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, confounding, 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, a combination of fMRI and 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 non-confounded, deterministic or stochastic, influences. Formal comparisons of a Bayesian causal model, sensitive to both confounding and stochasticity, and a purely error-driven algorithm, sensitive only to stochasticity, revealed selective neural representations of uncertainty across several regions, including the DMPFC and insula. 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 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). 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 neutrally, 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 prefrontal

cortex (DMPFC), anterior insula, thalamus 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, and which of, these neural regions also implement signals associated with causal confounding. In the current study, participants were scanned with functional MRI while observing and judging the influences of various putative causes with confounded or non-confounded, 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 state of the outcome based on that particular configuration, then confounding does not pose a problem. Indeed, most individual, non-confounded, 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 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" (PH) algorithm, which relates the "associability" of a cue to a weighted

average of the absolute prediction error on previous trials involving that cue. Since 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 predictive strength of a cause is reflected in the entropy of the posterior distribution over its possible strengths, which depends on the variance of the effect variable, but also on the independent occurrence of alternative causes. While 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 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.

Materials & Methods

Participants. 20 healthy normal volunteers (mean age = 20.9 ± 2.4, range: 18-27, 12 females) participated in the study. One participant was excluded prior to any analyses due to excessive head movement (> 6mm), 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. 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 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.

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 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 0.5 but unlike medicine S, never occurred in combination with any other medicine, so that contrasting S with s' identifies differences due to compound presentation.

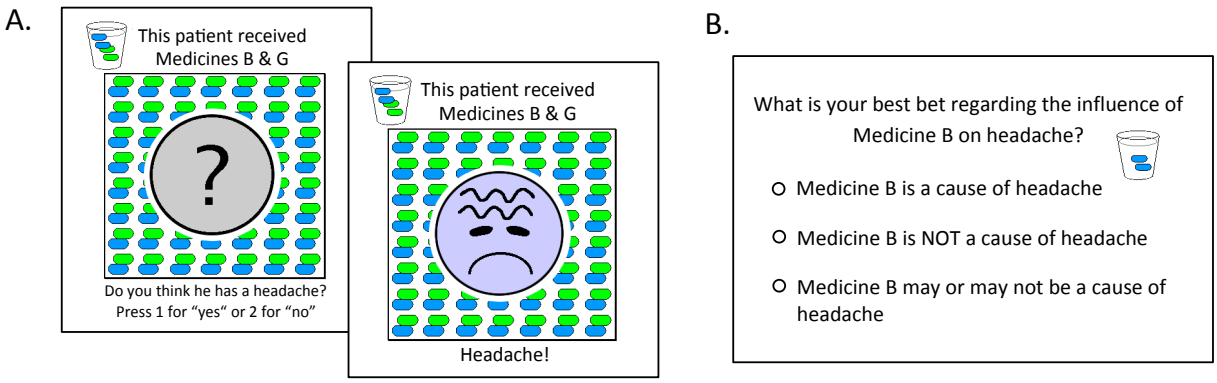

Figure 1. Task illustration. A. Prediction and outcome screens from a feedback-based learning trial. Screens were separated by a 2 sec. pause during which the word "Wait" (not shown in figure) was displayed. B. Causal query regarding the individual influence of a particular medicine on headache.

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 (2000ms) 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).

Each of 6 distinct trial types was presented 8 times in each of 4 sessions, for a total of 192 trials, with sessions separated by 2-minute breaks, during which the scanner was turned off. Each session was further divided into 4, non-delineated, blocks, with each type of trial occurring once in random order in each block. All trials were separated by a jittered 4-second ITI. 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 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 an ambivalent "may or may not be a cause of headache" (Figure 1B).

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:

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

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:

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

where $n\text{-}1$ references 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., Esberg & 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, 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, in which reasoners make inferences over causal structures potentially responsible for the observed data (Griffiths & Tenenbaum, 2005; Lu et al., 2008; Liljeholm, 2015). 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 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).

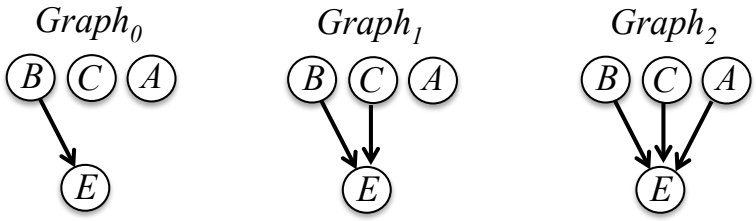

Figure 2. Possible causal structures involving a background cause, B, a target cause, C, an alternative cause, A, and an effect, E. A link to the effect always exists for B, and may exist for neither C nor A ($Graph_0$), C only ($Graph_1$), or both C and A ($Graph_2$).

