Neuroscience

Inferring illness causes recruits the animacy semantic network

Reviewed Preprint
v1 • December 19, 2024
Not revised

Miriam Hauptman, Marina Bedny

Department of Psychological & Brain Sciences, Johns Hopkins University, Baltimore, USA

eLife Assessment

This **valuable** study investigates the neural basis of causal inference of illness, suggesting that it relies on semantic networks specific to living things in the absence of a generalized representation of causal inference across domains. However, the evidence remains **incomplete** due to some unjustified design and analysis choices. Moreover, the authors do not fully exploit the potential of multivariate fMRI analyses to rigorously test their main hypothesis.

https://doi.org/10.7554/eLife.101944.1.sa3

Abstract

Inferring the causes of illness is universal across human cultures and is essential for survival. Here we use this phenomenon as a test case for understanding the neural basis of implicit causal inference. Participants (n=20) undergoing fMRI read two-sentence vignettes that encouraged them to make causal inferences about illness or mechanical failure (causal control) as well as non-causal vignettes. All vignettes were about people and were matched on linguistic variables. The same participants performed localizers: language, logical reasoning, and mentalizing. Inferring illness causes selectively engaged a portion of precuneus (PC) previously implicated in the semantic representation of animates (e.g., people, animals). This region was near but not the same as PC responses to mental states, suggesting a neural mind/body distinction. No cortical areas responded to causal inferences across domains (i.e., illness, mechanical), including in individually localized language and logical reasoning networks. Together, these findings suggest that implicit causal inferences are supported by content-specific semantic networks that encode causal knowledge.

Introduction

A distinguishing feature of human cognition is our ability to reason about complex cause-effect relationships, particularly when causes are hidden (Tooby & DeVore, 1987, Lagnado et al., 2007; Rottman, Ahn, & Luhmann, 2011; Muentener & Schulz, 2014; Sloman & Lagnado, 2015; Goddu & Gopnik, 2024). Inferring the causes of illness provides a classic example (Keil et al., 1999; Waldmann, 2000; Meder & Mayrhofer, 2017; Legare & Shtulman, 2018). When reading something like, *Hugh sat by sneezing passengers on the subway. Now he has a case of*

COVID, we naturally infer a causal relationship between crowded spaces and the invisible transmission of infectious disease. Here we investigate the neurocognitive mechanisms that support such automatic inferences.

Inferring illness causes is a universal yet culturally variable phenomenon (Ackerknecht, 1982; Foster, 1976; Legare & Gelman, 2008; Lock & Nguyen, 2010). Young children reason about the causes of illness even prior to formal schooling, for instance attributing illness to contamination or contact with a sick person (Springer & Ruckel, 1992; Kalish, 1996, 1997; Keil et al., 1999; Raman & Gelman, 2005; Legare & Gelman, 2008; Legare, Wellman, & Gelman, 2009). During development and throughout life, we acquire cultural knowledge about the invisible forces that bring about illness, from pathogen transmission to divine retribution (Notaro, Gelman, & Zimmerman, 2001; Raman & Winer, 2004; Lynch & Medin, 2006; Legare & Gelman, 2008; Legare et al., 2012). In many societies, designated 'healers' become experts in diagnosing and curing disease (Foster, 1976; Ackerknecht, 1982; Norman et al., 2006; Lightner, Heckelsmiller, & Hagen, 2021) and non-experts routinely infer the causes of illness in themselves and others (e.g., *how did my friend get COVID?*).

One hypothesis is that causal inferences about illness depend on content-specific semantic representations in the 'animacy network' (see preregistration *https://osf.io/cx9n2/*) (Fairhall & Caramazza, 2013b; Deen & Freiwald, 2022). This hypothesis is consistent with the prominent view that semantic knowledge is organized into distinct causal frameworks, so-called 'intuitive theories' (e.g., Wellman & Gelman, 1992; Gopnik & Meltzoff, 1997; Tenenbaum et al., 2007; Gerstenberg & Tenenbaum, 2017). Illness exclusively affects living things, and semantic networks that represent living things may contain causal information. The causal mechanisms that lead to illness are related to our intuitions about bodily function and are distinguishable from intuitions about inanimate objects (Wellman & Gelman, 1992; Keil et al., 1999; Inagaki & Hatano, 2006). For example, people but not machines transmit diseases to each other through physical contact. Thinking about living things (animates) as opposed to non-living things (inanimate objects/places) depends on partly dissociable neural systems (e.g., Warrington & Shallice, 1984; Hillis & Caramazza, 1991; Caramazza & Shelton, 1998; Farah & Rabinowitz, 2003). Recent evidence suggests that the precuneus (PC) in particular responds preferentially to images and words referring to living things (i.e., people and animals) across a variety of tasks, such as semantic categorization and similarity judgment (Fairhall & Caramazza, 2013a, 2013b; Fairhall et al., 2014; Peer et al., 2015; Wang et al., 2016; Silson et al., 2019; Rabini, Ubaldi, & Fairhall, 2021; Deen & Freiwald, 2022; Aglinskas & Fairhall, 2023; Hauptman, Elli, et al., 2023). Whether the PC is sensitive to causal inferences about illness is not known.

The PC is also active during mental state inferences (Saxe & Kanwisher, 2003; Saxe et al., 2006). We therefore tested whether inferences about illness (bodies) as opposed to mental states (minds) recruit the same or different subregions of the PC by localizing the 'mentalizing' network (Dodell-Feder et al., 2010; Dufour et al., 2013).

A non-mutually exclusive hypothesis is that inferring illness causes depends on neurocognitive mechanisms that support causal inferences regardless of their content, e.g., causal inferences about bodies (e.g., illness), minds, or mechanical objects (e.g., mechanical failure). A large body of behavioral evidence has shown that children and adults use similar cognitive principles to infer causality across domains (e.g., Cheng & Novick, 1992; Waldmann & Holyoak, 1992; Pearl, 2000; Gopnik et al., 2001; Steyvers et al., 2003; Gopnik et al., 2004; Schulz & Gopnik, 2004; Rehder & Burnett, 2005; Lagnado et al., 2007; Rottman & Hastie, 2014; Davis & Rehder, 2020). For instance, children as young as 3 years old correctly discard or 'screen off' conditionally independent variables when identifying the causes of both biological and psychological events (i.e., allergic reactions, fear responses) (Schulz & Gopnik, 2004). The deployment of content-invariant probabilistic knowledge during causal inference suggests that a common brain network could enable causal inferences about any domain, including illness.

Frontotemporal language processing mechanisms themselves could support causal inferences across domains that occur during language comprehension. Language has a rich capacity to convey causal information (Tooby & DeVore, 1987; Pinker, 2003; Solstad & Bott, 2017) and has been suggested to play an important role in some aspects of combinatorial thought (e.g., Spelke, 2003; 2022). Alternatively, causal inference may depend on general logical deduction mechanisms (Goldvarg & Johnson-Laird, 2001; Barbey & Patterson, 2011; Khemlani et al., 2014; Operskalski & Barbey, 2017). Finally, it is possible that causal inferences are supported by a dedicated 'causal engine' that is distinct from language processing or logical reasoning (Gopnik et al., 2004; Saxe & Carey, 2006; Tenenbaum et al., 2007; Carey, 2011).

To our knowledge, no prior study has examined the neural basis of implicit causal inferences about illness. A handful of prior experiments investigated neural responses during explicit causality judgments that were collapsed across semantic domains (e.g., biological, mechanical, mental state inference). For example, Kuperberg et al. (2006) asked participants to rate the causal relatedness of three-sentence stories and observed higher responses to causal stories in left frontotemporal cortex (Kuperberg et al., 2006). Frontotemporal responses in both hemispheres have been observed in other experiments using single words, symbols, and passages as stimuli, but across studies no consistent neural signature of causal inference has emerged (Ferstl & von Cramon, 2001; Satpute et al., 2005; Fugelsang & Dunbar, 2005; Chow et al., 2008; Fenker et al., 2009; Mason & Just, 2011; Prat et al., 2011; Kranjec et al., 2012; Pramod, Chomik-Morales, et al., 2024). Nearly all prior experiments use explicit tasks, and, in many cases, causal trials were more difficult. Some of the observed effects may therefore reflect linguistic or executive load rather than cognitive processes specific to causal inferences. Additionally, prior studies did not localize language or logical reasoning networks in individual participants, making it difficult to assess the involvement of these systems (e.g., Fedorenko et al., 2010; Monti et al., 2009).

In the current fMRI study, we use an implicit task to capture automatic causal inferences that unfold during language comprehension (Black & Bern, 1981; Keenan et al., 1984; Trabasso & Sperry, 1985; Myers et al., 1987; Duffy et al., 1990). We treat causal inferences about illness as a case study of a circumscribed yet highly familiar and motivationally relevant domain (Keil et al., 1999; Legare & Shtulman, 2018). We asked participants to read two-sentence vignettes (e.g., "Hugh sat by sneezing passengers on the subway. Now he has a case of COVID"). The first sentence described a potential cause and the second sentence a potential effect. Participants performed a covert task of detecting 'magical' catch-trial vignettes.

Causal inferences about illness were compared to two control conditions: i) causal inferences about mechanical failure (e.g., "Jake dropped all of his things on the subway. Now he has a shattered phone") and ii) vignettes that contained illness-related language but were causally unrelated (e.g., "Lynn dropped all of her things on the subway. Now she has a case of COVID"). This combination of control conditions allowed us to test jointly for sensitivity to content domain and causality. Critically, all vignettes, including mechanical ones, described events involving people, such that responses to causal inferences about illness in the animacy network could not be explained by the presence of animate agents in the stories. Non-causal vignettes were constructed by shuffling the causes/effects across conditions and were therefore matched in linguistic content to the causal vignettes. A separate group of participants rated the causal relatedness of all vignettes prior to the experiment. We predicted that illness inferences would activate the PC relative to both the mechanical inference (causal non-illness) and non-causal control conditions. We also localized language and logical reasoning networks in each participant to test the alternative but not mutually exclusive prediction that the language and/or logical reasoning networks respond preferentially to causal inference regardless of content domain (i.e., illness, mechanical).

Method

Open science practices

The methods and analysis of this experiment were pre-registered prior to data collection (*https://osf.io/cx9n2/*).

Participants

Twenty adults (7 females, 13 males, 25-37 years old, *M* = 28.7 years ± 3.2 SD) participated in the study. Participants either had or were pursuing graduate degrees (*M* = 8.8 years of post-secondary education). Two additional participants were excluded from the final dataset due to excessive head motion (> 2 mm) and an image artifact. One participant in the final dataset exhibited excessive head motion (> 2 mm) during 1 run of the language/logic localizer task that was excluded from analysis. All participants were screened for cognitive and neurological disabilities (self-report). Participants gave written informed consent and were compensated $30 per hour. The study was reviewed and approved by the Johns Hopkins Medicine Institutional Review Boards.

Causal inference experiment

Stimuli

Participants read two-sentence vignettes in 4 conditions, 2 causal and 2 non-causal (**Figure 1C**). Each vignette focused on a single agent, specified by a proper name in the initial sentence and by a pronoun in the second sentence. The first sentence described something the agent did or experienced and served as the potential cause. The second sentence described the potential effect (e.g., "Kelly shared plastic toys with a sick toddler at her preschool. Now she has a case of chickenpox"). *Illness-Causal* vignettes elicited inferences about biological causes of illness, including pathogen transmission, exposure to environmental toxins, and genetic mutations (see Supplementary Table 1 for a full list of the types of illnesses included in our stimuli).

Mechanical-Causal vignettes elicited inferences about physical causes of structural damage to personally valuable inanimate objects (e.g., houses, jewelry). Two non-causal conditions used the same sentences as in the *Illness-Causal* and *Mechanical-Causal* conditions but in a shuffled order: illness cause with mechanical effect (*Non-Causal Illness First)* or mechanical cause with illness effect (*Non-Causal Mechanical First*). Explicit causality judgments collected from a separate group of online participants (n=26) verified that the both causal conditions *Illness-Causal*, *Mechanical-Causal*) were more causally related than both non-causal conditions, $t(25) = 36.97$, $p < .0001$. In addition, *Illness-Causal* and *Mechanical-Causal* items received equally high causality ratings, $t(25) = –0.64$, $p = 0.53$ (see Appendix 1 for details).

Illness-Causal and *Mechanical-Causal* vignettes were constructed in pairs such that each member of a given pair shared parallel or near-parallel phrase structure. All conditions were also matched (pairwise t-tests, all *p*s > 0.3, no statistical correction) on multiple linguistic variables known to modulate neural activity in language regions (e.g., Pallier, Devauchelle, & Dehaene, 2011; Shain, Blank et al., 2020). These included number of characters, number of words, average number of characters per word, average word frequency, average bigram surprisal (Google Books Ngram Viewer, *https://books.google.com/ngrams/*), and average syntactic dependency length (Stanford Parser; de Marneffe, MacCartney, & Manning, 2006). Word frequency was calculated as the negative log of a word's occurrence rate in the Google corpus between the years 2017-2019. Bigram surprisal was calculated as the negative log of the frequency of a given two-word phrase in the

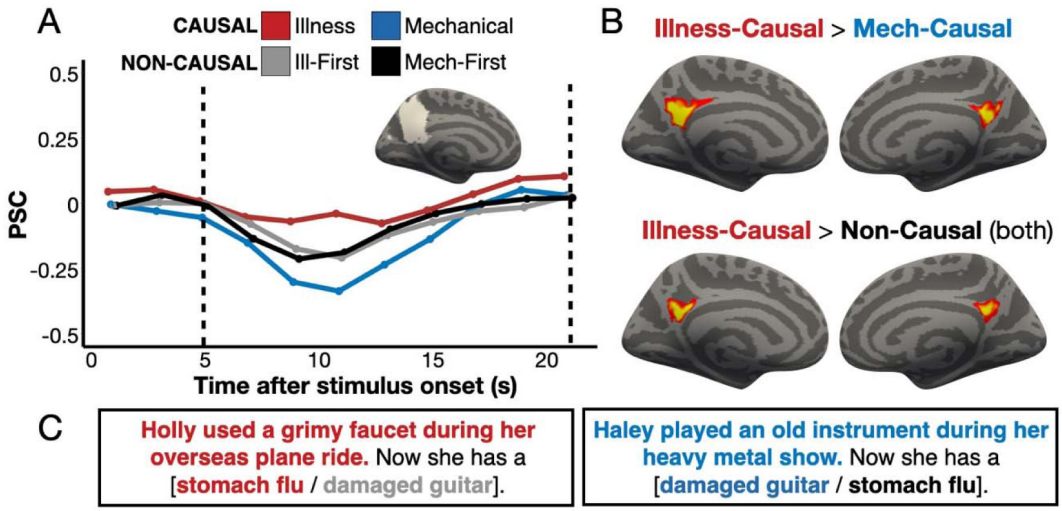


Figure 1.

Responses to illness inferences in the precuneus (PC).

Panel A: Percent signal change (PSC) for each condition among the top 5% *Illness-Causal* > *Mechanical-Causal* vertices in a left PC mask (Dufour et al., 2013) in individual participants, established via a leave-one-run-out analysis. Panel B: Whole-cortex results (one-tailed) for *Illness-Causal* > *Mechanical-Causal* and *Illness-Causal* > *Non-Causal* (both versions of non-causal vignettes), corrected for multiple comparisons ($p < .05$ FWER, cluster-forming threshold $p < .01$ uncorrected). Vertices are color coded on a scale from p=0.01 to p=0.00001. Panel C: Example stimuli. 'Magical' catch trials similar in meaning and structure (e.g., "Sadie forgot to wash her face after she ran in the heat. Now she has a cucumber nose") enabled the use of a semantic 'magic detection' task.

