Neural Correlates of Causal Confounding

Mimi Liljeholm

Abstract

■ As scientists, we are keenly aware that if putative causes perfectly covary, the independent influence of neither can be discerned—a "no confounding" constraint on inference, fundamental to philosophical and statistical perspectives on causation. Intriguingly, a substantial behavioral literature suggests that naïve human reasoners, adults and children, are tacitly sensitive to causal confounding. Here, a combination of fMRI and computational cognitive modeling was used to investigate neural substrates mediating such sensitivity. While being scanned, participants observed and judged the influences of various putative causes with confounded or nonconfounded, deterministic or stochastic, influences. During judgments requiring generalization of causal knowledge from a feedback-based learning context to a transfer probe, activity in the dorsomedial pFC was better accounted for by a Bayesian causal model, sensitive to both confounding and stochasticity, than a purely error-driven algorithm, sensitive only to stochasticity. Implications for the detection and estimation of distinct forms of uncertainty, and for a neural mediation of domain-general constraints on causal induction, are discussed. ■

INTRODUCTION

Consider two scenarios involving a target cause, *C*, an alternative cause, *A*, and the presence (+) or absence (−) of some effect. In the first scenario, the effect occurs when *C* and *A* are presented in combination, but not when *A* is presented alone [*A*−, *AC*+]. In the second scenario, the *A*− trials are removed, so that both *C* and *A* occur only in combination with each other [*AC*+]. How would your judgment about the influence of *C* differ across these two scenarios? Philosophers, scientists, and statisticians alike recognize that perfect covariation of *C* and *A* in the second scenario occludes the independent influence of each putative cause, rendering any judgment about their respective effects fraught with uncertainty. Importantly, naïve human reasoners also appear to, tacitly, apply this domain-general constraint on causal induction, as evidenced by behavioral data showing sensitivity to confounding in a wide range of causal and predictive judgments, by children as well as adults (Liljeholm, 2015; Schulz & Bonawitz, 2007; Schulz, Gopnik, & Glymour, 2007; Meder, Hagmayer, & Waldmann, 2006; Kushnir & Gopnik, 2005; Spellman, 1996). Very little is known, however, about the neural substrates mediating sensitivity to, and uncertainty associated with, causal confounding. The current study aims to identify such neural computations.

A second source of uncertainty in causal and predictive inference that has been thoroughly investigated, both behaviorally and neurally, is the stochasticity (or variance) of the outcome variable, which is greatest when the probability distribution over possible outcome states is uniform. As with causal confounding, behavioral sensitivity to outcome stochasticity is well established (e.g., Holt & Laury, 2002). Moreover, several neuroimaging studies have implicated the dorsomedial pFC (DMPFC), the anterior insula, the thalamus, and the dorsolateral pFC (DLPFC) in the neural representation of outcome stochasticity (Abler, Herrnberger, Grön, & Spitzer, 2009; Grinband, Hirsch, & Ferrera, 2006; Huettel, Song, & McCarthy, 2005; Volz, Schubotz, & von Cramon, 2003). An important open question is whether and which of these neural regions also implement signals associated with causal confounding. In the current study, participants were scanned with fMRI while observing and judging the influences of various putative causes with confounded or nonconfounded, deterministic or stochastic, influences.

Note that if a particular set of confounded causes always occur in the same configuration and the only goal is to predict the outcome based on that particular configuration, then confounding does not pose a problem. Indeed, most individual, nonconfounded causes can presumably themselves be broken into a set of always co-occurring and thus confounded elements. The critical question, thus, is whether it is necessary to tease apart the influences of individual elements, such as when a confounded cause suddenly occurs on its own or in a novel combination with some other cause. In the current study, participants were required both to make predictions about the outcome given a particular recurring configuration of confounded causes and, in other

University of California, Irvine

https://doi.org/10.1162/jocn_a_01479

instances, to make explicit judgments about the individual influences of those same causes. Only in the latter case would uncertainty due to confounding be warranted.

Formally, two dominant approaches to predictive uncertainty can be discerned in the psychological literature. First, in associative learning theory, an error-driven representation of uncertainty about a cue's predictive strength is captured by the "Pearce–Hall" algorithm, which relates the "associability" of a cue to a weighted average of the absolute prediction error on previous trials involving that cue. Because the frequency and size of prediction errors increases with unpredictability, this quantity is proportional to the stochasticity of the outcome. Conversely, in Bayesian causal models, uncertainty about the strength of a cause is reflected in the entropy of the posterior distribution over its possible strengths, which depends not only on the variance of the effect variable but also on the independent occurrence of alternative causes. Although both the associative and Bayesian causal model predicts sensitivity to stochasticity, only the causal model accounts for uncertainty due to causal confounding. Here, a combination of neuroimaging and computational cognitive modeling was used to dissociate neural signals scaling with error-driven and causal uncertainty. In particular, judgments requiring generalization of causal knowledge from a feedback-based learning context to a transfer probe were expected to elicit strong neural responses to both stochasticity and confounding.

METHODS

Participants

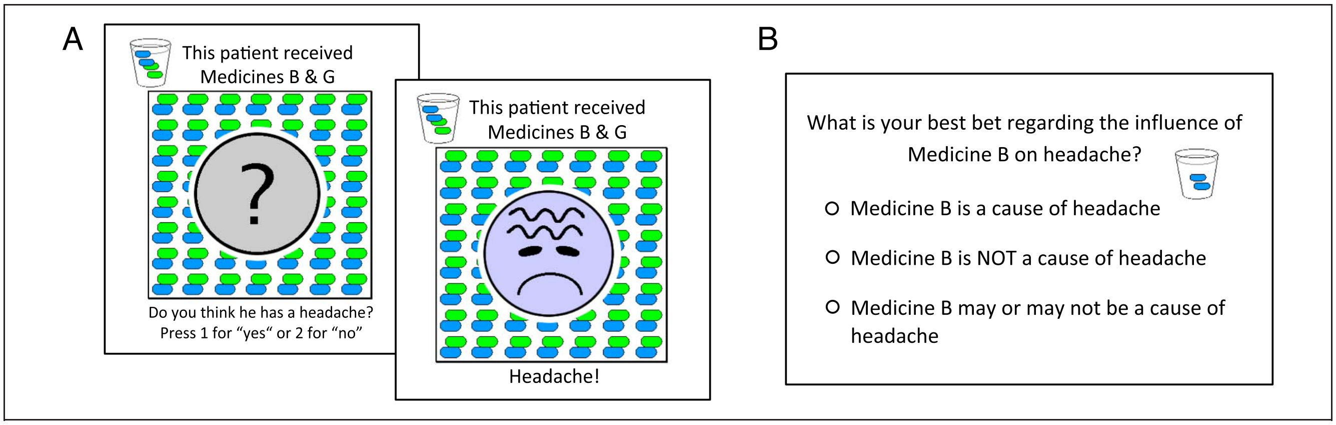
Twenty healthy volunteers (mean age = 20.9 ± 2.4 years, range = 18–27 years, 12 women) participated in the study. A power analysis performed on data from a pilot study on error-driven uncertainty, described in detail below, indicated that a sample size of 16 would yield a power of 0.9 at a Gaussian random field theory-corrected threshold of 0.05 in ROIs. One participant was excluded before any analyses because of excessive head movement (>6 mm), leaving a sample size of 19. The volunteers were preassessed to exclude those with a history of neurological or psychiatric illness. All participants gave informed consent, and the study was approved by the institutional review board of the University of California, Irvine.

Task and Procedure

Participants were scanned with fMRI while performing a causal induction task in which they assumed the role of a research scientist assessing the influence of various allergy medicines on headache—a potential side effect. At the beginning of the study, participants were instructed that each medicine could either produce headache or have no influence on headache (i.e., there were no preventive causes) and further that the influence of a given medicine might be stochastic, so that even if that medicine was indeed a cause of headache, it may still not produce headache every time it was administered. Three target medicines, C, D, and S, mnemonically labeled here to indicate "confounded," "deterministic," and "stochastic" influences, respectively, occurred only in combination with some alternative medicine during feedback-based learning. Specifically, with +, −, and ± respectively indicating a 1.0, 0, and .5 probability of the effect and with lower case letters indicating nontarget causes, there were five different types of medicine treatments: e−, s'±, De+, Se±, and Cc$'$+. None of the allergy patients had headache in the absence of any medicine, which was explicitly stated in the initial instructions as well as apparent on several "no medicine" (n−) trials.