Each graph has a set of parameters θ, which are strengths w_i associated with causal links in each graph (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 was modeled using only n- trials in G_0. Estimates of elemental (non-target) 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:

$$
\begin{aligned}
P\left(d \middle| w_B, w_C, w_A, G_2\right) = & \begin{pmatrix} N\left(c^-, a^-\right) \\ N\left(e^+, c^-, a^-\right) \end{pmatrix} \times \begin{pmatrix} N\left(c^+, a^-\right) \\ N\left(e^+, c^+, a^-\right) \end{pmatrix} \times \begin{pmatrix} N\left(c^-, a^+\right) \\ N\left(e^+, c^-, a^+\right) \end{pmatrix} \times \begin{pmatrix} N\left(c^+, a^+\right) \\ N\left(e^+, c^+, a^+\right) \end{pmatrix} w_B^{N\left(e^+, c^-, a^-\right)} \\
& (1-w_B)^{N\left(e^-, c^-, a^-\right)} [w_B + w_C - w_B w_C]^{N\left(e^+, c^+, a^-\right)} [1 - w_B - w_C + w_B w_C]^{N\left(e^-, c^+, a^-\right)} [w_B + w_A - w_B w_A]^{N\left(e^+, c^-, a^+\right)} [1 - w_B - w_A + w_B w_A]^{N\left(e^-, c^-, a^+\right)} \\
& [w_B + w_C + w_A - w_B w_C - w_B w_A - w_C w_A + w_B w_C w_A]^{N\left(e^+, c^+, a^+\right)} [1 - w_B - w_C - w_A + w_B w_C + w_B w_A + w_C w_A - w_B w_C w_A]^{N\left(e^-, c^+, a^+\right)}.
\end{aligned}
\tag{3}
$$

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 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 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,

$$P(w_C \mid d, G_2) = \int_0^1 \int_0^1 \frac{P(d \mid w_B, w_C, w_A, G_2) P(w_B, w_C, w_A \mid G_2)}{P(d)} dw_B \, dw_A, \tag{4}$$

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 $P(w_C|d)$:

$$H(w_C) = -\int_0^1 P(w_C \mid d) \ln P(w_C \mid d) \, dw_C. \tag{5}$$

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

$$\bar{w}_C = \int_0^1 w_C P(w_C \mid d) \, dw_C. \tag{6}$$

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 "Bayesian surprise" (Itti & Baldi, 2006) was computed as:

$$KL(P\|Q) = \int_0^1 p(w_C \mid d) \log \frac{p(w_C \mid d)}{q(w_C \mid d)} dw_C. \quad (7)$$

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, while 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 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 (BIC) was used for model selection. The model-derived probabilities of ambivalent causal judgments are plotted in Figure 3.

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. All effects were reported at a whole brain corrected $p < 0.05$, using cluster size thresholding (CST) to adjust for multiple comparisons (Forman et al., 1995). Using an individual voxel probability threshold of $p=0.001$ indicated that

using a minimum cluster size of 63 MNI transformed voxels resulted in an overall significance of $p < 0.05$.

For each subject, two general linear models (GLM) were specified, one for the causal and one for the error-driven computational model. In each 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, while outcome-screen onsets were modulated by trial-by-trial absolute 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 while 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, non-modeled, prior beliefs. Moreover, in each model, two separate regressors modeled the motor responses on each feedback-based and causal query trial, 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.

To directly contrast error-driven and causal accounts, Bayesian model selection (BMS) 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. First-level Bayesian estimation procedure was used to compute a log-model evidence map for every subject and each GLM and inferences at the group level were modeled by applying a random effects approach (Rosa et al., 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 0.95. 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 0.95: 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.

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 (*CI*) and effect sizes (Cohen's *d*) are reported for all comparisons. The mean proportions of ambivalent, "may or may not be a cause" judgments, averaged across probes for each medicine,

are shown in Figure 3, together with the model-derived probability of an ambivalent judgment based on causal and error-driven uncertainty respectively.