Google corpus divided by the frequency of the first word of the phrase (see Appendix 2 for details). All conditions were matched for all linguistic variables across the first sentence, second sentence, and the entire vignette.

Procedure

We used a 'magic detection' task to encourage participants to process the meaning of the vignettes without making explicit causality judgments. Participants saw 'magical' catch trials that closely resembled the experimental trials but were fantastical (e.g., "Sadie forgot to wash her face after she ran in the heat. Now she has a cucumber nose."). On each trial, participants indicated via button press whether 'something magical' occurred in the vignette (Yes/No). Both sentences in a vignette were presented simultaneously for 7 s, one above the other, followed by a 12 s inter-trial interval. Each participant saw 38 trials per condition plus 36 'magical' catch trials (188 total trials) in one of two versions, counterbalanced across participants, such that individual participants did not see the same sentence in both causal and non-causal vignettes. The two stimulus versions had similar meanings but different surface forms (e.g., "Luna stood by coughing travelers on the train…" vs. "Hugh sat by sneezing passengers on the subway…").

The experiment was divided into 6 10-minute runs containing a similar number of trials per condition per run presented in a pseudorandom order. Specifically, vignettes from the same experimental condition repeated no more than twice consecutively, vignettes that were constructed to share a similar phrase structure never repeated within a run, vignettes that referred to the same illness never repeated consecutively, and vignettes from each condition, including catch trials, were equally distributed in time across the course of the experiment.

Language/logic localizer experiment

A localizer task was used to identify the language and logic networks in each participant. The task had three conditions: language, logic, and math. In the language condition, participants judged whether two visually presented sentences, one in active and one in passive voice, shared the same meaning. In the logic condition, participants judged whether two logical statements were consistent (e.g., *If either not Z or not Y then X* vs. *If not X then both Z and Y*). In the math condition, participants judged whether the variable *X* had the same value across two equations (for details see Liu et al., 2020). Trials lasted 20 s (1 s fixation + 19 s display of stimuli) and were presented in an event-related design. Participants completed 2 9-minute runs of the task, with trial order counterbalanced across runs and participants. Following prior studies, the language network was identified in individual participants by contrasting *language* > *math* and the logic network by contrasting *logic* > *language* (Monti et al., 2009; Kanjlia et al., 2016; Liu et al., 2020).

Mentalizing localizer experiment

An additional localizer task was used to identify the mentalizing network in each participant (Saxe & Kanwisher, 2003; Dodell-Feder et al., 2011; *http://saxelab.mit.edu/use-our-efficient-false-belief-localizer*). In this task, participants read 10 mentalizing stories (e.g., a protagonist has a false belief about an object's location) and 10 physical stories (physical representations depicting outdated scenes, e.g., a photograph showing an object that has since been removed) before answering a true/false comprehension question. We used the mentalizing stories from the original localizer but created new stimuli for the physical stories condition. Our physical stories incorporated more vivid descriptions of physical interactions and did not make any references to human agents. They were also linguistically matched to the mentalizing stories to reduce linguistic confounds (see Shain et al., 2022). Specifically, we matched physical and mentalizing stories (pairwise t-tests, all *p*s > 0.3, no statistical correction) for number of characters, number of words, average number of characters per word, average syntactic dependency length, average word frequency, and average bigram surprisal, as was done for the causal inference vignettes. A comparison of both localizers in 3 pilot participants can be found in Supplementary Figure 9.

Trials were presented in an event-related design, with each one lasting 16 s (12 s stories + 4 s comprehension question) followed by a 12 s inter-trial interval. Participants completed 2 5-minute runs of the task, with trial order counterbalanced across runs and participants. The mentalizing network was identified in individual participants by contrasting *mentalizing stories* > *physical stories* (Saxe & Kanwisher, 2003; Dodell-Feder et al., 2011).

Data acquisition

Whole-brain fMRI data was acquired at the F.M. Kirby Research Center of Functional Brain Imaging on a 3T Phillips Achieva Multix X-Series scanner. T1-weighted structural images were collected in 150 axial slices with 1 mm isotropic voxels using the magnetization-prepared rapid gradient-echo (MP-RAGE) sequence. T2*-weighted functional BOLD scans were collected in 36 axial slices (2.4 2.43 mm voxels, TR = 2 s). Data were acquired in one experimental session lasting approximately 120 minutes. All stimuli were visually presented on a rear projection screen with a Cambridge Research Systems BOLDscreen 32 UHD LCD display (image resolution = 1920 x 1080) using custom scripts written in PsychoPy3 (*https://www.psychopy.org/*, Peirce et al., 2019). Participants viewed the screen via a front-silvered, 45° inclined mirror attached to the top of the head coil.

fMRI data preprocessing and general linear model (GLM) analysis

Preprocessing included motion correction, high-pass filtering (128 s), mapping to the cortical surface (Freesurfer), spatially smoothing on the surface (6 mm FWHM Gaussian kernel), and prewhitening to remove temporal autocorrelation. Covariates of no interest included signal from white matter, cerebral spinal fluid, and motion spikes.

For the main causal inference experiment, the GLM modeled the four main conditions (*Illness-Causal, Mechanical-Causal, Non-Causal Illness First, Non-Causal Mechanical First*) as well as the 'magical' catch trials during the 7 s display of the vignettes after convolving with a canonical hemodynamic response function and its first temporal derivative. For the language/logic localizer experiment, a separate predictor was included for each of the three conditions (*language, logic, math*), modeling the 20 s duration of each trial. For the mentalizing localizer experiment, a separate predictor was included for each condition (*mentalizing stories, physical stories*), modeling the 16 s display of each story and corresponding comprehension question.

For each task, runs were modeled separately and combined within subject using a fixed-effects model (Dale, Fischl, & Sereno, 1999; Smith et al., 2004). Group-level random-effects analyses were corrected for multiple comparisons across the whole cortex at $p < .05$ family-wise error rate (FWER) using a nonparametric permutation test (cluster-forming threshold $p < .01$ uncorrected) (Winkler et al., 2014; Eklund, Nichols, & Knutsson, 2016; Eklund, Knutsson, & Nichols, 2019). A control analysis of the causal inference task that modeled participant response time and number of people in each vignette revealed similar results to a model without the covariates added. We therefore report only the results of the model without the covariates.

Individual-subject ROI analysis (univariate)

We defined individual-subject functional ROIs in the PC, TPJ, language (frontal and temporal masks), and logic networks. In an exploratory analysis, we defined individual-subject fROIs in anterior medial VOTC. For all analyses, PSC was extracted and averaged over the entire duration of the trial (17 s total), allowing 4 s to account for the hemodynamic lag.

Illness inference ROIs were created in left and right PC group search spaces (Dufour et al., 2013) using an iterated leave-one-run-out procedure, which allowed us to perform sensitive individual-subjects analysis while avoiding double-dipping (Vul & Kanwisher, 2011). In each participant, we identified the most illness inference-responsive vertices in left and right PC search spaces in 5 of

the 6 runs (top 5% of vertices, *Illness-Causal* > *Mechanical-Causal*). We then extracted percent signal change (PSC) for each condition compared to rest in the held-out run (*Illness-Causal, Mech-Causal, Non-Causal Illness First, Non-Causal Mechanical First*), averaging the results across all iterations. We performed the same analysis using left and right TPJ search spaces (Dufour et al., 2013). We used the same approach to create mechanical inference ROIs in left and right anterior medial VOTC search spaces from a previous study on place word representations (Hauptman, Elli, et al., 2023). All aspects of this analysis were the same as those described above, except that the most mechanical inference-responsive vertices (top 5%, *Mechanical-Causal* > *Illness-Causal*) were selected.

Mentalizing ROIs were created by taking the most mentalizing-responsive vertices (top 5%) in bilateral PC and TPJ search spaces (Dufour et al., 2013) using the *mentalizing stories* > *physical stories* contrast from the mentalizing localizer. Language ROIs were identified by taking the most language-responsive vertices (top 5%) in left frontal and temporal language areas (search spaces: Fedorenko et al., 2010) using the *language* > *math* contrast from the language/logic localizer. A logic-responsive ROI was identified by taking the most logic-responsive vertices (top 5%) in a left frontoparietal network (search space: Liu et al., 2020) using the *logic* > *language* contrast. In each ROI, we extracted PSC for all conditions in the causal inference experiment.

ROI MVPA

We performed MVPA (PyMVPA toolbox; Hanke et al., 2009) to test whether patterns of activity in the PC distinguished illness inferences from mechanical inferences. In each participant, we identified the top 300 vertices most responsive to causal inference across domains (i.e., both *Illness-Causal* + *Mech-Causal* > Rest) in a left PC mask (Dufour et al., 2013).

For each vertex in each participant's ROIs, we obtained one observation per condition per run (z-scored beta parameter estimate of the GLM). A linear support vector machine (SVM) was then trained on data all but one of the runs and tested on the left-out run in a cross-validation procedure. Classification accuracy was averaged across all permutations of the training/test splits. We compared classifier performance within each ROI to chance (50%; one-tailed test).

Significance was evaluated against an empirically generated null distribution using a combined permutation and bootstrap approach (Schreiber & Krekelberg, 2013; Stelzer et al., 2013). In this approach, t-statistics obtained for the observed data are compared against an empirically generated null distribution. We report the t-values obtained for the observed data and the nonparametric p-values, where p corresponds to the proportion of the shuffled analyses that generated a comparable or higher t-value.

The null distribution was generated using a balanced block permutation test by shuffling condition labels within run 1000 times for each subject (Schreiber & Krekelberg, 2013). Then, a bootstrapping procedure was used to generate an empirical null distribution for each statistical test across participants by sampling one permuted accuracy value from each participant's null distribution 15,000 times (with replacement) and running each statistical test on these permuted samples, thus generating a null distribution of 15,000 statistical values for each test (Stelzer et al. 2013).

Searchlight MVPA

We used a whole-brain SVM classifier to decode among *Illness-Causal* vs. *Mechanical-Causal*, *Illness-Causal* vs. *Non-Causal Mechanical First*, *Illness-Causal* vs. *Non-Causal Illness First*, and *Causal* (both) vs. *Non-Causal* (both) over the whole cortex using a 10 mm radius spherical searchlight (according to geodesic distance, to better respect cortical anatomy over Euclidean distance; Glasser et al., 2013). This yielded for each participant 4 classification maps indicating the classifier's accuracy in a neighborhood surrounding every vertex. Individual subject

searchlight accuracy maps were then averaged, and the resulting group-wise map was thresholded using the PyMVPA implementation of the 2-step cluster-thresholding procedure described in Stelzer et al. (2013) (Hanke et al., 2009). This procedure permutes block labels within participant to generate a null distribution within subject (100 times) and then samples from these (10,000) to generate a group-wise null distribution (as in the ROI analysis). The whole-brain searchlight maps are then thresholded using a combination of vertex-wise threshold ($p < 0.001$ uncorrected) and cluster size threshold (FWER $p < 0.05$, corrected for multiple comparisons across the entire cortical surface).

Results

Behavioral results

Accuracy on the magic detection task was at ceiling (M = 97.9% ± 2.2 SD) and there were no significant differences across the 4 main experimental conditions (*Illness-Causal*, *Mechanical-Causal*, *Non-Causal Illness First*, *Non-Causal Mechanical First*), $F(3,57) = 2.39$, $p = .08$. A one-way repeated measures ANOVA evaluating response time revealed a main effect of condition, $F(3,57) = 32.63$, $p < .0001$, whereby participants were faster on *Illness-Causal* trials (M = 4.73 ± 0.81 SD) compared to *Non-Causal Illness First* (M = 5.33 s ± 0.85 SD) and *Non-Causal Mechanical First* (M = 5.27 s ± 0.89 SD) trials. There were no differences in response time between *Mechanical-Causal* (M = 5.15 s ± 0.88 SD) and any other conditions. Performance on the localizer tasks was similar to previously reported studies that used these paradigms (see Appendix 3 for details).

Inferring illness causes recruits animacy-responsive PC

In whole-cortex analysis, left and right PC were the only regions to show a preference for causal inferences about illness over both mechanical inferences and causally unrelated sentences ($p < .05$, corrected for multiple comparisons; **Figure 1B**). PC responses during illness inferences overlap with previously reported responses to people-related concepts (Fairhall & Caramazza, 2013b; Supplementary Figure 2). In individual-subject fROI analysis, we similarly found that inferring illness causes activated the left and right PC more than inferring causes of mechanical failure (leave-one-run-out analysis; left: $F(1,19) = 28.69$, $p < .0001$; right: $F(1,19) = 5.14$, $p = .04$; **Figure 1A**). Illness inferences also activated left PC more than illness-related language that was not causally related (i.e., average of both non-causal conditions, $F(1,19) = 13.23$, $p < .01$). MVPA performed in PC fROIs and across the whole cortex similarly revealed that illness inferences and mechanical inferences produced spatially distinguishable neural patterns in left PC ($t(19) = 3.50$, permuted $p < .001$, Supplementary Table 2).

Inferring the causes of mechanical failure relative to both illness inferences and the non-causal conditions activated bilateral anterior medial ventral occipitotemporal-cortex (VOTC) (**Figure 4B**). This anterior medial VOTC region overlaps with the so-called anterior parahippocampal place area (PPA) (Epstein & Kanwisher, 1998; Weiner et al., 2017) and is engaged during memory and verbal tasks related to physical spaces (Baldassano et al., 2013; Fairhall et al., 2014; Silson et al., 2019; Häusler et al., 2022; Hauptman, Elli, et al., 2023). In individual-subject fROI analysis, we similarly found that mechanical inferences activated left anterior medial VOTC more than illness inferences (leave-one-run-out analysis; left: $F(1,19) = 16.05$, $p < .001$) and more than the non-causal conditions (left: $F(1,19) = 17.46$, $p < .0001$, **Figure 4A**)

Inferring illness causes is dissociable from other types of inference about animates within the PC

Within left PC, responses to illness inferences were spatially dissociable from responses to other types of inferences about animate entities (i.e., mental state inferences) in individual participants, with illness responses located more inferiorly (**Figure 2**, Supplementary Figure 3). In an

exploratory analysis, we quantified this effect by identifying the peak vertices for illness inferences (illness inferences > mechanical inferences) and mentalizing (mentalizing stories > physical stories) in individual participants within a left PC mask (Dufour et al., 2013). We then compared the z-coordinates of these peaks and observed a significant difference across participants, $F(1,19) = 13.52$, $p < .01$.

The most illness-responsive vertices in left PC exhibited a strong preference for mentalizing, $F(1,19) = 38.65$, $p < .0001$ (Supplementary Figure 8). The most mentalizing-responsive vertices in left PC also exhibited a preference for illness inferences over mechanical inferences, $F(1,19) = 6.48$, $p = .02$. However, this effect was weaker than the preference for illness inferences observed among the most illness inference-responsive vertices (leave-one-run-out analysis) in this region, $F(1,38) = 6.56$, $p = .01$ (Supplementary Figure 7). Together, these results suggest that illness inferences are carried out by a partially distinctive subset of the PC vertices captured by the mentalizing stories vs. physical stories contrast.

Responses to illness inferences and mentalizing in right TPJ provide further evidence of a partial dissociation between illness inferences and mental state inferences. In right TPJ, where the strongest univariate preference for mentalizing is classically observed (e.g., Saxe & Kanwisher, 2003; see Supplementary Figure 1 for results from our localizer), the most mentalizing-responsive vertices did not exhibit a robust univariate preference for illness inferences in individual-subject fROI analysis, $F(1,19) = 3.40$, $p = .08$. Patterns of activity in the same vertices similarly did not distinguish between illness inferences and mechanical inferences, $t(19) = 0.94$, permuted $p = .19$. These findings are consistent with the fact that right TPJ did not exhibit a preference for illness inferences in whole-cortex analysis ($p < .05$, corrected for multiple comparisons; Supplementary Figure 4).

