Review

Causal inference in perception

Ladan Shams[1] and Ulrik R. Beierholm[2]

[1] Department of Psychology, University of California, Los Angeles, CA, 90095-1563, USA
[2] Gatsby Computational Neuroscience Unit, University College London, Alexandra House, 17 Queen Square, London, WC1N 3AR, UK

Until recently, the question of how the brain performs causal inference has been studied primarily in the context of cognitive reasoning. However, this problem is at least equally crucial in perceptual processing. At any given moment, the perceptual system receives multiple sensory signals within and across modalities and, for example, has to determine the source of each of these signals. Recently, a growing number of studies from various fields of cognitive science have started to address this question and have converged to very similar computational models. Therefore, it seems that a common computational strategy, which is highly consistent with a normative model of causal inference, is exploited by the perceptual system in a variety of domains.

Traditional contexts for studying causal inference

The process of inferring whether or not an event A is caused by another event B is often referred to as causal inference [1]. Causal inference has been studied within fields as diverse as philosophy [2], statistics [1,3], machine learning [4–9] and psychology [10–12]. However, the question of how humans perform causal inference has been traditionally studied in the context of reasoning and cognition. For example, it has been extensively studied how the causal relationships between diseases and the factors that cause them are inferred by human subjects [11,12]. Here we argue that causal inference is also an important problem in perceptual processing. The perceptual system has to solve the problem of causal inference at any given moment in one or more processes. In addition, correct causal inference is arguably crucial to the survival of the organism as incorrect inference can for example lead to the misidentification and mislocalization of a predator or a prey, as described below.

Defining causal inference

Every system that makes an estimate about unobserved variables based on observed variables performs inference. For example, in all three models depicted in Figure 1 one can perform inference about hidden variables (white nodes) based on observed variables (blue nodes). The model in panel (a) is an example of systems in which inference about a variable s can be performed using two or more observations (x_1 and x_2). The inference process in this scenario amounts to cue integration [13]. For example, if s is the direction of a sound source, it can give rise to two cues: interaural level difference (x_1) and interaural temporal difference (x_2). In this type of scenario, there is only one cause for the observations, and the goal of inference is to quantify the value of the cause based on the two observations. Therefore, this process does not determine whether or not x_1 and x_2 were caused by s (i.e. causal inference) but what the parameters of the cause are. The model in panel (b) is an example of systems in which two or more sources can influence the value of an observed variable [14,15]. In these systems, inference can be made about unobserved variables (e.g. s_1: reflectivity of a surface, and s_2: illuminant) using measurement on observed variables (e.g. x: lightness of a surface). Although there are now two sources/causes that influence x, the inference process still does not determine whether s_1 or s_2 caused x but to what degree each contributed to the value of x. In contrast to (a) and (b), the system depicted in panel (c) makes inference about whether s_1 or s_2 caused x. In this scenario, there are two qualitatively different and mutually exclusive causes (e.g. dark surface vs. shadow) possible for the observed variable x (e.g. dark image region), and the inference process chooses between the two. We refer to this process as 'causal inference.' We, henceforth, refer to such inference problems that involve choosing between distinct and mutually exclusive causal structures as causal inference, and focus on studies of this form of inference.

Glossary

Bayes' rule: Specifies that the probability of variable s (having observed x) is the normalized product of the likelihood and prior: $p(s|x) = p(x|s)\,p(s)/\,p(x)$. Intuitively, this can be interpreted as follows: one's current knowledge about an unknown variable, s, is a combination of one's previous knowledge and one's new data x.
Bayesian inference: Statistical inference of an unknown variable using Bayes' rule.
Graphical model: A graph-based representation of the statistical relationship between random variables [48]. For an example see Figure 1a. Each node represents a random variable and each arrow represents the statistical dependency of one variable on another.
Heavy-tailed distributions: There are different definitions for heavy-tailed distributions. In line with the literature reviewed here, we define heavy-tailed distribution as a probability distribution whose tails are fatter than those of a normal distribution. Heavy-tailed distributions therefore include Laplacian, mixture of two Gaussians, and Student t distributions.
Likelihood: A conditional probability $p(x|s)$ describing the probability of obtaining a certain observation x, provided a given event occurs or a source s exists. It specifies knowledge about how the data are generated.
Maximum likelihood estimation: For a given dataset, and assuming a probability distribution model, the maximum likelihood estimation chooses parameter values that maximize the likelihood of obtaining the dataset given the model.
Posterior: The probability of a random scene/source/event s, after taking into account the observation x. Posterior probability is also a conditional probability ($p(s|x)$) and is derived using Bayes' rule.
Prior: A probability distribution, $p(s)$, representing the prior knowledge or expectation about the hidden variable s.

Corresponding author: Shams, L. (ladan@psych.ucla.edu).

1364-6613/$ – see front matter doi:10.1016/j.tics.2010.07.001 Trends in Cognitive Sciences 14 (2010) 425–432

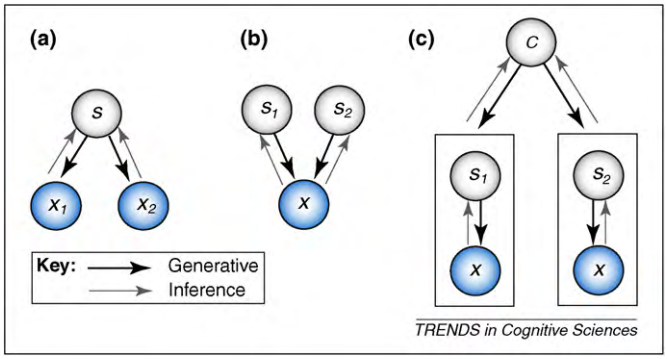

Figure 1. Three graphical models (see Glossary) corresponding to three different classes of inference. Observed and hidden variables are denoted by blue and white nodes, respectively. The black arrows represent the direction of generative process, and white arrows represent the direction of inference. **(a)** An example of a generative model where one source (s) can give rise to two or more observed variables (x_1 and x_2). For example, if s is the direction of a sound source, it can give rise to two cues: interaural level difference (x_1) and interaural temporal difference (x_2). The inference would involve combining these two cues to estimate the direction of sound, i.e. a parameter of the cause not the identity of the cause. **(b)** An example of a generative model in which there are two (s_1 and s_2) or more causes for a given observed variable (x). For example, the lightness of a surface (x) is affected by both reflectivity of the surface (s_1) and the intensity of the light source (s_2). Although there are two 'causes' for the lightness of the surface, the inference process does not address whether one or the other gave rise to lightness but to what degree each of them contributed to the lightness. **(c)** An example of a generative model that includes two or more 'models' or causal structures or hypotheses (represented by the boxes). In such systems, the inference process determines the probability of each causal structure. This process is sometimes referred to as structural inference or model inference. For example, if x represents a dark image region, it is either caused by a dark surface (s_1) or by a shadow (s_2). Hidden variable C determines which scenario/model gave rise to the dark image region. In this case, the inference process determines *whether* x was caused by s_1 or s_2. We refer to this type of inference as *causal inference*.

