Fig. 2*B*; blue and red). Because the amount of separation between targets varied randomly from trial to trial, subjects could not know ahead of time which response pattern was ideal. This forced them to adopt some kind of behavioral strategy to arbitrate between these two response patterns (Fig. 2*B*, *left* versus *right*).

To characterize the localization component of the task under different behavioral strategies, we considered three possible models, two that incorporate causal inference and one that does not. The first causal inference model was a Bayes optimal strategy, in which the observer combined the potential localization estimates (i.e., fused or separate) according to the posterior probability of causal structure, $\Pr(c = 1|x_A, x_V)$, which we refer to as model averaging (MA) (Körding et al. 2007). The second causal inference model, which we refer to as model selection (MS), implemented a heuristic decision rule (Rohe and Noppeney 2015b). Instead of weighting the fused and separate estimates according to posterior probability, the model implemented a threshold decision and simply selected whichever causal structure was most likely (i.e., selecting the fusion strategy when $\Pr(c{=}1|x_A, x_V){>}0.5$, and otherwise using the segregated strategy).

It is possible that subjects perform causal inference when determining the number of saccades to make, but that this causal inference does not affect the actual locations reported. For instance, subjects could determine whether one or two saccades were required and then simply rely on the segregated unisensory estimates to direct saccades (rather than fusing the two estimates when making a single saccade). We therefore compare the two CI models against a default probabilistic fusion (PF) model, which does not incorporate information from the causal judgment in its generated saccade locations. Instead, the subject is modeled as choosing between fused and segregated response distributions randomly at some fixed rate, defined by p_{common}. This includes the possibility for either always-fuse or always-segregate response patterns. Importantly, all of these models implement the same rule for choosing to make either one or two saccades on a given trial (the Bay model for unity judgment). This means that each model was compared only based on how well it captured the distribution of location reports, rather than the ratio of single to dual saccade reports.

To provide an intuitive understanding of each of these models, it is helpful to consider the differences in predicted response distributions for each. Both CI strategies (Bay-MA

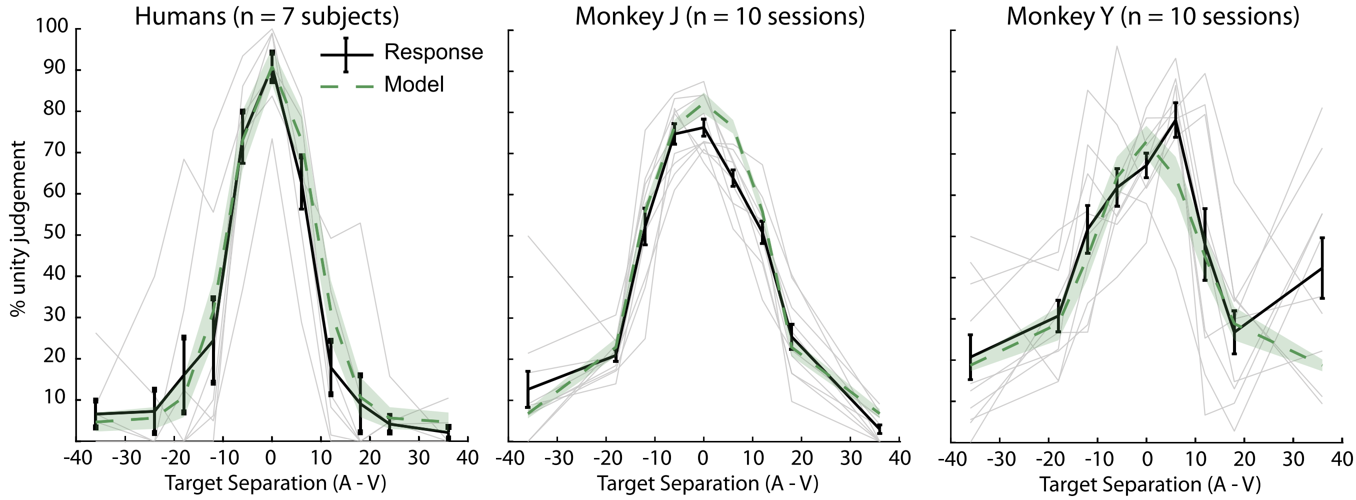


Fig. 3. Unity judgment as a function of target disparity. Human subjects (n = 7, *left*) and monkey subjects (n = 10 experimental sessions, *middle*, *right*) demonstrated a pronounced preference for making one saccade when targets were close together, rather than well separated. These judgments were well fit by a Bayesian model of causal inference (jointly fit Bay-MA model, green curves). Gray lines show mean responses for individual subjects (humans) or experimental sessions (monkeys). Error bars reflect SE across subjects (human) or across sessions (monkey). Shaded region represent SE of model predictions across subjects/sessions. Bay, idealized Bayesian observer; MA, model averaging.

and Bay-MS) predict that the subjects will make more fused saccades when the targets are close together (purple distribution in Fig. 2*B*) but then transition to segregated auditory- and visual-guided saccades when they are well separated (red and blue distributions in Fig. 2*B*). These models differ mainly in how this transition will occur, as the Bay-MA model will transition smoothly between the fused and segregated estimates, while the Bay-MS model will have a sharper boundary. Conversely, the Bay-PF model will have a fixed ratio of the fused and segregated distributions regardless of target separation. In practice, this would most likely take the form of a pure segregation strategy, where the subject makes a single visual or auditory guided saccade when the targets are close together, rather than fusing the two estimates to make an intermediate fused saccade (i.e., saccades will always be drawn from blue or red distributions in Fig. 2*B*, and never the purple distribution).