No evidence for a content-invariant preference for causal inference in logic or language networks

Neither the logic nor the language network exhibited elevated neural responses during causal inferences. Language regions in frontotemporal cortex responded more to non-causal than causal vignettes (frontal search space: $F(1,19) = 23.91$, $p < .0001$; temporal search space: $F(1,19) = 4.31$, $p = .05$, **Figure 3**, Supplementary Figure 6). A similar pattern was observed in the logic network, $F(1,19) = 3.88$, $p = .07$ (**Figure 3**). These effects likely reflect the greater difficulty associated with integrating unrelated sentences. In whole-cortex analysis, no shared regions emerged across both causal contrasts (i.e., illness inferences > non-causal and mechanical inferences > non-causal). Although MVPA searchlight analysis identified several areas where patterns of activity distinguished between causal and non-causal vignettes, all of these regions showed a preference for non-causal vignettes in univariate analysis (Supplementary Figure 5).

Discussion

We find that an inferior precuneus (PC) region previously implicated in thinking about animates is preferentially active when participants infer causes of illness. The PC responded more to causal inferences about illness compared to vignettes that contained illness-related language but were causally unrelated. The PC also responded more to causal inferences about illness compared to causal inferences about mechanical failure. In the current study, we did not find any cortical areas that responded to implicit causal inferences across content domains. These findings suggest that implicit causal inferences about illness during language comprehension draw upon a content-specific semantic network for representing animacy.

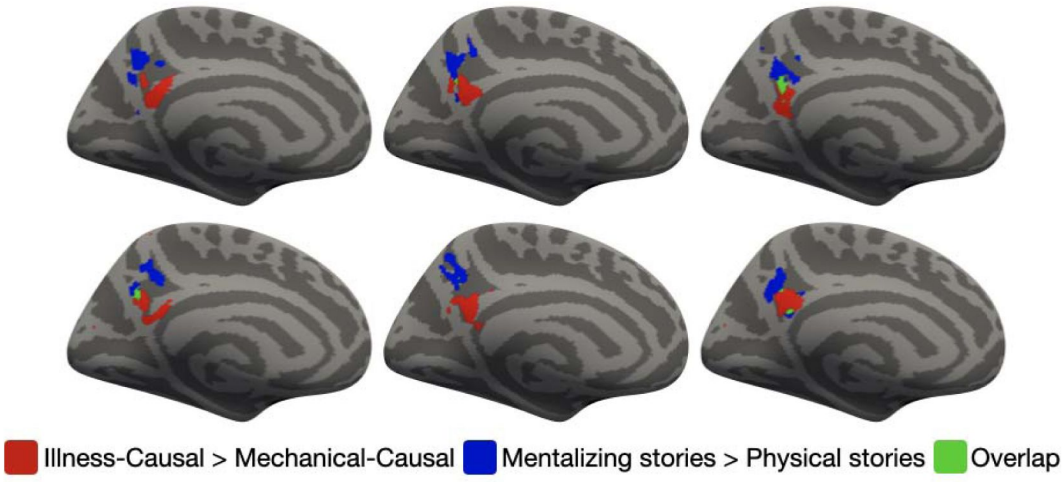


Figure 2.

Spatial dissociation between responses to illness inferences and mental state inferences in the precuneus (PC).

The left medial surface of 6 individual participants were selected for visualization purposes. The locations of the top 10% most responsive vertices to *Illness-Causal* > *Mechanical-Causal* in a PC mask (Dufour et al., 2013) are shown in red. The locations of the top 10% most responsive vertices to *mentalizing stories* > *physical stories* (mentalizing localizer) in the same PC mask are shown in blue. Overlapping vertices are shown in green.

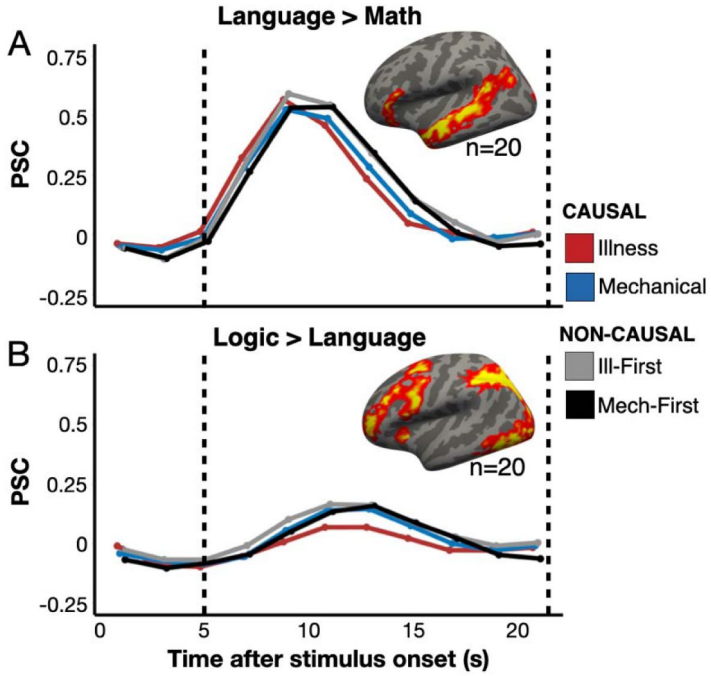


Figure 3.

Individual-subjects analysis of language– and logic-responsive vertices.

Panel A: percent signal change (PSC) for each condition among the top 5% most language-responsive vertices (*language* > *math*) in a temporal language network mask (Fedorenko et al., 2010). Results from a frontal language mask (Fedorenko et al., 2010) can be found in Supplementary Figure 6. Panel B: PSC among the top 5% most logic-responsive vertices (*logic* > *language*) in a logic network mask (Liu et al., 2020). Group maps for each contrast of interest (one-tailed) are corrected for multiple comparisons ($p < .05$ FWER, cluster-forming threshold $p < .01$ uncorrected). Vertices are color coded on a scale from p=0.01 to p=0.00001.

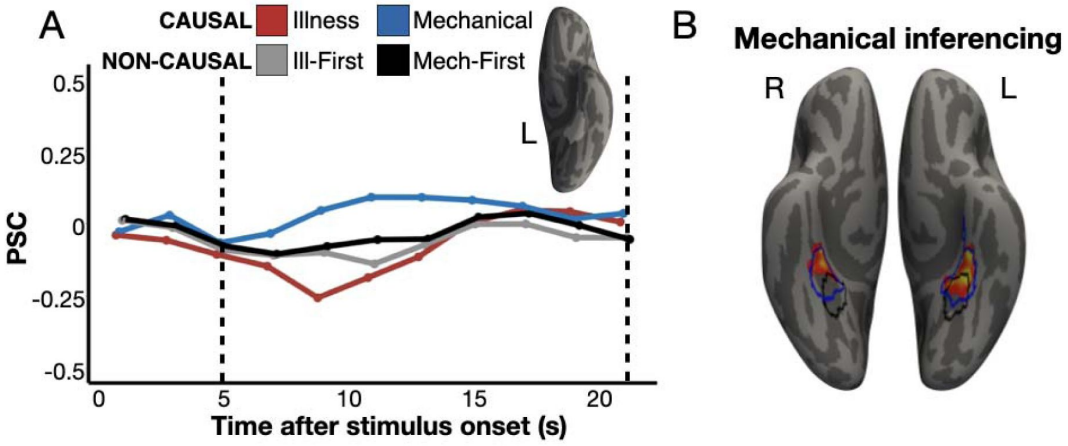


Figure 4.

Responses to mechanical inferences in anterior medial ventral occipito-temporal cortex (VOTC).

Panel A: Percent signal change (PSC) for each condition among the top 5% *Illness-Causal* > *Mechanical-Causal* vertices in a left anterior medial VOTC mask (Hauptman, Elli, et al., 2023) in individual participants, established via a leave-one-run-out analysis. Panel B: The intersection of two whole-cortex contrasts, *Mechanical-Causal* > *Illness-Causal* and *Mechanical-Causal* > *Non-Causal*, FWER cluster-correction for multiple comparisons ($p < .05$ FWER, cluster-forming threshold $p < .01$ uncorrected). Vertices are color coded on a scale from p=0.01 to p=0.00001. Similar to PC responses to illness inferences, anterior medial VOTC is the only region to emerge across both mechanical inference contrasts. The average PPA location from separate study involving perceptual place stimuli (Weiner et al., 2017) is overlaid in black. The average PPA location from separate study involving verbal place stimuli (Hauptman, Elli, et al., 2023) is overlaid in blue.

Previous work has implicated the PC in the representation of animate entities, such as people and animals (Fairhall & Caramazza, 2013a, 2013b; Fairhall et al., 2014; Peer et al., 2015; Wang et al., 2016; Silson et al., 2019; Rabini, Ubaldi, & Fairhall, 2021; Deen & Freiwald, 2022; Aglinskas & Fairhall, 2023; Hauptman, Elli, et al., 2023). The present results are consistent with these findings and expand upon them by showing that inferring the causes of illness, a phenomenon that is exclusive to animates, activates the PC. Thus, the PC exhibits sensitivity to causal inferences about processes specific to animates (e.g., illness) beyond the mere mention of animate entities.

The finding that the PC is sensitive to causal inferences across sentences is also consistent with prior evidence that the PC is involved in discourse-level processes and responds to semantic information across long timescales during narrative comprehension (e.g., Hasson et al., 2008; Lerner et al., 2011; Lee & Chen, 2022). PC responses observed during narrative comprehension could be driven by causal inferences related to animacy, since narratives are rich in information about the behavior of animate agents. Likewise, PC involvement in episodic memory could be related to animacy-related inferential processes (DiNicola, Braga, & Buckner, 2020; Ritchey & Cooper, 2020). Whether the PC is uniquely relevant to *causal* inferences compared to other types of inferences (e.g., temporal, referential; Graesser et al., 1994) about animates remains to be tested.

The results of the present study suggest that causal knowledge is embedded in semantic networks that represent animates. This hypothesis is consistent with evidence from developmental psychology that causal knowledge is central to our understanding of animacy. For example, preschoolers intuit that animates but not inanimate objects get sick, need nourishment to grow and live, and can die (e.g., Rosengren et al., 1991; Kalish, 1996; Gutheil, Vera, & Keil, 1998; Raman & Gelman, 2005; see Inagaki & Hatano, 2004; Opfer & Gelman, 2011 for reviews). Early emerging intuitions about the biological world are present across cultures and are sometimes referred to as 'intuitive biology' (Keil, 1992; Wellman & Gelman, 1992; Hatano & Inagaki, 1994; Simons & Keil, 1995; Atran, 1998; Keil et al., 1999; Coley, Solomon, & Shafto, 2002; Medin & Atran, 2004). The current study suggests that such knowledge is encoded in the animacy semantic network. In future work, it will be important to test whether the inferior PC is sensitive to causal knowledge about biological processes beyond illness, such as growth, inheritance, and reproduction. Another open question concerns how cultural expertise in causal reasoning about illness (e.g., medical expertise) influences representations in the PC.

Our findings additionally suggest that inferences about the biological and mental properties of animates are linked yet neurally separable. We find that inferring illness causes recruits a subregion of the PC that neighbors but is distinct from peak PC responses to mental state inferences (Saxe & Kanwisher, 2003; Saxe et al., 2006). The neural distinction between body and mind is consistent with developmental work showing that even young children provide different causal explanations for biological vs. psychological processes (Springer & Keil, 1991; Callanan & Oakes, 1992; Wellman & Gelman, 1992; Inagaki & Hatano, 1993; 2004; Keil, 1994; Hickling & Wellman, 2001; Medin et al., 2010; cf. Carey, 1985; see also Medin & Atran, 2004). For example, when asked why blood flows to different parts of the body, 6-year-olds endorse explanations referring to bodily function, e.g., "because it provides energy to the body" and not to mental states "because we want it to flow" (Inagaki & Hatano, 1993). At the same time, animate entities have a dual nature: they have both bodies and minds (Opfer & Gelman, 2011; Spelke, 2023). The current findings are consistent with the notion of distinct but partially overlapping systems for biological and mentalistic knowledge.

In addition to responses to illness inferences in the PC, we find that mechanical inferences activate an anterior portion of left medial ventral occipito-temporal cortex (VOTC), a region that has been previously implicated in supporting abstract (i.e., non-perceptual) place representations (Baldassano et al., 2013; Fairhall et al., 2014; Silson et al., 2019; Häusler et al., 2022;

Hauptman, Elli, et al., 2023). This finding illustrates a neural double dissociation between biological and mechanical causal knowledge and suggests that neural systems representing semantic knowledge are sensitive to causal information in their respective domain. Our neuroscientific evidence coheres with the 'intuitive theories' proposal, according to which semantic knowledge is organized into causal frameworks that serve as 'grammars for causal inference' (Tenenbaum et al., 2007; Wellman & Gelman, 1992; Gopnik & Meltzoff, 1997; Gerstenberg & Tenenbaum, 2017; see also Boyer 1995; Barrett, Cosmides, & Tooby, 2007; Cosmides & Tooby, 2013; Bender, Beller, & Medin, 2017). It is worth noting that many real-world causal inferences, including inferences about illness, combine knowledge from multiple domains (e.g., intentional bewitchment and viral infection together cause illness; Lynch & Medin, 2006; Legare & Gelman, 2008; Legare & Shtulman, 2018). Such causal inferences might recruit multiple semantic networks.

In the current study, we failed to find a neural signature of domain-general causal inference. That is, no brain region responded more to causal than non-causal vignettes across domains. The language network responded more to non-causal than causal vignettes, perhaps because of greater integration demands associated with the non-causal condition. This result aligns with evidence that the language network is specialized for sentence-internal processing and is not sensitive to inferences at the discourse level (Fedorenko & Varley, 2016; Jacoby & Fedorenko, 2020; Blank & Fedorenko, 2020). Interestingly, responses to causal inference in semantic networks (i.e., PC, anterior medial VOTC) were stronger in the left hemisphere. The left lateralization of responses to causal inference may enable efficient interfacing with the language system during discourse comprehension. The frontoparietal logical reasoning network similarly did not exhibit a domain-general preference for causal over non-causal vignettes. This finding coheres with prior evidence that the logic network supports inferences involving symbolic, 'content-free' stimuli, such as *If X then Y = If not Y then X* (Monti et al., 2009; Feng et al., 2021). In whole cortex analysis, we also did not find any region that showed a general preference for causal inferences. Our results suggest that implicit causal inferences do not depend on domain-general causal inference machinery.

These findings do not rule out the possibility that domain-general inference mechanisms contribute to causal inference under some circumstances. Causal inferences are a highly varied class. Here we focused on implicit causal inferences that unfold spontaneously during language comprehension. A centralized mechanism could still enable causal inferences during more complex explicit tasks, even explicit inferences about the causes of illness. The vignettes used in the current study stipulate illness causes, allowing participants to reason from causes to effects. By contrast, illness reasoning performed by medical experts proceeds from effects to causes and can involve searching for potential causes within highly complicated and interconnected causal systems (Schmidt, Norman, & Boshuizen, 1990; Norman et al., 2006; Meder & Mayrhofer, 2017). Domain-general mechanisms may also be relevant to learning novel causal relationships, such as identifying new illness causes or inferring causal powers of unfamiliar objects (e.g., 'blicket detectors'; Gopnik et al., 2001). Future neuroimaging work can help test these possibilities.