Causal inference as a core problem in perception

Many of the problems that the human perceptual system has to solve almost continuously involve causal inference. A clear example of causal inference in perception is auditory scene analysis. In nature, there are almost always multiple sounds that reach our ears at any given moment. For example, while walking in the forest, we might simultaneously hear the chirps of a bird, the sound of a stream of water, our own footsteps, the wind blowing and so on. The ears receive a complex wave that is the sum of multiple complex sound waves from different sources. To make sense of the events and objects in the environment, the auditory system has to infer which sound components were caused by the same source and should be combined and which components were caused by different sources and should be segregated.

Similarly, many basic perceptual functions in vision involve causal inference. An obvious example here is grouping (in the service of object formation) and segmentation. Given the ubiquity of occlusion and other kinds of noise in visual scenes and retinal images, visual information is always fragmented and noisy, and the visual system has to determine which parts of the retinal image correspond to the same object and should be grouped together and which parts correspond to different objects and should not be grouped together. In addition to the problem of perceptual organization, the visual system has to solve another type of problem that involves causal inference. To determine how to interpret some cues, the system has to infer which kind/class of source gave rise to the cue. For example, to interpret a motion cue for structure, the nervous system has to determine whether the object is rigid or nonrigid; to interpret the texture cue for slant it needs to infer whether the texture is isotropic or homogeneous and so on [16].

These kinds of inference problems exist in all sensory modalities. However, in addition to these problems within each modality, there is a prominent problem of causal inference across sensory modalities. At any given moment, an individual typically receives multiple sensory signals from the different modalities and needs to determine which of these signals originated from the same object and should be combined and which originated from different objects and should not be bound together. For example, when walking in the forest, the nervous system has to infer that the odor of dampness, the sound, and the image of the stream correspond to the same object (stream), the odor of a flower and the image of the flower correspond to the same object, the sound of chirp is caused by a bird, and not any other possible combination of signals and sources. This problem can be challenging even in the simplest scenario with only one visual stimulus and one auditory stimulus. For example, when hearing the roar of a lion while seeing a furry object in the foliage, the perceptual system needs to infer whether these signals are caused by the same animal, or different animals (a second occluded lion could be making the roar). We will return to this simple scenario of one visual stimulus (x_V) and one auditory stimulus (x_A) in our discussion of models of causal inference later.

The problem of perceptual organization has been studied extensively within each sensory modality, although not explicitly in the framework of causal inference. The problem of multisensory causal inference has not been studied until recently, but there has been a surge of interest in this question over the past few years. Below, we discuss some recent models of human perception in these domains. We start with a discussion of multisensory processing because this area provides perhaps the most unequivocal examples of causal inference in perception.

Causal inference in multisensory perception

Until recently, models of multisensory perception, and cue combination at large, all assumed that different signals are all caused by the same source [17–19], and they modeled how the nervous system would combine the different cues for estimating a physical property under this condition of assumed unity (single cause, C = 1). Therefore, these models did not consider the general condition in which multiple sensory signals (e.g. auditory or visual signals) can have multiple causes (e.g. multiple lions). For simplicity, we will focus our discussion henceforth on a situation with two sensory signals, x_V and x_A (Figure 1b). Although the majority of previous models (e.g. [20,21]) have used maximum likelihood estimation (see Glossary) to model cue integration, we frame the discussion in terms of Bayesian decision theory [5,13,22] because it is a more general framework, not disregarding priors (see Glossary) and decision strategy (Box 1). Under the assumption of a common cause, the estimate of the source $\hat{s}_{C=1}$ is obtained by a weighted average of the estimate of each modality and the prior, provided that the signals are normally distributed. (Maximum likelihood estimation disregards the prior

Box 1. Decision-making strategy

An important component of perceptual processing is the decision strategy. The same sensory abilities in the same individual can result in very different patterns of behavior in different tasks or even in the same task but under different pay-off conditions. The reason for this is that the sensory estimates are chosen so as to maximize a certain utility, for example to help with a successful grasp of an object, or to identify correctly who is speaking, or to determine correctly from where the voice is coming regardless of the speaker's identity and so on. Bayesian decision theory provides a normative framework for how optimal decisions would be chosen for an observer who uses previously acquired knowledge about the environment (i.e. prior). Bayesian inference provides an inferred distribution of the possible values over a hidden variable s (i.e. the posterior distribution; see Glossary). Which decision to make or action to take strongly depends on the task at hand. The best decision is one that maximizes the utility or equivalently, minimizes the cost/loss $L(s)$ for the given task. Typical loss functions in modeling human behavior include the squared error cost function $L(\hat{s}) = (\hat{s} - s)^2$ or the 'all or nothing' cost function $L(\hat{s}) = 1 - \delta(\hat{s} - s)$. Choosing the mean and the max of the posterior distribution of s minimizes these loss functions, respectively.

When dealing with a hierarchical model such as the HCI model, further options become possible, for example does the cost function also become dependent on the causal structure C, i.e. $L = L(s,C)$ as opposed to $L(s)$? In the version of HCI proposed by Körding *et al.* [28], C is an irrelevant variable to the observer (i.e. $L = L(s)$) and thus it is marginalized. This strategy is known as *model averaging*, $L(s) = \sum_C L(s, C)$ (see Figure 2c caption). However, if subjects are explicitly asked to report their perceived causal structure (e.g. was there 1 or 2 sources, see [49]), or for whatever reason the correct decision about causal inference is important in and of itself, then the correct choice of causal structure can enter into the utility function, i.e. $L = L(s,C)$. This will result in a strategy that would choose the most likely causal structure, and choose the estimate of sources strictly according to the most likely structure C. This strategy is known as *model selection* ($c = \text{argmin}(L(C, s))$).

resulting in a weighted average of the two sensory estimates.)

