Animacy semantic network supports implicit causal inferences about illness

Miriam Hauptman and Marina Bedny

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

Corresponding Author
Miriam Hauptman
mhauptm1@jhu.edu

Abstract

Inferring the causes of illness is a culturally universal example of causal thinking. We tested the hypothesis that implicit causal inferences about biological processes (e.g., illness) depend on the animacy semantic network. Participants (n=20) undergoing fMRI read two-sentence vignettes that elicited causal inferences across sentences, either about the emergence of illness or about the mechanical breakdown of inanimate objects, in addition to noncausal control vignettes. All vignettes were about people and were linguistically matched. The same participants performed localizer tasks: language, logical reasoning, and mentalizing. Inferring illness causes, relative to all control conditions, selectively engaged a portion of the precuneus (PC) previously implicated in the semantic representation of animates (e.g., people, animals). Neural responses to causal inferences about illness were adjacent to but distinct from responses to mental state inferences, suggesting a neural mind/body distinction. We failed to find evidence for domain-general responses to causal inference. 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 not directly perceptible (Tooby & DeVore, 1987, Lagnado et al., 2007; Rottman, Ahn, & Luhmann, 2011; Muentener & Schulz, 2014; Sloman & Lagnado, 2015; Goddu & Gopnik, 2024). 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 by studying causal inferences about illness.

Adults have rich, culturally-specific causal 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; Legare & Shtulman, 2017). In many societies, designated 'healers' become experts in diagnosing and treating disease (Foster, 1976; Ackerknecht, 1982; Norman et al., 2009; Lightner, Heckelsmiller, & Hagen, 2021). Non-expert adults routinely infer the causes of illness in themselves and others (e.g., *how did my friend get COVID?*). Even young children think about illness in systematic ways, reflecting their burgeoning commonsense understanding of the biological world (Wellman & Gelman, 1992; Keil, 1992; Inagaki & Hatano, 2006). Young children attribute illness to contaminated food, contact with a sick person, and parental inheritance (Springer & Ruckel, 1992; Kalish, 1996, 1997; Keil et al., 1999; Notaro et al., 2001; Raman & Winer, 2004; Raman & Gelman, 2005; Legare & Gelman, 2008; Legare, Wellman, & Gelman, 2009; DeJesus, Venkatesh, & Kinzler, 2021).

Illness affects living things (e.g., people and animals) rather than inanimate objects (e.g., rocks, machines, houses). Thinking about living things (animates) as opposed to non-living things (inanimate objects/places) recruits partially distinct neural systems (e.g., Warrington & Shallice, 1984; Hillis & Caramazza, 1991; Caramazza & Shelton, 1998; Farah & Rabinowitz, 2003). The precuneus (PC) is part of the 'animacy' semantic network and responds preferentially to living things (i.e., people and animals), whether presented as images or words (Devlin et al., 2002; 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., 2025). By contrast, parts of the visual system (e.g., fusiform face area) that respond preferentially to animates do so primarily for images (Kanwisher et al., 1997; Grill-Spector et al., 2004; Noppeney et al., 2006; Mahon et al., 2009; Konkle & Caramazza, 2013; Connolly et al., 2016; see Bi et al., 2016 for a review). We hypothesized that the PC represents causal knowledge relevant to animates and tested the prediction that it would be activated during implicit causal inferences about illness, which rely on such knowledge (preregistration: https://osf.io/6pnqg).

We also compared neural responses to causal inferences about the body (i.e., illness) and inferences about the mind (i.e., mental states). Both types of inferences are about animate entities, and some developmental work suggests that children use the same set of causal principles to think about bodies and minds (Carey, 1985, 1988). Other evidence suggests that by early childhood, young children have distinct causal knowledge about the body and the mind

(Springer & Keil, 1991; Callanan & Oakes, 1992; Wellman & Gelman, 1992; Inagaki & Hatano, 1993; 2004; Keil, 1994; Hickling & Wellman, 2001; Medin et al., 2010). For instance, preschoolers are more likely to view illness as a consequence of biological causes, such as contagion, rather than psychological causes, such as malicious intent (Springer & Ruckel, 1992; Raman & Winer, 2004; see also Legare & Gelman, 2008). The neural relationship between inferences about bodies and minds has not been fully described. The 'mentalizing network', including the PC, is engaged when people reason about agents' beliefs (Saxe & Kanwisher, 2003; Saxe et al., 2006; Saxe & Powell, 2006; Dodell-Feder et al., 2011; Dufour et al., 2013). We localized this network in individual participants and measured its neuroanatomical relationship to the network activated by illness inferences.