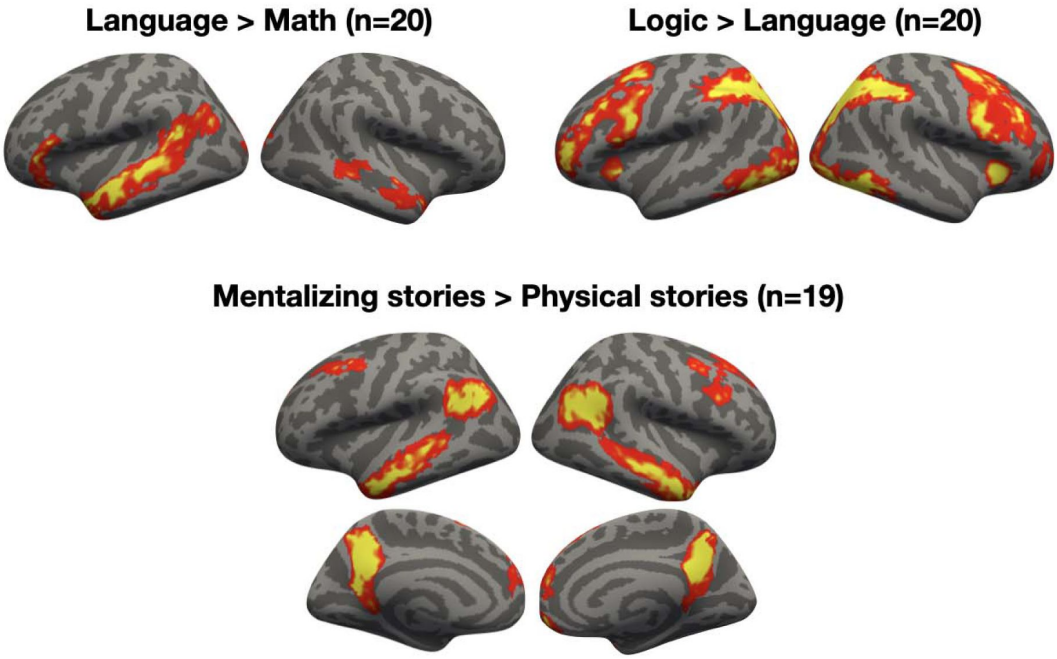
Supplementary figures and tables

Supplementary Figure 1.

Functional localization of language (Liu et al., 2020), logical reasoning (Liu et al., 2020), and mentalizing (Dodell-Feder et al., 2011) networks.

Group maps for each contrast of interest (one-tailed) are corrected for multiple comparisons (p < .05 FWER, cluster-forming threshold p < .01 uncorrected). Vertices are color coded on a scale from p=0.01 to p=0.00001.

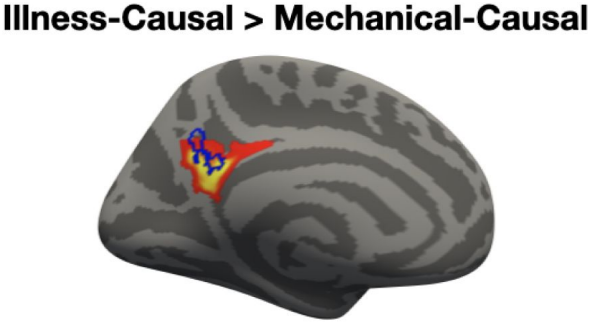
Supplementary Figures

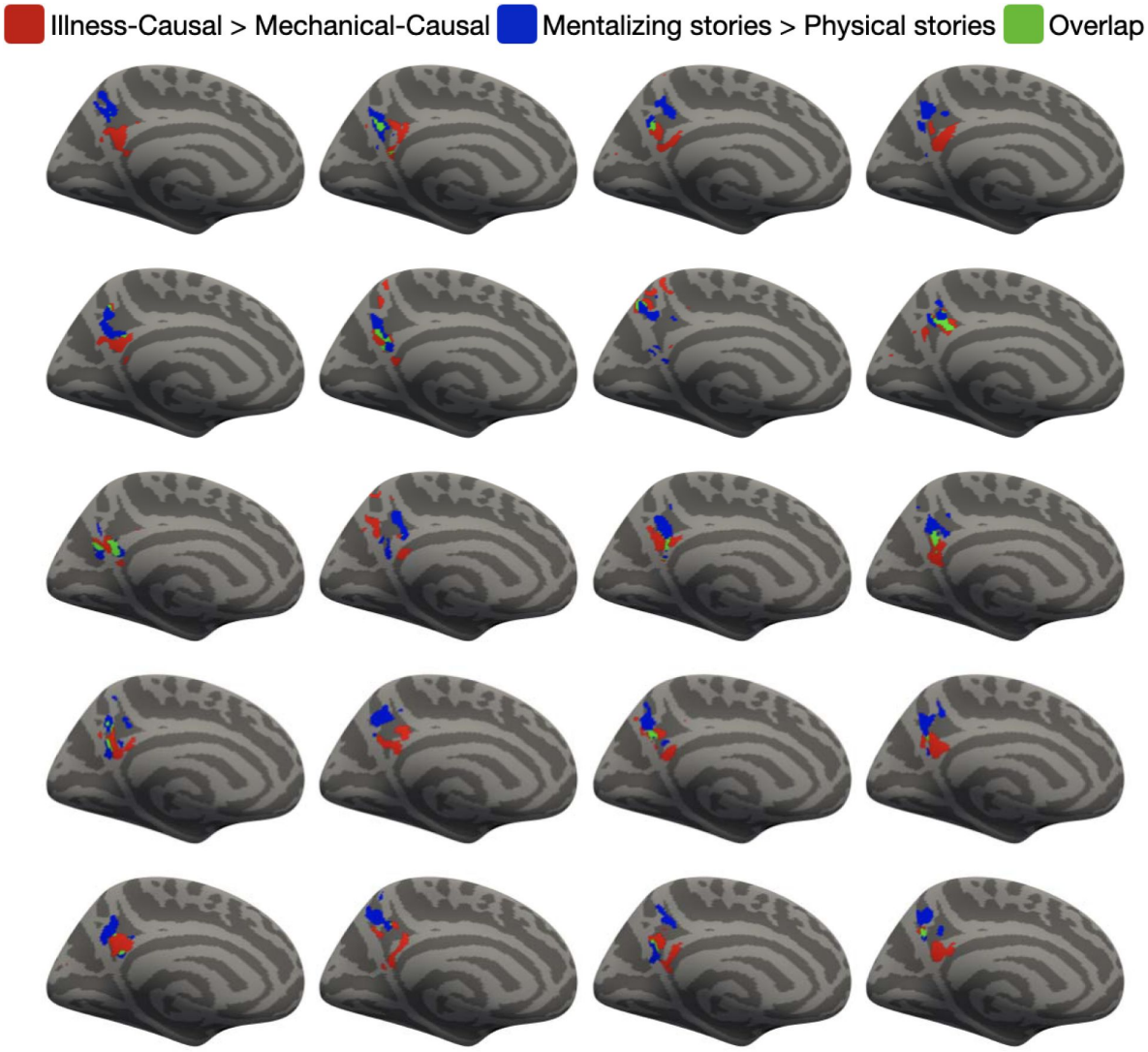


Supplementary Figure 2.

Overlap between left precuneus (PC) responses to illness inferences in the current study and people stimuli in a separate study (Fairhall & Caramazza, 2013b).

The average location from a separate study comparing people and place concepts (Fairhall & Caramazza, 2013b) is overlaid in blue on the response to illness inferences observed in the current study. Group map (one-tailed) is corrected for multiple comparisons (p < .05 FWER, cluster-forming threshold p < .01 uncorrected). Vertices are color coded on a scale from p=0.01 to p=0.00001.

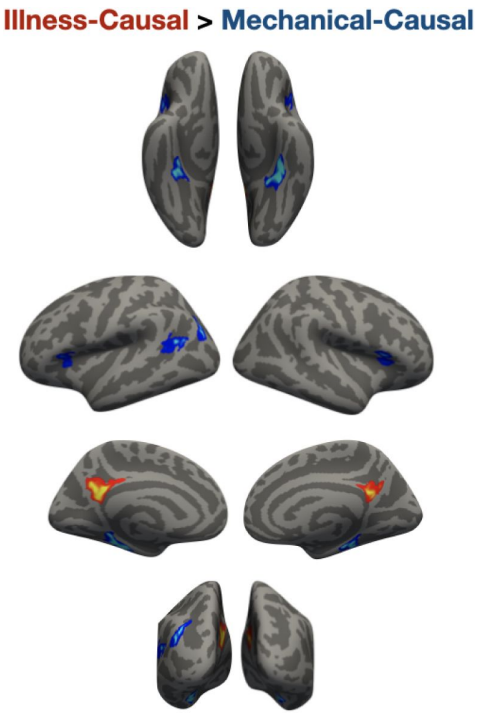




Supplementary Figure 3.

Spatial dissociation between responses to illness inferences and mental state inferences in the left precuneus (PC).

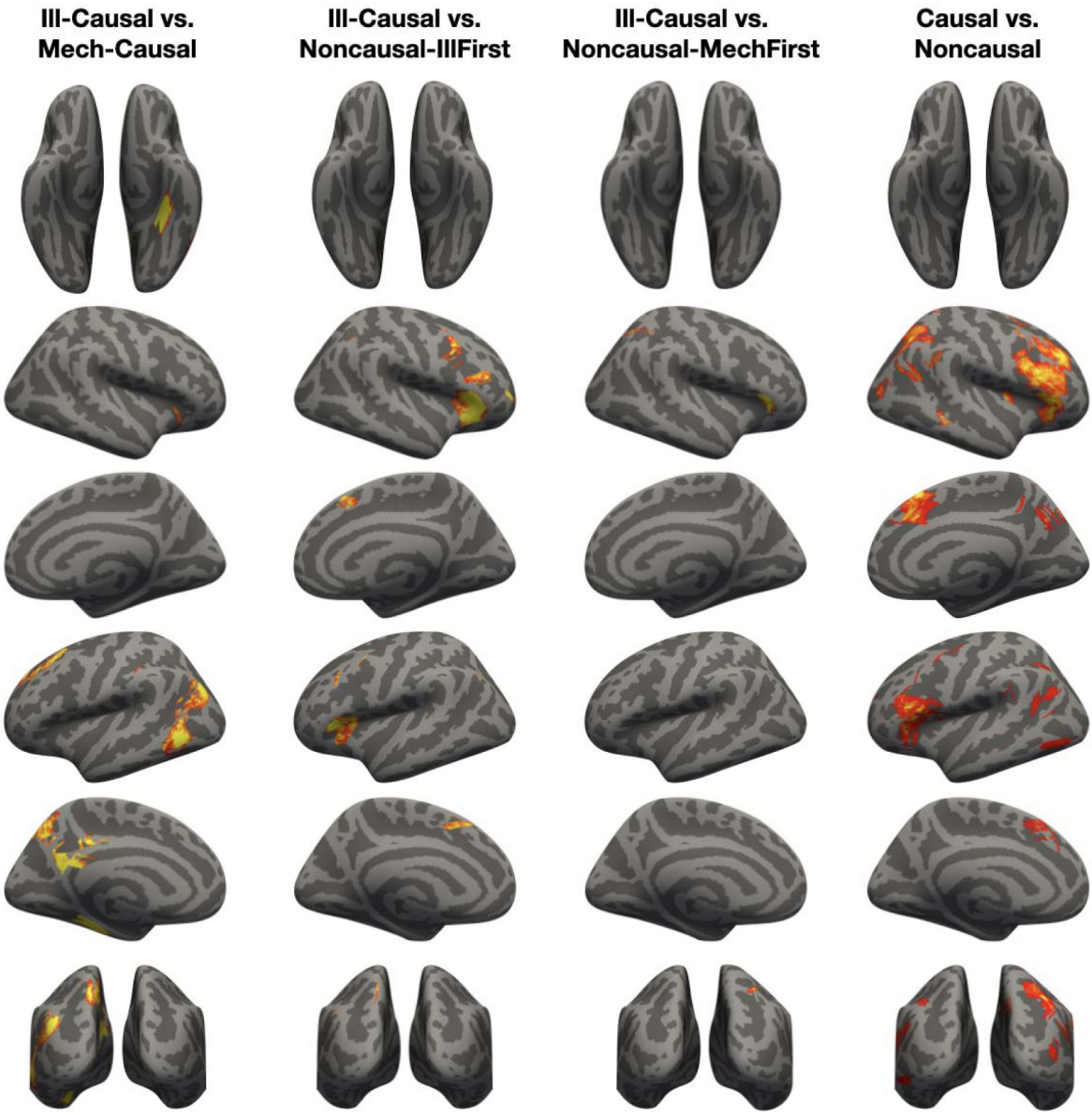
The left medial surface of all participants (n=20) is shown. The locations of the top 10% most responsive vertices to *Illness-Causal* > *Mechanical-Causal* in a PC mask (Dufour et al., 2013) are shown in red. The locations of the top 10% most responsive vertices to *mentalizing stories* > *physical stories* (mentalizing localizer) in the same PC mask are shown in blue. Overlapping vertices are shown in green.



Supplementary Figure 4.

Full whole-cortex results for *Illness-Causal* > *Mechanical-Causal*.

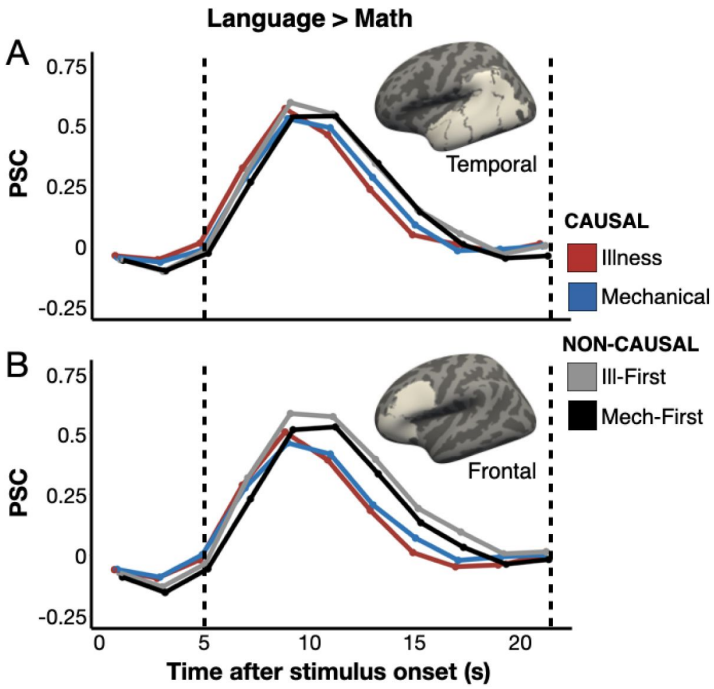
Group maps (two-tailed) are corrected for multiple comparisons ($p < .05$ FWER, cluster-forming threshold $p < .01$ uncorrected). Vertices are color coded on a scale from p=0.01 to p=0.00001.



Supplementary Figure 5.

Searchlight MVPA group maps.