A model that did allow independent sources as well as a common source for two sensory signals was proposed by Shams *et al.* [23]. This Bayesian model accounted for auditory–visual interactions ranging from fusion to partial integration and segregation in a numerosity judgment task. In this model, two sources, s_A and s_V, were allowed, and the joint prior probability of the two sources (i.e. $p(s_A, s_V)$) captured the interaction/binding between the two modalities, and resulted in the full spectrum of interactions. Similar models that did not assume a single cause and used joint priors to characterize the interaction between two modalities were later shown to account for other perceptual tasks. Bresciani *et al.* [24] and Rowland *et al.* [25] used a Gaussian ridge (in contrast to the nonparametric joint prior used by Shams *et al.* [23]) to capture the binding between two modalities accounting for auditory-haptic interactions in a numerosity judgment task and in a cats' spatial localization task, respectively. Roach *et al.* used a mixture of two parametric components in their joint prior to capture the interactions between hearing and vision in a rate judgment task [26]. A three-dimensional Gaussian prior capturing the interaction among three modalities was used by Wozny *et al.* to account for auditory–visual–tactile numerosity judgments [27].

Körding *et al.* [28] demonstrated that a hierarchical Bayesian model that explicitly performs causal inference accounts well for the auditory–visual spatial judgments of observers. By showing that priors are encoded independently of the likelihoods in this task, Beierholm *et al.* [29] provided further evidence that the human nervous system could indeed be performing Bayesian causal inference in this task (see [22] for a discussion of tests of Bayesian inference as a model of perceptual process). This model has also been shown to explain within-modality and cross-modality oddity detection in a unified fashion [30]. The same causal inference framework has also recently been used to make predictions about the optimal time window of auditory–visual integration [31]. Patterns of human motor adaptation have also been shown to be consistent with predictions of this type of causal inference model (Box 2).

This model, which we will henceforth refer to as Hierarchical Causal Inference (HCI) model (Figure 1c), performs causal inference explicitly by computing the probability of each possible causal structure (single cause vs. independent causes). As a simple example of multisensory perception, let us assume that there is a visual signal x_V (e.g. neural representation of the image of a lion) and an auditory signal x_A (e.g. neural representation of the roar of a lion) being processed by the nervous system. These two signals could have been caused by the same object, $s = s_A = s_V$ (e.g. lion 1) or they might have been caused by two independent objects, s_V and s_A (e.g. lion 1 and lion 2). The probability of these two signals having a common cause ($C = 1$) can be calculated using Bayes' Rule

Box 2. Causal inference in action

Causal inference is not confined to cognitive and perceptual processing; it seems to also play an important role in processes relevant to action. For example, in deciding to correct motor errors, the motor system has to first determine the source of the error (e.g. the deviation from the target in reaching). In many motor tasks, the errors can stem from a variety of sources. Some of these sources are related to the nervous system whereas others are not. The motor system should correct for all errors that are due to the motor system itself, but ignore unstructured errors that are due to, for example stochasticity in the environment. In a recent study by Wei and Körding [50], the participants received visual feedback after reaching for a target in a virtual environment. The magnitude of the error conveyed by feedback was manipulated from trial to trial. They found a nonlinear relation between the magnitude of error and the degree of subsequent correction in motor behavior observed, with the largest correction occurring for medium error range. This nonlinear behavior was explained well by a model similar to the HCI model (see text) that tries to infer the source of error. If the error is large (relative to the variability in motor behavior) then it does not get attributed to the motor system, and is instead attributed to an external source (such as an experimenter); if the error is small, then the error is attributed to the motor system but the degree of necessary correction is small accordingly. The largest correction occurs for the largest size error that can still be attributed to the motor system. This behavior is highly similar to the nonlinear relationship between cue interaction and degree of discrepancy between the signals as discussed in the text (also see Figure 3a, right). In a related study, Berniker and Körding [51] examined motor adaptation using a more general model that explicitly assumed a changing environment and a motor system that undergoes physiological changes (such as fatigue etc.). In such a paradigm, the nervous system has to infer which errors are due to the changing properties of the world and which are due to changes in the motor system over time. A similar causal inference model was shown to account well for the behavior of subjects in several studies in which participants' movements are perturbed through external means.

(see Glossary), and depends essentially on how similar the sensations x_V and x_A are to each other, and the prior expectation of a common cause. It has been recently shown that altering the prior expectation of a common cause (i.e. $p(C = 1)$) can strongly affect the degree of integration of two signals [32]. If the goal of the nervous system is to minimize a mean squared error cost function of the visual or auditory estimates (Box 1), then the optimal estimate of the source(s) transpires to be a weighted average of the optimal estimate for the scenario of common cause, and the optimal estimate for the scenario of independent causes. This is a surprisingly intuitive result: When the causal structure of the stimuli is not known (i.e. in general), the physical property of interest (in this example, the location of stimuli) is estimated by taking into account the estimates of both causal structures, each weighted by their respective probability (Figure 1c).

Model inference in unisensory perception

Yuille, Bülthoff and colleagues [13,33] were the first to point out that the perceptual system is faced with a model inference problem. They made the observation that the visual system makes specific assumptions in relation to subclasses of objects/surfaces in interpreting many visual cues. For example, the motion cue for extracting structure is interpreted differently depending on whether the object is rigid or nonrigid, and the shading cue for shape is interpreted differently depending on whether the object has a lambertian or specular reflectance. Yuille and colleagues proposed that different assumptions about objects compete in finding the best explanation of the retinal image [13,33]. They showed that competitive priors in a Bayesian Decision Theory framework can characterize this process well. In this competitive prior model, the prior distribution is a mixture of different priors, each corresponding to a subclass of the object property relevant to the interpretation of a cue (e.g. rigid vs. nonrigid for motion cue; or lambertian vs. specular for shading cue, etc.). Estimating the object property would involve selecting the model (e.g. lambertian vs. specular) that would best explain the sensory data. This is the case even if the task of the observer is only to estimate the property of interest (shape or depth, etc.) and not to explicitly estimate the object class (lambertian vs. specular).