An alternative hypothesis is that domain-general neural mechanisms, separate from semantic networks, support causal inferences across domains. Children and adults make causal inferences across a wide range of domains and use similar cognitive principles (e.g., 'screening off') when doing so (e.g., Saxe & Carey, 2006; Tenenbaum et al., 2007; Carey, 2011; 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). Prior neuroscience work has hypothesized that the frontotemporal language network may support a broad range of causal inferences during comprehension (Kuperberg et al., 2006; Mason & Just, 2011; Prat et al., 2011; see also Spelke, 2003; 2022; Pinker, 2003). Alternatively, causal inference could depend on frontoparietal mechanisms that also support other types of reasoning, such as logical deduction (Goldvarg & Johnson-Laird, 2001; Barbey & Patterson, 2011; Khemlani et al., 2014; Operskalski & Barbey, 2017). Finally, it has been suggested that causal inferences are supported by a dedicated 'causal engine' in prefrontal cortex that supports all and only causal inferences across domains (Pramod, Chomik-Morales, et al., 2023). We tested these alternative hypotheses in the specific case of implicit causal inferences that unfold naturally during language comprehension (Black & Bern, 1981; Keenan et al., 1984; Trabasso & Sperry, 1985; Myers et al., 1987; Duffy et al., 1990).

Most prior studies investigating causal inference used explicit causality judgment tasks (Ferstl & von Cramon, 2001; Satpute et al., 2005; Fugelsang & Dunbar, 2005; Kuperberg et al., 2006; Fenker et al., 2010; Kranjec et al., 2012; Pramod, Chomik-Morales, et al., 2023). For example, Kuperberg et al. (2006) asked participants to rate the causal relatedness of three-sentence stories and observed higher responses to causally related stories in left frontotemporal cortex. Studies of implicit causal inference find frontotemporal and frontoparietal responses (Chow et al., 2008; Mason & Just, 2011; Prat et al., 2011). Across these prior studies, no consistent neural signature of causal inference has emerged. Importantly, in many studies, causal trials were more difficult, and/or linguistic variables were not matched across causal and noncausal conditions. As a result, some of the observed effects may reflect linguistic or executive load. In addition, almost no prior studies localized language or logical reasoning networks in individual participants, making it is difficult to assess the involvement of these systems (e.g., Fedorenko et al., 2010; Monti et al., 2009); cf. Pramod, Chomik-Morales, et al., 2023). Most prior work also did not distinguish between causal inferences about different semantic domains known to depend on partially distinct neural networks, e.g., biological, mechanical, or mental state inferences (cf. Mason & Just, 2011; Pramod, Chomik-Morales, et al., 2023). If such inferences recruit partially distinct neural systems, their neural signatures might have been missed.

In the current experiment, 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. Such causally connected sentences arise frequently in naturalistic discourse (Singer, 1994; Graesser et al., 1994). Participants performed a covert task of detecting 'magical' catch trial vignettes that encouraged them to attend to the meaning of the critical vignettes while reading as naturally as possible. We chose an orthogonal foil detection task rather than an explicit causal judgment task to investigate automatic causal inferences during reading and to unconfound such processing as much as possible from explicit decision-making processes. Analogous foil detection paradigms have been used to study sentence processing and word recognition (e.g., Pallier et al., 2011; Dehaene-Lambertz et al., 2018).

