Neural Correlates of Causal Confounding

Mimi Liljeholm

Department of Cognitive Sciences, University of California, Irvine

Corresponding author:

Mimi Liljeholm

Department of Cognitive Sciences

2312 Social and Behavioral Sciences Gateway

University of California

Irvine, CA, 92697-5100

E-mail: m.liljeholm@uci.edu

Keywords: Causal induction, Uncertainty, Bayesian inference, prediction-errors, fMRI

Abstract

As scientists, we are keenly aware that if putative causes perfectly co-vary, 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, we used a combination of cognitive modeling and fMRI to investigate neural substrates mediating such sensitivity. While being scanned, participants observed and judged the influences of various fictitious allergy medicines on headache. Formal comparisons of a Bayesian causal model, sensitive to both confounding and stochasticity, and a purely error-driven algorithm, sensitive only to stochasticity, revealed dissociable neural representations of predictive uncertainty across several regions, including the insula, dlPFC and hippocampus. 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 relevant 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 co-variation of C and A in the second scenario occludes the independent influences 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 (Spellman, 1996; Kushnir and Gopnik, 2005; Meder et al., 2006; Schulz and Bonawitz, 2007; Schulz et al., 2007; Liljeholm, 2015). In contrast, very little is known 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 anterior insula, dorsomedial

prefrontal cortex (DMPFC) and dorsolateral prefrontal cortex (DLPFC) in the neural representation of outcome stochasticity (Volz et al., 2003; Huettel et al., 2005; Grinband et al., 2006; Abler et al., 2009). An important open question is whether, or which of, these neural regions also encode uncertainty associated with causal confounding.

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 cues predictive strength is captured by the "Pearce-Hall" (PH) algorithm, which relates the "associability", α, of a cue to a weighted average of the absolute prediction error on previous trials involving that cue:

$$\alpha_n = \gamma\left|\lambda_{n-1} - \sum V_{n-1}\right| + \left(1-\gamma\right)\alpha_{n-1} \qquad (1)$$

where ΣV_{n-1} is the expectation about the outcome summed over all present cues, λ_{n-1} is the observed state of the outcome on that trial, and γ is a free parameter accounting for the weighting of the previous trial relative to preceding ones. The estimated quantity, α, is proportional to the stochasticity of the outcome.

Conversely, in Bayesian Causal Models (BCM), uncertainty about the predictive strength of a cause is reflected in the entropy of the posterior distribution over its possible strengths, which depends on the variance in the effect variable, but also on the independent occurrence of putative causes. Specifically, for a scenario involving two candidate causes (e.g., C and A in the example above), reasoners make

judgments based on possible causal graphs (see Figure 1A) in which neither, either or both C and A cause the effect, with all graphs assuming a constantly present background cause of the effect, B. The likelihood of a particular graph given the observed data is derived using a particular parameterization that specifies how causes combine to influence their effects. The most commonly used parameterization for generative causes is the noisy-OR (Cheng, 1997; Glymour, 1998). For the graph in which both C and A exert a generative influence on the effect, e, the probability of e is:

$$P\left(e^{+}\middle|b,c,a;w_B,w_C,w_A\right)=1-\left(1-w_B^b\right)\times\left(1-w_C^c\right)\times\left(1-w_A^a\right) \quad (2)$$

where c and a are 0 or 1 depending on the absence vs. presence of the candidate and alternative cause, b, the background cause, is always 1 (i.e., present), and w_B, w_C and w_A are the respective strengths of putative.

While both the associative and Bayesian causal model predicts sensitivity to stochasticity, only the causal model accounts for uncertainty due to causal confounding. A critical aspect of causal confounding is that, provided a sufficient sample of trials with a particular confounded configuration, there is no logical or empirical reason to expect a prediction about the outcome given that stimulus configuration to be inaccurate. This property uniquely distinguishes causal confounding, not just from stochasticity, which scales linearly with outcome prediction errors, but also from other sources of uncertainty, such as ambiguity and volatility. Of particular interest, therefore, is whether uncertainty due to causal confounding is discernable during feedback-based learning, as participants make predictions, and receive feedback, about the occurrence of the effect given a confounded cause. Here, a combination of neuroimaging and cognitive modeling is used to dissociate neural

signals scaling with error-driven uncertainty from those encoding uncertainty associated with causal confounding.

Materials & Methods

Participants. 20 healthy normal volunteers (mean age = 20.9 ± 2.4, range: 18-27, 12 females) participated in the study. In our previous work (e.g., Liljeholm et al., 2011; Liljeholm et al., 2013), this sample size has proven sufficient to generate effects based on a univariate model-based fMRI analysis at appropriate corrected thresholds. One participant was excluded prior to any analyses due to excessive head movement (> 5mm), leaving a sample size of 19. The volunteers were pre-assessed to exclude those with a history of neurological or psychiatric illness. All subjects gave informed consent and the study was approved by the Institutional Review Board of the University of California, Irvine.

Task & procedure. Participants were scanned with functional MRI 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. The study focused on three target medicines, C, D and S, mnemonically labeled here to indicate “confounded”, “deterministic” and “stochastic” influences respectively, with each target medicine

occurring only in combination with some alternative medicine during feedback-based learning. Target conditions, relevant events and model predictions are listed in Table 1. With +, - and ± respectively indicating a 1.0, 0 and 0.5 probability of the effect, and with lower case letters indicating non-target causes, there were 5 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.

Table 1: Conditions, relevant events and model predictions about uncertainty during feedback-based learning. C, D and S respectively indicate confounded, deterministic and stochastic target causes; e and c' indicate non-target compound elements; +, - and ± respectively indicate a 1.0, 0.0 and 0.5 probability of the effect.

Condition	Relevant events	Target compounds	Uncertainty	
			Causal	Error-driven
Confounded	n-, Cc'+	Cc'	6.97	0.19
Deterministic	n-, e-, De+	De	4.26	0.24
Stochastic	n-, e-, Se±	Se	6.29	0.48

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 0.5 rather than 1.0; a

136 comparison of medicines S and D, therefore, identifies a difference between
137 deterministic and stochastic causation. Finally, medicine s' also produced headache
138 with a probability of 0.5 but unlike medicine S, never occurred in combination with
139 any other medicine; contrasting S with s' identifies differences due to compound
140 presentation.

141 On each trial of feedback-based learning (see Figure 1A), participants were
142 presented with an individual allergy patient and were told either that the patient had
143 not received any medicine or that a particular medicine, or combination of medicines,
144 had been administered. Medicines were color coded and labeled accordingly (i.e.,
145 Medicine "B" was blue) with color assignments counterbalanced across participants.
146 On the first screen, the allergy patient's state was obscured by a question mark and
147 participants were asked to indicate whether or not the patient had a headache,
148 pressing the ''Y'' key for ''yes'' and the ''N'' key for ''no". After a prediction was
149 made, there was a brief (2000ms) pause during which the word "wait" was displayed
150 on the screen, followed by a screen revealing the allergy patient's state (i.e., with or
151 without headache). Each of 6 distinct trial types was presented 8 times in each of 4
152 sessions, for a total of 192 trials, with sessions separated by 2-minute breaks, during
153 which the scanner was turned off. Each session was further divided into 4, non-
154 delineated, blocks, with each type of trial occurring once in random order in each
155 block. All trials were separated by a jittered 4-second ITI. In every other block,
156 starting with the 2nd block of the 1st session, each trial (except for n- trials) was
157 followed by a query regarding the causal influence of the target medicine presented
158 on that trial (i.e., C, D or S), or of medicine e or s' on non-target 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 “may or may not be a cause of headache” (see Figure 1B).