Note that medicines C and c' both occurred only in combination with one another, so that target medicine C was perfectly confounded with its compound counterpart. Medicine D also occurred only in combination with another medicine, e; however, "elemental" medicine e also occurred by itself, allowing for an estimation of the independent causal influence of medicine D across the constant presence of e (i.e., across e− and De+ trials). Medicine S was identical to medicine D, except that the probability of headache given the Se compound was .5 rather than 1.0; a comparison of medicines S and D, therefore, identifies a difference between deterministic and stochastic causation. Finally, medicine s' also produced headache with a probability of .5 but, unlike medicine S, never occurred in combination with any other medicine, so that contrasting S with s' identifies differences due to compound presentation. From a modeling perspective, an algorithm that relies solely on outcome variance to track uncertainty would only respond to medicines S and s'. In contrast, a computation that treats both confounding and stochasticity as sources of uncertainty should generate an increased signal to medicine C as well as S and s': These divergent predictions are respectively instantiated by the error-driven and Bayesian causal model specified in the subsequent section.

On each trial of feedback-based learning (see Figure 1A), participants were presented with an individual allergy patient and were told either that the patient had not received any medicine or that a particular medicine or combination of medicines had been administered. Medicines were color coded and labeled accordingly (i.e., Medicine "B" was blue), with color assignments counterbalanced across participants. On the first screen, the allergy patient's state was obscured by a question mark, and participants were asked to indicate whether or not the patient had a headache, pressing the "Y" key for "yes" and the "N" key for "no." After a prediction was made, there was a brief (2000 msec) pause during which the word "wait" was displayed on the screen, followed by a screen revealing the allergy patient's state (i.e., with or without headache).



This patient received Medicines B & G

?

Do you think he has a headache? Press 1 for "yes" or 2 for "no"

This patient received Medicines B & G

Headache!

What is your best bet regarding the influence of Medicine B on headache?

○ Medicine B is a cause of headache

○ Medicine B is NOT a cause of headache

○ Medicine B may or may not be a cause of headache

Figure 1. Task illustration. (A) Prediction and outcome screens from a feedback-based learning trial. Screens were separated by a 2-sec pause during which the word "Wait" (not shown in figure) was displayed. (B) Causal query regarding the individual influence of a particular medicine on headache.

Each of six distinct trial types was presented eight times in each of four sessions, for a total of 192 trials, with sessions separated by 2-min breaks, during which the scanner was turned off. Each session was further divided into four nondelineated blocks, with each type of trial occurring once in random order in each block. All trials were separated by a jittered 4-sec intertrial interval. In every other block, each trial (except for n− trials) was followed by a query regarding the individual causal influence of the target medicine present on that trial (i.e., C, D, or S) or of medicine e or s′ on nontarget trials. Specifically, participants were asked to choose between the following three options regarding the relevant medicine: "is a cause of headache," "is not a cause headache," and an ambivalent "May or may not be a cause of headache" (Figure 1B). Categorical response options, rather than the rating scales commonly employed in the causal learning literature (Liljeholm, 2015; Liljeholm & Cheng, 2009; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008), were used for two main reasons: First, because the limited number of response buttons in the scanner would have necessitated a scale slider, the initial positioning of which might bias ratings, and second, to reduce movement due to back-and-forth scale scrolling.

Computational Models

Two formal accounts of predictive strength and uncertainty were implemented. First, an error-driven associative learning algorithm updates predictive strength, on each trial, using the difference between the observed state of the outcome and the expected state based on all present cues:

$$V_{c,n+1} \leftarrow V_{c,n} + \alpha_{c,n}\left(\lambda - \sum_i V_i\right)_n \qquad (1)$$

where V is the predictive strength of a particular cue, c, ΣV_i is the summed strength of all cues present on trial n, and λ is the observed state of the outcome on that trial. The associability of cue c on trial n, $\alpha_{c,n}$, is defined as

$$\alpha_{c,n} = \gamma\left|\lambda - \sum_i V_i\right|_{n-1} + (1-\gamma)\alpha_{c,n-1} \qquad (2)$$

where $n - 1$ refers to the previous trial involving cue c and γ is a free parameter accounting for the weighting of that previous trial relative to preceding ones. In addition to scaling the influence of the prediction error on predictive strength, α is taken to reflect current levels of uncertainty about the cue's influence (e.g., Esber & Haselgrove, 2011). The strengths and uncertainties associated with each cue were initialized to 0.5. On feedback-based trials with target causes, which always occurred in compound with an alternative cause, the α of the target cause (always greater than or equal to that of its compound counterpart) was used to indicate uncertainty, whereas the summed strength, ΣV, was used to predict the outcome and to compute the prediction error.

A second formal account is provided by a Bayesian causal model, in which reasoners make inferences over causal structures potentially responsible for the observed data (Liljeholm, 2015; Lu et al., 2008; Griffiths & Tenenbaum, 2005). Here, as illustrated in Figure 2, three possible graphs (Graphs 0–2) were defined for a candidate cause, C, an alternative cause, A, and a constantly present background

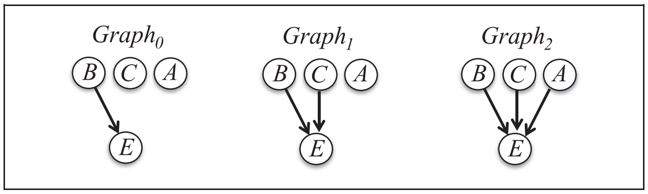


Figure 2. Possible causal structures involving a background cause, B; a target cause, C; an alternative cause, A; and an effect, E. A link to the effect always exists for B and may exist for neither C nor A (Graph_0), C only (Graph_1), or both C and A (Graph_2).

cause, *B,* where a causal link to the effect exists for neither *C* nor *A* (G_0), *C* only (G_1), or both *C* and *A* (G_2).

Each graph has a set of parameters θ, which are strengths w_i associated with causal links (i.e., w_B, w_C, and w_A for links associated with *B*, *C,* and *A,* respectively, in G_2). Sequential estimates of these parameters were modeled, for each candidate cause, using the smallest possible "focal set" (Cheng & Novick, 1992)—a set of events across which alternative causes can be assumed to occur with the same probability in the presence and absence of the relevant cause. Thus, estimates of the strength of the "no medicine" background cause were modeled using only n− trials in G_0. Estimates of elemental (nontarget) candidate causes s′ and e were respectively modeled based on trials with those causes and the "no medicine" trials (i.e., n− and s′± for candidate s′, and n− and e− for candidate e) using G_1. Finally, for each target candidate cause, which always occurred in compound with some alternative cause, estimation was modeled given G_2, using the trials relevant for the particular target cause (i.e., n− and Cc′ for target C; n−, e−, and De+ for target D; and n−, e−, and Se± for target S).