Recently, Knill has developed a model that achieves roust integration using model inference [34]. Robust integration refers to integration of cues in a manner in which a cue that is deemed to be unreliable/irrelevant does not get combined with other cues for the estimation of an environmental variable [35]. As with Yuille's model, this model is a Bayesian mixture model [16,36]. In one study, Knill investigated slant perception from figure compression cues and binocular disparity cues [34]. The figure compression cue for the type of stimuli used in the experiment (ellipses) is only informative if the object is a circle. Therefore, the weight of the compression (aspect ratio) cue for slant estimation depends on whether the object is a circle or an ellipse. The observers use the compression cue primarily when it is not in large conflict with the stereoscopic cue. When the conflict between the two estimates is large, the objects seem to be interpreted as ellipses and the compression cue is down-weighted and the stereoscopic cue dominates. Knill [34] showed that this and other patterns of behavior in this task are captured by a mixture model that assumes objects are either circular or ellipsoid. This Bayesian model assigns a greater weight to the model with the higher probability (which depends on the prior probability of the two models, as well as the sensory evidence for the model, as determined by the consistency with the other cue). Knill also showed that the human perception of planar orientation from texture cue is qualitatively consistent with this type of mixture model [16]. A further development in this line of research is to expand on the possible relations between the sources. For example, in vision, occlusion specifies that one object has a depth larger than another, a relation that can be inferred with a variant of HCI [37].

Model inference has recently been shown to account for motion direction perception in multiple tasks [38]. Observers' responses in a coarse discrimination task (left or right of a reference) as well as a subsequent response on a motion direction identification task were predicted well by a Bayesian model that computed the probability of each hypothesis (left vs. right) and estimated the direction of motion based on the more likely hypothesis.

The types of model inference discussed above can be thought of as causal inference, if we consider these mutually exclusive classes of objects (rigid vs. nonrigid, circle vs. ellipse, leftward vs. rightward movement) as causes for the sensory signal. Moreover, as we will discuss in a following section, this type of model inference is closely related computationally to the types of causal inference performed by the models of multisensory perception we discussed earlier.

Heavy-tailed distributions and causal inference

Another class of models that are computationally highly similar to the causal inference models discussed above is models utilizing heavy-tailed distributions (see Glossary). Some recent studies have suggested that heavy-tailed distributions explain human behavior in certain perceptual tasks better than commonly assumed Gaussian distributions. In a study of human motion perception, Stocker and Simoncelli [39] estimated the shape of a prior distribution from observers' judgments of speed, and found that the distribution had a tail that was heavier than that of a Gaussian distribution. More recently, Lu *et al.* [40] have shown that heavy-tailed distributions for slowness and smoothness priors can explain observers' perception of direction of coherent motion better than Gaussian priors. A heavy-tailed prior for motion allows the perceptual system to effectively ignore the prior if the visual data are highly discordant with the prior bias for slowness and smoothness of motion. Heavy-tailed likelihood or prior distributions have also been shown to be able to account for robust integration of visual cues [34,41,42] or postural cues [43].

What all of these tasks have in common is that they all involve a causal inference problem. For example, in the motion perception task, the perceptual system has to decide whether the two (or more) patches of local motion belong to the same object, and thus should be subjected to

Box 3. Causal inference in animal navigation

Studies of navigation in animals have indicated that most animals can exploit multiple sources of information (cues) for navigation. One source of information is called *path integration*, and refers to the ability to keep track of distance and direction of path traversed. Another cue that many animals are able to exploit is landmarks. It has been shown that rats, hamsters, honeybees and spiders are able to use both of these cues for navigation [52], and when the discrepancy between the two cues is small, a bigger weight is given to the landmark cue than path integration [53–55]. However, when the conflict between the two cues is large (90 or 180 degrees), the landmark cue seems to be ignored, and rats and hamsters seem to rely entirely on the path integration cue [54,56]. The dominance of the landmark cue in low-conflict conditions suggests that it is more reliable. Therefore, it seems odd that in conditions of large conflict the navigation system would switch to the less reliable cue and veto the more reliable cue. It has been proposed [52] that a sparse landmark cue is precise (and hence reliable), however ambiguous, because some landmarks (such as trees) are not unique, or some landmarks can move. To the contrary, the path integration cue is not very precise (and hence not very reliable), however, neither is it ambiguous. The navigation system of these animals seems to be involved in a process of causal inference, deciding whether the landmark cue corresponds to the target or to another location. If the landmark cue is consistent with another cue (path integration), it pushes the inference in favor of a common cause, and then the two cues get integrated (and the higher reliability of the landmark cue would result in a higher weight for it in the combination). If, however, the landmark cue is in large conflict with the path integration cue, it is inferred to correspond to another location, and does not get integrated with the path integration cue. In other words, the mixed prior model (either the same target or different targets) together with the narrow landmark likelihoods and the broad path integration likelihoods would predict exactly the kind of behavior that is observed in these animals. Note that there is a very strong parallel between this scheme of cue combination and the human multisensory integration [28] and robust integration in vision [34] described in the text.

the constraint of smoothness prior, or whether they belong to two different objects and thus can have completely different speeds and directions. In the cue combination tasks, the nervous system has to determine whether the two cues originate from the same object, and thus should be integrated, or whether they correspond to different objects and should not be integrated. The heavy-tailed prior or likelihood distributions in effect serve as mixture models allowing two different regimes, a unity regime represented by the narrow peak, and the independence regime represented by the wide tails. Therefore, models utilizing heavy-tailed distributions can be considered to be implicitly performing causal inference (Box 3).

Comparing different models

We now briefly discuss how the various models discussed so far relate to each other to examine computational similarities and differences across these models. The HCI model and Knill's mixture model are mathematically almost equivalent (Figure 3a). The main conceptual difference between the two models is that the HCI model infers whether the two sensory signals are caused by one or two objects, whereas in Knill's model it is assumed that the two sensory signals are caused by a single source, and an inference is made about the *class* (circle vs. ellipse) of the object. Hospedales *et al.*'s model [30] is the same as HCI.