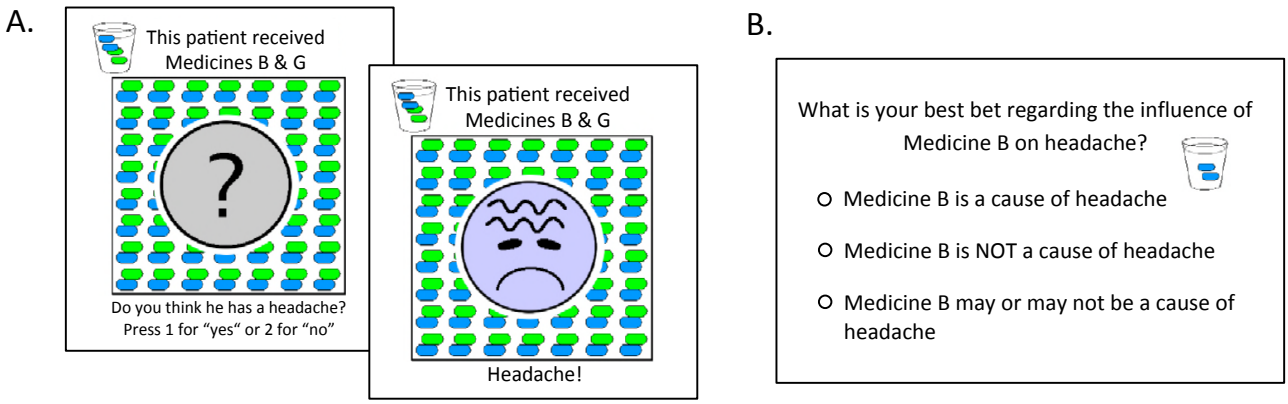

Figure 1. Task illustration. A. Prediction and outcome screens from a feedback-based learning trial. Screens were separated by a 2 sec. pause during which the word “Wait” (not shown in figure) was displayed. B. Causal query regarding the individual influence of a particular medicine on headache.

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:

$$V_{n+1} \leftarrow V_n + \alpha_n(\lambda_n - \sum V_n) \qquad (3)$$

where V is the predictive strength of a cue, and all other terms, including α_n, are defined as in Equation 1 above. In addition to scaling the influence of the prediction error on predictive strength, α is taken to reflect current levels of uncertainty about the cues influence. Note that, because of the Σ term in Equations 1 and 3, both the uncertainty (α), and strength (V) of a cue depend on the effects of other cues with

which it is paired. Specifically, in the current study, the variance in the occurrence of headache across e- and De+ trials, and on Se± trials, results in a constantly fluctuating strength of Medicine e and, consequently, in persistent prediction errors, and corollary changes to strength and uncertainty, on e-, as well as De+, trials (see model predictions in Table 1).

The error-driven approach further assumes that, during feedback-based learning, participants select "Yes"/"No" responses to questions about whether the outcome will be present using probabilities generated by a softmax distribution, in which a free parameter controls the influence of predictive strength (i.e., the strength with which a cue predicts the outcome) on choice behavior. Free parameters were fit to behavioral data by minimizing the negative log likelihood of observed responses for each individual. 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, while 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 (BCM), in which reasoners make inferences about which of several causal structures ($G1$-$G4$) is responsible for the observed data (Griffiths & Tenenbaum, 2005; Liljeholm, 2015). Support for a given target cause, C, is defined as the log likelihood ratio in favor of graphs in which there is a link between that cause and the effect (e.g., $G1$ and $G3$), over graphs in which there is no such link (e.g., $G0$ and $G2$).

$$\sup_{C\to E} = \log \frac{P(D\,|\,G1) + P(D\,|\,G3)}{P(D\,|\,G0) + P(D\,|\,G2)}. \tag{4}$$

In the current study, sequential inferences were modeled using the smallest possible *focal set* (Novick & Cheng, 1990) – that is, the smallest set of events relevant for making the inference. Thus, inferences about elemental (non-target) candidate causes s' and e were modeled based on trials with those respective causes and the "no causes" trials (i.e., n- and s'± for candidate s', and n- and e- for candidate e), using only $G0$ and $G1$. For each target candidate cause, which always occurred in compound with some alternative cause (see Table 1), inferences were modeled over all four graphs (as in Equation 4) 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). Each graph has a set of parameters θ, which are the causal strengths w of the background (w_B), target candidate cause (w_C) and alternative candidate cause (w_A) respectively. The likelihoods, $P(D|G_i)$, are obtained by averaging out the relevant unknown parameters, such that, for $G0$

$$P(D\,|\,G0) = \int_0^1 P(D\,|\,w_B, G0)P(w_B\,|\,G0)dw_B, \tag{5}$$

for $G1$ (or $G2$, replacing C with A)

$$P(D\,|\,G1) = \int_0^1\int_0^1 P(D\,|\,w_B, w_C, G1)P(w_B, w_C\,|\,G1)\,dw_B\,dw_C, \tag{6}$$

and for G3

$$P(D\,|\,G3) = \int_0^1\int_0^1\int_0^1 P(D\,|\,w_B, w_C, w_A, G3)P(w_B, w_C, w_A\,|\,G3)\,dw_B\,dw_C\,dw_A, \tag{7}$$

where $P(\theta|G_i)$ denotes the priors on causal strength parameters.

The likelihoods $P(D|\theta, G_i)$ are computed using the 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 $G3$ is:

$$P\left(D|w_B,w_C,w_A,G3\right)=\begin{pmatrix} N\left(c^-,a^-\right) \\ N\left(e^+,c^-,a^-\right)\end{pmatrix}\times\begin{pmatrix} N\left(c^+,a^-\right) \\ N\left(e^+,c^+,a^-\right)\end{pmatrix}\times\begin{pmatrix} N\left(c^-,a^+\right) \\ N\left(e^+,c^-,a^+\right)\end{pmatrix}\times\begin{pmatrix} N\left(c^+,a^+\right) \\ N\left(e^+,c^+,a^+\right)\end{pmatrix} w_B^{N\left(e^+,c^-,a^-\right)}$$
$$(1-w_B)^{N\left(e^-,c^-,a^-\right)}[w_B+w_C-w_Bw_C]^{N\left(e^+,c^+,a^-\right)}[1-w_B-w_C+w_Bw_C]^{N\left(e^-,c^+,a^-\right)}[w_B+w_A-w_Bw_A]^{N\left(e^+,c^-,a^+\right)}[1-w_B-w_A+w_Bw_A]^{N\left(e^-,c^-,a^+\right)} \quad (8)$$
$$[w_B+w_C+w_A-w_Bw_C-w_Bw_A-w_Cw_A+w_Bw_Cw_A]^{N\left(e^+,c^+,a^+\right)}[1-w_B-w_C-w_A+w_Bw_C+w_Bw_A+w_Cw_A-w_Bw_Cw_A]^{N\left(e^-,c^+,a^+\right)}.$$

where $\binom{n}{k}$ 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 $G0$, $G1$ and $G2$ 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 link to the effect in the relevant graph.