The likelihoods $P(d\,|\,\theta, G_i)$ were computed using a noisy OR parameterization (Cheng, 1997). Specifically, summarizing data *d* by contingencies $N(e, c, a)$, the frequencies of each combination of the presence versus absence of the effect, target cause, and alternative cause, the likelihood term for G_2 is shown in equation (3), where

$$\begin{aligned}P(d|w_B,w_c,w_A,G_2) = &\binom{N(c^-,a^-)}{N(e^+,c^-,a^-)}\times\binom{N(c^+,a^-)}{N(e^+,c^+,a^-)}\times\binom{N(c^-,a^+)}{N(e^+,c^-,a^+)}\times\binom{N(c^+,a^+)}{N(e^+,c^+,a^+)}w_B^{N(e^+,c^-,a^-)}\\ &\times(1-w_B)^{N(e^-,c^-,a^-)}[w_B+w_C-w_Bw_C]^{N(e^+,c^+,a^-)}[1-w_B-w_C+w_Bw_C]^{N(e-,c^+,a^-)}[w_B+w_A-w_Bw_A]^{N(e^+,c^-,a^+)}\\ &\times[1-w_B-w_A+w_Bw_A]^{N(e^-,c^-,a^+)}[w_B+w_C+w_A-w_Bw_C-w_Bw_A-w_Cw_A+w_Bw_Cw_A]^{N(e^+,c^+,a^+)}\\ &\times[1-w_B-w_C-w_A+w_Bw_C+w_Bw_A+w_Cw_A-w_Bw_Cw_A]^{N(e^-,c^+,a^+)}\end{aligned} \tag{3}$$

$\binom{n}{k}$ denotes the number of ways of picking *k* unordered outcomes from *n* possibilities. $N(c^+)$ indicates the frequency of events in which the target cause is present, with analogous definitions for the other *N*(.) terms. The likelihood terms for G_0 and G_1 are similarly specified, where frequencies *N*(.) are summed across the presence and absence of relevant events, and $w_i = 0$, for any cause that does not have a link to the effect in the relevant graph.

The marginal posterior distribution over strengths for a particular cause is obtained by applying Bayes' rule and integrating out the parameters for other causes in the graph, such that, for G_2,

$$\begin{aligned}&P(w_C|d,G_2)\\ &=\int_0^1\int_0^1\frac{P(d|w_B,w_C,w_A,G_2)P(w_B,w_C,w_A|G_2)}{P(d)}\mathrm{d}w_B\mathrm{d}w_A\end{aligned} \tag{4}$$

where $P(d\,|\,w_B,w_C,w_A,G_2)$ is the likelihood term, $P(w_B, w_C,w_A\,|\,G_2)$ refers to the prior probabilities of causal strength parameters, and $P(d)$ is the normalizing term, denoting the probability of the observed data.

Uncertainty about the strength of a particular cause, the focus of the study, was modeled as the Shannon entropy $H(w_c)$ of its marginal posterior distribution $P(w_C\,|\,d)$,

$$H(w_C) = -\int_0^1 P(w_C|d)\ln P(w_C|d)\mathrm{d}w_c \tag{5}$$

Point estimates of causal strength for each elemental and target cause were modeled as the mean of the relevant marginal posterior distribution (Liljeholm, 2015; Lu et al., 2008), given by

$$\bar{w}_C = \int_0^1 w_C P(w_C|d)\mathrm{d}w_c \tag{6}$$

The model was sequentially implemented on a trial-by-trial basis, such that, for each cause, likelihoods were computed on each trial of feedback-based learning that yielded information relevant to that cause, using only the data point provided on that trial, with the posterior on the previous trial being used as the prior to obtain the posterior on the current trial. On the first trial in which data relevant to a particular candidate cause was presented, the priors were assigned independent uniform distributions. Finally, as an analog to the model-free prediction error, the Kullback–Leibler (KL) divergence between the prior (*Q*) and posterior (*P*) on each trial, commonly referred to as the "Bayesian surprise" (Itti & Baldi, 2009), was computed as

$$\mathrm{KL}(P\,||\,Q) = \int_0^1 p(w_C|d)\log\frac{p(w_C|d)}{q(w_C|d)}\mathrm{d}w_C \tag{7}$$

As noted, target causes only occurred in compound with alternative causes during feedback-based learning. On such trials, the entropy and KL divergence of the marginal distribution for the target cause was used to indicate uncertainty and surprise, respectively, whereas predictions regarding the occurrence of the effect were generated using a noisy OR integration of causal strengths.

Both models assumed that, during feedback-based learning, participants selected "yes"/"no" responses to

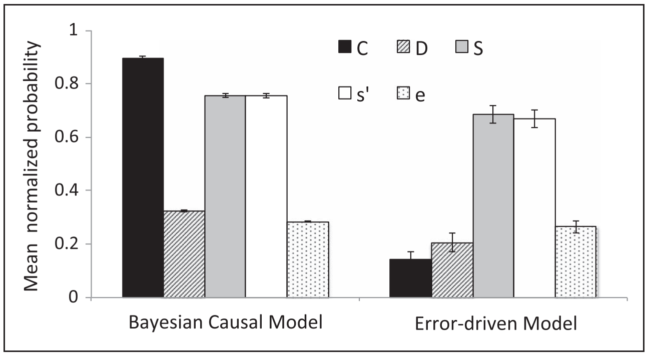


Figure 3. Mean (normalized) probabilities of uncertain "May or may not be a cause" judgments derived from the Bayesian causal model and error-driven algorithm, respectively. Confounded, deterministic, and stochastic target causes (C, D, & S) occurred only in compound with alternative causes during feedback-based learning, whereas nontarget causes occurred only individually (s′) or individually as well as in compound (e). Error bars = *SEM*.

questions about whether headache would occur using probabilities generated by a softmax distribution, in which a free noise parameter controls the influence of predictive strength (i.e., the strength with which the medicine(s) present on a given trail predicted headache; noisy-logical $P(e|w_i)$ and ΣV in causal and error-driven models, respectively) on choice behavior. An additional noise parameter controlled the influence of uncertainty (i.e., α and $H(w_c)$ in error-driven and causal models, respectively) on the proportion of ambivalent "May or may not be a cause" judgments during queries about the individual influence of each medicine. The first two instances of each trial type and the first causal judgment for each medicine were excluded from model fitting to eliminate transient noise due to task adjustment and to minimize the influence of assumptions about priors. Free parameters were fit to behavioral data by minimizing the negative log likelihood of observed responses for each individual, and the Bayesian information criterion was used for model selection.

The model-derived probabilities of uncertain "May or may not be a cause" judgments are plotted in Figure 3 for the causal and error-driven account of uncertainty, respectively. Note that, as was foreshadowed in the description of the study design, the Bayesian causal model predicts high levels of uncertainty for both confounded, C, and stochastic, S and s′, causes, whereas the error-driven algorithm predicts that only stochasticity will generate high levels of uncertainty.

Imaging Procedure and Analysis

A 3-T scanner (Phillips Achieva) was used to acquire structural T1-weighted images and T2*-weighted echoplanar images (repetition time = 2.65 sec, echo time = 30 msec, flip angle = 90°, 45 transverse slices, matrix = 64 × 64, field of view = 192 mm, thickness = 3 mm, slice gap = 0 mm) with BOLD contrast. The first four volumes of images were discarded to avoid T1 equilibrium effects. All remaining volumes were corrected for differences in the time of slice acquisition, realigned to the first volume, spatially normalized to the Montreal Neurological Institute (MNI) echoplanar imaging template, and spatially smoothed with a Gaussian kernel (8 mm, FWHM). High-pass filtering with a cutoff of 128 sec was used. All effects were reported at a whole-brain family-wise error cluster-corrected threshold of $p < .05$, calculated using the Statistical nonParametric Mapping toolbox (SnPM13; warwick.ac.uk/snpm; Nichols & Holmes, 2002), with 5000 permutations, no variance smoothing, and an uncorrected height threshold of $p < .001$.

To address collinearity due to the shared prediction by causal and error-driven accounts that uncertainty increases with increased stochasticity, the two algorithms were analyzed using separate first-level general linear models (GLM) and were directly contrasted at the group level using a Bayesian model selection. Specifically, for each participant, two GLMs were specified, one for the causal and one for the error-driven computational model. In each such GLM, three regressors respectively modeled (1) the onset of the prediction screen on each trial during feedback-based learning, (2) the onset of the outcome screen on each trial during feedback-based learning, and (3) the onsets of the screens soliciting judgments about the influences of individual medicines. In one GLM, prediction screen onsets and causal judgment onsets were modulated by trial-by-trial error-driven estimates of uncertainty and strength, whereas outcome screen onsets were modulated by trial-by-trial values of the prediction errors. In the second GLM, prediction screen and causal judgment onsets were modulated by trial-by-trial estimates of uncertainty and strength by the causal model, whereas outcome screen onsets were modulated by trial-by-trial estimates of Bayesian surprise. In both models, the first two instances of each trial type and corollary causal judgments were modeled by separate regressors to eliminate transient noise due to task adjustment and reduce the influence of nonmodeled prior beliefs. Moreover, in each model, two onset regressors respectively modeled the responses on feedback-based and causal query trials, three regressors indicated separate sessions, and six additional regressors accounted for the residual effects of head motion. Except for those indicating sessions and head motion, all regressors were convolved with a canonical hemodynamic response function. No orthogonalization was applied.