Yuille *et al.*'s competitive prior model [13,33] and Stocker and Simoncelli's [38] compute the probability of each hypothesis the same way as the models mentioned above. The main difference between Yuille *et al.*'s and Stocker and Simoncelli's models versus Knill's and the HCI models is in the loss function that results in a model selection scheme in the former as opposed to model averaging in the latter (Box 1).

HCI [28] is a hierarchical Bayesian model. By integrating out the variable C (Figure 2c), the HCI model can be recast as a special form of the nonhierarchical model of Shams *et al.* [23] ($p(s_A, s_V|x_A, x_V) = \frac{p(x_A|s_A)\,p(x_V|s_V)\,p(s_A,s_V)}{p(x_A,x_V)}$) where the prior over sources takes the form $p(s_A, s_V) = p(C = 1)p(s) + p(C = 2)p(s_A)p(s_V)$. This formulation makes the HCI model easy to compare with several other models that, although not explicitly formulated to study causal inference, are computationally very similar.

Although more restricted than Shams *et al.*'s [23] model, HCI has the advantage that it is more parsimonious (fewer parameters) and allows for a model selection strategy [42] (Box 1). Of course, if C is not marginalized, then HCI can in

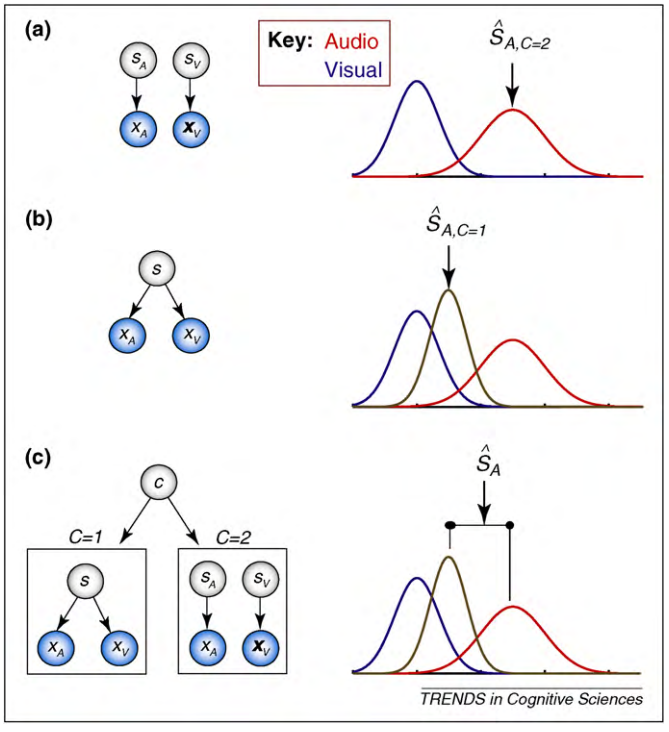

Figure 2. Different causal structures and their respective optimal estimates of location of an auditory source. The red curve represents auditory likelihood distribution $p(x_A|s_A)$ and the blue curve represents the visual likelihood $p(x_V|s_V)$. Here, for the sake of simplicity, we assume that the prior is uniform (i.e. uninformative), and both auditory and visual likelihoods have a normal distribution. **(a)** If the audio and visual signals x_A and x_V are assumed to have been caused by separate sources ($C = 2$), the signals should be kept separate. Thus, the best estimate of location of sound is only based on the auditory signal. The optimal estimate of location of sound $\hat{s}_{A,C=2}$ is therefore the mean of the auditory likelihood. **(b)** If the two signals are assumed to have been caused by the same object ($C = 1$), the two sensory signals should be combined by multiplying the two likelihoods that results in the brown distribution. Therefore, the optimal estimate of the location of sound $\hat{s}_{A,C=1}$ is the mean of this combination distribution. **(c)** In general, the observer does not know the causal structure of events in the environment, and only has access to the sensory signals; therefore there is uncertainty about the cause of the signals. In this case, the optimal estimate of the location of sound, $\hat{s}_A$, is a weighted average of the estimates corresponding to the two causal structures, the mean of the brown and the mean of the red distributions, each weighted by their respective probability $\hat{s}_A = p(C = 1|x_A, x_V)\hat{s}_{A,C=1} + p(C = 2|x_A, x_V)\hat{s}_{A,C=2}$.

addition make direct predictions about the judgments of causal structure.

In the models of multisensory integration presented in Roach *et al*. [26] and Sato *et al*. [44] (Fig. 3b), the choice of causal structures is not between a single source or independent sources, but instead between *correlated* sources and independent sources. In heavy-tailed models (Figure 3d,e), the heavy-tailed distribution can be thought of as a mixture of two distributions, a narrow one and a wide one, very similar to the prior distributions shown in Figure 3a,b. In Stocker and Simoncelli [39], Lu *et al*. [40] and Dokka *et al*. [43] it is the prior that is heavy tailed, whereas in Natarajan *et al*. [42] and Girshick and Banks [41] it is the likelihood function that is heavy-tailed. However, the effect on the posterior function and response behavior is the same as can be seen in Figure 3d,e. Therefore, all of these models represented in Figure 3a,b,d,e are computationally very similar and produce very similar quantitative predictions for the estimate of sources as seen in the last column.

