**Central contribution:** Monkeys and humans use similar strategies for multisensory causal inference.

**Key points of this paper:**

1. We developed a task that can be used by both monkeys and humans
2. Humans behavior is consistent with previous descriptions of causal inference in similar tasks
   1. Unity
   2. Localization
   3. Joint fit of both
3. Monkey behavior is consistent with human behavior as well as XYZ models of causal inference
4. (maybe) demonstrate that enough behavioral data can be collected in a single session for ephys
5. Compare with other typical models for validation

Blue text is outline: 1 informal sentence per section, then 1 informal sentence per paragraph, then fill in.

**Abstract**

Option A: The sensory scene is sampled by multiple senses, but optimally fusing such signals first requires ascertaining whether particular signals arise from the same or different underlying objects or events. Here we present results from a novel behavioral task in which monkeys and humans were required to localize visual and auditory stimuli by saccading to each distinct location that they perceived. When visual and auditory locations were widely separated, participants made two saccades, whereas when the two stimuli were presented at the same location, only one saccade was needed. Intermediate levels of separation produced mixed response patterns, i.e. single saccades to an intermediate position on some trials or two saccades to both the visual and auditory locations on others. The distribution of responses was well matched to predictions of a hierarchical model which incorporated a) the spatial distribution of responses on single modality trials involving those target locations, suggesting a Bayesian process incorporating prior knowledge of spatial uncertainty; and b) a distinct “same source vs. different sources” judgment. The results from this task are broadly consistent with prior work in humans involving button presses to report spatial locations of visual and auditory stimuli, and offer advantages for future neurophysiological studies both due to the demonstration of successful performance in monkeys and due to the fact that the task effectively has two readouts in one: an explicit categorical report of whether the participant perceived multiple locations as well as a continuous report of the exact position of the perceived stimuli.

Option B: The sensory scene is sampled by multiple senses, but optimally fusing such signals first requires ascertaining whether particular signals arise from the same or different underlying objects or events. Many prior studies (esp. in animals) have assumed fusion of crossmodal information, whereas more recent work in humans has begun to probe the source inference question. Here we present results from a novel behavioral task in which both monkeys and humans were required to localize visual and auditory stimuli by saccading to each distinct location that they perceived. When visual and auditory locations were widely separated, participants made two saccades, whereas when the two stimuli were presented at the same location, only one saccade was needed. Intermediate levels of separation produced mixed response patterns, i.e. single saccades to an intermediate position on some trials or two saccades to both the visual and auditory locations on others. The distribution of responses was well matched to predictions of a hierarchical model which incorporated a) the spatial distribution of responses on single modality trials involving those target locations, suggesting a Bayesian process incorporating prior knowledge of spatial uncertainty; and b) a distinct “same source vs. different sources” judgment. The results from this task are broadly consistent with prior work in humans involving button presses to report spatial locations of visual and auditory stimuli, and offer advantages for future neurophysiological studies both due to the demonstration of successful performance in monkeys and due to the fact that the task effectively has two readouts in one: an explicit categorical report of whether the participant perceived multiple locations as well as a continuous report of the exact position of the perceived stimuli.

**Intro – We need monkeys to do this kind of task, so we can poke into their brains and see what’s going on**

Information from multiple senses can be combined to improve accuracy of sensory inference, but this is only useful if all combined sensory streams reflect information about the same source.

Perception is inherently multisensory, with multiple discrete sensory streams being processed and combined to produce a coherent representation of the world that most effectively guides behavior. This is particularly obvious in conditions where information from one sensory modality can be used directly to reduce uncertainty about another, such as reading the lips of a speaker to improve language comprehension. However, this integration of sensory modalities is only beneficial if the two sensory streams originate from the same source in the external environment.

Humans solve this problem in a Bayes optimal manner, which has been thoroughly explored in the literature.

Determining whether two noisy observations originate from the same source is a form of causal inference (CI). Recent behavioral research has demonstrated that human observers perform this task in a manner consistent with an ideal Bayesian observer. This is a type of hierarchical model that compares the relative likelihood of the two causal scenarios (same source or different sources), as well as some prior expectation, in order to determine which causal scenario is most likely (figure ##). The resulting inference can then be used to either to make a judgement about the number of discrete sources (explicit causal inference) or influence how the sensory input is perceived (implicit causal inference).

Understanding of this phenomenon at the level of individual neurons has been challenging, because to date only human subjects have been used to look at CI, and we can’t easily record single neuron activity from them.