To directly contrast error-driven and causal accounts, Bayesian Model Selection (BMS) analyses was performed using a set of GLMs with the same onsets as those specified above, but with a single parametric modulator (e.g., error-driven uncertainty at prediction screen onsets) entered for each GLM. Thus, model comparisons isolated the relative contribution of a particular computational variable to neural activity. A first-level Bayesian estimation procedure was used to compute a log model evidence map for every participant and each GLM, and inferences at the group level were modeled by applying a random effects approach (Rosa, Bestmann, Harrison, & Penny, 2010) at every voxel of the log evidence data. Group-level exceedance probability maps, reflecting the probability that one model is more likely than the other, were thresholded to identify voxels in which the exceedance probability was greater than .95. Classical inferences were then assessed using BMS masks, within which a particular computational variable was more likely than its competing counterpart, at an exceedance probability greater than .95: For example, significant effects of error-driven uncertainty during the prediction period of feedback-based learning trials were reported only for those regions in which error-driven uncertainty during this event period was more likely than causal uncertainty during the same period, according to the BMS.

A Pilot Study on Error-driven Uncertainty

Previous research has implicated the anterior insula and dorsal medial frontal cortex in outcome stochasticity, using a range of decision-making tasks (Mohr, Biele, & Heekeren, 2010; Grinband et al., 2006; Huettel et al., 2005; Volz et al., 2003). To ensure that such effects, used here as a benchmark against which to assess neural representations of uncertainty due to confounding, also emerge in our causal learning task, a pilot study ($n = 10$) was conducted that was highly similar to that reported here, with the following exceptions: First, only stochasticity, not confounding, was manipulated across putative causes; second, both generative and preventive causal influences were included; and, finally, judgments of individual causal influences were solicited at the end of, rather than throughout, each scanning session and measured on a scale ranging from −100 (*strongly removes headache*) to +100 (*strongly produces headache*). The error-driven model of uncertainty was implemented as described above, fit to behavior during feedback-based learning, and regressed against the BOLD data. A single ROI was constructed from the anterior insula and dorsal medial frontal cortex using the "Willard" functional parcellation atlas (Richiardi et al., 2015). A power analysis performed with NeuroPower (neuropowertools.org; Durnez, Degryse, Seurinck, Moerkerke, & Nichols, 2015), using a screening threshold of $z = 2.3$, revealed that a sample size of 16 would yield a power of .9 to detect effects of error-driven predictive uncertainty in this ROI at a Gaussian random field-corrected alpha level of .05.

RESULTS

Behavioral Results

All statistical tests of behavioral data were planned comparisons, employing two-tailed t tests, and were calculated using $n = 19$. Confidence intervals (CIs) and effect sizes (Cohen's d) are reported for all comparisons. The distributions of causal judgments across response options are shown, with associated RTs, for each putative cause in Table 1. Note that the proportion of ambivalent "May or may not be a cause" judgments (third row in Table 1) were greatest for the confounded cause (C), intermediate, although still substantial, for the two stochastic causes (S and s′), and virtually absent for deterministic causes (D and e), exactly as predicted by the causal, but not the error-driven, model (cf. Figure 3; see Liljeholm, 2015, for similar results). Note also that the distribution of "Is a cause" and "Is not a cause" judgments differed markedly across stochastic and confounded causes, such that participants almost never select the latter option for the confounded cause, consistent with the normative increase in the likelihood that target cause C is causal given Cc+ trials but distributed their responses fairly evenly across these two options for stochastic causes, reflecting, perhaps, the outcome on trials immediately preceding each judgment.

Table 1. Mean Proportions of "Is a Cause," "Is Not a Cause," and "May or May Not Be a Cause" Judgments and Associated RTs, with Standard Deviations, for Target Causes C, D, and S and Nontarget Causes s′ and e′

	C	*D*	*S*	*s′*	*e*
Proportions					
Is a cause	0.36 ± 0.33	0.84 ± 0.25	0.38 ± 0.25	0.40 ±0.27	0.02 ± 0.03
Is not a cause	0.03 ± 0.04	0.06 ± 0.09	0.23 ± 0.11	0.30 ± 0.17	0.96 ± 0.05
May be a cause	0.61 ± 0.35	0.10 ± 0.21	0.39 ± 0.31	0.30 ± 0.28	0.02 ± 0.04
RTs					
Is a cause	2.92 ± 1.88	2.52 ± 1.23	4.03 ± 3.62	2.82 ± 2.59	2.86 ± 2.11
Is not a cause	4.26 ± 2.69	3.94 ± 3.36	4.19 ± 3.64	2.72 ± 1.23	2.44 ± 1.15
May be a cause	3.79 ± 2.68	3.76 ± 1.46	2.99 ± 1.94	2.59 ± 1.32	4.08 ± 4.78

Planned comparisons revealed that the mean proportion of ambivalent causal judgments were significantly lower for the deterministic (D) target cause than for both confounded (C), $t(18) = 6.78$, $p < .001$, CI [0.35, 0.67], $d = 1.79$, and stochastic (S), $t(18) = 4.09$, $p < .001$, CI [0.14, 0.44], $d = 1.11$, target causes. The difference between the confounded and stochastic target cause was also significant, with the mean proportion of ambivalent judgments being greater for C than for S, $t(18) = 2.65$, $p < .05$, CI [0.0.05, 0.40], $d = 0.68$. Recall that all target causes, including S, occurred only in compound with alternative causes during feedback-based learning. The proportion of ambivalent judgments for the two stochastic causes, one always occurring in isolation (s′) and the other always in compound (S), was marginally significant, $p = .08$.

Judgment RTs, averaged across response options for each cause, were significantly faster for judgments about the deterministic target cause D than stochastic target cause S, $t(18) = 2.20$, $p < .05$, CI [0.03, 1.02], $d = 0.35$, but did not differ significantly between D and C, nor between C and S, $p > .1$. Moreover, RTs did differ significantly across the two stochastic causes, being significantly slower for target cause S, which never occurred in isolation during feedback-based learning, than for cause s′, $t(18) = 3.42$, $p < .005$, CI [0.29, 1.21], $d = 0.52$.

With respect to model fitting, the mean best fitting values of the free parameters for the error-driven model were 0.35 ± 0.27 for the weighting parameter, 3.25 ± 0.95 for the prediction noise parameter, and 1.08 ± 0.88 for the judgment noise parameter. The mean best fitting values of the free parameters for the causal model were

Table 2. Nonmasked Significant, Whole-brain-corrected Neural Effects of the Causal Model of Uncertainty, Strength, and Surprise during Causal Judgments and Feedback-based Learning, with Peak MNI Coordinates and Cluster Sizes at $p < .001$

Contrast	*Region*	*Peak MNI*	*Cluster Size*
Judgment uncertainty	DMPFC	−8, 28, 46	1900
	R IPG	50, −52, 40	292
	DLPFC	42, 34, 30	201
Prediction uncertainty	Fusiform	28, −72, −12	7195
	DMPFC	−10, 16, 52	1387
	R insula	−30, 18, 4	726
	L insula	34, 30, −4	470
	Thalamus	6, −22, 2	867
Prediction strength	Calcarine/occipital	18, −96, 4	236
Prediction strength (−)	Lingual gyrus	14, −62, −2	2232
	Superior frontal gyrus	−26, −8, 58	458
	Superior parietal lobule	−28, −50, 62	304
	Inferior parietal gyrus	−46, −26, 36	336
	SMA	8, 2, 52	297
Surprise	Right insula	36, 18, −2	652
	L insula	−28, 22, −6	482
	DMPFC	4, 46, 34	1146
	L DLPFC	−46, 4, 42	192
	R DLPFC	44, 24, 30	325
Surprise (−)	Supramarginal gyrus	−50, −26, 38	1858
	Caudate	18, 32, 0	505
	Mid cingulum	−10, −30, 42	472
	Postcentral gyrus	50, −20, 48	2399
	Parahippocampal	34, −32, −12	227
	Precuneus	16, −52, 14	221

R = right; L = left; (−) = negative correlation.