Causal strength estimates for each elemental and target cause were modeled as the mean of the marginal posterior distribution $P(w_C|D)$ from the most likely graph in which a link existed between the cause and the effect. The marginal posterior distribution is obtained by applying Bayes' rule, such that, if for example $G3$ had the highest likelihood, then

$$P(w_C \mid D,G3)=\int_0^1\int_0^1\frac{P(D \mid w_B,w_C,w_A,G3)P(w_B,w_C,w_A \mid G3)}{P(D)}dw_B\,dw_A, \quad (9)$$

where $P(D|w_B,w_C,w_A,G3)$ is the likelihood term, $P(w_B,w_C,w_A|G3)$ refers, again, to the prior probabilities of causal strength parameters, and $P(D)$ is the normalizing term, denoting the probability of the observed data. The mean of w_C is given by

$$\bar{w}_C = \int_0^1 w_C P(w_C \mid D)\, dw_C. \tag{10}$$

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

$$H(w_C) = -\int_0^1 P(w_C \mid D) \log P(w_C \mid D)\, dw_C. \tag{11}$$

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, and using the posterior on the previous trial as the prior 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, $Beta(1,1)$, 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 was computed as:

$$KL(P\|Q) = \int_0^1 p(w_C \mid D) \log \frac{p(w_C \mid D)}{q(w_C \mid D)}\, dw_C. \tag{12}$$

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, while predictions regarding the occurrence of the effect were generated using the noisy-OR integration of strengths specified in Equation 2.

Imaging procedure and analysis. A 3 Tesla scanner (Phillips Achieva) was used to acquire structural T1-weighted images and T2*-weighted echoplanar images (repetition time = 2.65 s; echo time = 30 ms; 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, full width at half-maximum). High-pass filtering with a cutoff of 128 s was used.

For each subject, a general linear model (GLM) was specified, with regressors in each session modeling 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. Prediction-screen onsets and causal judgment onsets were modulated, in order, by trial-by-trial error-driven and causal estimates of uncertainty and strength. Outcome-screen onsets were modulated by trial-by-trial error-driven and causal estimates of surprise. No orthogonalization was applied. Two separate regressors modeled the motor responses on each feedback-based and causal query trial, respectively, and six additional regressors accounted for the residual effects of head motion. All regressors were convolved with a canonical

hemodynamic response function. To contrast error-driven and causal accounts, disjunction tests were performed, assessing neural effects that were significant, based on whole brain correction, given one (e.g., the causal) account but that did not emerge, in either a positive or negative direction, given the alternative (e.g., error-driven) account even at an uncorrected threshold of $p<0.1$.

To address loss of power due to multicollinearity, Bayesian model selection (BMS) analyses were used to further arbitrate between error-driven and causal accounts. Here, for each subject, two separate GLMs were specified, one for the error-driven approach and another for the causal approach. Thus, these GLMs were identical to that described above, except that only error-driven variables were entered as parametric modulators in the first GLM, and only "causal" variables in the second GLM. A first-level Bayesian estimation procedure (SPM 12) was used to compute a log-model evidence map for every subject and each GLM (Penny et al., 2005). Inferences at the group level were modeled by applying a random effects approach (Rose et al., 2010) at every voxel of the log evidence data falling within a particular functional mask, specified based on the results of classical estimation (e.g., inferences about uncertainty during prediction screens were constrained by a mask of all voxels that reached significance using whole-brain (CST) correction with either the error-driven or causal GLM). Group-level exceedance probability maps, reflecting the probability that one model is more likely than the others, were thresholded to identify regions in which the exceedance probability was greater than 0.95.

While it effectively addresses colinearity between causal and error-driven models, the BMS pertains to *all predictions* made by a particular model, discriminating neither

297 between different computational variables (i.e., uncertainty, strength and surprise) nor
298 between different experimental events (i.e., feedback-based learning vs. explicit
299 causal judgments). Given this limitation, we constrain our reporting of BMS results in
300 the text to whether they confirm the disjunction results (for completeness, a full lists of
301 all BMS results are provided in Supplementary Tables 1-4), and focus our discussion
302 on regions that are identified across both disjunction and BMS tests. To further
303 facilitate the interpretation of disjunction results, we plot neural responses to each
304 target cause, as well as to increasing values of computational variables, during
305 feedback-based learning, extracting effect sizes from 8mm spheres centered on peak
306 MNI coordinates, using rfxplot (Glascher, 2009).

307 **Results**

308 **Behavioral results**

309 All statistical tests of behavioral data were planned comparisons, employing two-tailed
310 t-tests, and were calculated using n=19. Confidence intervals (*CI*) and effect sizes
311 (Cohen's *d*) are reported for all comparisons.

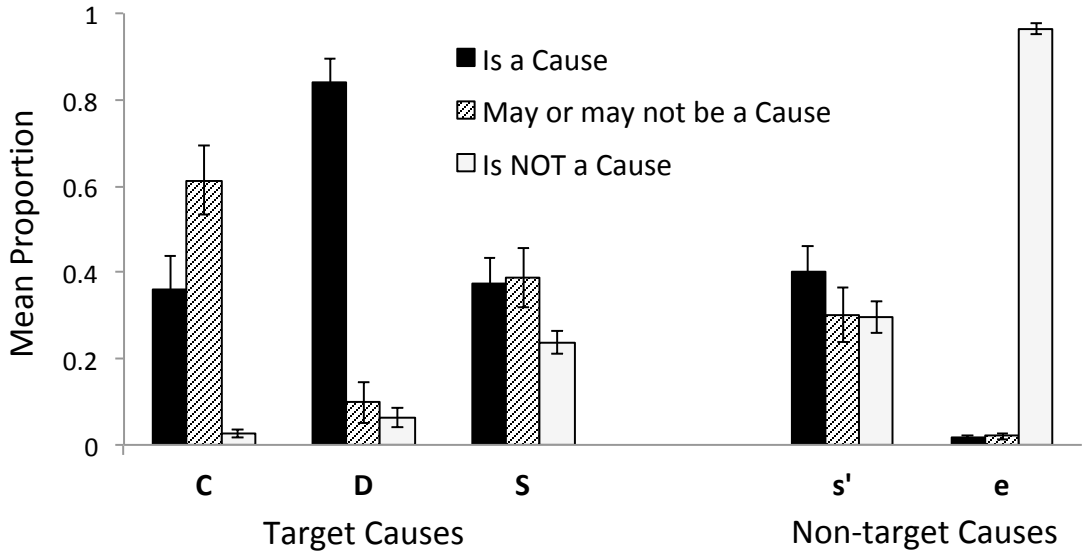

Figure 2. Mean proportions of "Is cause", "Is not a cause" and ambivalent "May or may not be cause" judgments about the individual influence of each medicine on headache. Target causes (C, D & S) occurred only in compound with alternative causes during feedback-based learning, while non-target causes occurred only individually (s'), or individually as well as in compound (e).

The mean proportions of "is a cause", "is not a cause" and, ambivalent, "may or may not be a cause" judgments, averaged across multiple probes for each probed medicine, are shown in Figure 2. Planned comparisons of target causes revealed that the mean proportion of ambivalent judgments were significantly lower for the deterministic (D) medicine than for both confounded (C), t(18)=6.78, $p<0.001$, *CI*=[0.35 0.67], $d=1.79$, and stochastic (S), t(18)=4.09, $p<0.001$, *CI*=[0.14 0.44], $d=1.11$, medicines. The difference between the confounded and stochastic medicine was also significant, with the mean proportion of ambivalent judgments being greater for C than for S t(18)=2.65, $p<0.05$, *CI*=[0.0.05 0.40], $d=0.68$. Recall that all target medicines, including S, occurred only in compound with alternative medicines, controlling for cue complexity across target comparisons. The proportion of ambivalent judgments for the two stochastic causes, one always occurring on its own (s') and the

325 other always in a compound (S) during feedback based learning, was only marginally
326 significant, *p=0.08*, suggesting that cue complexity did not strongly influence
327 perceived causal ambiguity.