We fit the models using a maximum-likelihood strategy, estimating the parameters that provided the best fit for the behavioral data (number of saccades or location of saccades, for unity judgment and localization models, respectively) for each condition. In addition to fitting to each of these separate task components independently, we also fit models jointly to both components of the task (i.e., maximizing likelihood for both the number of saccades and the location of saccades using the same set of parameters). When fitting the joint models, the unity judgment component was assumed to follow the Bay strategy, while the localization component was varied between the three possibilities described above. For illustration purposes only the jointly fit Bay-MA model is shown in Figs. 3 and 4, though all models are compared quantitatively in Fig. 5.

Unity Judgment

To determine whether subjects were performing causal inference in our task, we first analyzed the explicit portion of the response: whether the subject made one or two saccades. We found that subjects were much more likely to make one saccade when the targets were coincident or close together and much more likely to make two saccades when the targets were well separated (Fig. 3). This means that the observers were not performing pure fusion (always integrating stimuli), nor pure segregation (always treating the stimuli as independent), but instead adopting a strategy that depended on target separation. Importantly, humans and monkeys showed qualitatively similar performance on this component of the task (Fig. 3, *left* versus *middle* and *right*). This response pattern was well described by an ideal Bayesian observer model of causal inference for all subjects (Fig. 3, Bay-MA model, green) (Körding et al. 2007). These results indicate that monkeys understood and perform the explicit causal inference component of the task and that their behavior is well described by a causal inference model previously only applied to human behavior.

Localization

We next sought to determine the effects of causal inference on the localization of stimuli. When targets were presented at a single location, subjects overwhelmingly made a single saccade to that location (Fig. 4, *left*, black bars). Conversely, when the targets were well separated, subjects accurately reported the location of both the visual and auditory sources (Fig. 4, *right*, red and blue bars). At intermediate locations, subjects responded with a mixture of fused single-saccade trials and separated double-saccade trials (Fig. 4, *middle*). Like in the unity judgment component of the task, these response distributions were well fit by an ideal observer model performing Bayesian causal inference with model averaging (solid green line). These results demonstrate that our task recapitulates both the explicit (ratio between single- and double-saccade trials) and implicit (biases in localization between single- and dual-saccade trials) components of causal inference in both humans and monkeys.

Model Comparison

Finally, we performed a quantitative model comparison between potential behavioral strategies for each component of the task (unity judgment and localization), as well as for