Table 3. Nonmasked Significant, Whole-brain-corrected Neural Effects of the Error-driven Model of Uncertainty, Strength, and Surprise during Feedback-based Learning, with Peak MNI Coordinates and Cluster Sizes at $p < .001$

Contrast	*Region*	*Peak MNI*	*Cluster Size*
Prediction uncertainty	Calcarine/lingual gyrus	14, −88, 0	3116
	SMA	10, 20, 48	153
	L insula	−24, 22, 2	65
	R insula	34, 28, −2	109
Prediction uncertainty (−)	Middle temporal gyrus	−46, −62, 20	341
	Middle temporal gyrus	−56, −22, −10	84
	Ventromedial pFC	−2, 30, −14	157
	Middle temporal gyrus	46, -58, 22	64
Prediction strength	Lingual gyrus	−20, −78, −10	2801
Prediction strength (−)	Superior parietal lobule	−18, −52, 64	407
	Supramarginal gyrus	−60, −28, 30	390
	Precentral gyrus	26, −12, 50	240
	Supramarginal gyrus	58, −38, 28	728
	Middle temporal gyrus	−54, −60, 2	603
	Precentral gyrus	−22, −14, 64	383
	Thalamus	−12, −10, −4	214
Prediction error	R insula	38, 18, −4	912
	L insula	−40, 20, 4	945
	Precentral gyrus	−42, −2, 40	713
	Anterior cingulate	4, 38, 30	2021
	Inferior parietal gyrus	−28, −52, 42	246
	DLPFC	44, 28, 26	885
Prediction error (−)	Middle occipital gyrus	−42, −74, 26	735
	Fusiform gyrus	−34, −42, −12	253
	Postcentral gyrus	54, −26, 52	906

R = right; L = left; (−) = negative correlation.

4.13 ± 1.44 for the prediction noise parameter and 1.03 ± 1.05 for the judgment noise parameter. The mean Bayesian information criterion was significantly lower, indicating superior performance, for the causal model (220.00 ± 44.15) than for the error-driven model (237.09 ± 35.58), $t(18) = 2.51$, $p < .05$, CI [2.81, 31.37], $d = 0.43$.

Neuroimaging Results

All results reported below survived inclusive masking with voxels identified by Bayesian model selection analyses. A complete list of the nonmasked significant effects of each model is provided in Tables 2 and 3. Bar graphs show effect sizes extracted from 8-mm spheres centered on peak MNI coordinates, using rfxplot (Gläscher, 2009).

Uncertainty Signals during Causal Judgments

During causal queries, as participants made judgments about the causal influences of individual medicines, selective and significant effects of the causal model of uncertainty emerged in the medial and lateral superior frontal gyrus and the right DLPFC (Figure 4). Note that the effect size of BOLD responses (bar graphs in Figure 4) across putative causes were greatest for the confounded cause (C), intermediate for stochastic causes (S and s′), and smallest for deterministic causes (D and e), exactly as predicted by the causal, but not the error-driven, model (cf. Figure 3). To further probe these effects, RTs were included as a parametric modulator of judgment uncertainty in first-level, individual-subject, models and as a between-subject covariate in group-level analyses.

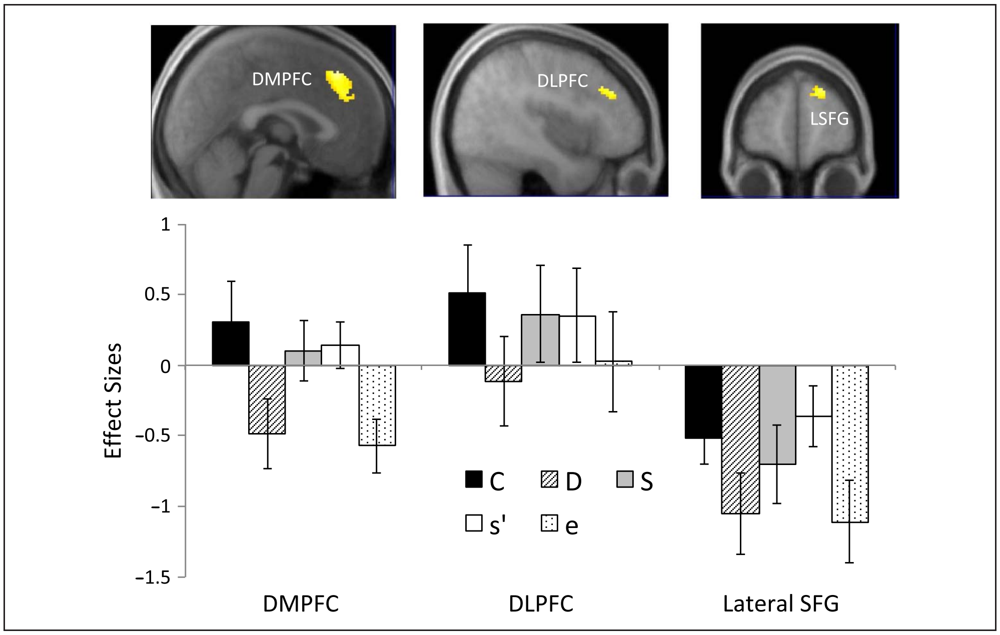


Figure 4. Selective effects of the causal model of uncertainty during judgments about the causal influences of individual medicines, showing effects in the DMPFC and DLPFC and the lateral superior frontal gyrus (LSFG). Bar plots show effect sizes extracted from spheres centered on peak coordinates for each queried medicine. Error bars = *SEM*.

Neither of these additional analyses revealed any modulatory influence by RTs on the neural effects of the causal model of judgment uncertainty. No effects of the error-driven measure of uncertainty during causal judgments survived whole-brain correction.

Uncertainty Signals during Prediction

During the prediction period of feedback-based learning trials, the error-driven, but not causal, measure of uncertainty was significantly positively correlated with activity throughout the inferior occipital and lingual gyri, as well as with activity in the SMA and the right anterior insula (Figure 5). Moreover, activity that was negatively correlated with error-driven predictive uncertainty was observed in the posterior middle temporal gyrus.

These effects of the error-driven account of uncertainty appear relatively robust, as they closely mirror those from the pilot study, in which, at an uncorrected threshold of 0.005, extensive (cluster size > 100) effects of error-driven predictive uncertainty emerged in the dorsal medial frontal cortex, the bilateral anterior insula, and the lingual gyrus during feedback-based learning. In contrast to the widespread effects of error-driven predictive uncertainty