It should be noted that all of these models (Figure 3a,b,d,e) involve 'complex' priors (referring to mixture of priors or unstructured priors). This feature distinguishes these models from models that use a single parametric (typically Gaussian) prior distribution. For example, Bresciani *et al*. [24] used a Gaussian distribution (Figure 3c) to model the difference between causes ($p(s_A, s_V) = N(s_A - s_V, \sigma)$), and Wozny *et al*. [27] used a multivariate Gaussian distribution as prior for three sources ($p(s_A, s_V, s_T)$). Because the product of Gaussians is itself a

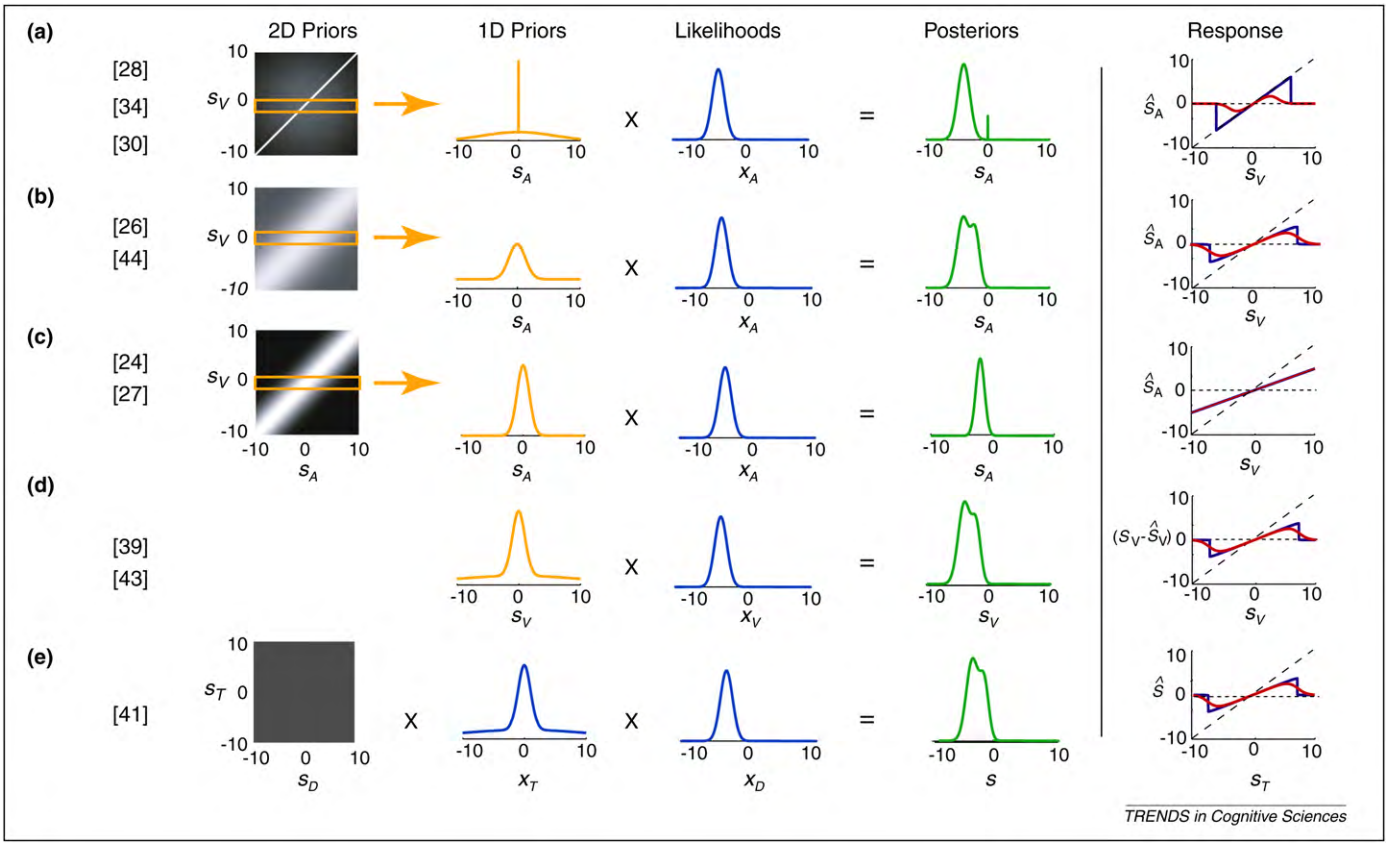

Figure 3. Comparison of different models. The prior distributions for different models are shown in the leftmost two columns. For simplicity, for models with a 2-dimensional prior, we only focus on one dimension (e.g. location of sound), and show the one dimensional prior for a specific value of the second dimension (e.g. $s_V = 0$), i.e. a slice from the two dimensional prior. This is shown in the second column. Likelihood functions are shown in the third column. Multiplying the priors and likelihoods creates the posterior distribution (e.g. $p(s_A|x_A, x_V)$) shown in the fourth column. A response (e.g. $\hat{s}_A$) is generated by taking the arg-max or mean of the posterior distribution. The relationship between the physical source (e.g. s_V) and the perceptual estimate (e.g. $\hat{s}_A$) is shown in the rightmost column. Red line represents model-averaging for each model, blue is model selection (Box 1), dotted line is the response for the independent model (C = 2), dashed line for the full fusion (C = 1). All models, except for those in **(c)**, exhibit a nonlinear response behavior (both red and blue curves): the sources of information get combined so long as the discrepancy between them is not large. In this regime, larger discrepancies result in a larger interaction. However, once the discrepancy becomes large, the sources no longer get combined, and the interaction goes back to zero. The models listed for each row are qualitatively similar, although small differences can exist. For example, Knill [34] uses a log-gaussian (as opposed to Gaussian) as one element of the mixed prior. Wozny *et al*. [27] use a three-dimensional Gaussian. **(a)** These are mixture models with the prior distribution composed of two components. The spike represents one hypothesis (common cause in [28] and [30]; and circle hypothesis in [34]) and the isotopic Gaussian represent another hypothesis (e.g. independent causes or ellipse shape; see text for explanation). The rightmost panel represents the response $\hat{s}_A$ (e.g. perceived auditory location or the perceived slant) for $s_A = 0$ (representing a specific location of sound or the slant as conveyed by binocular disparity) as a function of the other source, s_V (e.g. location of visual stimulus or the compression cue). As s_V gets farther from s_A (i.e. zero) the response gets farther from s_A, however when the discrepancy between s_V and s_A get large, the response goes back to s_A again (in effect ignoring s_V). **(b)** The prior is composed of two components, a Gaussian and a uniform distribution. The response behavior is qualitatively the same as that of models in **(a)**. **(c)** The prior is a Gaussian ridge, together with the Gaussian likelihood, this results in a Gaussian posterior, and a linear response behavior. **(d)** The prior is a heavy-tailed distribution. This results in a nonlinear response behavior similar to those of **(a)** and **(b)**. The prior has a mode at zero speed. The rightmost panel shows response (perceived speed) as a function of true visual speed. As the visual sensation gets farther from the expected speed (zero), the error first increases, and then it decreases as the discrepancy between the two gets too large. **(e)** In this model, the prior is uniform, however one of the likelihood functions is a heavy-tailed distribution, shown in this case in the second column. This likelihood for the texture cue is combined with the Gaussian likelihood of another cue (binocular disparity) shown in the third column. The rightmost panel shows perceived slant as a function of texture signal, for a disparity cue consistent with zero slant. Comparison of the posterior and response in this model with those of **(a)**, **(b)** and **(d)** shows that a heavy-tailed prior and heavy-tailed likelihood have the same effect on the posterior and thus, the response behavior.