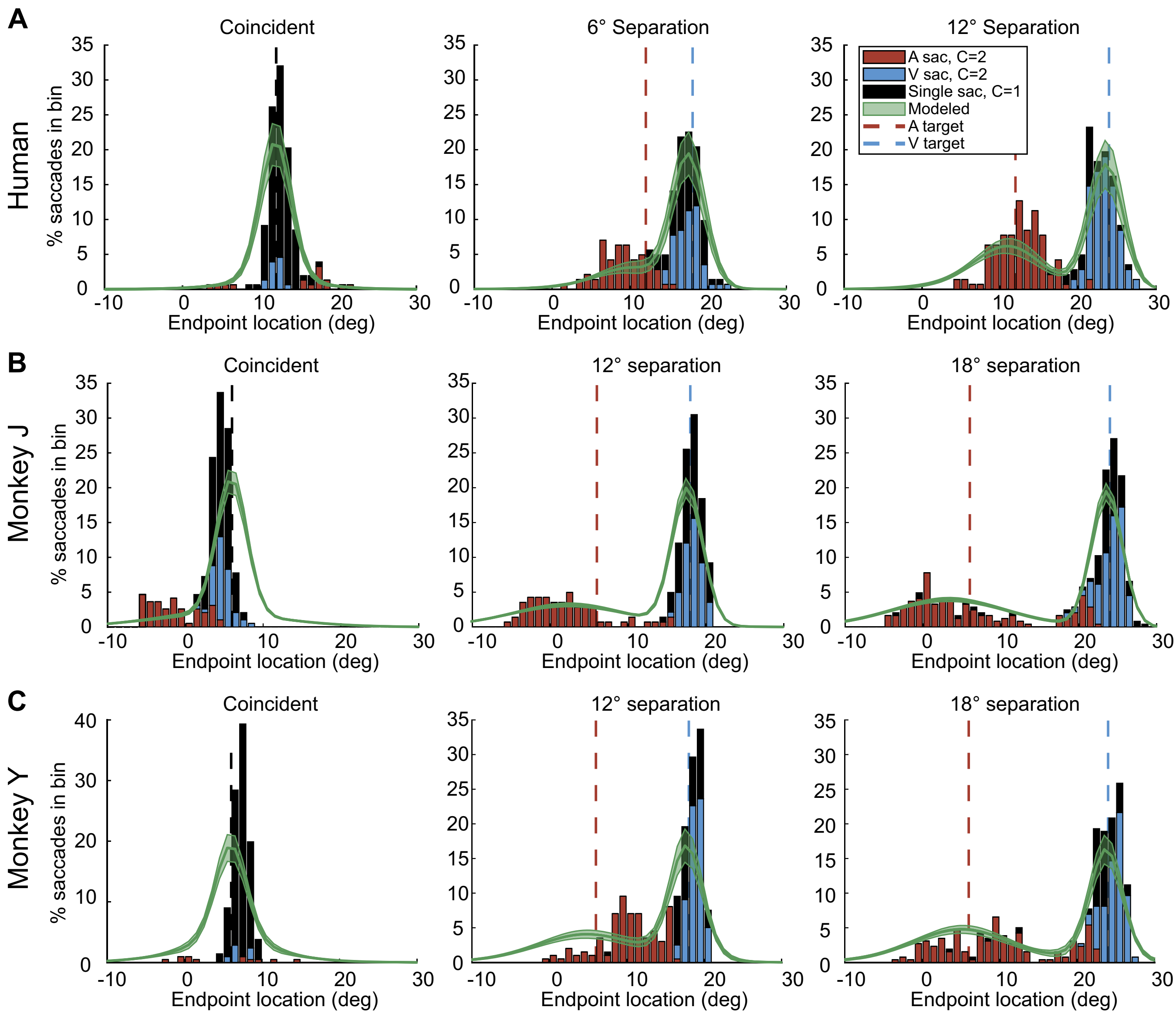


Fig. 4. Localization of stimulus sources. Reported location is shown combined across single- and dual-saccade trials for three example conditions, with single-saccade trials labeled in black while dual-saccade trials contribute to both auditory (red; A) and visual (blue; V) distributions and are well described by a single causal inference (CI) model (jointly fit Bay-MA model). At coincident target locations, subjects usually reported with a single saccade at that location (*left*, black bars). As the targets move further apart in space, subjects gradually shift from an integration strategy to complete segregation (*middle* and *right*). Humans (*A*) and monkeys (*B* and *C*) had similar behavioral performance, though monkeys had a more pronounced auditory bias and worse localization accuracy (consistent with differences seen in unisensory trials, Fig. 1, *C* and *D*). Shaded region reflects SE of model predictions across subjects (human) or sessions (monkeys). Bay, idealized Bayesian observer; MA, model averaging; sac, saccade.

models fit jointly to both components. We performed a Bayesian random-effects model comparison to determine which of the tested models provided the best fit for the observed data. In Fig. 5, we report the protected exceedance probability (P_{exc}, i.e., the probability that a given model is more likely than any other, corrected for chance) as well as the posterior model frequency for reference (Daunizeau et al. 2014; Rigoux et al. 2014). First, we compared the Bay model, which implements Bayesian causal inference, with the null PF model in the unity judgment case (Fig. 5, left). We found that the Bay model provided much better fits for both human ($P_{exc} = 0.98$) and monkey ($P_{exc} = 0.76$) subjects. Put another way, this indicates that a Bayesian causal inference strategy was ~49 times and ~3 times more likely to be the most representative model of behavior compared with a probabilistic fusion strategy for humans and monkeys, respectively. These results again confirm that both species were taking cue disparity into account when reporting common or separate causes, rather than simply responding according to some fixed guessing strategy.

We compared the two different CI strategies for localization (MA and MS) as well as the non-CI probabilistic PF strategy (Fig. 5, *middle*). Both species were much better fit by one of the CI models, though the preferred strategy differed between species. The model-averaging observer provided the best fit for human subjects ($P_{exc} = 0.73$), while providing worse fits for monkeys ($P_{exc} = 0.14$). Conversely, monkeys were better fit by a model selection strategy ($P_{exc} = 0.71$) than humans ($P_{exc} = 0.18$). For both species, the probabilistic fusion model