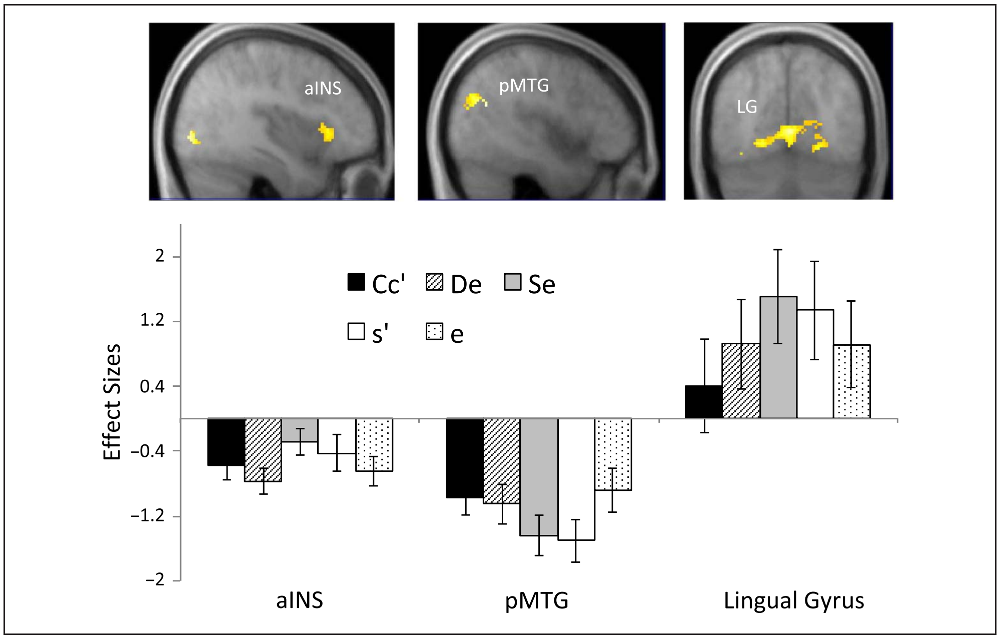


Figure 5. Selective effects of the error-driven model of uncertainty during the prediction period of feedback-based learning trials, showing effects in the right anterior insula (aINS), posterior middle temporal gyrus (pMTG), and lingual gyrus (LG). Bar plots show effect sizes extracted from spheres centered on peak coordinates for each type of medicine trial. Error bars = *SEM*.

during feedback-based learning, selective and significant effects of the causal measure of uncertainty emerged only in the right dorsal thalamus.

Strength and Surprise Signals

Although uncertainty in causal inferences, due to confounding or stochasticity, is of primary interest here, estimates of the strength with which a cause or a combination of causes predicts the effect, and of the surprise at the outcome on each trial of feedback-based learning, were also modeled for completeness. During the prediction period of feedback-based learning trials, the causal measure of predictive strength was selectively and significantly positively correlated with activity in the calcarine and superior occipital gyrus, as well as negatively correlated with activity in the left lateral superior frontal and precentral gyri (Figure 6). In contrast, a selective and significant negative correlation with the error-driven measure of predictive strength emerged in the posterior middle temporal gyrus. No significant effects of either the causal or error-driven account of predictive strength were found during causal judgments.

For measures of surprise during the outcome period of feedback-based learning trials (Figure 7), the absolute prediction error generated by the error-driven account was selectively and significantly positively correlated with activity in the bilateral anterior insula, the medial and lateral dorsal pFC, and the inferior parietal gyrus. In addition, significant negative correlations with this measure were identified in the middle occipital gyrus, the fusiform gyrus, the superior parietal lobule, the precentral gyrus, and the calcarine sulcus. In contrast, a selective and significant negative correlation with the causal measure of (Bayesian) surprise was observed in the anterior caudate.

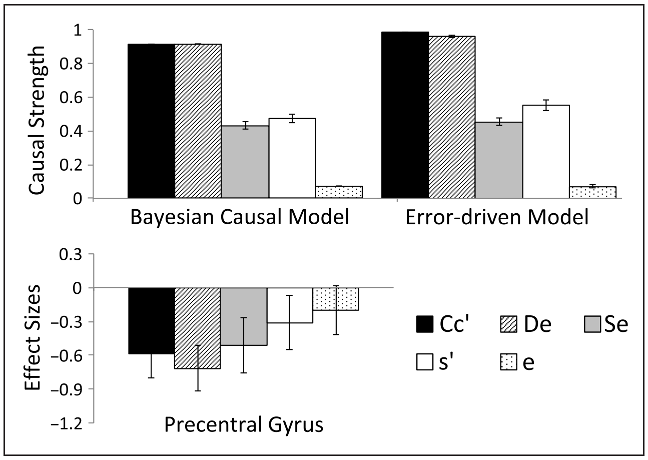


Figure 6. Strength predictions derived from the causal and error-driven model (top), together with effect sizes (bottom) extracted from spheres centered on peak coordinates in the precentral gyrus, for each type of medicine trial during feedback-based learning. Error bars = *SEM*.

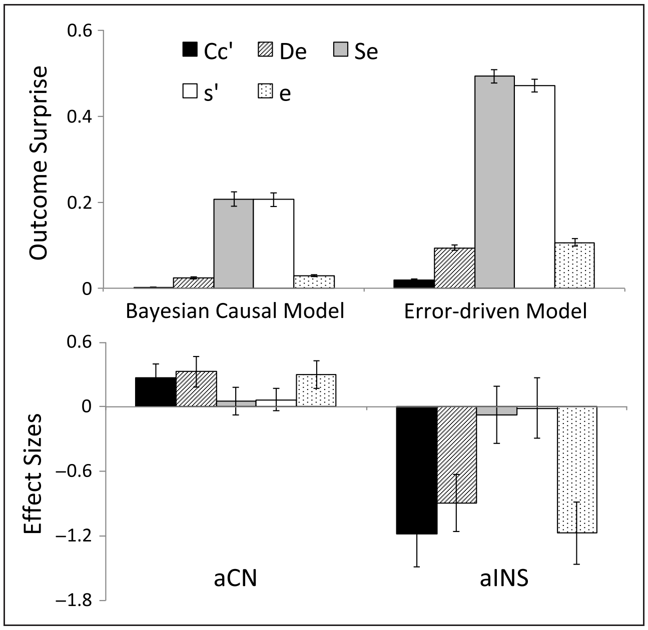


Figure 7. Outcome surprise signals derived from the causal and error-driven model (top), together with effect sizes (bottom) extracted from spheres centered on peak coordinates in the anterior caudate nucleus (aCN) and anterior insula (aINS), for each type of medicine trial during feedback-based learning. Error bars = *SEM*.

DISCUSSION

Sensitivity to the "no confounding" constraint—a requirement that putative causes do not covary, fundamental to philosophical and statistical perspectives on causation—has been reliably demonstrated in a range of behavioral judgments by children as well as adults (Liljeholm, 2015; Schulz & Bonawitz, 2007; Schulz et al., 2007; Meder et al., 2006; Kushnir & Gopnik, 2005; Spellman, 1996). The current study used fMRI and computational cognitive modeling to investigate neural substrates mediating uncertainty associated with causal confounding. While being scanned, participants studied and judged the influences of various putative causes with confounded or nonconfounded, deterministic or stochastic influences. During judgments requiring generalization of causal knowledge from a feedback-based learning context to a transfer probe, activity in the DMPFC was better accounted for by a Bayesian causal model, sensitive to both confounding and stochasticity, than a purely error-driven algorithm, sensitive only to stochasticity.

The DMPFC has been implicated in encoding uncertainty across a variety of tasks (Michael, de Gardelle, Nevado-Holgado, & Summerfield, 2015; Xue et al., 2008; Grinband et al., 2006; Volz et al., 2003). For example, scanning participants with fMRI as they classified stimuli according to a variable category boundary, Grinband et al. (2006) found that activity in the DMPFC scaled with the proximity of a stimulus to a category boundary. An important aspect of this and other demonstrations of the involvement of the

DMPFC in uncertainty is that the level of uncertainty is directly related to the degree of experienced errors: That is, the greater the uncertainty on a given trial, the more likely it is that the prediction on that trial will be incorrect, as indicated by explicit feedback. In contrast, during feedback-based learning in the current task, the outcome on confounded trials was deterministic. Consequently, the increased activity in the DMPFC, as participants generated judgments about the confounded cause, cannot be attributed to a history of errors. Instead, these results suggest that the DMPFC encodes a more abstract representation of uncertainty, divorced from the immediate consequences of choice, and irrespective of whether the specific source of uncertainty is confounding or stochasticity.