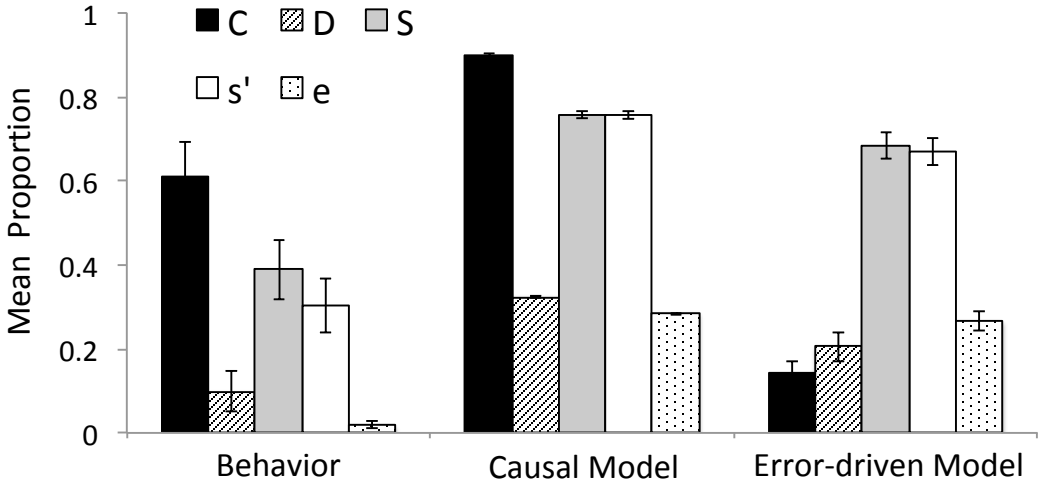

Figure 3. Mean proportions of ambivalent "May or may not be a cause" judgments, together with the mean (normalized) model-derived probabilities of such judgments based on causal and error-driven uncertainty respectively. Confounded, deterministic and stochastic target causes (C, D & S) occurred only 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). Error bars=SEM.

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 4.13±1.44 for the prediction noise parameter and 1.03±1.05 for the judgment noise parameter. The mean BIC was significantly lower, indicating superior performance, for the causal model (220.00±44.15) than for the error-driven model (237.09±35.58), $t(18)=2.51$, $p<0.05$, $CI=[2.81\ 31.37]$, $d=0.43$.

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

Neuroimaging results

All results reported below survived inclusive masking with voxels identified by Bayesian model selection analyses. A complete list of the non-masked significant effects of each model is provided in Tables 1-4. Bar graphs show effect sizes extracted from 8mm spheres centered on peak MNI coordinates, using rfxplot (Glascher, 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, the right dorsolateral prefrontal cortex and the left cerebellum

332 (Figure 4). No effects of the error-driven measure of uncertainty during causal
333 judgments survived whole-brain correction.