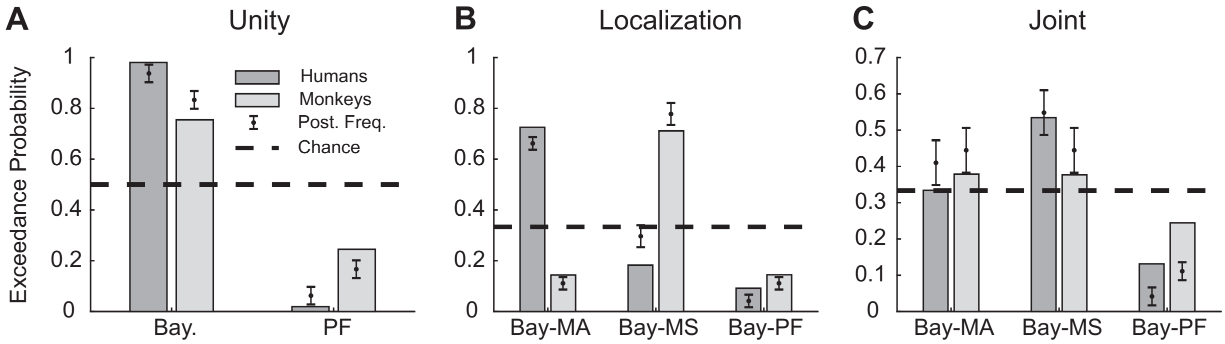


Fig. 5. Model comparison. Protected exceedance probabilities (bars) and expected posterior model frequencies (Post. Freq.; mean ± SD) are shown for the two unity judgment and three localization models across species. *A*: Bayesian causal inference (CI; Bay) model provides better fits of subjects' responses on the unity judgment component than a probabilistic fusion (PF) model. *B*: performance of two CI models [model averaging (MA) and model selection (MS)] compared with non-CI PF model, applied only to localization data. Both species are better fit by models incorporating CI (MA and MS models) than the non-CI model (PF), though the preferred model is different between species. *C*: same as *B*, but now comparing models fit jointly to unity judgment and localization responses. There is little evidence favoring either CI model over the other, though both outperform the non-CI PF model.

provided significantly worse fits (human subjects: $P_{exc} = 0.09$; monkey subjects: $P_{exc} = 0.15$) than the combination of CI strategies. Collapsing across CI models, some form of CI was ~10 times more likely than the PF model for humans, while for monkeys CI was ~6 times more likely. These results indicate that subjects from both species are incorporating causal inference into the localization component of the task, rather than only the unity judgment component.

When this same set of models were fit jointly to both the localization and unity judgment data, the species level difference between model fits disappeared and it was no longer possible to differentiate the two CI strategies (Fig. 5, *right*). Both of the CI strategies provided a better fit than the PF strategy. The PF strategy was ~7 times and ~3 times less likely than the combination of MA and MS models for humans and monkeys, respectively (human subjects: $P_{exc} = 0.13$; monkey subjects: $P_{exc} = 0.24$).

Model fit parameters for the jointly fit Bay-MA model were also compared (Table 1) and provided reasonable results that were broadly consistent across species. Fit perceptual noise values roughly matched those obtained from purely unisensory localization (Fig. 1) and were higher for the auditory stimulus (reflecting worse sensory acuity for this modality). Both monkeys and humans had relatively large standard deviation on their prior, indicating a weak biasing toward central reports. Humans displayed a slightly bias toward reporting only one target (p_{common}>0.5), but both species had priors that approximately reflected the true single-target rate of 0.5. Both monkeys and humans had relatively high lapse rates (with monkeys performing worse), suggesting that some of the single-saccade trials were better described as failed or aborted double-saccade trials rather than reflective of truly fused stimuli.

Table 1. *Model fit parameters for joint Bay-MA model*

Parameter	Description	Human	Monkeys
σ_A	Auditory perceptual noise	3.95° (0.60°)	6.09° (0.33°)
σ_V	Visual perceptual noise	1.47° (0.02°)	1.52° (0.09°)
σ_p	Standard deviation of central prior	23.17° (2.94°)	18.81° (2.40°)
p_{common}	Prior probability of common cause	0.60 (0.09)	0.52 (0.06)
λ	Lapse rate	0.09 (0.05)	0.25 (0.03)

Mean parameter values are shown, calculated across subjects (Human) or across subjects and experimental sessions (Monkeys). Parenthetical values reflect SE of parameter fits. Bay, idealized Bayesian observer; MA, model averaging.

In conclusion, our results show that CI models provided much better fits than non-CI models across unity judgment, localization, and combined joint data sets for both monkeys and humans. We did not find significant differences between the two versions of CI model tested (idealized MA or heuristic MS) when comparing the models on the jointly fit data, and so the exact causal inference strategy implemented in this task remains an open question. Most importantly, we found that goodness of model fits and parameter values were quite similar across species for the best performing Bay-MA causal inference model, suggesting that nonhuman primates provide a relevant model organism for the study of causal inference.

DISCUSSION

We have presented a novel behavioral paradigm for multisensory localization that provides rich perceptual readouts for both human and monkey subjects. We demonstrated that subjects are capable of localizing both auditory and visual stimuli on single trials where the sources of those stimuli are well separated in space. On trials where the stimuli were either coincident or separated by small amounts (6 to 12°), subjects often colocalized the stimuli, reporting them as sharing the same source, consistent with previous reports of audiovisual fusion or visual capture (Alais and Burr 2004; Hecht and Reiner 2009; Jack and Thurlow 1973; Thurlow and Jack 1973; Wallace et al. 2004). Importantly, subjects shifted from fusion to segregation as a function of target separation, as predicted by existing models of causal inference (Acerbi et al. 2018; Körding et al. 2007; Rohe and Noppeney 2015b; Wozny et al. 2010). Both human and monkey subject behavior was much better fit by models which incorporated some form of causal inference, as compared with models which spanned the possibilities from forced fusion to segregation (i.e., processing is completely separated by modality). Together these results demonstrate the effectiveness of this paradigm for eliciting causal inference judgments in both humans and monkeys and validate nonhuman primates as a model organism for studying the neural basis of causal inference.