In the current study, causal inferences about illness were compared to two control conditions: i) causal inferences about mechanical breakdown (e.g., "Jake dropped all of his things on the subway. Now he has a shattered phone.") and ii) illness-related language that was not causally connected (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. In other words, we tested the hypothesis that causal inferences about illness recruit the animacy semantic network. 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. As a further control, we included the number of people in each vignette as a covariate of no interest in our neural analysis. Noncausal vignettes were constructed by shuffling causes/effects across conditions and were therefore matched to the causal vignettes in linguistic content. A separate group of participants rated the causal relatedness of all vignettes prior to the experiment. In addition to the main causal inference experiment, we also localized language, logical reasoning, and mentalizing networks in each participant. Following prior work, we predicted that neural systems that support causal inference would exhibit increased activity during such inferences. Thus, our primary neural prediction was that animacy-responsive PC would respond more to causal inferences about illness compared to all other control conditions. We also used multivariate methods to investigate differences between conditions.

Method

Open science practices

The methods and analysis of this experiment were pre-registered prior to data collection (https://osf.io/6pnqg).

Participants

Twenty adults (7 women, 13 men, 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 noncausal (Figure 1D). 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 noncausal conditions used the same sentences as in the *Illness-Causal* and *Mechanical-Causal* conditions but in a shuffled order: illness cause with mechanical effect (*Noncausal-Illness First)* or mechanical cause with illness effect (*Noncausal-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 noncausal conditions, $t(25) = 36.97$, $p < .001$. In addition, *Illness-Causal* and *Mechanical-Causal* items received equally high causality ratings, $t(25) = -0.64$, $p = .53$ (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 et al., 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 frequency 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 (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). This semantic foil detection task encouraged participants to attend to the meaning of the critical vignettes while reading as naturally as possible. We required participants to press a button on every trial to ensure they were attending to the stimuli. Both sentences in a given 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 (152 trials) 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 noncausal 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 6-7 trials per condition per run presented in a pseudorandom order. Vignettes from the same experimental condition repeated no more than twice consecutively, vignettes that shared 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.

Mentalizing localizer experiment

To test the relationship between neural responses to inferences about the body and the mind, and to localize animacy regions, we used a localizer task 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, enabling us to use the mentalizing localizer as a localizer for animacy. The new physical stories were also linguistically matched to the mentalizing stories to reduce linguistic confounds (see Shain et al., 2023). Specifically, we matched physical and mentalizing stories (pairwise t-tests, all ps > 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 localizer versions in 3 pilot participants can be found in Supplementary Figure 15.

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).

Language/logic localizer experiment

To test for the presence of domain-general responses to causal inference in the language and logic networks (e.g., Kuperberg et al., 2006; Operskalski & Barbey, 2017), we used an additional localizer task. 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).

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. Functional T2*-weighted BOLD scans were collected using a gradient echo planar imaging (EPI) sequence with the following parameters: 36 sequential ascending axial slices, repetition time (TR) = 2 s, echo time (TE) = 0.03 s, flip angle = 70°, field of view (FOV) matrix = 76 × 70, slice thickness = 2.5 mm, inter-slice gap = 0.5, slice-coverage FH = 107.5, voxel size = 2.4 × 2.4 × 3 mm, PE direction = L/R, first order shimming. 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*, *Noncausal-Illness First, Noncausal-Mechanical First*) and 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. The GLM additionally included participant response time and number of people in each vignette as covariates of no interest. 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 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 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).

Individual-subject fROI analysis: univariate

We defined individual-subject functional ROIs (fROIs) in the PC and temporoparietal junction (TPJ) as well as in the language (frontal and temporal search spaces) and logic networks. In an exploratory analysis, we defined individual-subject fROIs in an anterior parahippocampal region (i.e., anterior PPA). For all analyses, percent signal change (PSC) was extracted and averaged over the entire duration of the trial (17 s total), starting at 4 s to account for hemodynamic lag.