334

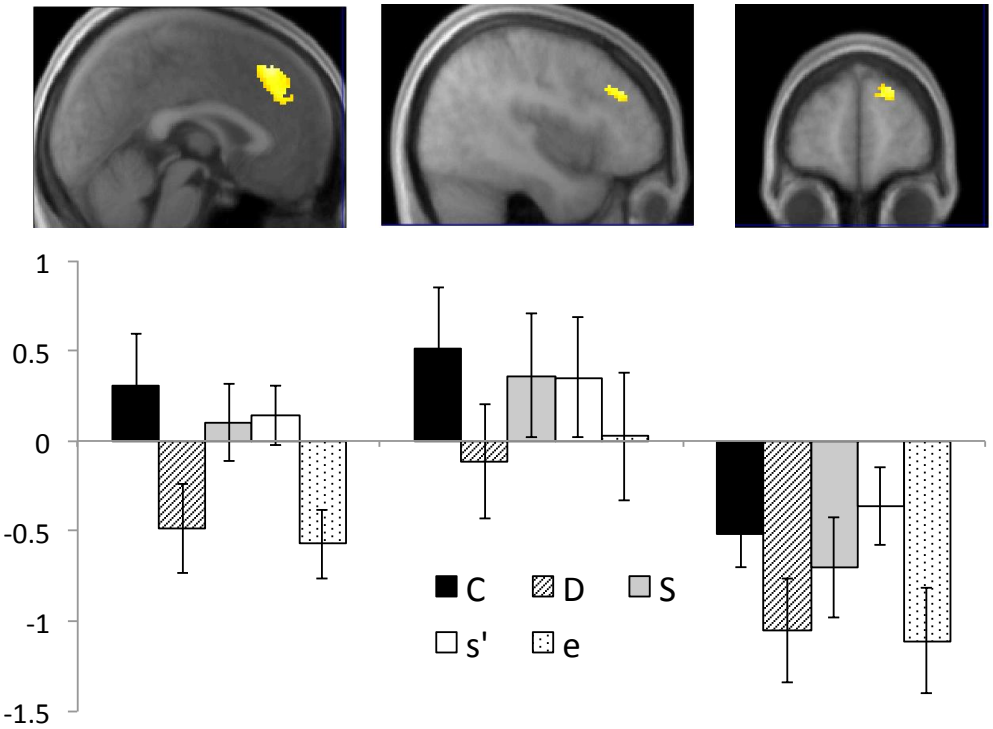

Figure 4. Selective effects of the causal model of uncertainty during judgments about the causal influences of individual medicines, showing effects in the DMPFC (left), DLPFC (middle) and lateral SFG (right). Bar plots show effect sizes (y-axis) extracted from spheres centered on peak coordinates for each queried medicine. Error bars=SEM.

335

336 *Uncertainty signals during prediction.* During the prediction period of feedback-based
337 learning trials, the error-driven, but not causal, measure of uncertainty was
338 significantly positively correlated with activity throughout the inferior occipital and
339 lingual gyri, as well as with activity in the supplementary motor area and right anterior
340 insula (Figure 5). Moreover, activity negatively correlated with error-driven predictive
341 uncertainty was observed in the posterior middle temporal gyrus. In contrast, during

342 the same period, selective and significant effects of the causal measure of uncertainty

343 emerged only in the right dorsal thalamus.

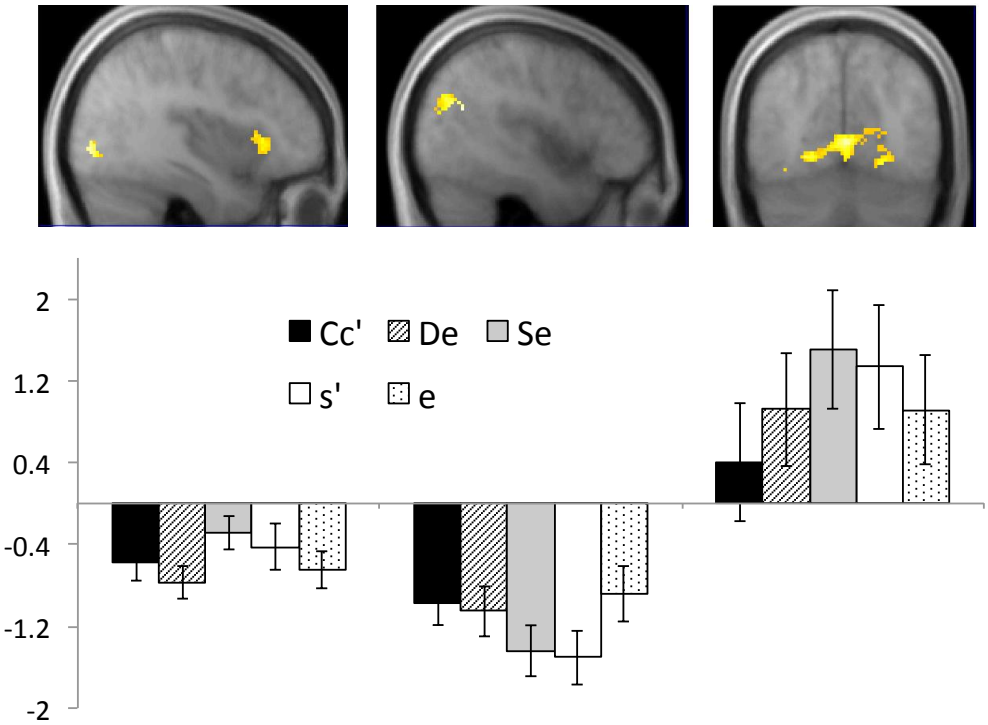

Figure 5. Selective effects of the error-driven model of uncertainty during the prediction period of feedback-based learning trials, showing effects in the right insula (left), MTG (middle) lingual gyrus (right). Bar plots show effect sizes (y-axis) extracted from spheres centered on peak coordinates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

344

345 *Strength & surprise signals.* While uncertainty in causal inferences, due to

346 confounding or stochasticity, is of primary interest here, estimates of the strength with

347 which a cause, or combination of causes, predicts the effect, and of the surprise at the

348 outcome on each trial of feedback-based learning, were also modeled for

349 completeness. During the prediction period of feedback-based learning trials, the

350 causal measure of predictive strength was selectively and significantly positively

correlated with activity in the superior occipital gyrus, as well as negatively correlated with activity in the left precentral gyrus (Figure 6). In contrast, a selective and significant negative correlation with the error-driven measure of predictive strength emerged in the posterior MTG. No significant effects of either the causal or error-driven account of predictive strength were found during causal judgments.