These results are consistent with a growing body of behavioral research in humans indicating that causal inference is a critical component of sensory processing (Acerbi et al. 2018; Chen and Spence 2017; Körding et al. 2007; Locke and Landy 2017; Mahani et al. 2017; Odegaard et al. 2015; Rohe and Noppeney 2015b; de Winkel et al. 2017). We found that relatively simple models of causal inference (with 5 or fewer parameters) provided good fits even in a complex behavioral paradigm, which had both a dual-report structure and a continuous reporting variable. We note that the success of these models in describing behavioral output does not imply a perfect description of perception. For instance, it is not possible for us to determine conclusively whether single saccades reflect that the auditory and visual stimuli have been "bound" into a single object (an assumption made by our CI models), or whether they are simply perceived as coming from the same location in space. Differentiating these possibilities requires further extensions of our paradigm, which can explicitly test for feature binding (Acerbi et al. 2018; Bizley et al. 2016b). Nevertheless, the prediction accuracy of the simple models considered here further supports the relevance of such models for the study of multisensory perception, as a small number of biologically relevant parameters offer significant predictive power.

Recent human neuroimaging studies have suggested that causal inference may be accomplished by subdividing the task into pieces (i.e., integration, segregation, etc.) and performing these computations in separate brain regions before combining them in some higher level brain region such as prefrontal cortex (Aller and Noppeney 2019; Bizley et al. 2016a; Cao et al. 2019; Mahani et al. 2017; Regenbogen et al. 2018; Rohe and Noppeney 2015a, 2016). This view of hierarchical neural processing is pleasingly consistent with the hierarchical nature of ideal observer models of causal inference (Fig. 2*A*). However, it is inconsistent with other research showing significant interaction between modalities even in primary sensory areas (Atilgan et al. 2018; Cappe and Barone 2005; Ibrahim et al. 2016; Iurilli et al. 2012), as well as numerous descriptions of multisensory integration in subcortical brain regions (Alvarado et al. 2007; Angelaki et al. 2009; Gruters et al. 2018; Kadunce et al. 1997; Meredith and Stein 1986; Porter et al. 2007; Stein et al. 2014; Wallace et al. 1998). It is possible that this conflict is due in part to the level of experimentation, as the former findings rely on human neuroimaging (fMRI and MEG), which is necessarily limited to large-scale changes in neural activity in cortical structures, while the latter studies principally investigated single neurons or small networks of neurons. A major advantage of the current study is that it employs a single behavioral paradigm for both monkeys and humans. This allows for direct comparison at the behavioral level, and this strategy can be extended in the future to resolve this disagreement at the neuronal level. Further experiments are needed to determine whether multisensory causal inference is truly a brainwide computation or whether it can be accomplished in smaller networks of individual neurons.

One unexpected finding in this work was that monkey and human subjects appeared to adopt different causal inference strategies for the localization task, with humans being best fit by model averaging while monkeys were best fit by model selection. This raises the intriguing possibility of cross-species differences in causal inference, but verifying such differences would require an alternative experimental emphasis. We designed this task to be compatible with monkey single-unit electrophysiology experiments, which imposes some limitations from a behavioral modeling perspective. Most importantly, to keep the total number of conditions reasonably small, we did not vary the sensory reliability for either stimulus type. This prohibits us from weighing in on the exact nature of the causal inference, whether Bayes optimal (i.e., model averaging, which minimizes localization error) or some heuristic alternative (i.e., probability matching or comparing to a fixed criterion, which may be simpler computationally). Recent work with a more thorough model comparison has shed some light on this subject, suggesting that a heuristic fixed-criterion model may provide better fits for human behavior (Acerbi et al. 2018). However, this question is by no means settled, and there is significant variability even between individual human subjects performing the same task (Odegaard and Shams 2016; Wozny et al. 2010). We therefore leave the question of which exact model of causal inference best describes behavior, and whether monkeys and humans implement identical or subtly different strategies, to future studies that can bring more statistical power to bear.

While using saccades as a continuous behavioral readout offers many advantages, this choice does impose certain important limitations. The first of these is that motor noise is conflated with sensory noise: that is, a portion of the variability in responses is due to errors in the motor output rather than uncertainty about the target location. Previous work has determined that the contribution of motor noise to saccade variance is only slightly smaller than the contribution of sensory noise (van Beers 2007), suggesting that this issue is not trivial. Therefore, our estimates of sensory noise may be higher than the actual estimates used by the subject's brain when performing causal inference, particularly for the visual targets, which are likely to have low sensory noise. This appears to be the case when comparing response distributions to those predicted by the model (Fig. 4, over dispersion of model distributions relative to histograms). However, because this issue affects all our models equally, it should not affect our model comparison results nor our conclusions.

Additionally, motor errors have been reported to result in "averaging saccades" even when the targets are easily discriminable as separate (Ottes et al. 1984, 1985). However, this type of averaged saccade response is typically seen in saccades made with very short response latencies after target onset ($<$300 ms for humans, $<$150 ms for monkeys) and does not typically occur in delay saccade paradigms such as the one described here (see, e.g., Kim and Basso, 2008). Therefore, this specific motor error is unlikely to be a major driving factor in our results.