328 **Neuroimaging results**

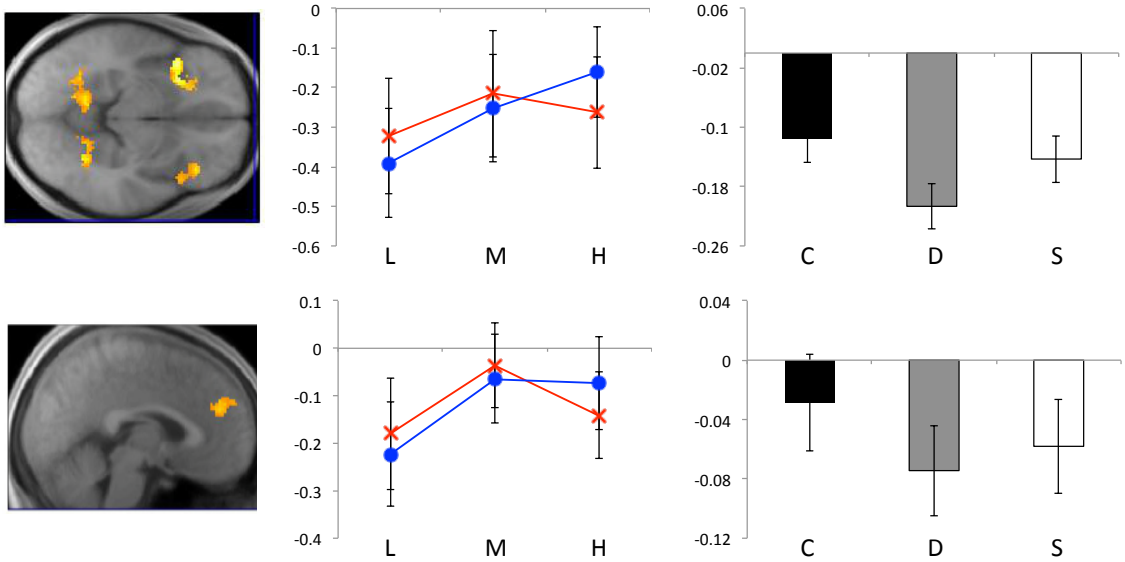

Figure 3. Disjunction results favoring the causal model of uncertainty during the prediction period of feedback-based learning trials, showing effects in the anterior insula (top) and DMPFC (bottom). Plotted values show effect sizes (y-axis) extracted from spheres centered on peak coordinates at low (L), medium (M) and high (H) levels of causal (blue) and error-driven (red) models (middle column), and for confounded (C), deterministic (D) and stochastic (S) compound causes (right column). Error bars=SEM.

329

330 *Uncertainty in causal prediction and judgment.* During feedback-based learning
331 (Figure 3 and Table 2), as participants generated a prediction about whether or not
332 headache would occur given a particular medicine or combination of medicines,
333 disjunction analyses revealed significant effects of the causal, but not the error-driven,
334 measure of uncertainty in the bilateral anterior insula, bilateral fusiform gyrus, anterior
335 cingulate and rostrolateral prefrontal cortex.

Table 2: Results of disjunction tests in favor of the causal model. Peak MNI coordinates and cluster sizes (>= 142) at p<0.005. R/L = Right/Left, (-) = negative correlation, ‘ ‘ = continuous with previous cluster.

Contrast	Area	x, y, z	Cluster size at p<0.005
Prediction Uncertainty	L. Fusiform Gyrus	-34, -50, -20	858
	L. Anterior Insula	-40, 16, -8	672
	L. Rostrolateral Prefrontal Cortex	-34, 52, 12	‘ ‘
	R. Fusiform Gyrus	-38, -50, -14	202
	R. Anterior Insula	38, 30, -6	166
	Anterior Cingulate Cortex	8, 44, 22	233
Prediction Strength	Calcarine/Lingual	14, -92, -4	1109
Surprise	R. Angular Gyrus	54, -60, 32	1564
	R. Middle Temporal Gyrus	64, -46, 10	‘ ‘
	R. Anterior Insula	38, 18, 0	1294
	R. Inferior Frontal Gyrus	34, 14, 30	‘ ‘
	L. Anterior Insula	-30, 22, 0	952
	Anterior Cingulate Cortex	8, 44, 26	733
	Ventral Tegmental Area	4, -22, -26	622
Surprise (-)	L. Caudate	-14, 28, -2	846
	R. Superior Parietal Lobule	20, -58, 60	1086
	R. Hippocampus	32, -38, 4	241
	L. Parahippocampus	-34, -42, 0	258
	R. Supramarginal Gyrus	58, -20, 38	243
Judgment Uncertainty	Dorsomedial Prefrontal Cortex	16, 58, 32	2507
	R. Dorsolateral Prefrontal Cortex	32, 22, 28	143

Conversely, the error-driven, but not causal, measure of predictive uncertainty (Figure 4 & Table 3) was significantly positively correlated with activity in large aspects of the lingual gyrus, the left inferior parietal lobule, and the dorsolateral prefrontal cortex, as well as negatively correlated with activity in the bilateral middle temporal gyrus (MTG),

the right temporal-parietal junction (TPJ) and ventromedial prefrontal cortex (vmPFC). Note, however, that the negative correlation in the vmPFC reflects increased activity in response to the confounded cause (bottom right of Figure 4), suggesting that this region is in fact selectively responding to causal confounding. The BMS analysis (Supplemental Table 1) confirmed the selective effects of the causal model in the bilateral insula, anterior cingulate and fusiform gyri, as well as the selective effects of the error-driven algorithm in the vmPFC, lingual gyrus and dorsolateral prefrontal cortex.