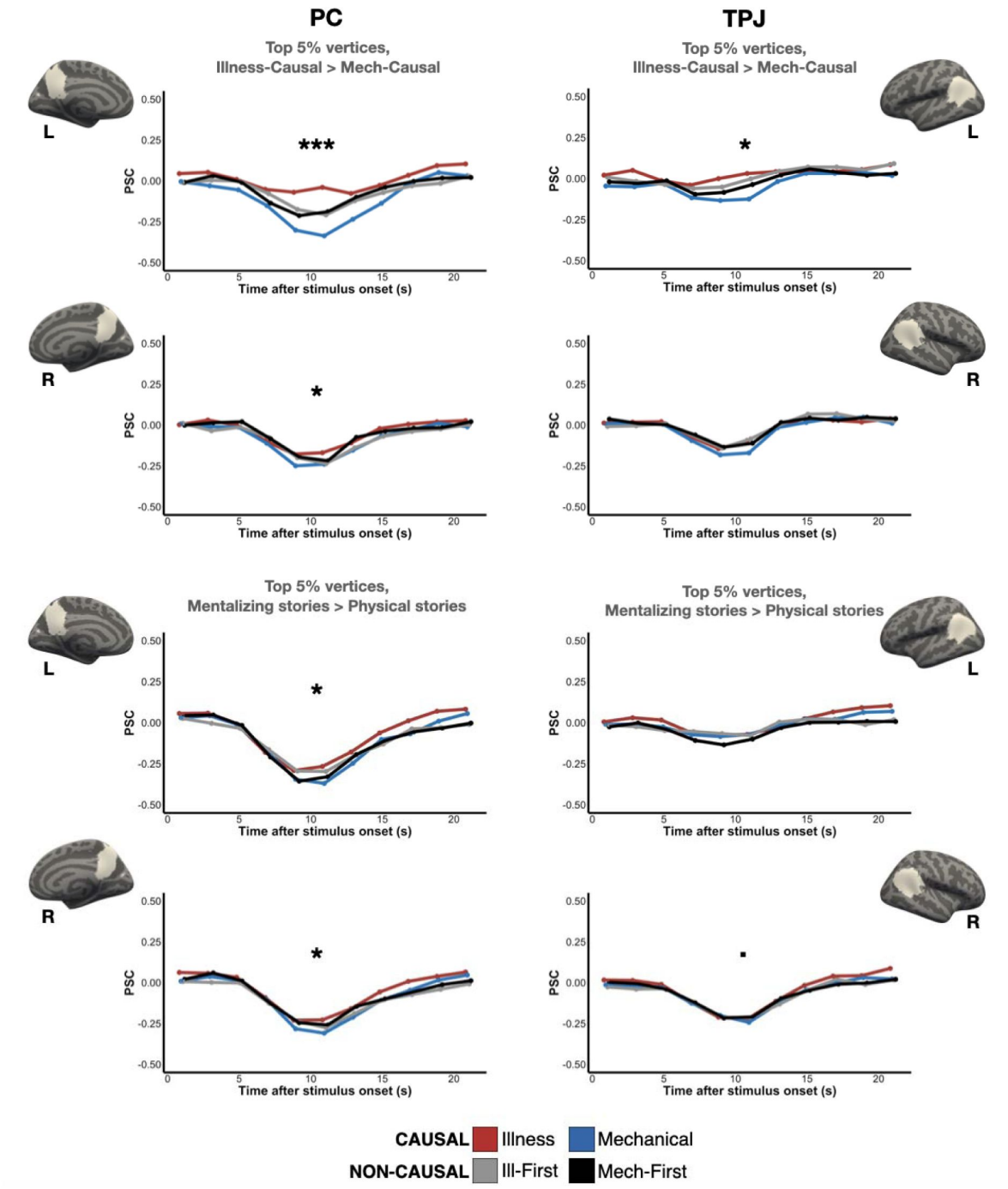
Whole-brain searchlight maps were thresholded using a combination of vertex-wise threshold ($p < 0.001$ uncorrected) and cluster size threshold (FWER $p < 0.05$, corrected for multiple comparisons across the entire cortical surface). Vertices are color coded on a scale from 55-65% decoding accuracy.



Supplementary Figure 6.

Responses to causal inference in the language network.

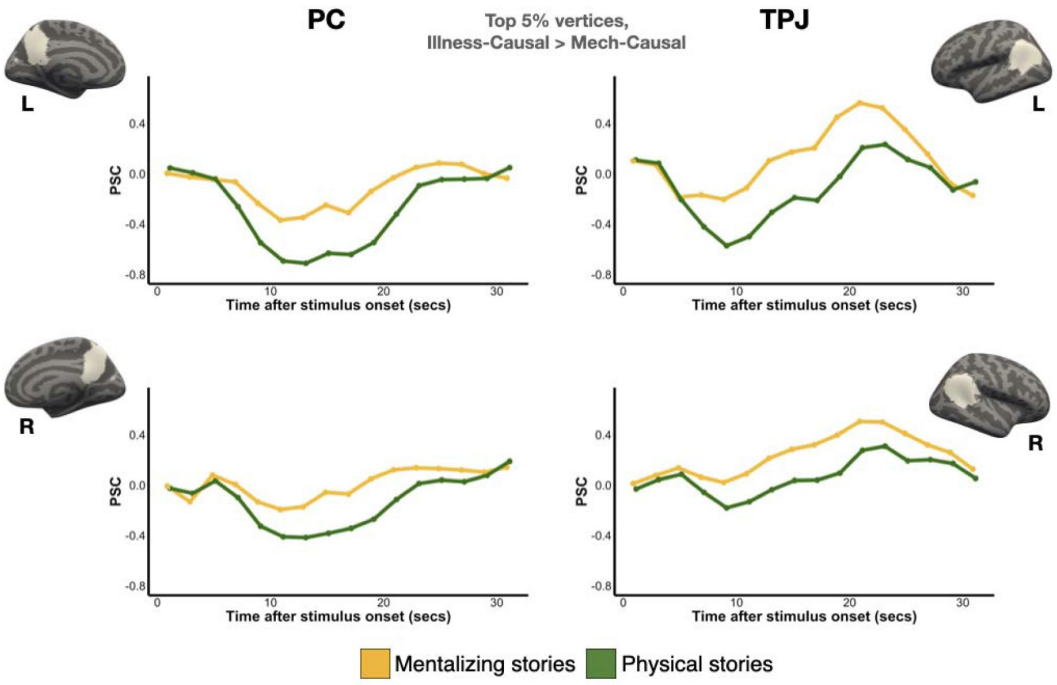
Panel A: Percent signal change (PSC) for each condition among the top 5% most language-responsive vertices (*language* > *math*) in a temporal language network mask (Fedorenko et al., 2010). Panel B: The same results in a frontal language mask (Fedorenko et al., 2010).



Supplementary Figure 7.

Responses to illness inferences in bilateral PC and TPJ.

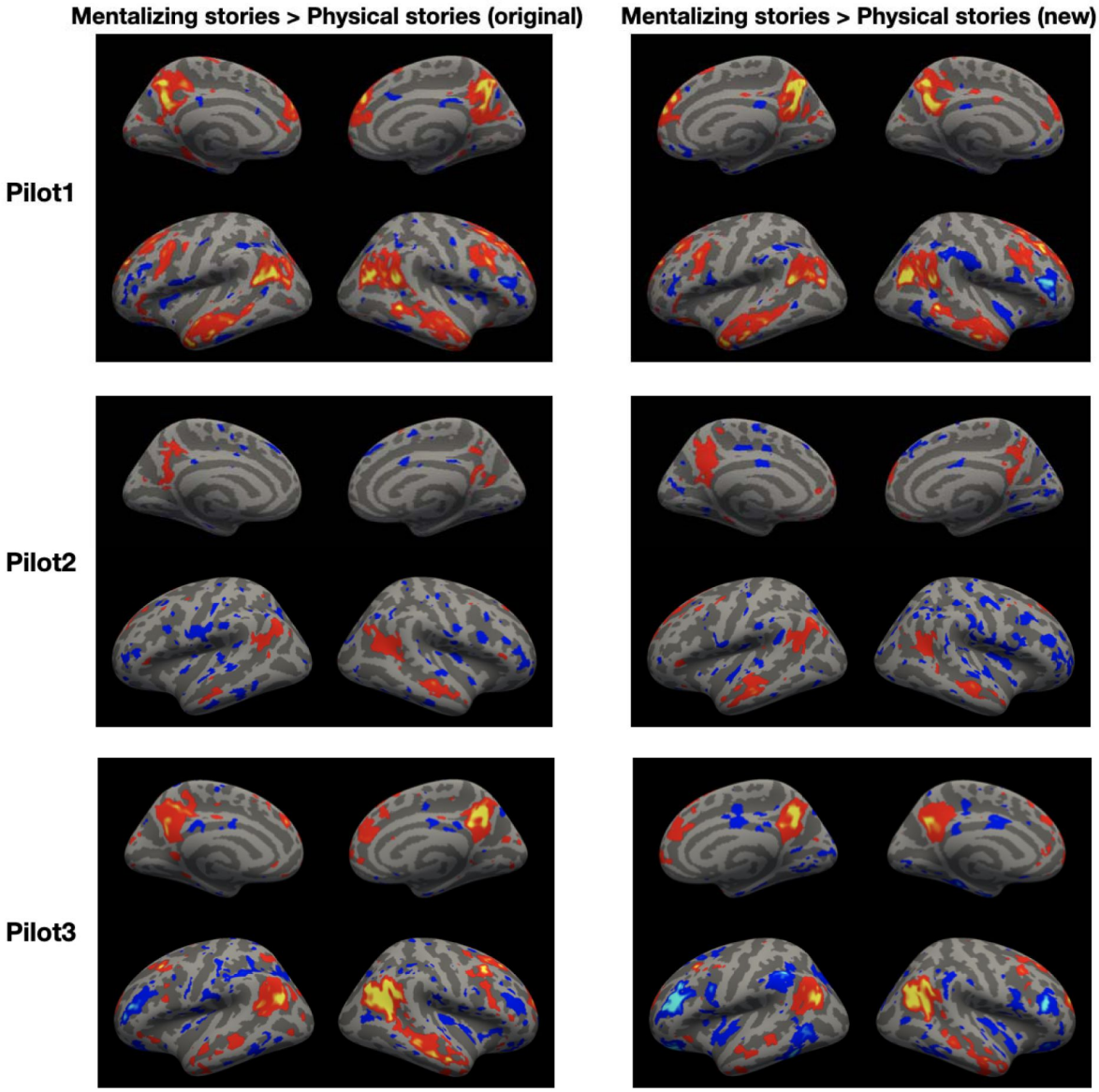
Top 4 plots: percent signal change (PSC) for each condition among the top 5% *Illness-Causal* > *Mechanical-Causal* vertices in bilateral PC and TPJ masks (Dufour et al., 2013) in individual participants, established via a leave-one-run-out analysis. Bottom 4 plots: PSC for each condition among the top 5% *mentalizing stories* > *physical stories* vertices in the same masks. We hypothesized that the PC would exhibit a preference for illness inferences and report all other responses for completeness (see preregistration *https://osf.io/cx9n2/*). Significance codes for *Illness-Causal* > *Mechanical-Causal* comparison: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1.



Supplementary Figure 8.

Responses to mentalizing in illness-responsive vertices in bilateral PC and TPJ.

Percent signal change (PSC) for mentalizing stories and physical stories (mentalizing localizer) was extracted from the top 5% *Illness-Causal* > *Mechanical-Causal* vertices in bilateral PC and TPJ masks (Dufour et al., 2013) in individual participants. The difference between mentalizing stories and physical stories was significant (all ps < .01) across all analyses.



Supplementary Figure 9.

Comparison of mentalizing localizers used in previous work and in the current study, in 3 pilot participants.

The mentalizing localizer in the current study used the same mentalizing stories as in previous work (Dodell-Feder et al., 2010) but contained new physical stories that included more vivid physical description and did not refer to animate agents. Group maps are shown at $p < .01$ uncorrected.

Illness type	Number of trials
acne	2
allergies	1
anemia	1
asthma	1
blood cancer	1
brain cancer	1
breast cancer	1
cancer (unspecified)	1
chickenpox	1
cold	2
COVID	2
epilepsy	1
flu	3
food poisoning	1
GI inflammation	2
GI virus	1
heart disease	3
high blood pressure	1
HIV-AIDS	1
liver disease	2
lung cancer	1
lung disease	2
malaria	1
pneumonia	1
skin cancer	2
throat cancer	1
Type 2 diabetes	1

Supplementary Table 1

Illness types present in the stimulus set.

Search space	Contrast	Accuracy	t	Permuted p	Bonferroni adj. p
Left TPJ	ment_vs_phys	61.70%	2.72	0.0073	0.068
Right TPJ	ment_vs_phys	52.90%	0.94	0.194	1
Left PC	ment_vs_phys	61.30%	2.93	0.0062	0.043
Right PC	ment_vs_phys	61.30%	3.09	0.0042	0.03
Left TPJ	caus_vs_rest	55.80%	1.61	0.0695	0.624
Right TPJ	caus_vs_rest	55%	1.43	0.0771	0.844
Left PC	caus_vs_rest	63.70%	3.5	0.0009	0.012
Right PC	caus_vs_rest	52.90%	1	0.1663	1
Logic	logic_vs_lang	67.10%	8.73	0	0
Language	lang_vs_math	61.20%	3.27	0.0038	0.02

Supplementary Table 2

MVPA results in individual-subject functional ROIs. Each ROI was created by selecting the top 300 vertices for each contrast in each search space.

Accuracy refers to classifier performance against chance (50%) for *Illness-Causal* vs. *Mechanical-Causal*. Permuted and Bonferroni-corrected (across ROIs) p-values are reported. Ment_vs_phys: *mentalizing stories* > *physical stories* (mentalizing localizer). Caus_vs_rest: *Illness-Causal* + *Illness-Mechanical* > Rest. Logic_vs_lang: *logic* > *language* (language/logic localizer). Lang_vs_math: *language* > *math* (language/logic localizer).

Appendix 1: Online experiment protocol

Prior to the fMRI experiment, we collected explicit causality judgments from a separate group of online participants (n=30). Each online participant read all vignettes from the causal inference experiment (152 vignettes) in addition to 12 filler vignettes that were designed to be either maximally causally related or unrelated (164 vignettes total), one vignette at a time. Their task was to judge the extent to which it was possible that the event described in the first sentence of each vignette caused the event described in the second sentence on a 4-point scale (1 = not possible; 4 = very possible). 4 participants were excluded on the basis of inaccurate responses on the filler trials (i.e., difference between average ratings for maximally causally related and maximally causally unrelated vignettes <2). Among the 26 remaining participants, 12 read vignettes from Version A and 14 read vignettes from Version B of the experiment. To eliminate erroneous responses, we first excluded trials with RTs 2.5 SD outside their respective condition means within participants, and then excluded trials with outlier RTs (more than 1.5 IQR below Q1 or more than 1.5 IQR above Q3) across participants (approximately 5% of all trials excluded in total). We found that both causal conditions (*Illness-Causal*, *Mechanical-Causal*) were more causally connected than both non-causal conditions, $t(25) = 36.97$, $p < .0001$ (causal: $M = 3.51 \pm 0.78$ SD, non-causal: $M = 1.10 \pm 0.45$ SD). In addition, *Illness-Causal* and *Mechanical-Causal* items received equally high causality ratings, $t(25) = –0.64$, $p = 0.53$ (*Illness-Causal*: $M = 3.49 \pm 0.77$ SD, *Mechanical-Causal*: $M = 3.53 \pm 0.79$ SD).

Appendix 2: Details on measuring linguistic variables

All conditions were matched (pairwise t-tests, all *p*s > 0.3, no statistical correction) on multiple linguistic variables known to modulate neural activity in language regions (e.g., Pallier, Devauchelle, & Dehaene, 2011; Shain, Blank et al., 2020). These included number of characters, number of words, average number of characters per word, average word frequency, average bigram surprisal (Google Books Ngram Viewer, *https://books.google.com/ngrams/*), and average syntactic dependency length (Stanford Parser; de Marneffe, MacCartney, & Manning, 2006). Sentences that were incorrectly parsed by the automatic syntactic parser (i.e., past participle adjectives parsed as verbs) were corrected by hand. Word frequency was calculated as the negative log of a word's occurrence rate in the Google corpus between the years 2017-2019. Bigram surprisal was calculated as the negative log of the frequency of a given two-word phrase in the Google corpus divided by the frequency of the first word of the phrase.