Gaussian, such a scheme leads to a linear combination of cues and does not predict complete segregation of signals (Figure 3c). Although in experimental settings a complete segregation might not occur frequently (these models have accounted for data very well), in real life with multiple signals potentially caused by entirely disparate sources the complete segregation of signals (i.e. no interaction whatsoever) might occur more frequently.

Concluding remarks

As described earlier, there are several studies from a diverse number of cognitive science subfields including vision science [13,33,34,38], multisensory integration [23,24,26–30], robotics and machine learning [44,45] that have proposed very similar computational solutions to the problem of causal inference. These theories have also been tested against human behavior showing that the human perceptual system seems to be endowed with mechanisms that can perform causal inference in a fashion consistent with normative models [23,24,26–30,38]. Another noteworthy aspect of some of these studies is that a common computational strategy seems to be used in different perceptual domains. For example, Knill [34] and Körding *et al.* [28] have independently developed two normative models to account for perceptual phenomena. These models are essentially computationally equivalent, and they account very well for two very different perceptual functions: perception of slant of a surface based on inference about the shape of the object (type of cause), and perception of space based on inference about causal structure of the sources of multisensory stimuli.

The computational strategies used for solving the causal inference problem in perception could have evolutionary roots. In Box 3 we discuss a problem in animal (rats, hamsters and bees) navigation that remarkably parallels some problems of causal inference that we discussed here. A computational strategy similar to the normative solutions discussed here in the context of human perception seems to be employed by these animals in solving this problem. On the other hand, a modeling study has recently shown that causal inference in multisensory perception can arise through reinforcement learning from interactions with the environment [46]. Indeed, this kind of simple reward-based learning is a mechanism that is shared across species; therefore, this computational strategy could also develop as a result of learning. It has been proposed that norepinephrine mediates the inference about causal structure uncertainty (the identity of a cause) in the brain [47]. However, further theoretical and experimental research is needed to unravel the neural mechanisms of causal inference see also Box 4.

We believe that the fact that different research groups from different fields of cognitive science have converged on similar computational models for a diverse set of perceptual functions is remarkable, and speaks to the importance of the causal inference framework for understanding perception.

Box 4. Questions for future research

- How would causal inference work in more complex/realistic conditions in which the perceptual system has to choose from many possible causal structures as opposed to two or three? Would optimal performance become computationally too expensive to achieve by the nervous system? Are there heuristics or constraints that the nervous system can utilize to achieve optimal inference in such conditions?
- How do subjective priors (the *a priori* biases that an observer exhibits in their responses) compare to objective priors (reflecting the statistics of the environment)? Are they typically consistent, that is are subjective priors always ecologically valid?
- What is the loss function employed in perceptual causal inference? Is it consistent with the loss function(s) used in cognitive tasks? Are different loss functions used for different tasks and conditions and if so, why?
- Does unisensory causal inference take place before multisensory causal inference or vice versa, or do all perceptual causal inferences occur in parallel?
- How context-dependent is perceptual causal inference? How easy is it to learn or modify the causal inference process for new contexts and environments?
- Does causal inference occur explicitly or implicitly in perception in natural settings (when observers are not probed to report their perceived causal structure)? Does the nervous system commit to a certain causal structure at any stage of perceptual processing?

Acknowledgements

We thank Stefan Schaal, Konrad Körding, Alan Yuille and the three anonymous reviewers for their insightful comments on the manuscript. LS was supported by UCLA Faculty Grants Program and Faculty Career Award, and UB was supported by the Gatsby Charitable Foundation and the Marie Curie FP7 Programme.

References

1 Pearl, J. (2009) Causal inference in statistics: an overview. *Statistics Surveys* 3, 96–146
2 Hume, D. (1960) *A Treatise on Human Nature*, (1739 edn), Clerendon Press
3 Holland, P.W. (1986) Statistics and Causal Inference. *J. Am. Stat. Assoc.* 81, 945–960
4 Bell, A.J. and Sejnowski, T.J. (1995) An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Comput.* 7, 1129–1159
5 Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*, Springer
6 Jordan, M.I., ed. (1998) *Learning in Graphical Models*, Kluwer Academic Publishers
7 Koller, D. and Friedman, N. (2009) *Probabilistic Graphical Models: Principles and techniques*, MIT Press
8 Pearl, J. Causal Inference. *J. Mach. Learn. Res.* (in press)
9 Spirtes, P. (2010) Introduction to causal inference. *J. Mach. Learn. Res.* 11, 1643–1662
10 Griffiths, T.L. *et al.* (2008) Bayesian models of cognition. In *Cambridge Handbook of Computational Cognitive Modeling* (Sun, R., ed.), Cambridge University Press
11 Lu, H. *et al.* (2008) Bayesian generic priors for causal learning. *Psychol. Rev.* 115, 955–984
12 Novick, L.R. and Cheng, P.W. (2004) Assessing Interactive Causal Influence. *Psychol. Rev.* 111, 455–485
13 Yuille, A.L. and Bülthoff, H.H. (1996) Bayesian decision theory and psychophysics. In *Perception as Bayesian Inference* (Knill, D.C. and Richards, W., eds), pp. 123–161, Cambridge University Press
14 Battaglia, P.W. *et al.* (2010) Within- and cross-modal distance information disambiguates visual size-change perception. *PLoS Comput. Biol.* 6 (3), e1000697
15 Kersten, D. *et al.* (2004) Object perception as Bayesian inference. *Annu. Rev. Psychol.* 55, 271–304
16 Knill, D.C. (2003) Mixture models and the probabilistic structure of depth cues. *Vision Res.* 43, 831–854
17 Ernst, M.O. and Bülthoff, H.H. (2004) Merging the senses into a robust percept. *Trends Cogn. Sci.* 8, 162–169
18 Witten, I.B. and Knudsen, E.I. (2005) Why seeing is believing: merging auditory and visual worlds. *Neuron* 48, 489–496