At the neural level, it has been suggested that this inference may be accomplished in a similarly hierarchical fashion, with lower level sensory cortices representing unisensory information (segregated), other brain regions reflecting combined stimuli (integrated), and finally culminating in complete causal inference at the level of pre-frontal cortex or intraparietal sulcus [refs]. This view of hierarchical neural processing is pleasingly consistent with the hierarchical nature of Bayesian causal inference. However, it is inconsistent with other research that shows significant interaction between modalities even in primary sensory and subcortical regions [refs]. It’s possible that this conflict is due in part to the level of experimentation, with the previously discussed causal inference findings relying on human neuroimaging (fMRI and MEG) which cannot detect multisensory effects occurring at the level of single neurons. This question is most easily answered using animal models, but it is not yet known if animals can be trained to perform the type of causal inference task commonly used in these experiments, or if non-human animals even perform causal inference at all.

Here we create a behavioral paradigm that requires both implicit and explicit causal inference, which can be used in both monkeys and humans, and find that monkeys appear to apply the same type of Bayesian causal inference strategy seen in humans. (validating them as a potential model organism for the study of causal inference at the single neuron and neural circuit level)

We developed a new behavioral paradigm which requires both explicit and implicit causal inference, and can be used in both humans and non-human primates. We find that human subjects perform the task in a Bayes optimal manner, consistent with performance in other similar tasks. Additionally, we find that monkeys can perform this task at similar levels and that their behavior reflects similarly optimal causal inference strategies, with some important differences. Specifically… To our knowledge, this is the first demonstration of multisensory causal inference in a non-human species.

The dual nature of our task also allows collection of sufficient data within a single experimental session, critical for pairing this behavior with electrophysiological experiments in animals. Together these results suggest that non-human primates are a viable model organism for beginning to understand the neural basis of causal inference at the single cell level.

**Results – Monkeys can do it, and seem to be basically doing it like humans with some caveats**

We have created a dual causal inference task, one which requires that subjects report both the number and unique stimuli as well as their position, which relies on both auditory and visual stimulus information. (figure 1, behavioral paradigm description)

Because of this dual task design, subjects are essentially performing two tasks simultaneously: 1 to identify whether it is more likely that there are one or two targets; and one where they must localize the source of both the auditory and visual stimuli. (schematic CI description?)

We have modeled this by adapting several forms of ideal observer models that have been previously applied to human behavioral experiments, including both optimal and heuristic decision rules. Because it is not strictly necessary that subjects use the same causal inference strategy for both localization and unity judgement, we combined models in a factorial manner. (fig 2, model descriptions and predictions)

We compare the predictions of various models fit to both human and monkey subjects, first for the **unity judgement case**, then for the **localization case**, and finally for the **joint condition** (which more exactly represents the task subjects are being asked to perform) figs 3-5

We find that both humans and monkeys are well fit by models of causal inference which have the following features:…

Some differences are apparent between human and monkey subjects, in particular it seems like monkeys have more pronounced bias. These might be associated with learned priors over actual stimulus distributions. We addressed this by implementing alternative priors which better match the task design and find that the fits are slightly better.

**Discussion – We demonstrated that monkeys can do this task, and have some speculation about neural implementation**

Here we demonstrate that monkeys can be trained to perform a multisensory causal inference task, and do so in a manner consistent with human models of Bayesian observers.

In particular, these results are consistent with xxx result from the human literature, and extend that result to apply to macaque monkeys as well.

Some advantages of this dual task design are that it allows for rapid data collection that simultaneously captures implicit and explicit features of causal inference. This is critical for single unit recording as it is practically impossible to combine data across multiple recording sessions and ensure that the same units are being recorded.

This work is important because it will allow us to link our understanding of multisensory processing at the neuronal level with our deepening understanding of causal inference at the behavioral level. For instance, we might make these predictions for what is happening at the neural level…

**Methods**

**Behavioral paradigm**

We created a novelmultisensory task closely related to tasks commonly used in the literature [refs]. This paradigm uses a dual task design, where subjects are reporting both a causal judgement (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 XXX m from a row of speakers and LEDs located on the horizontal plane. Eye movements were monitored via magnetic eye coil (Riverbend) or video eye tracker (Eyelink XXX). While fixating at a central point, subjects were presented with either a light (green LED), sound (white noise), or both at one of 8 visual (+- 6-24 degrees in 6 degree increments) or 4 auditory (+- 6 and 24 degrees) 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 degrees visual paired with -+ 6 or 24 degrees auditory) 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 required to make sequential saccades to each target in any order. The timing of the task was such that subjects must make both saccades in rapid succession, and so cannot adopt a strategy of waiting until the reward is delivered (or not) before making a decision about the second saccade.