Illness inference fROIs were created in bilateral PC and TPJ 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 statistical non-independence (Vul & Kanwisher, 2011). In each participant, we identified the most illness inference-responsive vertices in bilateral PC and TPJ search spaces separately in 5 of the 6 runs (top 5% of vertices, *Illness-Causal > Mechanical-Causal*). We then extracted PSC for each condition compared to rest in the held-out run (*Illness-Causal*, *Mechanical-Causal*, *Noncausal-Illness First, Noncausal-Mechanical First*), averaging the results across all iterations. We used the same approach to create mechanical inference fROIs in bilateral anterior PPA search spaces from a previous study on place word representations (Hauptman, Elli, et al., 2025). 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 fROIs were created by selecting 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 fROIs were identified by selecting 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 fROI was identified by selecting 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 fROI, we extracted PSC for all conditions in the causal inference experiment.

Individual-subject fROI analysis: multivariate

We performed MVPA (PyMVPA toolbox; Hanke et al., 2009) to test whether patterns of activity in the PC, TPJ, language network, and logic network distinguished illness inferences from mechanical inferences. In each participant, we identified the top 300 vertices most responsive to the mentalizing localizer (*mentalizing stories > physical stories*) in bilateral PC and TPJ search

spaces (Dufour et al., 2013). We also identified the top 300 vertices most responsive to language (*language* > *math*) in a left language network search space (Fedorenko et al., 2010) and the top 300 vertices most responsive to logical reasoning (*logic* > *language*) in a left logic network search space (Liu et al., 2020).

In an exploratory analysis, we performed MVPA to test whether patterns of activity in the left PC and in the language and logic networks distinguished causal from noncausal vignettes. To avoid statistical non-independence, we defined additional fROIs in the left PC for the purposes of this analysis. In each participant, we identified the top 300 vertices most responsive to the critical conditions over rest (*Illness-Causal* + *Mechanical-Causal* + *Noncausal-Illness First* + *Noncausal-Mechanical First* > *Rest*) in a left PC search space (Dufour et al., 2013).

For each vertex in each participant's fROIs, 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 fROI 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 linear support vector machine classifier to test decoding between all pairs of causal and noncausal conditions (i.e., *Illness-Causal* vs. *Mechanical-Causal*, *Illness-Causal* vs. *Noncausal-Mechanical First*, *Illness-Causal* vs. *Noncausal-Illness First*, *Mechanical-Causal* vs. *Noncausal-Mechanical First*, and *Mechanical-Causal* vs. *Noncausal-Illness First*) across 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 5 classification maps indicating the classifier's accuracy in a neighborhood surrounding every vertex. Individual subject searchlight accuracy maps were then averaged within analysis, and the resulting group-wise maps were 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 fROI analysis). The whole-brain searchlight maps are then thresholded using a combination of vertex-wise threshold ($p < .001$ uncorrected) and cluster size threshold (FWER $p < .05$, corrected for multiple comparisons across the entire cortical surface).

Results

Behavioral results

Accuracy on the magic detection task was at ceiling ($M = 97.9\% \pm 2.2$ SD) and there were no significant differences across the 4 main experimental conditions (*Illness-Causal*, *Mechanical-Causal*, *Noncausal-Illness First, Noncausal-Mechanical First*), $F_{(3,57)} = 2.39$, $p = .08$. A one-way repeated measures ANOVA evaluating response time revealed a main effect of condition, $F_{(3,57)} = 32.63$, $p < .001$, whereby participants were faster on *Illness-Causal* trials ($M = 4.73 \pm 0.81$ SD) compared to *Noncausal-Illness First* ($M = 5.33$ s $\pm$ 0.85 SD) and *Noncausal-Mechanical First* ($M = 5.27$ s $\pm$ 0.89 SD) trials. There were no differences in response time between the *Mechanical-Causal* condition ($M = 5.15$ s $\pm$ 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 full behavioral results).

Inferring illness causes recruits animacy-responsive PC

We found distinctly localized neural responses to causal inferences about illness relative to both mechanical causal inferences and noncausal vignettes. A bilateral precuneus (PC) region previously implicated in thinking about animate entities (i.e., people and animals) responded preferentially to causal inferences about illness over both mechanical causal inferences and causally unrelated sentences in whole-cortex analysis ($p < .05$, corrected for multiple comparisons; Figure 1C) and in individual-subject overlap maps (Supplementary Figures 6 and 7). PC responses during illness inferences overlapped with previously reported responses to people-related concepts (Fairhall & Caramazza, 2013b; Supplementary Figure 2).