Several neuroeconomic studies (Pushkarskaya, Smithson, Joseph, Corbly, & Levy, 2015; Levy, Snell, Nelson, Rustichini, & Glimcher, 2010; Bach, Seymour, & Dolan, 2009; Huettel, Stowe, Gordon, Warner, & Platt, 2006; Krain, Wilson, Arbuckle, Castellanos, & Milham, 2006; Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005) have contrasted gambles involving stochasticity with "ambiguous" gambles, for which outcome probabilities are completely unknown, with some finding greater activity in the DMPFC in response to ambiguity than stochasticity (e.g., Hsu et al., 2005). One might argue that, in the current study, the influence of the confounded medicine is unknown, just as the outcome probabilities of ambiguous gambles in previous studies were unknown. But what does it mean to "know" something? Critically, all target medicines occurred only in compound with alternative medicines during feedback-based learning: Consequently, when presented individually during causal queries, each was equally novel and, presumably, equally unknown. In other words, differences in ambiguity between confounded and other target causes cannot be attributed to a generalization decrement based on changes in stimulus features. Moreover, ambiguity due to small samples and uncertainty due to confounding call for very different plans of exploratory action: Whereas in the former case, repeated observation of the same stimulus can resolve the uncertainty, the latter case requires an intervention that unconfounds the stimulus configuration. Further work is needed to assess the overlap between neural representations of confounding and ambiguity.

In addition to stochasticity and ambiguity, the DMPFC has been heavily implicated in cognitive control (Shenhav, Botvinick, & Cohen, 2013; Taren, Venkatraman, & Huettel, 2011; Egner, 2009; Fellows & Farah, 2005; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004), raising the question of whether the currently observed effects in this region might reflect similar processes. In particular, judgments about individual causal influences may be associated with efforts to retrieve instances of the queried cause occurring in isolation from episodic memory—a search that is futile and thus presumably more effortful for confounded causes. However, such retrieval-induced control demands are unlikely to account for the current effects in the DMPFC for several reasons: First, effort associated with episodic memory retrieval does not address responses to stochasticity—in contrast, the causal uncertainty account explains why DMPFC responses are greater for both confounded and stochastic causes, relative to nonconfounded and deterministic causes. It may be argued, of course, that stochasticity elicits additional control processes mediated by the DMPFC, such as, for example, response conflict (Kerns et al., 2004; Ridderinkhof et al., 2004; Botvinick, Braver, Barch, Carter, & Cohen, 2001). Recall, however that none of the target causes (C, S, and D) were presented in isolation during feedback-based learning and that, consequently, all target causes should involve some level of searching for nonpresented information during causal judgments. If DMPFC activity reflected this search process as well as response conflict due to stochasticity, one would expect a difference between neural responses to causes S and s', both of which are stochastic, but only one of which, s', was presented in isolation during feedback-based learning, eliminating the need to search for an unexperienced memory. Notably, although RTs (Table 1), commonly used as a measure of cognitive load, are indeed consistent with the notion of an increased control demand for cause S over s', the DMPFC responses (Figure 4) are not. Finally, and perhaps most conclusively, the inclusion of RTs as a parametric modulator of judgment uncertainty in first-level models and as a covariate in group analyses did not have any impact on the neural effects of causal judgment uncertainty. Taken together, these aspects of the results make causal uncertainty the more parsimonious explanation of the DMPFC effects observed here.

As with the DMPFC and DLPFC, the anterior insula has been implicated in outcome stochasticity or "risk" in several previous neuroimaging studies (e.g., d'Acremont, Fornari, & Bossaerts, 2013; Preuschoff, Quartz, & Bossaerts, 2008; Huettel et al., 2005; Volz et al., 2003; Critchley, Mathias, & Dolan, 2001). Here, the dorsal anterior right insula was selectively recruited by the error-driven model of uncertainty during the prediction period of feedback-based learning trials. Plots of effect sizes for each type of trial (in Figure 5) suggest that these responses were indeed driven by stochasticity-based uncertainty about the state of the outcome on any given trial. Moreover, the fact that, unlike the DMPFC, activity in the anterior insula did not scale with either an error-driven or causal measure of uncertainty during judgments about individual causal influences indicates that representations in this region pertain primarily to the immediate consequences of choice, rather than generalizable knowledge. Consistent with this interpretation, the bilateral anterior insula also responded selectively to stochastic conditions during the outcome period of feedback-based learning trials, suggesting that this region is recruited by both the anticipation and detection of errors. It is difficult to discern based on the current results whether the unsigned error signals observed in the anterior insula are

Bayesian or model-free, they are consistent with a substantial literature implicating this region in expectancy violations across a wide range of tasks (Bastin et al., 2017; Allen et al., 2016; Metereau & Dreher, 2012; Klein et al., 2007).

Selective effects of the error-driven account of predictive uncertainty were also identified throughout large portions of the posterior lingual gyrus. In a recent neuroimaging study, Causse et al. (2013) assessed decision-making under uncertainty using an aviation task in which participants had to either land a plane or abort landing, given certain (100% or 0%) versus uncertain (50%) information regarding landing conditions. Consistent with the current results, they found greater activity in the posterior lingual gyrus during high than during low uncertainty conditions. Our results are also consistent with those of Payzan-LeNestour, Dunne, Bossaerts, and O'Doherty (2013), who found that the lingual gyrus scaled specifically with uncertainty due to outcome stochasticity, as opposed to uncertainty due to changes in the statistical structure of the environment or due to ambiguity. Payzan-LeNestour et al. argued that the lingual gyrus might not typically emerge in studies assessing outcome stochasticity because those studies did not dissociate stochasticity from other uncertainty components. A similar case can be made here, that the inclusion of an alternative, generative, model that captures additional aspects of uncertainty may have increased sensitivity to detect stochasticity-specific signals in the lingual gyrus.

An important aspect of the current study is the direct comparison of a Bayesian inference model with an error-driven, model-free algorithm. Although not uncommon (Prévost, McNamee, Jessup, Bossaerts, & O'Doherty, 2013; Payzan-LeNestour & Bossaerts, 2011; Courville, Daw, & Touretzky, 2006; Hampton, Bossaerts, & O'doherty, 2006), the contrasting of these quite different computational frameworks somewhat complicates interpretation. In particular, it raises the question of whether selective effects of the causal account of uncertainty indeed reflect sensitivity to confounding or instead some extraneous, perhaps incidental, feature of the different modeling frameworks. Here, a significant clue is provided by neural and behavioral responses to each queried putative cause, shown in Figure 4 and Table 1, respectively, which clearly correspond to the increased uncertainty associated with confounded and stochastic causes predicted by the causal model (cf. Figure 3). It is worth noting that, far from incidental, the failure of the error-driven account to predict sensitivity to confounding is intrinsic to the model-free approach, because of its lack of a representation of the independence of causal influences. Critically, a configural solution (e.g., Pearce, 1987, 2002), such that individually queried causes are treated as non-overlapping with the compounds in which they occur during feedback-based learning, would fail to account for both behavioral and neural differences between confounded and deterministic target causes, both of which occurred only in compound with alternative causes during feedback-based learning.

Although the effects of causal uncertainty in the DMPFC likely reflect sensitivity to causal confounding, other neuroimaging results are clearly unrelated and, in some cases, largely incidental to the specific model implementation. In particular, as can be seen in Figures 6 and 7, the ordinal mean values derived from causal and error-driven accounts across trial types during feedback-based learning are identical for both strength and surprise measures, suggesting that model selection may reflect more granular differences. For example, because of the ΣV term in Equations 1 and 2 of the error-driven account, the variance in the occurrence of the outcome across e−, De+, and Se± trials results in a constantly fluctuating strength and persistent prediction errors, even on deterministic e− and De+ trials. In contrast, in the implementation of the causal model, such non-conditional variance is eliminated by the use of a focal set (e.g., Spellman, 1996; Cheng & Novick, 1992). Behaviorally, the focal set assumption is strongly supported by the lack of variability in judgments about non-target cause e. This cause, which is paired with the outcome on half of the trials in which it occurs, but never when occurring in isolation, is almost exclusively judged to be noncausal, suggesting that participants conditioned their inferences about its influence on trials across which alternative causes could be assumed to occur with constant probabilities. Nonetheless, at the neural level, several regions may instead have been tracking the unconditional outcome variance computed by the error-driven model.