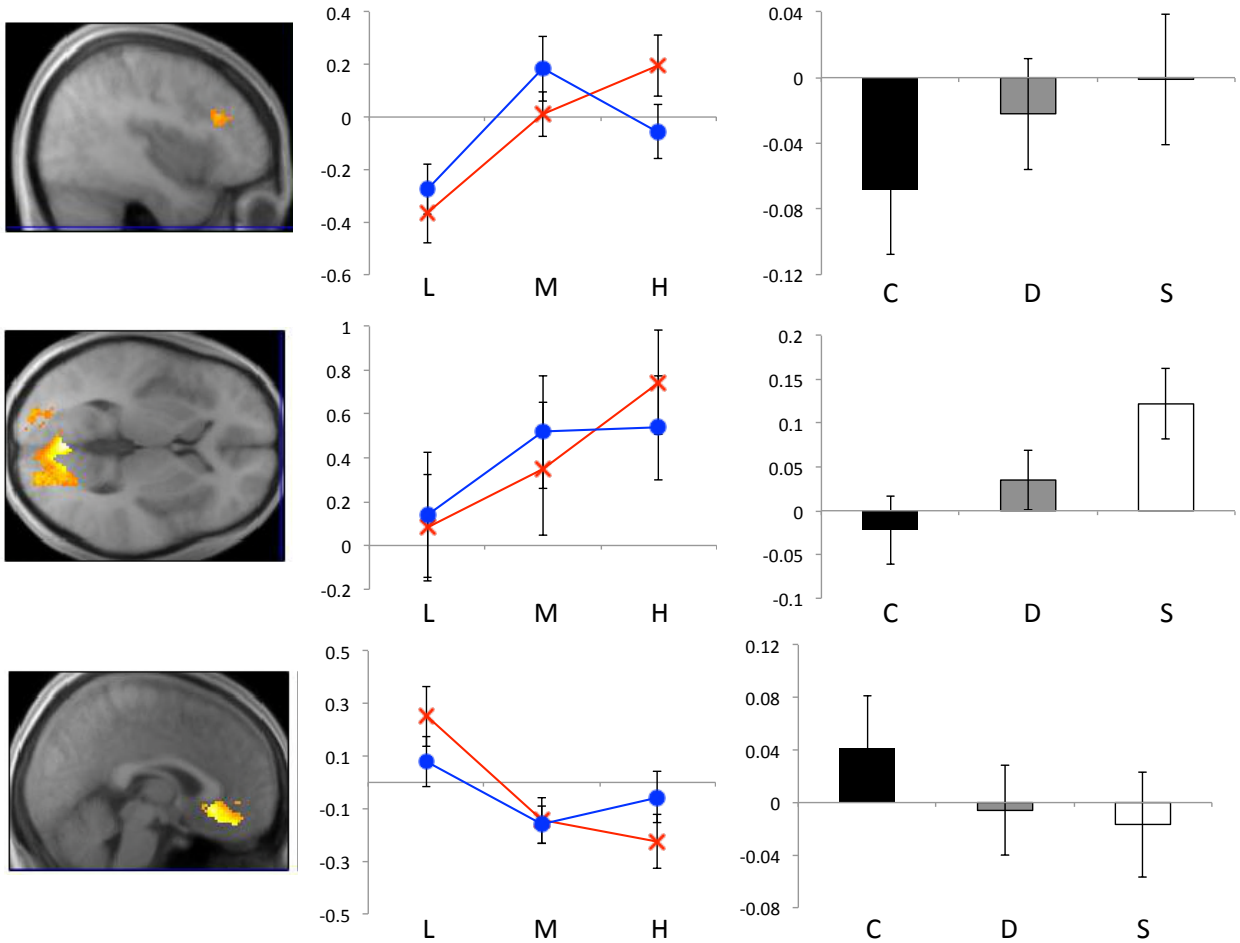

Figure 4. Disjunction results favoring the error-driven model of uncertainty during the prediction period of feedback-based learning trials, showing effects in the DLPFC (top), lingual gyrus (middle) and vmPFC (bottom). Plotted values show effect sizes (y-axis) extracted from spheres centered on peak coordinates at low (L), medium (M) and high (H) levels of causal (blue) and error-driven (red) models (middle column), and for confounded (C), deterministic (D) and stochastic (S) compound causes (right column). Error bars=SEM.

349

350 During causal queries, as participants made judgments about the causal influences of
351 individual medicines, disjunction analyses revealed extensive effects of the causal
352 measure of uncertainty throughout the DMPFC, as well as more modest effects in the
353 right DLPFC, while a selective negative correlation with the error-driven measure
354 again emerged in the vmPFC. BMS analyses (Supplemental Table 2) confirmed that
355 the causal model provided a better account of activity throughout the DMPFC.

Table 3: Results of disjunction tests in favor of the error-driven algorithm. Peak MNI coordinates and cluster sizes (>= 142) at p<0.005. R/L = Right/Left, (-) = negative correlation.

Contrast	Area	x, y, z	Cluster size at p<0.005
Prediction Uncertainty	Lingual Gyrus	0, -72, 2	1556
	L. Inferior Parietal Lobule	-32, -60, 44	177
	R. Dorsolateral Prefrontal Cortex	44, 26, 26	143
Prediction Uncertainty (-)	Ventromedial Prefrontal Cortex	2, 26, -10	471
	L. Middle Temporal Gyrus	-52, -12, -22	450
	L. Middle Temporal Gyrus	-44, -66, 22	503
	R. Middle Temporal Gyrus	44, -62, 20	340
	R. Middle Temporal Gyrus	60, -54, 2	177
	R. Temporal Parietal Junction	50, -38, 24	175
Surprise	Posterior Cingulate Cortex	-10, -32, 24	149
	R. Inferior Temporal Gyrus	40, -54, -8	151
Judgment Uncertainty (-)	Ventromedial Prefrontal Cortex	-10, 36, -10	154

356

357 *Measures of surprise.* While uncertainty in causal inferences, due to confounding or
358 stochasticity, is of primary interest here, estimates of the strength of the relationship
359 between a cause and effect, and of the surprise at the outcome on each trial, were

also modeled from both a causal and error-driven perspective. We focus our reporting and discussion of results on the measures of surprise, rather than strength: First, because causal and error-driven measures of strength were highly correlated (mean *r* = 0.87) and, second, because, to the extent that they differed, the divergence is largely incidental – using a linear, rather than noisy-OR, integration function in the Bayesian causal model would render strength measures virtually identical across causal and error-driven models, while preserving differences in uncertainty due to stochasticity vs. confounding (see discussion). For completeness, significant results of all comparisons of strength measures are listed in relevant tables (i.e., Table 2 for disjunction tests and Supplemental Table 3 for model selection).

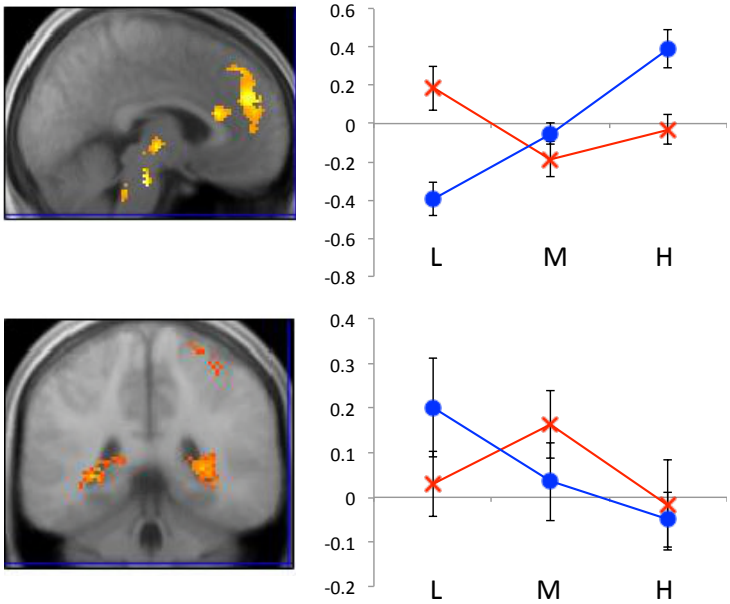

Figure 5. Disjunction results favoring the causal model of surprise, in the dorsomedial prefrontal cortex (top) and hippocampus (bottom). Line graphs show effect sizes (y-axis) extracted from spheres centered on peak coordinates at low (L), medium (M) and high (H) levels of causal (blue) and error-driven (red) models of surprise.

For measures of surprise, disjunction tests identified activity positively correlated with the causal model (i.e., the information theoretic distance between the prior and posterior strength distribution on each trial) in the bilateral anterior insula, anterior cingulate cortex and right middle temporal, inferior frontal and angular gyri, and the

375 ventral tegmental area. Activity negatively correlated with the causal measure of
376 surprise was revealed in the bilateral hippocampus, caudate and right superior and
377 inferior parietal lobules. In contrast, selective effects of the model-free prediction error
378 emerged in the posterior cingulate cortex and right inferior temporal gyrus. BMS
379 analyses (Supplemental Table 4) confirmed the selective effects of causal surprise in
380 the bilateral insula, anterior cingulate cortex, bilateral hippocampus and ventral
381 tegmental area.

382 **Discussion**

383 Sensitivity to the "no confounding" constraint – a requirement that putative causes do
384 not co-vary, fundamental to philosophical and statistical perspectives on causation –
385 has been reliably demonstrated in a range of behavioral judgments by children as well
386 as adults (Spellman, 1996; Kushnir and Gopnik, 2005; Meder et al., 2006; Schulz and
387 Bonawitz, 2007; Schulz et al., 2007; Liljeholm, 2015). The current study combined
388 computational cognitive modeling with fMRI to investigate neural substrates encoding
389 uncertainty associated with causal confounding. While being scanned, participants
390 studied the influences of various fictitious allergy medicines on headache – a side
391 effect. Three target medicines with respectively confounded (C), deterministic (D) and
392 stochastic (S) influences each occurred only in compound with an alternative medicine
393 during feedback-based learning, ruling out contingency and cue complexity as sources
394 of any effects of causal confounding. Moreover, unlike stochasticity, uncertainty
395 associated with confounding during feedback-based learning could not be attributed to
396 experienced or anticipated errors. Formal model comparisons indicated that a
397 Bayesian causal model, sensitive both to stochasticity and confounding, provided a

398 better account of activity in the anterior insula and dorsomedial prefrontal cortex
399 (DMPFC) than did an error-driven, model-free, algorithm sensitive only to
400 stochasticity; conversely, the error-driven model selectively elicited activity in the right
401 DLPFC and posterior lingual gyrus.