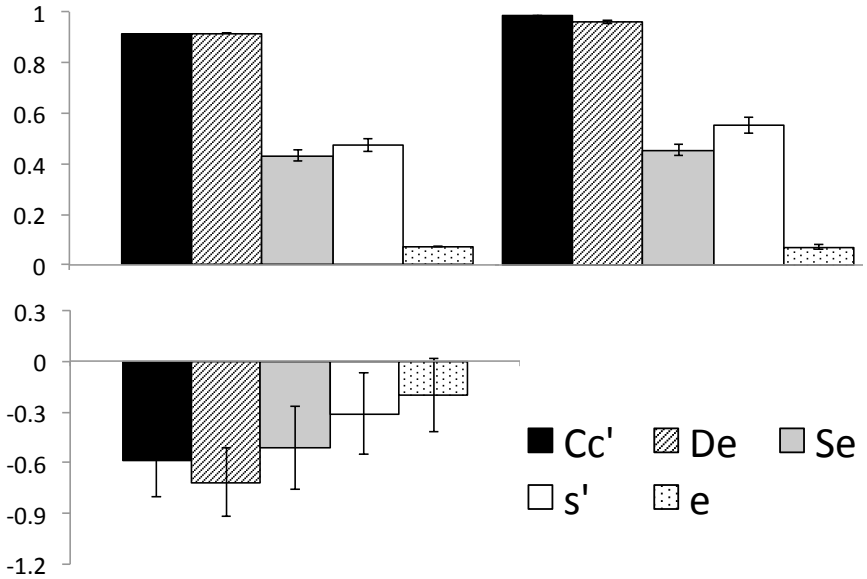

Figure 6. Strength predictions derived from the causal (top left) and error-driven (top right) model, together with effect sizes extracted from spheres centered on peak coordinates in the precentral gyrus, for each type of medicine trial during feedback-based learning. Error bars=SEM.

For measures of surprise during the outcome period of feedback-based learning trials, 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 prefrontal cortex, the inferior parietal gyrus, the cerebellum, the ventral tegmental area, and the right thalamus. In addition, significant negative correlations with this measure were identified in the middle occipital gyrus, fusiform

gyrus, the superior parietal lobule, precentral gyrus and calcarine sulcus. In contrast, a selective and significant negative correlation with the causal measure of (Bayesian) surprise was observed in the anterior caudate.

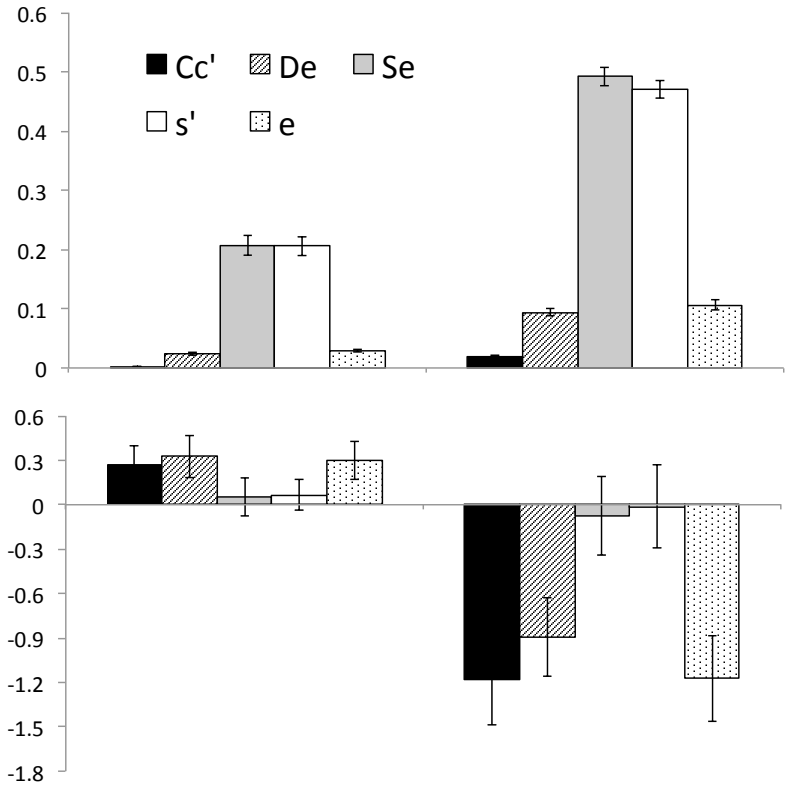

Figure 7. Outcome surprise signals derived from the causal (top left) and error-driven (top right) model, together with effect sizes extracted from spheres centered on peak coordinates in the anterior caudate (bottom left) and anterior insula (bottom right), for each type of medicine trial during feedback-based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