Relative to illness inferences and noncausal vignettes, inferring the causes of mechanical breakdown in inanimate entities activated bilateral anterior parahippocampal regions (i.e., anterior PPA), suggesting a double dissociation between illness and mechanical inferences (Figure 4C) (Epstein & Kanwisher, 1998; Weiner et al., 2018). This anterior PPA region is engaged during memory/verbal tasks about physical spaces (Baldassano et al., 2013; Fairhall et al., 2014; Silson et al., 2019; Steel et al., 2021; Häusler et al., 2022; Hauptman, Elli, et al., 2025).

In individual-subject fROI analysis (leave-one-run-out), we similarly found that inferring illness causes activated the PC more than inferring causes of mechanical breakdown (repeated measures ANOVA, condition (*Illness-Causal*, *Mechanical-Causal*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 19.18$, $p < .001$, main effect of hemisphere, $F_{(1,19)} = 0.3$, $p = .59$, condition x hemisphere interaction, $F_{(1,19)} = 27.48$, $p < .001$; Figure 1A). This effect was larger in the left than in the right PC (paired samples t-tests; left PC: $t_{(19)} = 5.36$, $p < .001$, right PC: $t_{(19)} = 2.27$, $p = .04$). Illness inferences also activated the PC more than illness-related language that was not causally connected (repeated measures ANOVA, condition (*Illness-Causal*, *Noncausal-Illness First*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 4.66$, $p = .04$, main effect of

hemisphere, $F_{(1,19)} = 2.51$, $p = .13$, condition x hemisphere interaction, $F_{(1,19)} = 8.07$, $p = .01$; repeated measures ANOVA, condition (*Illness-Causal*, *Noncausal-Mechanical First*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 4.38$, $p = .05$; main effect of hemisphere, $F_{(1,19)} = 1.17$, $p = .29$; condition x hemisphere interaction, $F_{(1,19)} = 17.89$, $p < .001$; Figure 1A). Both effects were significant only in the left PC (paired samples t-tests; *Illness-Causal* vs. *Noncausal-Illness First*, left PC: $t_{(19)} = 2.77$, $p = .01$, right PC: $t_{(19)} = 1.28$, $p = .22$; *Illness-Causal* vs. *Noncausal-Mechanical First,* left PC: $t_{(19)} = 3.21$, $p = .005$, right PC: $t_{(19)} = 0.5$, $p = .62$).

We also observed increased activity for illness inferences compared to mechanical inferences in the temporoparietal junction (TPJ) (leave-one-run-out individual-subject fROI analysis; repeated measures ANOVA, condition (*Illness-Causal*, *Mechanical-Causal*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 5.33$, $p = .03$, main effect of hemisphere, $F_{(1,19)} = 1.02$, $p = .33$, condition x hemisphere interaction, $F_{(1,19)} = 4.24$, $p = .05$; Supplementary Figure 13). This effect was significant only in the left TPJ (paired samples t-tests; left TPJ: $t_{(19)} = 2.64$, $p = .02$, right TPJ: $t_{(19)} = 1.13$, $p = .27$). Unlike the PC, the TPJ did not show a preference for illness inferences compared to illness-related language that was not causally connected (repeated measures ANOVA, condition (*Illness-Causal*, *Noncausal-Illness First*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 0.006$, $p = .94$, main effect of hemisphere, $F_{(1,19)} = 2.19$, $p = .16$, condition x hemisphere interaction, $F_{(1,19)} = 1.27$, $p = .27$; repeated measures ANOVA, condition (*Illness-Causal*, *Noncausal-Mechanical First*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 0.73$, $p = .41$; main effect of hemisphere, $F_{(1,19)} = 1.24$, $p = .28$; condition x hemisphere interaction, $F_{(1,19)} = 3.34$, $p = .08$; Supplementary Figure 13).