402 The anterior insula has been implicated in encoding outcome stochasticity across a
403 variety of tasks (e.g., Volz et al., 2003; Huettel et al., 2005; d'Acremont et al., 2013;
404 Preuschoff et al., 2008; Critchley et al., 2001; Grinband et al., 2006). For example,
405 scanning participants with fMRI as they classified stimuli according to a variable
406 category boundary, Grinband et al. (2006) found that activity in the anterior insula, as
407 well as the DMPFC, scaled with categorization uncertainty – the proximity of a
408 stimulus to a category boundary. An important aspect of this and other
409 demonstrations of the role of the anterior insula in encoding outcome stochasticity is
410 that the uncertainty is directly related to experienced or inferred errors: that is, the
411 greater the outcome stochasticity on a trial, the more likely it is that the participants
412 prediction, or decision, on that trial will be incorrect. In contrast, during feedback-
413 based learning in the current task, the outcome on confounded trials was
414 deterministic. Consequently, unlike previous work, the increase in anterior insular and
415 DMPFC activity as participants generated predictions about the outcome on trials with
416 the confounded cause cannot be attributed to the experience or anticipation of errors.
417 This strongly suggests that uncertainty representations in these regions may
418 implement generative models of predictive knowledge, dissociable from the evaluation
419 of immediate consequences of choice.

Activity in the vmPFC selectively responded to the confounded compound during feedback-based learning. As with the insula and DMPFC, comparisons with other target causes rule out stimulus complexity (since all target causes occurred only in compound with alternative causes during feedback-based learning) as well as the deterministic nature of the outcome (since that was also the case for target cause D) as sources of the effect. Moreover, the lack of a response to the stochastic target cause suggests that this region is sensitive to the presence of confounding, rather than to uncertainty per se. Notably, causal confounding is an abstract concept, irreducible to either elemental contingency or contiguity. In a study by Kumaran et al (2009), in which the conjunction of two stimuli, or of a stimulus and a location, predicted a weather (sun vs. rain) outcome, the vmPFC, and its interaction with the hippocampus, was implicated in the acquisition and transfer of conceptual stimulus-outcome associations (see Bowman and Zeithamova, 2018, for related results). Here, particularly in light of our hippocampal findings discussed below, we conjecture that the vmPFC and hippocampus are similarly, albeit in this case futilely, attempting to extract generalizable knowledge from confounded compound trials.

In addition to scaling with causal uncertainty, activity in the anterior insula and anterior cingulate cortex, together with the bilateral hippocampus, correlated with the Kullback-Liebler divergence between the posterior and prior distributions, a variable termed “Bayesian surprise” (Itti & Baldi, 2006). Notably, surprise at the outcome of a particular trial during feedback-based learning is fundamentally related to how the two modeling approaches treat uncertainty. In the error-driven model, the uncertainty (or associability) of a cue is proportional to the history of prediction errors associated with

443 that cue; in other words, the greater the long run average surprise, the greater the
444 uncertainty. In contrast, in the causal model, while a high level of surprise (i.e., large
445 KL divergence between the posterior and prior distribution) on a given trial, due either
446 to outcome stochasticity or to a change in the underlying statistics of the environment,
447 may result in greater uncertainty (i.e., greater entropy of the posterior distribution) a
448 confounded deterministic cause generates very low levels of surprise, after a few
449 initial acquisition trials, but is still associated with very high levels of uncertainty.
450 Another notable distinction between the models contrasted here is that surprise in the
451 causal model is unsigned; conversely, in the error-driven model, the unexpected
452 presence and absence of the effect respectively yield a positive and negative
453 prediction error. Thus, in the current study, selective effects of the causal measure of
454 surprise may reflect a neural implementation of a full generative model, consistent with
455 the effects observed for uncertainty, or simply an unsigned prediction error. Further
456 work is needed to arbitrate between these possibilities.

457 Several previous studies have implicated the DLPFC, together with the anterior insula
458 and DMPFC, in encoding uncertainty related to outcome stochasticity (Huettel et al.,
459 2005; Abler et al., 2009; Causse et al., 2013). The specificity of this area with respect
460 to our comparison of error-driven and causal sources of uncertainty is somewhat
461 unequivocal; In particular, based on the disjunction tests, the right DLPFC was
462 selectively recruited by the error-driven measure of uncertainty during feedback-based
463 learning, but then significantly modulated by the causal measure of uncertainty during
464 judgments about the existence of causal relationships. A possible explanation for
465 this pattern of results is that the right DLPFC reflects the influence of uncertainty on

behavior – recall that uncertainty about the causal influence of the confounded target cause is only behaviorally relevant during explicit judgments about its causal status; confounded trials during feedback-based learning deterministically and unambiguously predict the presence of the effect. Indeed, several previous studies have used neural disruption techniques to demonstrate that the right DLPFC modulates behavioral responses to uncertainty (Knoch et al., 2006; Fectau et al., 2007; Ye et al., 2015). Nonetheless, we caution the reader that only the selective effects of the error-driven model during feedback-based learning were observed across disjunction and model selection analyses.

Selective effects of the error-driven account of predictive uncertainty were also identified throughout large portions of the right posterior lingual gyrus. In a recent neuroimaging study, Causse et al., (2013) assed 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%) vs. uncertain (50%) information regarding landing conditions. Consistent with the current results, they found greater activity in the right posterior lingual gyrus, as well as in the DLPFC, during high than during low uncertainty conditions. Our results are also consistent with those of Payzan-LeNestour et al. (2016), 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 small samples. 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.

Several neuroeconomic studies have contrasted gambles involving “risk” vs. “ambiguity”, corresponding respectively to stochastic and unknown probabilities (Hsu et al., 2005; Huettel et al., 2006; Bach et al., 2009; Levy et al., 2010). For example, when flipping a coin, you might know that the probability of heads is 0.5 (risk) or, in the case of a biased coin, you may have no knowledge at all about the probability of heads (ambiguity) – in either case, subjectively, heads and tails are equally likely. One might argue that, in the current study, the influence of the confounded medicine is unknown, just as the outcome probabilities of ambiguous gambles are 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 independently occurring causes cannot be attributed to a generalization decrement based on changes in stimulus features. Moreover, as with stochasticity, ambiguity, but not confounding, should elicit high levels of error expectation during outcome prediction. Finally, ambiguity and confounding call for very different plans of action: in the case of ambiguity, but not confounding, a simple increase in sample size (i.e., repeated observation of the same contingency) can resolve the uncertainty. Further work is needed to assess the overlap between neural representations of confounding and ambiguity.

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, predictive uncertainty can emerge in the presence of a deterministic, unchanging and well-sampled contingency. 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.

References

Abler, B., Herrnberger, B., Grön, G., & Spitzer, M. (2009). From uncertainty to reward: BOLD characteristics differentiate signaling pathways. BMC neuroscience, *10*(1), 154.

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*(6), 1648-1656.

Barthelmé, S., & Mamassian, P. (2009). Evaluation of objective uncertainty in the visual system. *PLoS computational biology*, *5*(9), e1000504.