As can be seen in Figures 6 and 7, the ordinal mean values across trial types derived from causal and error-driven accounts are identical for both strength and surprise measures, suggesting that model-selection may reflect more granular differences. For example, because of the Σ term in Equations 1 and 2 of the error-driven account, the variance in the occurrence of the outcome across e- and De+ trials, and on 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., Cheng & Novick, 1992; Spellman, 1996). Moreover, in the causal, but not the error-driven, model, whenever there is more than one link in a graph, the occurrence of the outcome has a lesser impact on the marginal posterior of any given cause than does its non-occurrence; this is because, for generative causes, the occurrence of the outcome may be due to any cause present on the trial, including the constantly present background cause, while the non-occurrence can be unambiguously attributed to a lack of influence by all present causes. Since these differences between the error-driven and causal model are largely incidental to the specifics of model implementation they will not be discussed further.

Discussion

Sensitivity to the "no confounding" constraint – a requirement that putative causes do not co-vary, fundamental to philosophical and statistical perspectives on causation – has been reliably demonstrated in a range of behavioral 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). The current study used fMRI 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 non-confounded, deterministic or stochastic, influences. Formal comparisons of a Bayesian causal model, sensitive to both confounding and stochasticity, and a purely error-driven algorithm, sensitive only

395 to outcome variance, revealed selective neural representations of uncertainty across
396 several regions, including the DMPFC and insula.

397 The DMPFC has been implicated in encoding uncertainty across a variety of tasks
398 (Volz et al., 2003; Huettel et al., 2005; Grinband et al., 2006; Xue et al., 2008; Michael
399 et al., 2015). For example, scanning participants with fMRI as they classified stimuli
400 according to a variable category boundary, Grinband et al. (2006) found that activity in
401 the DMPFC scaled with the proximity of a stimulus to a category boundary. An
402 important aspect of this and other demonstrations of the involvement of the DMPFC in
403 uncertainty is that the level of uncertainty is directly related to the degree of
404 experienced errors: that is, the greater the uncertainty on a given trial, the more likely
405 it is that the prediction on that trial will be incorrect, as indicated by explicit feedback.
406 In contrast, during feedback-based learning in the current task, the outcome on
407 confounded trials was deterministic. Consequently, the increased activity in the
408 DMPFC, as participants generated judgments about confounded cause, cannot be
409 attributed to a history of errors. Instead, these results suggest that the DMPFC
410 encodes a more abstract representation of uncertainty, divorced from the immediate
411 consequences of choice, and irrespective of whether the specific source of uncertainty
412 is confounding or stochasticity.

413 Several neuroeconomic studies (Hsu et al., 2005; Krain et al., 2006; Huettel et al.,
414 2006; Bach et al., 2009; Levy et al., 2010; Pushkarskaya et al., 2015) have contrasted
415 gambles involving stochasticity with “ambiguous” gambles, for which outcome
416 probabilities are completely unknown, with some finding greater activity in the DMPFC
417 in response to ambiguity than stochasticity (e.g., Hsu et al., 2005). One might argue

418 that, in the current study, the influence of the confounded medicine is unknown, just
419 as the outcome probabilities of ambiguous gambles in previous studies were
420 unknown. But what does it mean to "know" something? Critically, all target medicines
421 occurred only in compound with alternative medicines during feedback-based
422 learning: consequently, when presented individually during causal queries, each was
423 equally novel and, presumably, equally unknown. In other words, differences in
424 ambiguity between confounded and other target causes cannot be attributed to a
425 generalization decrement based on changes in stimulus features. Moreover, ambiguity
426 due to small samples and uncertainty due to confounding call for very different plans
427 of exploratory action: while in former case, repeated observation of the same stimulus
428 can resolve the uncertainty, the latter case requires an intervention that unconfounds
429 the stimulus configuration. Further work is needed to assess the overlap between
430 neural representations of confounding and ambiguity.