In contrast to the animacy-responsive PC, the anterior PPA showed the opposite pattern, responding more to mechanical inferences than illness inferences (leave-one-run-out individual-subject fROI analysis; repeated measures ANOVA, condition (*Mechanical-Causal*, *Illness-Causal*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 17.93$, $p < .001$, main effect of hemisphere, $F_{(1,19)} = 1.33$, $p = .26$, condition x hemisphere interaction, $F_{(1,19)} = 7.8$, $p = .01$; Figure 4A). This effect was significant only in the left anterior PPA (paired samples t-tests; left anterior PPA: $t_{(19)} = 4$, $p < .001$, right anterior PPA: $t_{(19)} = 1.88$, $p = .08$). The anterior PPA also showed a preference for mechanical inferences compared to mechanical-related language that was not causally connected (repeated measures ANOVA, condition (*Mechanical-Causal*, *Noncausal-Illness First*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 14.81$, $p = .001$, main effect of hemisphere, $F_{(1,19)} = 1.81$, $p = .2$, condition x hemisphere interaction, $F_{(1,19)} = 7.35$, $p = .01$; repeated measures ANOVA, condition (*Mechanical-Causal*, *Noncausal-Mechanical First*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 11.31$, $p = .003$; main effect of hemisphere, $F_{(1,19)} = 3.34$, $p = .08$; condition x hemisphere interaction, $F_{(1,19)} = 4$, $p = .06$; Figure 4A). Similar to the PC, both effects were larger in the left than in the right hemisphere (post-hoc paired samples t-tests; *Illness-Causal* vs. *Noncausal-Illness First*, left anterior PPA: $t_{(19)} = 3.85$, $p = .001$, right anterior PPA: $t_{(19)} = 2.22$, $p = .04$; *Illness-Causal* vs. *Noncausal-Mechanical First,* left anterior PPA: $t_{(19)} = 3.59$, $p = .002$, right anterior PPA: $t_{(19)} = 1.19$, $p = .25$).

In summary, we found distinctly localized responses to illness and mechanical causal inferences. Inferring illness causes preferentially recruited the animacy semantic network, particularly the PC.

Illness inferences and mental state inferences elicit spatially dissociable responses

Illness inferences and mental state inferences elicited spatially dissociable responses. In whole-cortex analysis, illness inferences recruited the PC bilaterally, with larger responses observed in the left hemisphere (Figure 1, see also fROI analysis showing left-lateralization above). By contrast, and in accordance with prior work (e.g., Saxe & Kanwisher, 2003), mental state inferences recruited a broader network, including not only bilateral PC, but also bilateral TPJ, superior temporal sulcus (STS), and medial and superior prefrontal cortex (Supplementary Figure 1).

Within the left PC, responses to illness inferences were located ventrally to mental state inference responses (Figure 2, Supplementary Figure 3). The z-coordinates of individual-subject activation peaks for illness inferences and mental state inferences were significantly different (repeated measures ANOVA, $F_{(1,19)} = 13.52$, $p = .002$). In addition, the size of the illness inference effect (*Illness-Causal* > *Mechanical-Causal*) was larger in illness-responsive vertices (leave-one-run-out individual-subject fROI analysis) than in mentalizing-responsive vertices in the left PC (individual-subject fROI analysis; repeated measures ANOVA, $F_{(1,19)} = 24.72$, $p < .001$, Supplementary Figure 13). These results suggest that illness inferences and mental state inferences are carried out by neighboring but partially distinct subsets of the PC.

No univariate evidence for domain-general responses to implicit causal inference

Prior neuroscience studies hypothesizing the existence of a domain-general 'causal engine' have predicted that the language network and/or domain-general executive systems (e.g., the logic network) should show elevated activity during causal inference across domains. In the current study, neither the language nor the logic network exhibited elevated neural responses during causal inferences relative to linguistically matched sentence pairs that were not causally connected. Language regions in frontotemporal cortex responded more to noncausal than causal vignettes (frontal search space: repeated measures ANOVA, $F_{(1,19)} = 23.91$, $p < .001$; temporal search space: repeated measures ANOVA, $F_{(1,19)} = 4.31$, $p = .05$; Figure 3, Supplementary Figure 8). The logic network likewise responded marginally more to noncausal vignettes, likely reflecting greater difficulty associated with integrating unrelated sentences (repeated measures ANOVA, $F_{(1,19)} = 3.88$, $p = .07$; Figure 3).