This calculation uses a log base of 2 in order to express surprisal in terms of "bits" that the first word provides in the context of the phrase. We used *bigram* surprisal as our surprisal measure to maximize the number of n-grams that had an entry in the corpus. Even so, 64 out of the 1515 total bigrams (4%) did not have an entry in the corpus and were therefore assigned the highest surprisal value among the rest of the bigrams (see Willems et al., 2016).

Appendix 3: Behavioral results from the localizer tasks

Accuracy on the language/logic localizer task was significantly lower for the logic task compared to both the language and math tasks (logic: $M = 67.5\% \pm 14.0$ SD, math: $M = 93.8\% \pm 6.4$ SD, language: $M = 98.1\% \pm 5.8$ SD; $F(2,38) = 60.38$, $p < .0001$). Similarly, response time was slowest on the logic

task, followed by math and then language (logic: M = 8.78 s ± 1.88 SD, math: M = 6.20 s ± 1.37 SD, language: M = 5.18 s ± 1.53 SD; $F(2,38) = 44.28$, $p < .0001$).

Accuracy on the mentalizing localizer task was not different across the mentalizing stories and physical stories conditions (mentalizing: 83.50% ± 15.7 SD, physical: 90.50% ± 12.3 SD; $F(1,19) = 2.73$, $p = .12$). However, response time for the mentalizing stories was significantly slower (mentalizing: 3.46 s ± 0.55 SD, physical: 3.11 s ± 0.56 SD; $F(1,19) = 16.59$, $p < .001$).

References

1. Ackerknecht E. H (1982) **A short history of medicine** Johns Hopkins University Press

2. Aglinskas A., Fairhall S. L (2023) **Similar representation of names and faces in the network for person perception** *NeuroImage* **274** https://doi.org/10.1016/j.neuroimage.2023.120100

3. Atran S (1998) **Folk biology and the anthropology of science: Cognitive universals and cultural particulars** *Behavioral and Brain Sciences* **21**:547–569 https://doi.org/10.1017/S0140525X98001277

4. Baldassano C., Beck D. M., Fei-Fei L (2013) **Differential Connectivity Within the Parahippocampal Place Area** *NeuroImage* **75**:228–237 https://doi.org/10.1016/j.neuroimage.2013.02.073

5. Barbey A., Patterson R (2011) **Architecture of Explanatory Inference in the Human Prefrontal Cortex** *Frontiers in Psychology* **2** https://doi.org/10.3389/fpsyg.2011.00162

6. Barrett H. C., Cosmides L., Tooby J, Gangestad S. W., Simpson J. A. (2007) **The Hominid Entry into the Cognitive Niche** *The evolution of mind: Fundamental questions and controversies* The Guilford Press :241–248

7. Bender A., Beller S., Medin D. L, Waldmann M. (2017) **Causal cognition and culture** *The Oxford Handbook of Causal Reasoning* Oxford University Press

8. Black J. B., Bern H (1981) **Causal coherence and memory for events in narratives** *Journal of Verbal Learning and Verbal Behavior* **20**:267–275 https://doi.org/10.1016/S0022-5371(81)90417-5

9. Blank I. A., Fedorenko E (2020) **No evidence for differences among language regions in their temporal receptive windows** *NeuroImage* **219** https://doi.org/10.1016/j.neuroimage.2020.116925

10. Boyer P, Sperber D., Premack D., Premack A. J. (1995) **Causal understandings in cultural representations** *Causal cognition: A multidisciplinary debate* Oxford University Press

11. Callanan M. A., Oakes L. M (1992) **Preschoolers' questions and parents' explanations: Causal thinking in everyday activity** *Cognitive Development* **7**:213–233 https://doi.org/10.1016/0885-2014(92)90012-G

12. Caramazza A., Shelton J. R (1998) **Domain-Specific Knowledge Systems in the Brain: The Animate-Inanimate Distinction** *Journal of Cognitive Neuroscience* **10**:1–34 https://doi.org/10.1162/089892998563752

13. Carey S (1985) **Conceptual change in childhood** Bradford Books

14. Carey S. (2011) **The origin of concepts** Oxford University Press

15. Cheng P., Novick L (1992) **Covariation in natural causal induction** *Psychological Review* **99**:365–382 https://doi.org/10.1037//0033-295X.99.2.365

16. Chomik-Morales J., Kanwisher N., Schulz L., Pramod R. T (2024) **Using fMRI to explore causal reasoning and abstract relational reasoning in the left lateral prefrontal cortex in the human brain** *Proceedings of the 46th Annual Meeting of the Cognitive Science Society*

17. Chow H. M., Kaup B., Raabe M., Greenlee M. W (2008) **Evidence of fronto-temporal interactions for strategic inference processes during language comprehension** *NeuroImage* **40**:940–954 https://doi.org/10.1016/j.neuroimage.2007.11.044

18. Coley J. D., Solomon G. E. A., Shafto P, Kahn P., Kellert S. (2002) **The development of folkbiology: A cognitive science perspective on children's understanding of the biological world** *Children and nature: Psychological, sociocultural and evolutionary investigations* Cambridge, MA: MIT Press :65–91

19. Cosmides L., Tooby J (2013) **Unraveling the enigma of human intelligence: Evolutionary psychology and the multimodular mind** *The evolution of intelligence* Psychology Press :145–198

20. Dale A. M., Fischl B., Sereno M. I (1999) **Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction** *NeuroImage* **9**:179–194 https://doi.org/10.1006/nimg.1998.0395

21. Davis Z. J., Rehder B (2020) **A Process Model of Causal Reasoning** *Cognitive Science* **44** https://doi.org/10.1111/cogs.12839

22. Deen B., Freiwald W. A (2022) **Parallel systems for social and spatial reasoning within the cortical apex** *bioRxiv* https://doi.org/10.1101/2021.09.23.461550

23. DiNicola L. M., Braga R. M., Buckner R. L (2020) **Parallel distributed networks dissociate episodic and social functions within the individual** *Journal of Neurophysiology* **123**:1144–1179 https://doi.org/10.1152/jn.00529.2019

24. Dodell-Feder D., Koster-Hale J., Bedny M., Saxe R (2011) **fMRI item analysis in a theory of mind task** *NeuroImage* **55**:705–712 https://doi.org/10.1016/j.neuroimage.2010.12.040

25. Duffy S. A., Shinjo M., Myers J. L (1990) **The effect of encoding task on memory for sentence pairs varying in causal relatedness** *Journal of Memory and Language* **29**:27–42 https://doi.org/10.1016/0749-596X(90)90008-N

26. Dufour N., Redcay E., Young L., Mavros P. L., Moran J. M., Triantafyllou C., Gabrieli J. D. E., Saxe R (2013) **Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism** *PLoS ONE* **8** https://doi.org/10.1371/journal.pone.0075468

27. Eklund A., Knutsson H., Nichols T. E (2019) **Cluster failure revisited: Impact of first level design and physiological noise on cluster false positive rates** *Human Brain Mapping* **40**:2017–2032 https://doi.org/10.1002/hbm.24350

28. Eklund A., Nichols T. E., Knutsson H (2016) **Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates** *Proceedings of the National Academy of Sciences* **113**:7900–7905 https://doi.org/10.1073/pnas.1602413113

29. Epstein R., Kanwisher N (1998) **A cortical representation of the local visual environment** *Nature* **392**:598–601 https://doi.org/10.1038/33402

30. Fairhall S. L., Caramazza A (2013) **Brain Regions That Represent Amodal Conceptual Knowledge** *Journal of Neuroscience* **33**:10552–10558 https://doi.org/10.1523/JNEUROSCI.0051-13.2013

31. Fairhall S. L., Caramazza A (2013) **Category-selective neural substrates for person– and place-related concepts** *Cortex* **49**:2748–2757 https://doi.org/10.1016/j.cortex.2013.05.010

32. Fairhall S. L., Anzellotti S., Ubaldi S., Caramazza A (2014) **Person– and Place-Selective Neural Substrates for Entity-Specific Semantic Access** *Cerebral Cortex* **24**:1687–1696 https://doi.org/10.1093/cercor/bht039

33. Farah M. J., Rabinowitz C (2003) **Genetic and Environmental Influences on the Organisation of Semantic Memory in the Brain:is "Living Things" an Innate Category?** *Cognitive Neuropsychology* **20**:401–408 https://doi.org/10.1080/02643290244000293

34. Fedorenko E., Varley R (2016) **Language and thought are not the same thing: Evidence from neuroimaging and neurological patients** *Annals of the New York Academy of Sciences* **1369**:132–153 https://doi.org/10.1111/nyas.13046

35. Fedorenko E., Hsieh P.-J., Nieto-Castañón A., Whitfield-Gabrieli S., Kanwisher N (2010) **New Method for fMRI Investigations of Language: Defining ROIs Functionally in Individual Subjects** *Journal of Neurophysiology* **104**:1177–1194 https://doi.org/10.1152/jn.00032.2010

36. Feng W., Wang W., Liu J., Wang Z., Tian L., Fan L (2021) **Neural Correlates of Causal Inferences in Discourse Understanding and Logical Problem-Solving: A Meta-Analysis Study** *Frontiers in Human Neuroscience* **15** https://doi.org/10.3389/fnhum.2021.666179

37. Fenker D. B., Schoenfeld M. A., Waldmann M. R., Schuetze H., Heinze H.-J., Duezel E (2010) **"Virus and Epidemic": Causal Knowledge Activates Prediction Error Circuitry** *Journal of Cognitive Neuroscience* **22**:2151–2163 https://doi.org/10.1162/jocn.2009.21387

38. Ferstl E. C., von Cramon D. Y. (2001) **The role of coherence and cohesion in text comprehension: An event-related fMRI study** *Cognitive Brain Research* **11**:325–340 https://doi.org/10.1016/S0926-6410(01)00007-6

39. Foster G. M (1976) **Disease Etiologies in Non-Western Medical Systems** *American Anthropologist* **78**:773–782 https://doi.org/10.1525/aa.1976.78.4.02a00030

40. Fugelsang J. A., Dunbar K. N (2005) **Brain-based mechanisms underlying complex causal thinking** *Neuropsychologia* **43**:1204–1213 https://doi.org/10.1016/j.neuropsychologia.2004.10.012

41. Graesser A. C., Singer M., Trabasso T (1994) **Constructing inferences during narrative text comprehension** *Psychological Review* **101**:371–395 https://doi.org/10.1037/0033-295X.101.3.371

42. Gerstenberg T., Tenenbaum J. B, Waldmann M. (2017) **Intuitive theories** *The Oxford Handbook of Causal Reasoning* Oxford University Press

43. Glasser M. F., Sotiropoulos S. N., Wilson J. A., Coalson T. S., Fischl B., Andersson J. L., ..., Wu-Minn HCP Consortium (2013) **The minimal preprocessing pipelines for the Human Connectome Project** *Neuroimage* **80**:105–124

44. Goddu M. K., Gopnik A (2024) **The development of human causal learning and reasoning** *Nature Reviews Psychology* :1–21

45. Goldvarg E., Johnson-Laird P. (2001) **Naive causality: A mental model theory of causal meaning and reasoning** *Cognitive Science* **25**:565–610 https://doi.org/10.1207/s15516709cog2504_3

46. Gopnik A., Meltzoff A. N. (1997) **Words, thoughts, and theories** The MIT Press

47. Gopnik A., Glymour C., Sobel D. M., Schulz L. E., Kushnir T., Danks D. (2004) **A Theory of Causal Learning in Children: Causal Maps and Bayes Nets** *Psychological Review* **111**:3–32 https://doi.org/10.1037/0033-295X.111.1.3

48. Gopnik A., Sobel D. M., Schulz L. E., Glymour C (2001) **Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation** *Developmental Psychology* **37**:620–629 https://doi.org/10.1037/0012-1649.37.5.620

49. Gutheil G., Vera A., Keil F. C (1998) **Do houseflies think? Patterns of induction and biological beliefs in development1** *Cognition* **66**:33–49 https://doi.org/10.1016/S0010-0277(97)00049-8

50. Hasson U., Yang E., Vallines I., Heeger D. J., Rubin N (2008) **A Hierarchy of Temporal Receptive Windows in Human Cortex** *The Journal of Neuroscience* **28**:2539–2550 https://doi.org/10.1523/JNEUROSCI.5487-07.2008

51. Hatano G., Inagaki K (1994) **Young children's naive theory of biology** *Cognition* **50**:171–188 https://doi.org/10.1016/0010-0277(94)90027-2

52. Hauptman M., Elli G., Pant R., Bedny M (2023) **Neural specialization for 'visual' concepts emerges in the absence of vision** *bioRxiv*

53. Häusler C. O., Eickhoff S. B., Hanke M (2022) **Processing of visual and non-visual naturalistic spatial information in the** *Scientific Data* **9** https://doi.org/10.1038/s41597-022-01250-4

54. Hickling A. K., Wellman H. M (2001) **The emergence of children's causal explanations and theories: Evidence from everyday conversation** *Developmental Psychology* **37**:668–683 https://doi.org/10.1037/0012-1649.37.5.668

55. Hillis A., Caramazza A (1991) **Category-specific naming and comprehension impairment: A double dissociation** *Brain: A Journal of Neurology* **114**:2081–2094 https://doi.org/10.1093/brain/114.5.2081

56. Inagaki K., Hatano G (1993) **Young Children's Understanding of the Mind-Body Distinction** *Child Development* **64**:1534–1549 https://doi.org/10.2307/1131551

57. Inagaki K., Hatano G (2004) **Vitalistic causality in young children's naive biology** *Trends in Cognitive Sciences* **8**:356–362 https://doi.org/10.1016/j.tics.2004.06.004

58. Inagaki K., Hatano G (2006) **Young Children's Conception of the Biological World** *Current Directions in Psychological Science* **15**:177–181 https://doi.org/10.1111/j.1467-8721.2006.00431.x

59. Jacoby N., Fedorenko E (2020) **Discourse-level comprehension engages medial frontal Theory of Mind brain regions even for expository texts** *Language, Cognition and Neuroscience* **35**:780–796 https://doi.org/10.1080/23273798.2018.1525494

60. Kalish C (1997) **Preschoolers' understanding of mental and bodily reactions to contamination: What you don't know can hurt you, but cannot sadden you** *Developmental psychology* **33**

61. Kalish C. W (1996) **Preschoolers' understanding of germs as invisible mechanisms** *Cognitive Development* **11**:83–106 https://doi.org/10.1016/S0885-2014(96)90029-5

62. Kanjlia S., Lane C., Feigenson L., Bedny M (2016) **Absence of visual experience modifies the neural basis of numerical thinking** *Proceedings of the National Academy of Sciences* **113**:11172–11177 https://doi.org/10.1073/pnas.1524982113

63. Keenan J. M., Baillet S. D., Brown P (1984) **The effects of causal cohesion on comprehension and memory** *Journal of Verbal Learning and Verbal Behavior* **23**:115–126 https://doi.org/10.1016/S0022-5371(84)90082-3

64. Keil F. C, Gunnar M. R., Maratsos M. (1992) **The origins of an autonomous biology** *Modularity and constraints in language and cognition* Psychology Press

65. Keil F. C, Hirschfeld L. A., Gelman S. A. (1994) **The birth and nurturance of concepts by domains: The origins of concepts of living things** *Mapping the mind: Domain specificity in cognition and culture* Cambridge University Press :234–254