431 As with the DMPFC and DLPFC, the anterior insula has been implicated in outcome
432 stochasticity, or "risk", in several previous neuroimaging studies (e.g., Volz et al.,
433 2003; Huettel et al., 2005; d'Acremont et al., 2013; Preuschoff et al., 2008; Critchley et
434 al., 2001). Here, the dorsal anterior right insula was selectively recruited by the error-
435 driven model of uncertainty during the prediction period of feedback-based learning
436 trials. Plots of effect sizes for each type of trial (in Figure 5) suggest that these
437 responses were indeed driven by stochasticity–based uncertainty about the state of
438 the outcome on any given trial. Moreover, the fact that, unlike the DMPFC, activity in
439 the anterior insula did not scale with either an error-driven or causal measure of
440 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. While, as noted, 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 (Klein et al., 2007; Metereau et al., 2012; Allen et al., 2016; Bastin et al., 2016).

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%) vs. 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 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 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.

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.

Acknowledgements

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

References

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

Allen, M., Fardo, F., Dietz, M. J., Hillebrandt, H., Friston, K. J., Rees, G., & Roepstorff, A. (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*(6), 1648-1656.

Baetens, K., Ma, N., Steen, J., & Van Overwalle, F. (2013). Involvement of the mentalizing network in social and non-social high construal. *Social cognitive and affective neuroscience*, *9*(6), 817-824.

Baetens, K. L., Ma, N., & Van Overwalle, F. (2017). The dorsal medial prefrontal cortex is recruited by high construal of non-social stimuli. *Frontiers in behavioral neuroscience*, *11*, 44.

Bastin, J., Deman, P., David, O., Gueguen, M., Benis, D., Minotti, L., ... & Kahane, P. (2016). Direct recordings from human anterior insula reveal its leading role within the error-monitoring network. *Cerebral Cortex*, bhv352.

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. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365.

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.

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.

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 B: Biological Sciences*, *278*(1718), 2553-2561.

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.

Glymour C (1998) Psychological and normative theories of causal power and the probabilities of causes. In: Uncertainty in artificial intelligence (F. CG, S. M, eds): Kaufmann.

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

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.

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.

Klein, T. A., Endrass, T., Kathmann, N., Neumann, J., von Cramon, D. Y., & Ullsperger, M. (2007). Neural correlates of error awareness. *Neuroimage*, *34*(4), 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*(1), 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*(2), 1036-1047.

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

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological review*, *115*(4), 955.

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.

Metereau, E., & Dreher, J. C. (2012). Cerebral correlates of salient prediction error for different rewards and punishments. *Cerebral Cortex*, *23*(2), 477-487.

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.

Rosa, M. J., Bestmann, S., Harrison, L., & Penny, W. (2010). Bayesian model selection maps for group studies. *Neuroimage*, *49*(1), 217-224.

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.

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

Volz KG, Schubotz RI, von Cramon DY. (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*(5), 1019-1027.

Figure Captions

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.

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 ($Graph_2$).

Figure 3. Mean proportions of ambivalent "May or may not be a cause" judgments, together with the mean (normalized) model-derived probabilities of such judgments based on causal and error-driven uncertainty respectively. Confounded, deterministic and stochastic 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). Error bars=SEM.

Figure 4. Selective effects of the causal model of uncertainty during judgments about the causal influences of individual medicines, showing effects in the DMPFC (left), DLPFC (middle) and lateral SFG (right). Bar plots show effect sizes (y-axis) extracted from spheres centered on peak coordinates for each queried medicine. Error bars=SEM.

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 insula

612 (left), MTG (middle) and lingual gyrus (right). Bar plots show effect sizes (y-axis)
613 extracted from spheres centered on peak coordinates for each type of medicine trial.
614 Error bars=SEM.

615 *Figure 6.* Strength predictions derived from the causal (top left) and error-driven (top
616 right) model, together with effect sizes extracted from spheres centered on peak
617 coordinates in the precentral gyrus, for each type of medicine treatment during the
618 prediction period of feedback-based learning trials. Error bars=SEM.

619 *Figure 7.* Outcome surprise signals derived from the causal (top left) and error-driven
620 (top right) model, together with effect sizes extracted from spheres centered on peak
621 coordinates in the anterior caudate (bottom left) and anterior insula (bottom right), for
622 each type of medicine treatment, during the outcome period of each feedback-based
623 learning trial. Error bars=SEM.

624
625

Table 1: Non-masked significant, whole-brain corrected, neural effects of the causal model of uncertainty during causal judgments and feedback-based learning, with peak MNI coordinates and cluster sizes (k). R/L = Right/Left, (-) = negative correlation.