The second limitation imposed by saccadic reporting is that there is a natural lower bound on target separation that can be realistically indicated using two saccades. Both humans and monkeys make near-constant corrective microsaccades ($<$1–2° in amplitude), which are necessarily differentiated from voluntary saccades in this task (Otero-Millan et al. 2008). This makes it essentially impossible for subjects to indicate two separate targets if they perceive them as separate but closer than ~3° apart, as any smaller magnitude of saccade would be interpreted as corrective. This is not expected to influence our results, as all target pairs used in this experiment have a separation of at least 6° (or are coincident), and at this smallest separation value subjects still overwhelmingly report a single percept. However, it is worth considering in similar paradigms that might use sensory stimuli with less inherent sensory noise.

Aside from these limitations, our task has many advantages over similar, previously reported CI tasks. Localizing and orienting to sensory stimuli through saccadic eye movements is an innate behavior for primates, which facilitates comparison across species. Using saccades also allows for a continuous report of perceptual location, rather than button presses or two-alternative forced-choice paradigms, which necessarily limit the number of potential responses and provide an indirect mapping between perception and response. Our approach therefore allows for a more thorough characterization of behavior from trial to trial, while leveraging a natural behavior to speed training and data collection compared with other continuous report tasks (Wallace et al. 2004).

Another advantage is that the dual-report design (requiring both a unity judgment and localization) and interleaved nature of the task is more consistent with how sensory inference is performed in the natural environment. Subjects must attend to and act on sensory information of different modalities from moment to moment. They cannot simply rely on a strategy such as focusing only on visual or auditory inputs and neglecting all others (as with a blocked trial design). This is critical for understanding the neural basis of this operation, as focusing attention on only one modality or region of space is likely to significantly influence neural responses (Goldberg and Wurtz 1972).

Most importantly, this task design allows for direct comparison between human and monkey behavioral subjects and can characterize important features of causal inference within a single experimental session. By demonstrating that monkeys and humans perform the task similarly, we validate an animal model aligned with the growing body of human behavioral research in this area. This will enable the use of much higher resolution recording techniques that are difficult or impossible to use in human subjects, which is critical for bridging the gap between our understanding of multisensory causal inference at the behavioral and neuronal levels.

ACKNOWLEDGMENTS

We thank Shawn Willett for helpful feedback during the preparation of this manuscript and Jeff Beck for useful discussion about various model features.

GRANTS

This work was supported by a National Defense Science and Engineering Graduate Fellowship (32 CFR 168a) from the Department of Defense and the American Society for Engineering Education to JTM and NIH R01 DC016363 to JMG.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

J.T.M. and J.M.G. conceived and designed research; J.T.M. performed experiments; J.T.M. analyzed data; J.T.M., J.P., and J.M.G. interpreted results of experiments; J.T.M. prepared figures; J.T.M. drafted manuscript; J.T.M., J.P., and J.M.G. edited and revised manuscript; J.T.M., J.P., and J.M.G. approved final version of manuscript.

REFERENCES

Acerbi L, Dokka K, Angelaki DE, Ma WJ. Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLOS Comput Biol* 14: e1006110, 2018. doi:10.1371/journal.pcbi.1006110.

Alais D, Burr D. The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol* 14: 257–262, 2004. doi:10.1016/j.cub.2004.01.029.

Aller M, Noppeney U. To integrate or not to integrate: temporal dynamics of hierarchical Bayesian causal inference. *PLoS Biol* 17: e3000210, 2019. doi:10.1371/journal.pbio.3000210.

Alvarado JC, Vaughan JW, Stanford TR, Stein BE. Multisensory versus unisensory integration: contrasting modes in the superior colliculus. *J Neurophysiol* 97: 3193–3205, 2007. doi:10.1152/jn.00018.2007.

Angelaki DE, Gu Y, DeAngelis GC. Multisensory integration: psychophysics, neurophysiology, and computation. *Curr Opin Neurobiol* 19: 452–458, 2009. doi:10.1016/j.conb.2009.06.008.

Atilgan H, Town SM, Wood KC, Jones GP, Maddox RK, Lee AKC, Bizley JK. Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. *Neuron* 97: 640–655.e4, 2018. doi:10.1016/j.neuron.2017.12.034.

Battaglia PW, Jacobs RA, Aslin RN. Bayesian integration of visual and auditory signals for spatial localization. *J Opt Soc Am A Opt Image Sci Vis* 20: 1391–1397, 2003. doi:10.1364/JOSAA.20.001391.

Bizley JK, Jones GP, Town SM. Where are multisensory signals combined for perceptual decision-making? *Curr Opin Neurobiol* 40: 31–37, 2016a. doi:10.1016/j.conb.2016.06.003.

Bizley JK, Maddox RK, Lee AKC. Defining auditory-visual objects: behavioral tests and physiological mechanisms. *Trends Neurosci* 39: 74–85, 2016b. doi:10.1016/j.tins.2015.12.007.