66. Keil F. C., Levin D. T., Richman B. A., Gutheil G, Medin D. L., Atran S. (1999) **Mechanism and explanation in the development of biological thought: The case of disease** *Folkbiology* MIT Press

67. Keil F. C., Levin D. T., Richman B. A., Gutheil G., Medin D. L., Atran S. (1999) **Mechanism and Explanation in the Development of Biological Thought: The Case of Disease** *Folkbiology* The MIT Press :285–320 https://doi.org/10.7551/mitpress/3042.003.0010

68. Khemlani S. S., Barbey A. K., Johnson-Laird P. N (2014) **Causal reasoning with mental models** *Frontiers in Human Neuroscience* **8** https://doi.org/10.3389/fnhum.2014.00849

69. Kranjec A., Cardillo E. R., Schmidt G. L., Lehet M., Chatterjee A (2012) **Deconstructing Events: The Neural Bases for Space, Time, and Causality** *Journal of Cognitive Neuroscience* **24**:1–16 https://doi.org/10.1162/jocn_a_00124

70. Lagnado D. A., Waldmann M. R., Hagmaye Y., Sloman S. A, Gopnik A., Schulz L. (2007) **Beyond Covariation: Cues to Causal Structure** *Causal Learning* Oxford University Press New York :154–172 https://doi.org/10.1093/acprof:oso/9780195176803.003.0011

71. Lee H., Chen J (2022) **Predicting memory from the network structure of naturalistic events** *Nature Communications* **13** https://doi.org/10.1038/s41467-022-31965-2

72. Legare C. H., Gelman S. A (2008) **Bewitchment, Biology, or Both: The Co-Existence of Natural and Supernatural Explanatory Frameworks Across Development** *Cognitive Science* **32**:607–642 https://doi.org/10.1080/03640210802066766

73. Legare C. H., Evans E. M., Rosengren K. S., Harris P. L (2012) **The Coexistence of Natural and Supernatural Explanations Across Cultures and Development** *Child Development* **83**:779–793 https://doi.org/10.1111/j.1467-8624.2012.01743.x

74. Legare C. H., Wellman H. M., Gelman S. A (2009) **Evidence for an explanation advantage in naïve biological reasoning** *Cognitive Psychology* **58**:177–194 https://doi.org/10.1016/j.cogpsych.2008.06.002

75. Legare C., Shtulman A (2017) **Explanatory Pluralism Across Cultures and Development** *Metacognitive Diversity: An Interdisciplinary Approach* https://doi.org/10.1093/oso/9780198789710.003.0019

76. Lerner Y., Honey C. J., Silbert L. J., Hasson U (2011) **Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story** *Journal of Neuroscience* **31**:2906–2915 https://doi.org/10.1523/JNEUROSCI.3684-10.2011

77. Lightner A. D., Heckelsmiller C., Hagen E. H (2021) **Ethnoscientific expertise and knowledge specialisation in 55 traditional cultures** *Evolutionary Human Sciences* **3** https://doi.org/10.1017/ehs.2021.31

78. Liu Y.-F., Kim J., Wilson C., Bedny M (2020) **Computer code comprehension shares neural resources with formal logical inference in the fronto-parietal network** *eLife* **9** https://doi.org/10.7554/eLife.59340

79. Lock M. M., Nguyen V. K (2010) **An anthropology of biomedicine** John Wiley & Sons

80. Lynch E., Medin D (2006) **Explanatory models of illness: A study of within-culture variation** *Cognitive Psychology* **53**:285–309 https://doi.org/10.1016/j.cogpsych.2006.02.001

81. Marneffe M.-C., MacCartney B., Manning C (2006) **Generating Typed Dependency Parses from Phrase Structure Parses** *Proc of LREC* **6**

82. Mason R. A., Just M. A (2011) **Differentiable cortical networks for inferences concerning people's intentions versus physical causality** *Human Brain Mapping* **32**:313–329 https://doi.org/10.1002/hbm.21021

83. Meder B., Mayrhofer R, Waldmann M. (2017) **Diagnostic reasoning** *The Oxford Handbook of Causal Reasoning* Oxford University Press

84. Medin D. L., Atran S (2004) **The Native Mind: Biological Categorization and Reasoning in Development and Across Cultures** *Psychological Review* **111**:960–983 https://doi.org/10.1037/0033-295X.111.4.960

85. Medin D., Waxman S., Woodring J., Washinawatok K (2010) **Human-centeredness is not a universal feature of young children's reasoning: Culture and experience matter when reasoning about biological entities** *Cognitive Development* **25**:197–207 https://doi.org/10.1016/j.cogdev.2010.02.001

86. Monti M. M., Parsons L. M., Osherson D. N (2009) **The boundaries of language and thought in deductive inference** *Proceedings of the National Academy of Sciences* **106**:12554–12559 https://doi.org/10.1073/pnas.0902422106

87. Muentener P., Schulz L (2014) **Toddlers infer unobserved causes for spontaneous events** *Frontiers in Psychology* **5** https://doi.org/10.3389/fpsyg.2014.01496

88. Myers J. L., Shinjo M., Duffy S. A (1987) **Degree of causal relatedness and memory** *Journal of Memory and Language* **26**:453–465 https://doi.org/10.1016/0749-596X(87)90101-X

89. Norman G. R., Grierson L. E. M., Sherbino J., Hamstra S. J., Schmidt H. G., Mamede S, Ericsson K. A., Hoffman R. R., Kozbelt A., Williams A. M. (2009) **Expertise in medicine and surgery** *The Cambridge handbook of expertise and expert performance* Cambridge University Press

90. Notaro P. C., Gelman S. A., Zimmerman M. A (2001) **Children's Understanding of Psychogenic Bodily Reactions** *Child Development* **72**:444–459 https://doi.org/10.1111/1467-8624.00289

91. Operskalski J. T., Barbey A. K, Waldmann M. (2017) **Cognitive neuroscience of causal reasoning** *The Oxford Handbook of Causal Reasoning* Oxford University Press

92. Opfer J. E., Gelman S. A., Goswami U. (2011) **Development of the animate-inanimate distinction** *The Wiley-Blackwell handbook of childhood cognitive development* Wiley-Blackwell

93. Pallier C., Devauchelle A.-D., Dehaene S (2011) **Cortical representation of the constituent structure of sentences** *Proceedings of the National Academy of Sciences* **108**:2522–2527 https://doi.org/10.1073/pnas.1018711108

94. Pearl J (2000) **Causality: models, reasoning, and inference** Cambridge University Press

95. Peer M., Salomon R., Goldberg I., Blanke O., Arzy S (2015) **Brain system for mental orientation in space, time, and person** *Proceedings of the National Academy of Sciences* **112**:11072–11077 https://doi.org/10.1073/pnas.1504242112

96. Peirce J., Gray J. R., Simpson S., MacAskill M., Höchenberger R., Sogo H., Kastman E., Lindeløv J. K (2019) **PsychoPy2: Experiments in behavior made easy** *Behavior Research Methods* **51**:195–203 https://doi.org/10.3758/s13428-018-01193-y

97. Pinker S, Christiansen M. H., Kirby S. (2003) **Language as an adaptation to the cognitive niche** *Language Evolution* Oxford University Press :16–37

98. Prat C. S., Mason R. A., Just M. A (2011) **Individual differences in the neural basis of causal inferencing** *Brain and Language* **116**:1–13 https://doi.org/10.1016/j.bandl.2010.08.004

99. Rabini G., Ubaldi S., Fairhall S (2021) **Combining Concepts Across Categorical Domains: A Linking Role of the Precuneus** *Neurobiology of Language* **2**:354–371 https://doi.org/10.1162/nol_a_00039

100. Raman L., Gelman S. A (2005) **Children's Understanding of the Transmission of Genetic Disorders and Contagious Illnesses** *Developmental Psychology* **41**:171–182 https://doi.org/10.1037/0012-1649.41.1.171

101. Raman L., Winer G. A (2004) **Evidence of more immanent justice responding in adults than children: A challenge to traditional developmental theories** *The British Journal of Developmental Psychology* **22**:255–274

102. Rehder B., Burnett R. C (2005) **Feature inference and the causal structure of categories** *Cognitive Psychology* **50**:264–314 https://doi.org/10.1016/j.cogpsych.2004.09.002

103. Ritchey M., Cooper R. A (2020) **Deconstructing the Posterior Medial Episodic Network** *Trends in Cognitive Sciences* **24**:451–465 https://doi.org/10.1016/j.tics.2020.03.006

104. Rosengren K. S., Gelman S. A., Kalish C. W., McCormick M (1991) **As Time Goes By: Children's Early Understanding of Growth in Animals** *Child Development* **62**:1302–1320 https://doi.org/10.1111/j.1467-8624.1991.tb01607.x

105. Rottman B. M., Hastie R (2014) **Reasoning about Causal Relationships: Inferences on Causal Networks** *Psychological Bulletin* **140**:109–139 https://doi.org/10.1037/a0031903

106. Rottman B. M., Ahn W. K., Luhmann C. C, Illari P., Russo F., Williamson J. (2011) **When and how do people reason about unobserved causes** *Causality in the Sciences* Oxford University Press

107. Satpute A. B., Fenker D. B., Waldmann M. R., Tabibnia G., Holyoak K. J., Lieberman M. D (2005) **An fMRI study of causal judgments** *European Journal of Neuroscience* **22**:1233–1238 https://doi.org/10.1111/j.1460-9568.2005.04292.x

108. Saxe R., Carey S (2006) **The perception of causality in infancy** *Acta Psychologica* **123**:144–165 https://doi.org/10.1016/j.actpsy.2006.05.005

109. Saxe R., Kanwisher N (2003) **People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind."** *NeuroImage* **19**:1835–1842 https://doi.org/10.1016/S1053-8119(03)00230-1

110. Saxe R., Moran J. M., Scholz J., Gabrieli J (2006) **Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects** *Social Cognitive and Affective Neuroscience* **1**:229–234 https://doi.org/10.1093/scan/nsl034

111. Schmidt H., Norman G., Boshuizen H (1990) **A Cognitive Perspective on Medical Expertise: Theory and Implications** *Academic Medicine* **65**:611–621

112. Schreiber K., Krekelberg B (2013) **The Statistical Analysis of Multi-Voxel Patterns in Functional Imaging** *PLoS ONE* **8** https://doi.org/10.1371/journal.pone.0069328

113. Schulz L. E., Gopnik A (2004) **Causal learning across domains** *Developmental Psychology* **40**:162–176 https://doi.org/10.1037/0012-1649.40.2.162

114. Shain C., Blank I. A., van Schijndel M., Schuler W., Fedorenko E. (2020) **fMRI reveals language-specific predictive coding during naturalistic sentence comprehension** *Neuropsychologia* **138** https://doi.org/10.1016/j.neuropsychologia.2019.107307

115. Shain C., Paunov A., Chen X., Lipkin B., Fedorenko E (2023) **No evidence of theory of mind reasoning in the human language network** *Cerebral Cortex* **33**:6299–6319 https://doi.org/10.1093/cercor/bhac505

116. Silson E. H., Steel A., Kidder A., Gilmore A. W., Baker C. I (2019) **Distinct subdivisions of human medial parietal cortex support recollection of people and places** *eLife* **8** https://doi.org/10.7554/eLife.47391

117. Simons D. J., Keil F. C (1995) **An abstract to concrete shift in the development of biological thought: The insides story** *Cognition* **56**:129–163 https://doi.org/10.1016/0010-0277(94)00660-D

118. Sloman S. A., Lagnado D (2015) **Causality in Thought** *Annual Review of Psychology* **66**:223–247 https://doi.org/10.1146/annurev-psych-010814-015135

119. Smith S. M. *et al.* (2004) **Advances in functional and structural MR image analysis and implementation as FSL** *NeuroImage* **23**:S208–S219 https://doi.org/10.1016/j.neuroimage.2004.07.051

120. Solstad T., Bott O, Waldmann M. (2017) **Causality and causal reasoning in natural language** *The Oxford Handbook of Causal Reasoning* Oxford University Press

121. Spelke E. S, Gentner D., Goldin-Meadow S. (2003) **What makes us smart? Core knowledge and natural language** *Language in mind* Cambridge, MA: MIT Press

122. Spelke E. S (2022) **What Babies Know: Core Knowledge and Composition Volume 1** New York, NY: Oxford University Press

123. Springer K., Keil F. C (1991) **Early Differentiation of Causal Mechanisms Appropriate to Biological and Nonbiological Kinds** *Child Development* **62**:767–781 https://doi.org/10.2307/1131176

124. Springer K., Ruckel J (1992) **Early beliefs about the cause of illness: Evidence against immanent justice** *Cognitive Development* **7**:429–443 https://doi.org/10.1016/0885-2014(92)80002-W

125. Stelzer J., Chen Y., Turner R (2013) **Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control** *NeuroImage* **65**:69–82 https://doi.org/10.1016/j.neuroimage.2012.09.063

126. Steyvers M., Tenenbaum J. B., Wagenmakers E.-J., Blum B (2003) **Inferring causal networks from observations and interventions** *Cognitive Science* **27**:453–489 https://doi.org/10.1207/s15516709cog2703_6

127. Tenenbaum J. B., Griffiths T. L., Niyogi S, Gopnik A., Schulz L. (2007) **Intuitive Theories as Grammars for Causal Inference** *Causal Learning* Oxford University Press New York :301–322 https://doi.org/10.1093/acprof:oso/9780195176803.003.0020

128. Tooby J., DeVore I, Kinzey W. G. (1987) **The reconstruction of hominid behavioral evolution through strategic modeling** *The evolution of human behavior: Primate models* Albany, NY: SUNY Press

129. Trabasso T., Sperry L. L (1985) **Causal relatedness and importance of story events** *Journal of Memory and Language* **24**:595–611 https://doi.org/10.1016/0749-596X(85)90048-8

130. Varley R (2014) **Reason without much language** *Language Sciences* **46**:232–244 https://doi.org/10.1016/j.langsci.2014.06.012

131. Waldmann M (2000) **Competition among Causes but Not Effects in Predictive and Diagnostic Learning** *Journal of Experimental Psychology. Learning, Memory, and Cognition* **26**:53–76 https://doi.org/10.1037//0278-7393.26.1.53

132. Wang X., Peelen M. V., Han Z., Caramazza A., Bi Y (2016) **The role of vision in the neural representation of unique entities** *Neuropsychologia* **87**:144–156 https://doi.org/10.1016/j.neuropsychologia.2016.05.007

133. Waldmann M. R., Holyoak K. J (1992) **Predictive and diagnostic learning within causal models: asymmetries in cue competition** *Journal of Experimental Psychology: General* **121**

134. Warrington E. K., Shallice T (1984) **Category specific semantic impairments** *Brain* **107**:829–853

135. Weiner K. S. *et al.* (2018) **Defining the most probable location of the parahippocampal place area using cortex-based alignment and cross-validation** *NeuroImage* **170**:373–384 https://doi.org/10.1016/j.neuroimage.2017.04.040

136. Wellman H. M., Gelman S. A (1992) **Cognitive development: Foundational theories of core domains** *Annual Review of Psychology* **43**:337–375 https://doi.org/10.1146/annurev.ps.43.020192.002005

137. Willems R. M., Frank S. L., Nijhof A. D., Hagoort P., van den Bosch A. (2016) **Prediction During Natural Language Comprehension** *Cerebral Cortex* **26**:2506–2516 https://doi.org/10.1093/cercor/bhv075

138. Winkler A. M., Ridgway G. R., Webster M. A., Smith S. M., Nichols T. E (2014) **Permutation inference for the general linear model** *NeuroImage* **92**:381–397 https://doi.org/10.1016/j.neuroimage.2014.01.060

Author information

Miriam Hauptman

Department of Psychological & Brain Sciences, Johns Hopkins University, Baltimore, USA
ORCID iD: 0000-0002-5903-1552

For correspondence: mhauptm1@jhu.edu

Marina Bedny

Department of Psychological & Brain Sciences, Johns Hopkins University, Baltimore, USA
ORCID iD: 0000-0002-2907-8042

Editors

Reviewing Editor
Roberto Bottini
University of Trento, Trento, Italy

Senior Editor
Yanchao Bi
Beijing Normal University, Beijing, China

Reviewer #1 (Public review):

Summary:

In this study, the authors aim to understand the neural basis of implicit causal inference, specifically how people infer causes of illness. They use fMRI to explore whether these inferences rely on content-specific semantic networks or broader, domain-general neurocognitive mechanisms. The study explores two key hypotheses: first, that causal inferences about illness rely on semantic networks specific to living things, such as the 'animacy network,' given that illnesses affect only animate beings; and second, that there might be a common brain network supporting causal inferences across various domains, including illness, mental states, and mechanical failures. By examining these hypotheses, the

authors aim to determine whether causal inferences are supported by specialized or generalized neural systems.