**Trial filtration and saccade detection**

Trials are included as long as the subject held fixation through the go cue, and then made at least one saccade, **without enforcing any restrictions on saccade accuracy**. For multi-stimulus trials, trials which ended less than 600 ms after the go cue (the minimum duration for a successfully completed trial, see timing section [xxx]) were also excluded. This was done to minimize the number of trials which ended before the subject’s full response could be reported, and ensure that single saccades were indicative of a unified percept rather than a lapse.

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

**Behavioral modeling**

We implemented a class of causal inference models which is common in human behavioral multisensory research. 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 [refs]. Different models of causal inference will then combine these two estimates according to specific rules, resulting in predictions that can be compared with behavior in our task.

Below we will briefly describe the important components of our models, and refer interested readers to [ref] for a much more thorough treatment of this class of models. 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 judgements about number of targets (unity task) and location of stimulus source(s).

*Location inference*

For all stimuli, internal representations are assumed to be corrupted by Gaussian noise, such that with the term denoting the actual location of the source of the respective stimulus and the term reflecting the sensory standard deviation (a free parameter). Estimates about stimulus locations for a given causal structure (c=1, common cause, eq 1; c=2, independent causes, eq 2) and internal representation can be computed via Bayes rule:

(1)

(2)

Where for the c=1 case the source 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,

(3)

which for the two cause case becomes,

(4)

where is the mean of the prior (often taken as 0) and is the prior standard deviation. This prior induces a compressive bias which is compatible with many psychophysical results. This prior has an additional advantage in that, because the prior and sensory likelihoods are Gaussian, the posterior is also Gaussian and can be computed analytically.

An alternative is that the subject may learn the actual stimulus distribution, rather than relying on this type of naïve compressive prior. This is particularly relevant for the monkey subjects who, unlike humans, will experience tens of thousands of trials over the course of training and experimentation. For this task we modeled the prior as either a uniform distribution over the exact target locations (perfectly empirical prior) or as a mixture of normals which grossly captures the task design of targets being concentrated either to the left or the right of fixation (heuristic empirical prior).

*Causal inference strategies*

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 either Bayesian principles, non-Bayesian heuristics, or strategies which do not actually implement causal inference at all (for example, 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. Instead we present **two(?)** forms of Bayesian causal inference common to the field, and contrast these with a null model which does not perform causal inference at all. This might need work, not really sure what models I’m talking about comparing it to.

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

(5)

Where p(c) reflects the prior probability of a common cause , which is left as a free parameter. Because there are only two possibilities for causal state in this paradigm, this can be written as

(6)

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

(7)

(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).

*Decision rule*

The final component of the observer model is the decision rule used to choose responses based on the above probability distributions. There are three common choices for this decision rule: model selection, probability matching, and model averaging (for the unity judgement task, only model selection and probability matching are relevant). Below we give an overview of these decision strategies and how they relate to subject responses in the unity judgement and localization components of the task.

For the unity judgement task, model selection corresponds with reporting whichever causal scenario has the highest posterior probability,

(9)

where represents the lapse rate (where the subject randomly makes a response), and is the Iverson bracket which is 1 when the statement inside is true and 0 otherwise.

For the probability matching case, the observer randomly chooses unity at a rate consistent with the posterior probability of the causal inference. In this case the probability of reporting unity is exactly equal to the posterior probability , except for the additional lapse parameter described above.

For the localization component of the task, subjects must arbitrate between the two potential models of location conditioned on number of causes (eqs. 3 and 4). This amounts to reweighting the two estimates according to some weight function that is dependent on the sensory response,

(10)

where defines the decision weight applied to the c=1 condition.

For the model selection strategy, this 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. A similar rule applies for the probability matching scenario, except that the weights are randomly set to 1 with probability and 0 otherwise. For model averaging, the weight is equivalent to the posterior probability , such that the expression in eq. 10 becomes equivalent to the optimal observer model commonly used in the literature [ref][may need to revisit this section for clarity].

*Comparing with behavioral data*

The above response estimates are dependent on internal variables and which are not accessible to the experimenter. To get distributions that can be compared with data, eq. 10 must be marginalized across the internal variables,

(11)

We compute this distribution using numerical integration for each of the 20 combinations of visual and auditory targets.

**Model fitting**

Models are fit using a maximum likelihood approach to determine the set of parameters which best explains the provided data. This is accomplished using [approach][ref]. Models are validated using a k-fold cross validation strategy, holding out 1/5th of trials for testing and training on the other 4/5th of the data. Total model likelihood used for model comparison is computed by summing likelihoods across all 5 folds of test datasets.