In whole-cortex univariate analysis, no shared regions responded more to causal than noncausal vignettes across domains. Two whole-cortex univariate contrasts comparing causal and noncausal conditions (*Illness-Causal* > *Noncausal-Mechanical First*, *Mechanical-Causal* > *Noncausal-Mechanical First*) revealed increased activity for the noncausal condition in bilateral prefrontal cortex. The same prefrontal areas that responded more to noncausal than causal stimuli also responded more when participants were slower to complete the task, suggesting that these responses reflect a non-specific difficulty effect (Supplementary Figure 4).

In summary, none of the predicted networks nor any regions across the whole cortex exhibited the predicted domain-general causal inference pattern, i.e., larger responses to all causal than all noncausal vignettes. These results suggest that implicit causal inferences, which draw upon a person's existing knowledge of relevant causes and effects, do not depend on domain-general neural mechanisms. These results leave open the possibility that domain-general systems support the explicit search for causal connections (see Discussion section).

Multivariate analysis

In searchlight MVPA performed across the whole cortex, illness inferences and mechanical inferences produced spatially distinguishable neural patterns in the left PC extending dorsally into the superior parietal lobule, as well as in left anterior PPA and lateral occipitotemporal cortex. A whole-cortex searchlight analysis that tested whether each causal condition could be decoded from each noncausal condition found no shared regions that exhibited significant decoding across all causal vs. noncausal comparisons (Supplementary Figure 9).

In individual-subject fROI decoding analyses, illness inferences and mechanical inferences produced spatially distinguishable neural patterns in the left PC, right PC, and left TPJ, as well as in language and logic networks (see Supplementary Figure 10, Supplementary Table 2 for full results). Note that these decoding results must be interpreted in light of the significant univariate differences observed across conditions that are reported above. Linear classifiers are highly sensitive to univariate differences (Coutanche, 2013; Kragel et al., 2012; Hebart & Baker, 2018; Woolgar et al., 2014; Davis et al., 2014; Pakravan et al., 2022). Successful decoding may be driven by univariate differences in the predicted direction (e.g., causal > noncausal) or in the opposite direction (e.g., noncausal > causal). In particular, given that both the language and the logic networks exhibited higher univariate responses to noncausal compared to causal vignettes, decoding results observed in these networks may be driven by univariate differences.

Discussion

Causal knowledge is embedded in higher-order semantic networks

We find that a semantic network previously implicated in thinking about animates, particularly the precuneus (PC), is preferentially engaged when people infer causes of illness compared to when they infer causes of mechanical breakdown or read causally unconnected sentences containing illness-related language. By contrast, mechanical inferences activate an anterior parahippocampal region previously implicated in thinking about and remembering places (Baldassano et al., 2013; Fairhall et al., 2014; Silson et al., 2019; Steel et al., 2021; Häusler et al., 2022; Hauptman, Elli, et al., 2025). This finding points to a neural double dissociation between biological and mechanical causal knowledge.

Previous work has implicated the PC in the representation of animate entities, i.e., 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., 2025). Here we show that the PC exhibits sensitivity to causal inferences about biological processes specific to animates, such as illness.

These findings are consistent with our preregistered hypotheses and suggest that causal knowledge about animate and inanimate entities is distributed across multiple distinct semantic networks. Further, our results suggest that the animacy semantic network supports biological causal knowledge. Future work should test whether the animacy network is sensitive to causal information beyond illness, including about growth, birth, and death. We hypothesize that changes in biological causal knowledge during development as well as cultural expertise in causal reasoning about illness (e.g., medical expertise) influences activity in the animacy network (Legare et al., 2012; Norman et al., 2009).