The authors observed that inferring illness causes selectively engaged a portion of the precuneus (PC) associated with the semantic representation of animate entities, such as people and animals. They found no cortical areas that responded to causal inferences across different domains, including illness and mechanical failures. Based on these findings, the authors concluded that implicit causal inferences are supported by content-specific semantic networks, rather than a domain-general neural system, indicating that the neural basis of causal inference is closely tied to the semantic representation of the specific content involved.

Strengths:

(1) The inclusion of the four conditions in the design is well thought out, allowing for the examination of the unique contribution of causal inference of illness compared to either a different type of causal inference (mechanical) or non-causal conditions. This design also has the potential to identify regions involved in a shared representation of inference across general domains.

(2) The presence of the three localizers for language, logic, and mentalizing, along with the selection of specific regions of interest (ROIs), such as the precuneus and anterior ventral occipitotemporal cortex (antVOTC), is a strong feature that supports a hypothesis-driven approach (although see below for a critical point related to the ROI selection).

(3) The univariate analysis pipeline is solid and well-developed.

(4) The statistical analyses are a particularly strong aspect of the paper.

Weaknesses:

Based on the current analyses, it is not yet possible to rule out the hypothesis that inferring illness causes relies on neurocognitive mechanisms that support causal inferences irrespective of their content, neither in the precuneus nor in other parts of the brain.

(1) The authors, particularly in the multivariate analyses, do not thoroughly examine the similarity between the two conditions (illness-causal and mechanical-causal), as they are more focused on highlighting the differences between them. For instance, in the searchlight MVPA analysis, an interesting decoding analysis is conducted to identify brain regions that represent illness-causal and mechanical-causal conditions differently, yielding results consistent with the univariate analyses. However, to test for the presence of a shared network, the authors only perform the Causal vs. Non-causal analysis. This analysis is not very informative because it includes all conditions mixed together and does not clarify whether both the illness-causal and mechanical-causal conditions contribute to these results.

(2) To address this limitation, a useful additional step would be to use as ROIs the different regions that emerged in the Causal vs. Non-causal decoding analysis and to conduct four separate decoding analyses within these specific clusters:
(a) Illness-Causal vs. Non-causal - Illness First;
(b) Illness-Causal vs. Non-causal - Mechanical First;
(c) Mechanical-Causal vs. Non-causal - Illness First;
(d) Mechanical-Causal vs. Non-causal - Mechanical First.
This approach would allow the authors to determine whether any of these ROIs can decode both the illness-causal and mechanical-causal conditions against at least one non-causal condition.

(3) Another possible analysis to investigate the existence of a shared network would be to run the searchlight analysis for the mechanical-causal condition versus the two non-causal

conditions, as was done for the illness-causal versus non-causal conditions, and then examine the conjunction between the two. Specifically, the goal would be to identify ROIs that show significant decoding accuracy in both analyses.

(4) Along the same lines, for the ROI MVPA analysis, it would be useful not only to include the illness-causal vs. mechanical-causal decoding but also to examine the illness-causal vs. non-causal conditions and the mechanical-causal vs. non-causal conditions. Additionally, it would be beneficial to report these data not just in a table (where only the mean accuracy is shown) but also using dot plots, allowing the readers to see not only the mean values but also the accuracy for each individual subject.

(5) The selection of Regions of Interest (ROIs) is not entirely straightforward:
In the introduction, the authors mention that recent literature identifies the precuneus (PC) as a region that responds preferentially to images and words related to living things across various tasks. While this may be accurate, we can all agree that other regions within the ventral occipital-temporal cortex also exhibit such preferences, particularly areas like the fusiform face area, the occipital face area, and the extrastriate body area. I believe that at least some parts of this network (e.g., the fusiform gyrus) should be included as ROIs in this study. This inclusion would make sense, especially because a complementary portion of the ventral stream known to prefer non-living items (i.e., anterior medial VOTC) has been selected as a control ROI to process information about the mechanical-causal condition. Given the main hypothesis of the study - that causal inferences about illness might depend on content-specific semantic representations in the 'animacy network' - it would be worthwhile to investigate these ROIs alongside the precuneus, as they may also yield interesting results.

(6) Visual representation of results:
In all the figures related to ROI analyses, only mean group values are reported (e.g., Figure 1A, Figure 3, Figure 4A, Supplementary Figure 6, Figure 7, Figure 8). To better capture the complexity of fMRI data and provide readers with a more comprehensive view of the results, it would be beneficial to include a dot plot for a specific time point in each graph. This could be a fixed time point (e.g., a certain number of seconds after stimulus presentation) or the time point showing the maximum difference between the conditions of interest. Adding this would allow for a clearer understanding of how the effect is distributed across the full sample, such as whether it is consistently present in every subject or if there is greater variability across individuals.

(7) Task selection:
(a) To improve the clarity of the paper, it would be helpful to explain the rationale behind the choice of the selected task, specifically addressing: (i) why an implicit inference task was chosen instead of an explicit inference task, and (ii) why the "magic detection" task was used, as it might shift participants' attention more towards coherence, surprise, or unexpected elements rather than the inference process itself.
(b) Additionally, the choice to include a large number of catch trials is unusual, especially since they are modeled as regressors of non-interest in the GLM. It would be beneficial to provide an explanation for this decision.

https://doi.org/10.7554/eLife.101944.1.sa2

Reviewer #2 (Public review):

Summary:

In this study, the authors hypothesize that "causal inferences about illness depend on content-specific semantic representations in the animacy network". They test this hypothesis in an fMRI task, by comparing brain activity elicited by participants' exposure to written situations

suggesting a plausible cause of illness with brain activity in linguistically equivalent situations suggesting a plausible cause of mechanical failure or damage and non-causal situations. These contrasts identify PC as the main "culprit" in a whole-brain univariate analysis. Then the question arises of whether the content-specificity has to do with inferences about animates in general, or if there are some distinctions between reasoning about people's bodies versus mental states. To answer this question, the authors localize the mentalizing network and study the relation between brain activity elicited by Illness-Causal > Mech-Causal and Mentalizing > Physical stories. They conclude that inferring about the causes of illness partially differentiates from reasoning about people's states of mind. The authors finally test the alternative yet non-mutually exclusive hypothesis that both types of causal inferences (illness and mechanical) depend on shared neural machinery. Good candidates are language and logic, which justifies the use of a language/logic localizer. No evidence of commonalities across causal inferences versus non-causal situations is found.

Strengths:

(1) This study introduces a useful paradigm and well-designed set of stimuli to test for implicit causal inferences.

(2) Another important methodological advance is the addition of physical stories to the original mentalizing protocol.

(3) With these tools, or a variant of these tools, this study has the potential to pave the way for further investigation of naïve biology and causal inference.

Weaknesses:

(1) This study is missing a big-picture question. It is not clear whether the authors investigate the neural correlates of causal reasoning or of naïve biology. If the former, the choice of an orthogonal task, making causal reasoning implicit, is questionable. If the latter, the choice of mechanical and physical controls can be seen as reductive and problematic.

(2) The rationale for focusing mostly on the precuneus is not clear and this choice could almost be seen as a post-hoc hypothesis.

(3) The choice of an orthogonal 'magic detection' task has three problematic consequences in this study:
(a) It differs in nature from the 'mentalizing' task that consists of evaluating a character's beliefs explicitly from the corresponding story, which complicates the study of the relation between both tasks. While the authors do not compare both tasks directly, it is unclear to what extent this intrinsic difference between implicit versus explicit judgments of people's body versus mental states could influence the results.
(b) The extent to which the failure to find shared neural machinery between both types of inferences (illness and mechanical) can be attributed to the implicit character of the task is not clear.
(c) The introduction of a category of non-interest that contains only 36 trials compared to 38 trials for all four categories of interest creates a design imbalance.

(4) Another imbalance is present in the design of this study: the number of trials per category is not the same in each run of the main task. This imbalance does not seem to be accounted for in the 1st-level GLM and renders a bit problematic the subsequent use of MVPA.

(5) The main claim of the authors, encapsulated by the title of the present manuscript, is not tested directly. While the authors included in their protocol independent localizers for mentalizing, language, and logic, they did not include an independent localizer for "animacy". As such, they cannot provide a within-subject evaluation of their claim, which is entirely

based on the presence of a partial overlap in PC (which is also involved in a wide range of tasks) with previous results on animacy.

https://doi.org/10.7554/eLife.101944.1.sa1

Reviewer #3 (Public review):

Summary:

This study employed an implicit task, showing vignettes to participants while a bold signal was acquired. The aim was to capture automatic causal inferences that emerge during language processing and comprehension. In particular, the authors compared causal inferences about illness with two control conditions, causal inferences about mechanical failures and non-causal phrases related to illnesses. All phrases that were employed described contexts with people, to avoid animacy/inanimate confound in the results. The authors had a specific hypothesis concerning the role of the precuneus (PC) in being sensitive to causal inferences about illnesses.

These findings indicate that implicit causal inferences are facilitated by semantic networks specialized for encoding causal knowledge.

Strengths:

The major strength of the study is the clever design of the stimuli (which are nicely matched for a number of features) which can tease apart the role of the type of causal inference (illness-causal or mechanical-causal) and the use of two localizers (logic/language and mentalizing) to investigate the hypothesis that the language and/or logical reasoning networks preferentially respond to causal inference regardless of the content domain being tested (illnesses or mechanical).

Weaknesses:

I have identified the following main weaknesses:

(1) Precuneus (PC) and Temporo-Parietal junction (TPJ) show very similar patterns of results, and the manuscript is mostly focused on PC (also the abstract). To what extent does the fact that PC and TPJ show similar trends affect the inferences we can derive from the results of the paper? I wonder whether additional analyses (connectivity?) would help provide information about this network.

(2) Results are mainly supported by an univariate ROI approach, and the MVPA ROI approach is performed on a subregion of one of the ROI regions (left precuneus). Results could then have a limited impact on our understanding of brain functioning.

(3) In all figures: there are no measures of dispersion of the data across participants. The reader can only see aggregated (mean) data. E.g., percentage signal changes (PSC) do not report measures of dispersion of the data, nor do we have bold maps showing the overlap of the response across participants. Only in Figure 2, we see the data of 6 selected participants out of 20.

(4) Sometimes acronyms are defined in the text after they appear for the first time.

https://doi.org/10.7554/eLife.101944.1.sa0

Author response:

We thank the editors and reviewers for their comments on our manuscript. We found the comments of the reviewers helpful and plan to add new text, analyses, and figures to answer some of the outstanding questions.

In response to the reviewers' comments, we will clarify the goal of the paper in the introduction: to test the hypothesis that causal knowledge (i.e., an intuitive theory of biology) is embedded in domain-preferring semantic networks (i.e., semantic animacy network). This work links developmental psychology work on intuitive theories and cognitive neuroscience.

As we will emphasize in the revised manuscript, the primary goal of the current paper is to test the claim that semantic networks encode causal knowledge, rather than to rule out the contribution of domain-general reasoning mechanisms to causal inference.

In response to the reviewers' suggestions, we will add multivariate and univariate whole-cortex analyses that provide further tests for domain-general causality responses. In particular, we will include new figures showing univariate responses to the mechanical inference condition over the non-causal control conditions as well as decoding between these conditions. The reviewers have also asked us to provide individual subject dispersion data. We appreciate this suggestion, and new figures will be added to display this information.

We will also perform additional analysis in the precuneus (PC) to look for shared responses to illness and mechanical inferences. In accordance with our hypotheses, we have shown that the PC responds preferentially to illness inferences. To address the reviewers' concerns about the selectivity of the PC to illness inferences, we will compare responses to i) illness inferences compared to the noncausal conditions and ii) mechanical inferences compared to the noncausal conditions in the PC to investigate the extent to which a shared response to causal inference across domains emerges in this region.

Critically, we find that the cortical areas that distinguish between causal and non-causal conditions in a 'domain general manner' (i.e., for both illness and mechanical inferences) are driven by higher responses to the non-causal condition. Moreover, these responses in prefrontal cortex and elsewhere overlap an RT predictor of neural activity, suggesting that they may reflect difficulty effects.

These results suggest that in the current task, signatures of causal inference are primarily found in domain-preferring semantic networks, rather than in domain-general fronto-parietal reasoning systems. We will provide additional discussion of the argument that the current results do not speak against the role of domain general systems across all types of causal reasoning. Instead, they suggest that the types of implicit causal inferences measured in the current study depend primarily on domain-preferring semantic networks.

The reviewers have asked us to analyze responses to causal inferences about illness in the fusiform face area (FFA). We will perform this analysis. However, we note that univariate and multivariate whole-cortex analyses that are already included in the paper did not identify lateral ventral occipito-temporal cortex as a key region involved in causal inferences about illness. Further, we do not have FFA localizer data in the current participants; therefore, the results cannot be interpreted to reflect activity in functionally defined FFA.

Two reviewers asked us to justify our choice of an implicit magic-detection task, which we will now do more clearly in the manuscript. This task was selected to ensure that participants were attending to the meaning of the vignettes. The goal of the current study was to investigate implicit causal inferences that routinely occur in language comprehension, e.g., when someone is reading a book. Past work has shown that explicitly judging the causality of causal and non-causal stimuli results in differences in response times across conditions (e.g.,

Kuperberg et al., 2006). In the current study, such judgments would also have introduced a confound between the behavioral decision and the condition of interest: the use of an explicit causal judgment task makes it impossible to know whether any observed neural differences between causal and non-causal conditions are simply due to differences in the selection of task responses. The selection of an orthogonal magic-detection task limits these confounds from complicating our interpretation of the neural data.

One of the reviewers asked us to justify the number of catch trials that we decided to include in our paradigm. Approximately 20% of the vignettes were “magical” vignettes (the same proportion as each of the 4 experimental conditions) to encourage participants to remain attentive throughout the task. Since these catch trials are excluded from analysis, their proportion is unlikely to influence the results of the study. We will clarify this in the manuscript.

A question was raised about the balance of trial numbers across conditions and across runs. To address this, we will include individual comparisons of each causal condition (n=36) with each non-causal condition (n=36; i.e., equal trial counts) where they are not already shown. With regard to runs, each condition is shown either 6 or 7 times per run (maximum difference of 1 trial between conditions), and the number of trials per condition is equal across the whole experiment: each condition is shown 7 times in two of the runs and 6 times four of the runs. This minor design imbalance is typical of fMRI experiments and is unlikely to impact the results. We will clarify this in the manuscript.

We believe that our planned revisions will strengthen the paper and highlight its contributions to our understanding of the neural basis of implicit causal inference.

https://doi.org/10.7554/eLife.101944.1.sa4