Contrast	Region	Peak MNI	Cluster Size
Judgment Uncertainty	DMPFC	-8, 28, 46	1900
	R. IPG	50, -52, 40	292
	L. IPG	-48, -60, 44	71
	Cerebellum	-38, -70, -40	111
	Cerebellum	-36, -68, -22	97
	Dorsolateral PFC	42, 34, 30	201
	Rostrolateral PFC	34, 52, 4	134
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
	Cerebellum	-2, -56, -40	82
	Mid Cingulate	6, -22, 30	104
	L. RLPFC	-36, 56, 16	70

Table 2: Non-masked significant, whole-brain corrected, neural effects of the causal model of strength and surprise during feedback-based learning, with peak MNI coordinates and cluster sizes (k). R/L = Right/Left, (-) = negative correlation.

Contrast	Region	Peak MNI	Cluster Size
Prediction Strength	Calcarine	18, -96, 4	236
	Superior Occipital	-14, -96, 12	120
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
	Precentral Gyrus	34, -10, 58	119
	Superior Parietal Lobule	34, -46, 60	63
	Middle Temporal Gyrus	-50, -48, 6	63
	Postcentral Gyrus	38, -24, 40	115
	Supp. Motor Area	8, 2, 52	297
	Rostrolateral PFC	-32, 42, 10	86
Surprise	Right Insula	36, 18, -2	652
	Left Insula	-28, 22, -6	482
	Dorsomedial PFC	4, 46, 34	1146
	Left Dorsolateral PFC	-46, 4, 42	192
	Right Dorsolateral PFC	44, 24, 30	325
	Thalamus	6, -26, 2	64
	Thalamus	10, -10, 8	65
Surprise (-)	Supramarginal Gyrus	-50, -26, 38	1858
	Caudate	18, 32, 0	505
	Mid Cingulum	-10, -30, 42	472
	Postcentral Gyrus	50, -20, 48	2399
	Fusiform Gyrus	-34, -40, -14	117
	Parahippocampal	34, -32, -12	227
	Precuneus	16, -52, 14	221

Table 3: Non-masked significant, whole-brain corrected, neural effects of the error-driven model of uncertainty and strength feedback-based learning, with peak MNI coordinates and cluster sizes (k). R/L = Right/Left, (-) = negative correlation.

Contrast	Region	Peak MNI	Cluster Size
Prediction Uncertainty	Calcarine	14, -88, 0	3116
	Supp. Motor Area	10, 20, 48	153
	Left Insula	-24, 22, 2	65
	Right Insula	34, 28, -2	109
	Lingual Gyrus	-16, -50, -2	63
	Precuneus	10, -66, 34	64
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
	Dorsomedial PFC	-2, 32, 42	92
	Dorsolateral PFC	50, 28, 30	76
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
	Middle Temporal Gyrus	-52, -4, -12	168
	Precentral Gyrus	-22, -14, 64	383
	Thalamus	-12, -10, -4	214
	Mid Cingulum	-4, -4, 50	138

Table 4: Non-masked significant, whole-brain corrected, neural effects of the error-driven model of surprise during feedback-based learning, with peak MNI coordinates and cluster sizes (k). R/L = Right/Left, (-) = negative correlation.

Contrast	Region	Peak MNI	Cluster Size
Prediction Error	Right Insula	38, 18, -4	912
	Left Insula	-40, 20, 4	945
	Precentral Gyrus	-42, -2, 40	713
	Anterior Cingulate	4, 38, 30	2021
	Inferior Parietal Gyrus	-28, -52, 42	246
	Dorsolateral PFC	44, 28, 26	885
	Cerebellum	-10, -82, -30	90
	Cerebellum	-28, -66, -32	118
	Ventral Tegmental Area	6, -24, -18	111
	Angular Gyrus	42, -52, 40	143
	Thalamus	10, -4, 6	137
Prediction Error (-)	Middle Occipital Gyrus	-42, -74, 26	735
	Fusiform Gyrus	-34, -42, -12	253
	Postcentral Gyrus	54, -26, 52	906
	Mid Cingulate	-10, -30, 44	157
	Mid Cingulate	14, -32, 48	103
	Inferior Parietal Gyrus	-50, -26, 40	90
	Superior Parietal Gyrus	-20, -46, 70	104
	Middle Occipital Gyrus	40, -78, 30	77
	Precuneus	-10, -52, 48	86
	Precentral Gyrus	34, -16, 66	118
	Parahippocampal	20, -20, -20	96
	Calcarine	-8, -62, 12	79