19 Knill, D.C. and Pouget, A. (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719
20 Alais, D. and Burr, D. (2004) The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* 14, 257–262
21 Ernst, M.O. and Banks, M.S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433
22 Maloney, L.T. and Mamassian, P. (2009) Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Vis. Neurosci.* 26, 147–155
23 Shams, L. *et al.* (2005) Sound-induced flash illusion as an optimal percept. *Neuroreport* 16, 1923–1927
24 Bresciani, J.P. *et al.* (2006) Vision and touch are automatically integrated for the perception of sequences of events. *J. Vis.* 6, 554–564
25 Rowland, B. *et al.* (2007) A Bayesian model unifies multisensory spatial localization with the physiological properties of the superior colliculus. *Exp. Brain Res.* 180, 153–161
26 Roach, N. *et al.* (2006) Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proc. Biol. Sci.* 273, 2159–2168
27 Wozny, D.R. *et al.* (2008) Human trimodal perception follows optimal statistical inference. *J. Vis.* 8, 1–11
28 Körding, K. *et al.* (2007) Causal inference in multisensory perception. *PLoS ONE* 2, e943
29 Beierholm, U. *et al.* (2009) Bayesian priors are encoded independently of likelihoods in human multisensory perception. *J. Vis.* 9, 1–9
30 Hospedales, T. and Vijayakumar, S. (2009) Multisensory oddity detection as Bayesian inference. *PLoS ONE* 4, e4205
31 Colonius, H. and Diederich, A. (2010) The optimal time window of visual-auditory integration: a reaction time analysis. *Front. Integr. Neurosci.* 4, Article 11
32 Helbig, H.B. and Ernst, M.O. (2007) Knowledge about a common source can promote visual-haptic integration. *Perception* 36, 1523–1533
33 Yuille, A.L. and Clark, J.J. (1993) Bayesian models, deformable templates and competitive priors. In *Spatial vision in humans and robots* (Harris, I. and Jenkin, M., eds), pp. 333–349, Cambridge University Press
34 Knill, D.C. (2007) Robust cue integration: A Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *J. Vis.* 7, 1–24
35 Landy, M.S. *et al.* (1995) Measurement and modeling of depth cue combination: In defense of weak fusion. *Vis. Res.* 35, 389–412
36 Jacobs, R.A. *et al.* (1991) Adaptive mixtures of local experts. *Neural Comput.* 3, 79–87
37 Stevenson, I.H. and Körding, K.P. (2009) Structural inference affects depth perception in the context of potential occlusion. In *Advances in Neural Informational Processing Systems* (Bengio, Y. *et al.*, eds), pp. 1777–1784, MIT Press
38 Stocker, A. and Simoncelli, E.P. (2008) A Bayesian model of conditioned perception. In *Advances in Neural Information Processing Systems* (Platt, John C. *et al.*, eds), pp. 1490–1501, MIT Press
39 Stocker, A.A. and Simoncelli, E.P. (2006) Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* 9, 578–585
40 Lu, H. *et al.* (2010) Recovering the functional form of the slow-and-smooth prior in global motion perception. *J. Vis.* 10 (7) (in press)
41 Girshick, A.R. and Banks, M.S. (2009) Probabilistic combination of slant information: Weighted averaging and robustness as optimal percepts. *J. Vis.* 9, 1–20
42 Natarajan, R. *et al.* (2009) Characterizing response behavior in multisensory perception with conflicting cues. In *Advances in Neural Information Processing Systems 21* (Koller, D. *et al.*, eds), pp. 1153–1160, MIT Press
43 Dokka, K. *et al.* (2010) Self versus environment motion in postural control. *PLoS Comput. Biol.* 19, e1000680
44 Sato, Y. *et al.* (2007) Bayesian inference explains perception of unity and ventriloquism aftereffect: Identification of common sources of audiovisual stimuli. *Neural Comput.* 19, 3335–3355
45 Hospedales, T. and Vijayakumar, S. (2008) Structure inference for Bayesian multisensory scene understanding. *IEEE Transac. Pattern Anal. Mach. Intell.* 30, 1–18
46 Weisswange, T.H. *et al.*, (2009) Can reinforcement learning explain the development of causal inference in multisensory integration? In *IEEE 8th International Conference on Development and learning*, pp. 263–270, IEEE
47 Yu, J.A. and Dayan, P. (2005) Uncertainty, neuromodulation, and attention. *Neuron* 46, 681–692
48 Pearl, J. (2009) *Causality: Models, reasoning, and inference*, (2nd edn), Cambridge University Press
49 Wallace, M.T. *et al.* (2004) Unifying mulitsensory signals across time and space. *Exp. Brain Res.* 158, 252–258
50 Wei, K. and Körding, K. (2009) Relevance of error: What drives motor adaptation? *J. Neurophysiol.* 101, 655–664
51 Berniker, M. and Körding, K. (2008) Estimating the sources of motor errors for adaptation and generalization. *Nat. Neurosci.* 11, 1454–1461
52 Cheng, K. *et al.* (2007) Bayesian integration of spatial information. *Psychol. Bull.* 133, 625–637
53 Etienne, A.S. and Jeffery, K.J. (2004) Path integration in mammals. *Hippocampus* 14, 180–192
54 Shettleworth, S.J. and Sutton, J.E. (2005) Multiple systems of spatial learning: Dead reckoning and beacon homing in rats. *J. Exp. Psychol. Anim. Behav. Proc.* 31, 125–141
55 Whishaw, I.Q. and Tomie, J. (1997) Piloting and dead reckoning dissociated by fimbria-fornix lesions in a rat food carrying task. *Behav. Brain Res.* 89, 87–97
56 Etienne, A.S. *et al.* (1990) The effect of a single light cue on homing behavior of the golden hamster. *Animal Behaviour* 39, 17–41