Beckers T, De Houwer J, Pineno O, Miller RR (2005) Outcome additivity and outcome maximality influence cue competition in human causal learning. Journal of experimental psychology Learning, memory, and cognition 31:238-249.

Been SL, Mitchell, C. J., Bouton, M. E., & Frohardt, R. (2003) Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. Memory & Cognition 31:133-142.

Bowman, C. R., & Zeithamova, D. (2018). Abstract memory representations in the ventromedial prefrontal cortex and hippocampus support concept generalization. *Journal of Neuroscience*, 2811-17.

Causse, M., Péran, P., Dehais, F., Caravasso, C. F., Zeffiro, T., Sabatini, U., & Pastor, J. (2013). Affective decision making under uncertainty during a plausible aviation task: An fMRI study. *NeuroImage*, *71*, 19-29.

Cheng P (1997) From covariation to causation: A causal power theory. Psycholgoical Review:367-405.

Cheng, P., Liljeholm, M., & Sandhofer, C. (2017) Analytic causal knowledge for constructing useable empirical causal knowledge: Two experiments on preschoolers. In G. Gunzelmann, A. Howes, T. Tendbrink, & E.J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1758-1763). Austin, TX: Cognitive Science Society.

Corlett, P. R., Aitken, M. R., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A., ... & Fletcher, P. C. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron*, *44*(5), 877-888.

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*(2), 537-545.

Chumbley, J. R., Flandin, G., Bach, D. R., Daunizeau, J., Fehr, E., Dolan, R. J., & Friston, K. J. (2012). Learning and generalization under ambiguity: an fMRI study. *PLoS computational biology*, *8*(1), e1002346.

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*(1), e1002895.

Fecteau, S., Knoch, D., Fregni, F., Sultani, N., Boggio, P., & Pascual-Leone, A. (2007). Diminishing risk-taking behavior by modulating activity in the prefrontal cortex: a direct current stimulation study. *Journal of Neuroscience*, *27*(46), 12500-12505.

Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R., Carpenter, T. A., Donovan, T., & Bullmore, E. T. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. Nature neuroscience, *4*(10), 1043.

Forman, S. D., Cohen, J. D., Fitzgerald, M., Eddy, W. F., Mintun, M. A., & Noll, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster‐size threshold. Magnetic Resonance in medicine, *33*(5), 636-647.

Gläscher, J. (2009) Visualization of group inference data in functional neuroimaging. Neuroinformatics, *7*(1), 73-82.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and error-driven reinforcement learning. Neuron, *66*(4), 585-595.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive psychology*, *51*(4), 334-384.

Grinband, J., Hirsch, J., & Ferrera, V. P. (2006). A neural representation of categorization uncertainty in the human brain. *Neuron*, *49*(5), 757-763.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. Journal of neuroscience, *28*(22), 5623-5630.

Harrison, L. M., Duggins, A., & Friston, K. J. (2006). Encoding uncertainty in the hippocampus. Neural Networks, *19*(5), 535-546.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American economic review*, *92*(5), 1644-1655.

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*(5), 765-775.

Huettel SA, Song AW, McCarthy G. (2005) Decisions under uncertainty: probabilistic context influences activity of prefrontal and parietal cortices. J Neurosci. 25:3304–3311

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*(5754), 1680-1683.

Itti, Laurent, and Pierre F. Baldi. "Bayesian surprise attracts human attention." *Advances in neural information processing systems*. 2006.

Knoch, D., Gianotti, L. R., Pascual-Leone, A., Treyer, V., Regard, M., Hohmann, M., & Brugger, P. (2006). Disruption of right prefrontal cortex by low-frequency repetitive transcranial magnetic stimulation induces risk-taking behavior. *Journal of Neuroscience*, *26*(24), 6469-6472.

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*(1), 477-484.

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the emergence of conceptual knowledge during human decision making. Neuron, *63*(6), 889-901.

Kushnir T, Gopnik A (2005) Young children infer causal strength from probabilities and interventions. Psychological science 16:678-683.

Le Pelley ME, Oakeshott SM, McLaren IP (2005) Blocking and unblocking in human causal learning. Journal of experimental psychology Animal behavior processes 31:56-70.

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*(2), 1036-1047.

Liljeholm, M. (2015). How multiple causes combine: independence constraints on causal inference. Frontiers in psychology, 6.

Liljeholm, M. & Cheng, P.W. (2007) When is a cause the "same"? Coherent generalization across contexts. Psychological science 18:1014-1021.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. Neuron, *38*(2), 339-346.

Meder B, Hagmayer Y, Waldmann M (2006) Understanding the Causal Logic of Confounds. In: Twenty-Eighth Annual Conference of the Cognitive Science Society, pp 579-584: Erlbaum.

Morris, R. W., Vercammen, A., Lenroot, R., Moore, L., Langton, J. M., Short, B., ... & Weickert, T. W. (2012). Disambiguating ventral striatum fMRI-related bold signal during reward prediction in schizophrenia. Molecular psychiatry, *17*(3), 280.

Cheng, P. W., & Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of personality and social psychology*, *58*(4), 545.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J.

(2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science, *304*(5669), 452–454. http://doi.org/10.1126/science.1094285

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*(1), 191-201.

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, *28*(11), 2745-2752.

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.

Schulz L, Bonawitz E (2007) Serious fun: preschoolers engage in more exploratory play when evidence is confounded. Developmental psychology 43:1045-1050.

Schulz L, Gopnik A, Glymour C (2007) Preschool children learn about causal structure from conditional interventions. Dev Sci 10:322-332.

Smith, K., Dickhaut, J., McCabe, K., & Pardo, J. V. (2002). Neuronal substrates for choice under ambiguity, risk, gains, and losses. *Management science*, *48*(6), 711-718.

Spellman B (1996) Acting as Intuitive Scientists: Contingency Judgments Are Made While Controlling for Alternative Potential Causes. Psychological science 7:337-342.

Turner, D. C., Aitken, M. R., Shanks, D. R., Sahakian, B. J., Robbins, T. W., Schwarzbauer, C., & Fletcher, P. C. (2004). The role of the lateral frontal cortex in

causal associative learning: exploring preventative and super-learning. Cerebral Cortex, *14*(8), 872-880.

Volz KG, Schubotz RI, von Cramon DY. (2003) Predicting events of varying probability: uncertainty investigated by fMRI. Neuroimage. ;19:271–280.

Waldmann MR (2001) Predictive versus diagnostic causal learning: evidence from an overshadowing paradigm. Psychonomic bulletin & review 8:600-608.

Wasserman EA, Berglan LR (1998) Backward blocking and recovery from overshadowing in human causal judgement: the role of within-compound associations. The Quarterly journal of experimental psychology B, Comparative and physiological psychology 51:121-138.

Ye, H., Chen, S., Huang, D., Wang, S., Jia, Y., & Luo, J. (2015). Transcranial direct current stimulation over prefrontal cortex diminishes degree of risk aversion. *Neuroscience letters*, *598*, 18-22.

673 Figure Captions

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

678 *Figure 2.* Mean proportions of "Is cause", "Is not a cause" and ambivalent "May or may
679 not be cause" judgments about the individual influence of each medicine on
680 headache. Target causes (C, D & S) occurred only in compound with alternative
681 causes during feedback-based learning, while non-target causes occurred only
682 individually (s'), or individually as well as in compound (e).

683 *Figure 3.* Disjunction results favoring the causal model of uncertainty during the
684 prediction period of feedback-based learning trials, showing effects in the anterior
685 insula (top) and DMPFC (bottom). Plotted values show effect sizes (y-axis) extracted
686 from spheres centered on peak coordinates at low (L), medium (M) and high (H) levels
687 of causal (blue) and error-driven (red) models (middle column), and for confounded
688 (C), deterministic (D) and stochastic (S) compound causes (right column). Error
689 bars=SEM.