As the mantra "covariation does not equal causation" implies, requirements for causal induction extend far beyond mere statistical regularity. Countless studies have demonstrated neural and behavioral effects of error-driven uncertainty, whether due to stochasticity, insufficient samples, or changes in the statistical structure of the environment. The current study reveals that, given violation of a basic boundary condition on causal inference, and in spite of the same constraints on perceptual generalization, differences in neural representations of uncertainty can emerge across equally deterministic, static, and well-sampled contingencies. These results contribute to a growing literature on the role of generative models in neural representations of uncertainty and shed light on a possible neural implementation of domain-general, a priori constraints on causal induction.

Acknowledgments

This work was supported by a start-up fund from the University of California, Irvine, to M. L.

Reprint requests should be sent to Mimi Liljeholm, Department of Cognitive Sciences, University of California, Irvine, 2312 Social and Behavioral Sciences Gateway, Irvine, CA 92697-5100, or via e-mail: m.liljeholm@uci.edu.

REFERENCES

Abler, B., Herrnberger, B., Grön, G., & Spitzer, M. (2009). From uncertainty to reward: BOLD characteristics differentiate signaling pathways. *BMC Neuroscience*, *10*, 154.

Allen, M., Fardo, F., Dietz, M. J., Hillebrandt, H., Friston, K. J., Rees, G., et al. (2016). Anterior insula coordinates hierarchical processing of tactile mismatch responses. *Neuroimage*, *127*, 34–43.

Bach, D. R., Seymour, B., & Dolan, R. J. (2009). Neural activity associated with the passive prediction of ambiguity and risk for aversive events. *Journal of Neuroscience*, *29*, 1648–1656.

Bastin, J., Deman, P., David, O., Gueguen, M., Benis, D., Minotti, L., et al. (2017). Direct recordings from human anterior insula reveal its leading role within the error-monitoring network. *Cerebral Cortex*, *27*, 1545–1557.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624–652.

Causse, M., Péran, P., Dehais, F., Caravasso, C. F., Zeffiro, T., Sabatini, U., et al. (2013). Affective decision making under uncertainty during a plausible aviation task: An fMRI study. *Neuroimage*, *71*, 19–29.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*, 365–382.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*, 294–300.

Critchley, H. D., Mathias, C. J., & Dolan, R. J. (2001). Neural activity in the human brain relating to uncertainty and arousal during anticipation. *Neuron*, *29*, 537–545.

d'Acremont, M., Fornari, E., & Bossaerts, P. (2013). Activity in inferior parietal and medial prefrontal cortex signals the accumulation of evidence in a probability learning task. *PLoS Computational Biology*, *9*, e1002895.

Durnez, J., Degryse, J., Seurinck, R., Moerkerke, B., & Nichols, T. E. (2015). Prospective power estimation for peak inference with the toolbox neuropower. In *Second Belgian neuroinformatics congress* (Vol. 9). Frontiers Media SA.

Egner, T. (2009). Prefrontal cortex and cognitive control: Motivating functional hierarchies. *Nature Neuroscience*, *12*, 821–822.

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: A model of attention in associative learning. *Proceedings of the Royal Society of London: Series B: Biological Sciences*, *278*, 2553–2561.

Fellows, L. K., & Farah, M. J. (2005). Is anterior cingulate cortex necessary for cognitive control? *Brain*, *128*, 788–796.

Gläscher, J. (2009). Visualization of group inference data in functional neuroimaging. *Neuroinformatics*, *7*, 73–82.

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

Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, *49*, 757–763.

Hampton, A. N., Bossaerts, P., & O'doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience*, *26*, 8360–8367.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*, 1644–1655.

Huettel, S. A., Song, A. W., & McCarthy, G. (2005). Decisions under uncertainty: Probabilistic context influences activity of prefrontal and parietal cortices. *Journal of Neuroscience*, *25*, 3304–3311.

Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., & Platt, M. L. (2006). Neural signatures of economic preferences for risk and ambiguity. *Neuron*, *49*, 765–775.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science*, *310*, 1680–1683.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*, 1295–1306.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science*, *303*, 1023–1026.

Klein, T. A., Endrass, T., Kathmann, N., Neumann, J., von Cramon, D. Y., & Ullsperger, M. (2007). Neural correlates of error awareness. *Neuroimage*, *34*, 1774–1781.

Krain, A. L., Wilson, A. M., Arbuckle, R., Castellanos, F. X., & Milham, M. P. (2006). Distinct neural mechanisms of risk and ambiguity: A meta-analysis of decision-making. *Neuroimage*, *32*, 477–484.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science*, *16*, 678–683.

Levy, I., Snell, J., Nelson, A. J., Rustichini, A., & Glimcher, P. W. (2010). Neural representation of subjective value under risk and ambiguity. *Journal of Neurophysiology*, *103*, 1036–1047.

Liljeholm, M. (2015). How multiple causes combine: Independence constraints on causal inference. *Frontiers in Psychology*, *6*, 1135.

Liljeholm, M., & Cheng, P. W. (2009). The influence of virtual sample size on confidence and causal-strength judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 157–172.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*, 955–984.

Meder, B., Hagmayer, Y., & Waldmann, M. (2006). Understanding the causal logic of confounds. In R. Sun (Ed.), *Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 579–584). Mahwah, NJ: Erlbaum.

Metereau, E., & Dreher, J. C. (2012). Cerebral correlates of salient prediction error for different rewards and punishments. *Cerebral Cortex*, *23*, 477–487.

Michael, E., de Gardelle, V., Nevado-Holgado, A., & Summerfield, C. (2015). Unreliable evidence: 2 sources of uncertainty during perceptual choice. *Cerebral Cortex*, *25*, 937–947.

Mohr, P. N., Biele, G., & Heekeren, H. R. (2010). Neural processing of risk. *Journal of Neuroscience*, *30*, 6613–6619.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*, 1–25.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology*, *7*, e1001048.

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., & O'Doherty, J. P. (2013). The neural representation of unexpected uncertainty during value-based decision making. *Neuron*, *79*, 191–201.

Pearce, J. M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, *94*, 61–73.

Pearce, J. M. (2002). Evaluation and development of a connectionist theory of configural learning. *Animal Learning & Behavior*, *30*, 73–95.

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, *28*, 2745–2752.

Prévost, C., McNamee, D., Jessup, R. K., Bossaerts, P., & O'Doherty, J. P. (2013). Evidence for model-based computations in the human amygdala during Pavlovian conditioning. *PLoS Computational Biology*, *9*, e1002918.

Pushkarskaya, H., Smithson, M., Joseph, J. E., Corbly, C., & Levy, I. (2015). Neural correlates of decision-making under ambiguity and conflict. *Frontiers in Behavioral Neuroscience*, *9*, 325.

Richiardi, J., Altmann, A., Milazzo, A. C., Chang, C., Chakravarty, M. M., Banaschewski, T., et al. (2015). Correlated gene expression supports synchronous activity in brain networks. *Science*, *348*, 1241–1244.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*, 443–447.

Rosa, M. J., Bestmann, S., Harrison, L., & Penny, W. (2010). Bayesian model selection maps for group studies. *Neuroimage*, *49*, 217–224.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*, 1045–1050.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*, 322–332.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*, 217–240.

Spellman, B. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science*, *7*, 337–342.

Taren, A. A., Venkatraman, V., & Huettel, S. A. (2011). A parallel functional topography between medial and lateral prefrontal cortex: Evidence and implications for cognitive control. *Journal of Neuroscience*, *31*, 5026–5031.

Volz, K. G., Schubotz, R. I., & von Cramon, D. Y. (2003). Predicting events of varying probability: Uncertainty investigated by fMRI. *Neuroimage*, *19*, 271–280.

Xue, G., Lu, Z., Levin, I. P., Weller, J. A., Li, X., & Bechara, A. (2008). Functional dissociations of risk and reward processing in the medial prefrontal cortex. *Cerebral Cortex*, *19*, 1019–1027.