Cao Y, Summerfield C, Park H, Giordano BL, Kayser C. Causal inference in the Multisensory Brain. *Neuron* 102: 1076–1087.e8, 2019. doi:10.1016/j.neuron.2019.03.043.

Cappe C, Barone P. Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *Eur J Neurosci* 22: 2886–2902, 2005. doi:10.1111/j.1460-9568.2005.04462.x.

Chen YC, Spence C. Assessing the role of the 'unity assumption' on multisensory integration: a review. *Front Psychol* 8: 445, 2017. doi:10.3389/fpsyg.2017.00445.

Cuppini C, Shams L, Magosso E, Ursino M. A biologically inspired neurocomputational model for audiovisual integration and causal inference. *Eur J Neurosci* 46: 2481–2498, 2017. doi:10.1111/ejn.13725.

Daunizeau J, Adam V, Rigoux L. VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLOS Comput Biol* 10: e1003441, 2014. doi:10.1371/journal.pcbi.1003441.

de Winkel KN, Katliar M, Bülthoff HH. Causal inference in multisensory heading estimation. *PLoS One* 12: e0169676, 2017. doi:10.1371/journal.pone.0169676.

Dokka K, DeAngelis GC, Angelaki DE. Multisensory integration of visual and vestibular signals improves heading discrimination in the presence of a moving object. *J Neurosci* 35: 13599–13607, 2015. doi:10.1523/JNEUROSCI.2267-15.2015.

Dokka K, Park H, Jansen M, DeAngelis GC, Angelaki DE. Causal inference accounts for heading perception in the presence of object motion. *Proc Natl Acad Sci USA* 116: 9060–9065, 2019. doi:10.1073/pnas.1820373116.

Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415: 429–433, 2002. doi:10.1038/415429a.

Fetsch CR, DeAngelis GC, Angelaki DE. Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nat Rev Neurosci* 14: 429–442, 2013. doi:10.1038/nrn3503.

Goldberg ME, Wurtz RH. Activity of superior colliculus in behaving monkey. II. Effect of attention on neuronal responses. *J Neurophysiol* 35: 560–574, 1972. doi:10.1152/jn.1972.35.4.560.

Gruters KG, Murphy DLK, Jenson CD, Smith DW, Shera CA, Groh JM. The eardrums move when the eyes move: a multisensory effect on the mechanics of hearing. *Proc Natl Acad Sci USA* 115: E1309–E1318, 2018. doi:10.1073/pnas.1717948115.

Hecht D, Reiner M. Sensory dominance in combinations of audio, visual and haptic stimuli. *Exp Brain Res* 193: 307–314, 2009. doi:10.1007/s00221-008-1626-z.

Ibrahim LA, Mesik L, Ji X-Y, Fang Q, Li H-F, Li Y-T, Zingg B, Zhang LI, Tao HW. Cross-modality sharpening of visual cortical processing through layer-1-mediated inhibition and disinhibition. *Neuron* 89: 1031–1045, 2016. doi:10.1016/j.neuron.2016.01.027.

Iurilli G, Ghezzi D, Olcese U, Lassi G, Nazzaro C, Tonini R, Tucci V, Benfenati F, Medini P. Sound-driven synaptic inhibition in primary visual cortex. *Neuron* 73: 814–828, 2012. doi:10.1016/j.neuron.2011.12.026.

Jack CE, Thurlow WR. Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Percept Mot Skills* 37: 967–979, 1973. doi:10.2466/pms.1973.37.3.967.

Judge SJ, Richmond BJ, Chu FC. Implantation of magnetic search coils for measurement of eye position: an improved method. *Vision Res* 20: 535–538, 1980. doi:10.1016/0042-6989(80)90128-5.

Kadunce DC, Vaughan JW, Wallace MT, Benedek G, Stein BE. Mechanisms of within- and cross-modality suppression in the superior colliculus. *J Neurophysiol* 78: 2834–2847, 1997. doi:10.1152/jn.1997.78.6.2834.

Kim B, Basso MA. Saccade target selection in the superior colliculus: a signal detection theory approach. *J Neurosci* 28: 2991–3007, 2008. doi:10.1523/JNEUROSCI.5424-07.2008.

Knill DC. Robust cue integration: a Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *J Vis* 7: 5, 2007. doi:10.1167/7.7.5.

Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L. Causal inference in multisensory perception. *PLoS One* 2: e943, 2007. doi:10.1371/journal.pone.0000943.

Lochmann T, Deneve S. Neural processing as causal inference. *Curr Opin Neurobiol* 21: 774–781, 2011. doi:10.1016/j.conb.2011.05.018.

Locke SM, Landy MS. Temporal causal inference with stochastic audiovisual sequences. *PLoS One* 12: e0183776, 2017 [Erratum in *PLoS One* 12: e0186922, 2017]. doi:10.1371/journal.pone.0183776.

Ma WJ, Rahmati M. Towards a neural implementation of causal inference in cue combination. *Multisens Res* 26: 159–176, 2013. doi:10.1163/22134808-00002407.