690 *Figure 4.* Disjunction results favoring the error-driven model of uncertainty during the
691 prediction period of feedback-based learning trials, showing effects in the DLPFC
692 (top), lingual gyrus (middle) and vmPFC (bottom). Plotted values show effect sizes (y-
693 axis) extracted from spheres centered on peak coordinates at low (L), medium (M)

694 and high (H) levels of causal (blue) and error-driven (red) models (middle column),
695 and for confounded (C), deterministic (D) and stochastic (S) compound causes (right
696 column). Error bars=SEM.

697 *Figure 5.* Disjunction results favoring the causal model of surprise during the outcome
698 period of feedback-based learning trials, showing effects throughout the DMPFC (top)
699 and bilateral hippocampus (bottom). Plotted values show effect sizes (y-axis)
700 extracted from spheres centered on peak coordinates at low (L), medium (M) and high
701 (H) levels of causal (blue) and error-driven (red) models (middle column), and for
702 confounded (C), deterministic (D) and stochastic (S) compound causes (right column).
703 Error bars=SEM.

704

Source data legends

Figure 2—source data 1

This spreadsheet contains the source data for the mean causal judgments plotted in Figure 2. The spreadsheet can be opened with Microsoft Excel or with open-source alternatives such as OpenOffice.

Figure 3—source data 1

This spreadsheet contains the source data for the mean effect sizes plotted in the middle and right columns of Figure 3. The spreadsheet can be opened with Microsoft Excel or with open-source alternatives such as OpenOffice.

Figure 4—source data 1

This spreadsheet contains the source data for the mean effect sizes plotted in the middle and right columns of Figure 4. The spreadsheet can be opened with Microsoft Excel or with open-source alternatives such as OpenOffice.

Figure 5—source data 1

This spreadsheet contains the source data for the mean effect sizes plotted in the right column of Figure 5. The spreadsheet can be opened with Microsoft Excel or with open-source alternatives such as OpenOffice.

Supplemental Table 1: Results of Bayesian Model Selection. Peak log-odds MNI coordinates and cluster sizes (>= 40) for group-level exceedance probability maps, thresholded at a probability (γ) > 0.95, assessing which model provided the best account of uncertainty during feedback-based learning. R/L = Right/Left, ‘ ’ = continuous with previous cluster.

Winning model	Area	x, y, z	Cluster size at γ > 0.95
Uncertainty during Prediction			
Causal	Anterior Cingulate Cortex	4, 48, 24	873
	Cerebellum	-36, -52, -30	677
	R. Anterior Insula	34, 18, -10	540
	L. Anterior Insula	-26, 16,8	534
	R. Calcarine	14, -86, 14	359
	R. Fusiform Gyrus	32, -40, -22	253
	L. Precentral Gyrus	-54, 10, 34	139
	Vermis	0, -62, -40	107
	Superior Frontal Gyrus	-12, 14, 58	55
	L. Dorsolateral Prefrontal Cortex	-36, 26, 22	51
Error-driven	Precuneus	-4 -44, 10	1097
	R. Lingual Gyrus	22, -84, -10	‘ ‘
	Middle Occipital	-18, -92, 18	165
	Ventral Striatum	-8, 2, -2	90
	Ventromedial Prefrontal Cortex	-2, 22, -8	86
	Posterior Cingulate	-2, -28, 22	71
	L. Angular Gyrus	-46, -72, 28	69
	Ventral Tegmental Area	-8, -20, -10	59
	L. Thalamus	-16, -10, 4	56
	R. Dorsolateral Prefrontal Cortex	44, 44, 22	43

Supplemental Table 2: Results of Bayesian Model Selection. Peak log-odds MNI coordinates and cluster sizes (>= 40) for group-level exceedance probability maps, thresholded at a probability (γ) > 0.95, assessing which model provided the best account of uncertainty during causal judgments. R/L = Right/Left, ' ' = continuous with previous cluster.

Winning model	Area	x, y, z	Cluster size at γ > 0.95
Uncertainty during Judgments			
Causal	Anterior Cingulate Cortex	4, 48, 24	1166
	Cerebellum	-24, -80, -36	219
	L. Anterior Insula	-36, 18, -8	164
	L. Putamen	-20, 0, -10	143
	L. Middle Temporal Gyrus	-52, -40, 0	76
	Supplementary Motor Cortex	-12, 14, 59	81
	R. Orbitofrontal Cortex	46, 44, -8	66
	Vermis	2, -62, -40	44
	Ventral Tegmental Area	0, -20, -20	53
	Premotor Cortex	32, 10, 52	51
Error-driven	Rostrolateral Prefrontal Cortex	24, 56, 8	163
	Precuneus	-6, -56, 20	118
	L. Orbitofrontal Cortex	-46, 44, -10	67
	R. Angular Gyrus	46, -66, 38	65
	Cerebellum	-20, -84, -26	57

Supplemental Table 3: Results of Bayesian Model Selection. Peak MNI coordinates and cluster sizes (>= 40) for group-level exceedance probability maps, thresholded at a probability (γ) > 0.95, assessing causal and error-driven accounts of predictive strength during feedback-based learning. R/L = Right/Left

Winning model	Area	x, y, z	Cluster size at $\gamma > 0.95$
Prediction Strength			
Causal	L. Precentral Gyrus	-26, -10, 58	1068
	L. Postcentral Gyrus	-36, -26, 46	842
	R. Superior Occipital Cortex	16, -76, 28	168
	L. Superior Occipital Cortex	-14, -80, 22	167
	Cerebellum	24, -52, -24	161
	R. Lingual Gyrus	12, -50, 6	144
	L. Middle Temporal Gyrus	-44, -52, 6	369
	Dorsomedial Prefrontal Cortex	0, 34, 36	173
	L. Putamen	-20, 2, -12	224
	R. Precentral Gyrus	32, -10, 62	93
	R. Superior Temporal Gyrus	62, -42, 20	56
	R. Postcentral Gyrus	38, -24, 38	56
Error-driven	R. Lingual Gyrus	6, -80, -12	509
	L. Middle Occipital Cortex	-18, -92, 18	159
	R. Supramarginal Gyrus	56, -36, 36	109
	R. Middle Temporal Gyrus	56, -62, 4	74
Judgment Strength			
Error-driven	R. Middle Frontal Gyrus	26, 28, 30	99

Supplemental Table 4: Results of Bayesian Model Selection. Peak MNI coordinates and cluster sizes (>= 40) for group-level exceedance probability maps, thresholded at a probability (γ) > 0.95, assessing causal and error-driven accounts of surprise during feedback-based learning. R/L = Right/Left

Winning model	Area	x, y, z	Cluster size at $\gamma > 0.95$
Surprise			
Causal	R. Anterior Insula	34, 18, -12	821
	L. Anterior Insula	-28, 10, -20	678
	R. Middle Temporal Gyrus	64, -24, -4	510
	Anterior Cingulate Cortex	4, 48, 24	490
	R. Hippocampus	26, -36, 8	348
	L. Precentral Gyrus	-24, -10, 68	267
	L. Postcentral Gyrus	-50, -22, 38	144
	L. Hippocampus	-22, 34, 2	96
	Ventral Tegmental Area	-2, -20, -22	79
	L. Cerebellum	-20, -78, -36	70
	L. Middle Occipital Gyrus	-32, -82, 22	46
Error-driven	R. Inferior Frontal Gyrus	34, 8, 30	89
	R. Lingual Gyrus	22, 86, -12	53
	L. Caudate	-8, 16, 4	49