Mahani MN, Sheybani S, Bausenhart KM, Ulrich R, Ahmadabadi MN. Multisensory perception of contradictory information in an environment of varying reliability: evidence for conscious perception and optimal causal inference. *Sci Rep* 7: 3167, 2017. doi:10.1038/s41598-017-03521-2.

Meredith M, Stein B. Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Res* 365: 350–354, 1986. doi:10.1016/0006-8993(86)91648-3.

Odegaard B, Shams L. The brain's tendency to bind audiovisual signals is stable but not general. *Psychol Sci* 27: 583–591, 2016. doi:10.1177/0956797616628860.

Odegaard B, Wozny DR, Shams L. Biases in visual, auditory, and audiovisual perception of space. *PLOS Comput Biol* 11: e1004649, 2015. doi:10.1371/journal.pcbi.1004649.

Otero-Millan J, Troncoso XG, Macknik SL, Serrano-Pedraza I, Martinez-Conde S. Saccades and microsaccades during visual fixation, exploration, and search: foundations for a common saccadic generator. *J Vis* 8: 21, 2008. doi:10.1167/8.14.21.

Ottes FP, Van Gisbergen JAM, Eggermont JJ. Metrics of saccade responses to visual double stimuli: two different modes. *Vision Res* 24: 1169–1179, 1984. doi:10.1016/0042-6989(84)90172-X.

Ottes FP, Van Gisbergen JAM, Eggermont JJ. Latency dependence of colour-based target vs nontarget discrimination by the saccadic system. *Vision Res* 25: 849–862, 1985. doi:10.1016/0042-6989(85)90193-2.

Porter KK, Metzger RR, Groh JM. Visual- and saccade-related signals in the primate inferior colliculus. *Proc Natl Acad Sci USA* 104: 17855–17860, 2007. doi:10.1073/pnas.0706249104.

Regenbogen C, Seubert J, Johansson E, Finkelmeyer A, Andersson P, Lundström JN. The intraparietal sulcus governs multisensory integration of audiovisual information based on task difficulty. *Hum Brain Mapp* 39: 1313–1326, 2018. doi:10.1002/hbm.23918.

Rigoux L, Stephan KE, Friston KJ, Daunizeau J. Bayesian model selection for group studies - revisited. *Neuroimage* 84: 971–985, 2014. doi:10.1016/j.neuroimage.2013.08.065.

Rohe T, Noppeney U. Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS Biol* 13: e1002073, 2015a. doi:10.1371/journal.pbio.1002073.

Rohe T, Noppeney U. Sensory reliability shapes perceptual inference via two mechanisms. *J Vis* 15: 22, 2015b. doi:10.1167/15.5.22.

Rohe T, Noppeney U. Distinct computational principles govern multisensory integration in primary sensory and association cortices. *Curr Biol* 26: 509–514, 2016. doi:10.1016/j.cub.2015.12.056.

Sato Y, Toyoizumi T, Aihara K. Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. *Neural Comput* 19: 3335–3355, 2007. doi:10.1162/neco.2007.19.12.3335.

Shams L, Beierholm UR. Causal inference in perception. *Trends Cogn Sci* 14: 425–432, 2010. doi:10.1016/j.tics.2010.07.001.

Stein BE, Stanford TR. Multisensory integration: current issues from the perspective of the single neuron. *Nat Rev Neurosci* 9: 255–266, 2008 [Erratum in *Nat Rev Neurosci* 9: 406, 2008]. doi:10.1038/nrn2331.

Stein BE, Stanford TR, Rowland BA. Development of multisensory integration from the perspective of the individual neuron. *Nat Rev Neurosci* 15: 520–535, 2014. doi:10.1038/nrn3742.

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. *J Acoust Soc Am* 26: 212–215, 1954. doi:10.1121/1.1907309.

Thurlow WR, Jack CE. Certain determinants of the "ventriloquism effect". *Percept Mot Skills* 36, 3_suppl: 1171–1184, 1973. doi:10.2466/pms.1973.36.3c.1171.

van Beers RJ. The sources of variability in saccadic eye movements. *J Neurosci* 27: 8757–8770, 2007. doi:10.1523/JNEUROSCI.2311-07.2007.

Wallace MT, Meredith MA, Stein BE. Multisensory integration in the superior colliculus of the alert cat. *J Neurophysiol* 80: 1006–1010, 1998. doi:10.1152/jn.1998.80.2.1006.

Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA. Unifying multisensory signals across time and space. *Exp Brain Res* 158: 252–258, 2004. doi:10.1007/s00221-004-1899-9.

Wei K, Körding K. Relevance of error: what drives motor adaptation? *J Neurophysiol* 101: 655–664, 2009. doi:10.1152/jn.90545.2008.

Witten IB, Knudsen EI. Why seeing is believing: merging auditory and visual worlds. *Neuron* 48: 489–496, 2005. doi:10.1016/j.neuron.2005.10.020.

Wozny DR, Beierholm UR, Shams L. Probability matching as a computational strategy used in perception. *PLOS Comput Biol* 6: e1000871, 2010. doi:10.1371/journal.pcbi.1000871.