Our findings are consistent with prior evidence from naturalistic paradigms showing that the PC is sensitive to discourse-level processes across sentences (e.g., Hasson et al., 2008; Lerner et al., 2011; Lee & Chen, 2022). We hypothesize that PC responses observed during naturalistic narrative comprehension are driven by causal inferences about animate agents, who are often the focus of narratives. Likewise, PC involvement in episodic memory could be related to animacy-related inferential processes (DiNicola, Braga, & Buckner, 2020; Ritchey & Cooper, 2020). Future work can test this hypothesis by comparing causal inferences about animate and inanimate entities in naturalistic contexts, such as films and verbal narratives (see Chen & Bornstein, 2024 for a review on causal inference in narrative comprehension).

We find that neural responses during inferences about biological and mental properties of animates are linked yet separable. Inferring illness causes recruits neural circuits that are adjacent to but distinct from responses to mental state inferences in the PC (Saxe & Kanwisher, 2003; Saxe et al., 2006). 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, e.g., "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, 2022). The current findings point to the existence of distinct but related neural systems for biological and mentalistic knowledge.

Our neuroimaging findings are consistent with evidence from developmental psychology suggesting that causal knowledge is central to human concepts starting early in development (Keil, 1992; Wellman & Gelman, 1992; Hatano & Inagaki, 1994; Springer & Keil, 1991; Simons & Keil, 1995; Atran, 1998; Keil et al., 1999; Coley, Solomon, & Shafto, 2002; Medin & Atran, 2004). According to the 'intuitive theories' account, semantic knowledge is organized into causal frameworks that serve as 'grammars for causal inference' (Tenenbaum et al., 2007; Wellman & Gelman, 1992; Gopnik & Meltzoff, 1997; Gopnik & Wellman, 2012; Gerstenberg & Tenenbaum, 2017; see also Boyer 1995; Barrett, Cosmides, & Tooby, 2007; Cosmides & Tooby, 2013; Bender, Beller, & Medin, 2017). For example, preschoolers intuit that animates but not inanimate objects get sick and need nourishment to grow and live (e.g., Rosengren et al., 1991; Kalish, 1996; Gutheil, Vera, & Keil, 1998; Raman & Gelman, 2005; see Inagaki & Hatano, 2004; Opfer & Gelman, 2011 for reviews). The present results suggest that such knowledge is encoded in higher-order semantic brain networks. By contrast, we failed to find sensitivity to

causal inference in portions of the ventral stream previously associated with the perception of animate agents (see Appendix 4, Supplementary Figure 8 for details). Sensitivity to causal information may be a distinguishing characteristic of higher-order, amodal semantic networks, as opposed to perceptual regions that are activated during semantic tasks (e.g., Martin & Chao, 2001; Thompson-Schill, 2003; Barsalou et al., 2003; Binder & Desai, 2011; Bi, 2021).

No evidence for domain-general neural responses during implicit causal inference

In the current study, participants read two sentence vignettes that either elicited causal inferences or were not causally connected. No brain regions responded more to causal inferences across domains compared to noncausal vignettes in this task. The language network responded more to noncausal than causal vignettes, possibly due to greater difficulty associated with processing the meaning of a sentence that does not follow from the prior context. Prior studies find that the language network is specialized primarily for sentence-internal processing (Fedorenko & Varley, 2016; Jacoby & Fedorenko, 2020; Blank & Fedorenko, 2020) and patients with agrammatic aphasia can make causal inferences about pictorial stimuli (Varley & Siegal, 2000; Varley, 2014). Together, these results suggest that the language system itself is unlikely to support implicit causal inference. Rather, during language comprehension, the language system interacts with semantic networks to enable such inferences (Simony et al., 2016; Yeshurun et al., 2017; Chang et al., 2022). Notably, in the current study, responses to causal inference in semantic networks were stronger in the left hemisphere. The left lateralization of such responses may enable efficient interfacing with the language system during comprehension.

We also failed to find evidence for the claim that the frontoparietal logical reasoning network, a domain-general executive system, supports implicit causal inferences. By contrast, the frontoparietal network responded more to noncausal than causal vignettes. Finally, we failed to observe elevated responses to causal inference across domains anywhere in the brain in whole-cortex analysis. A large swath of prefrontal cortex responded more to one noncausal condition (*Noncausal-Mechanical First*) compared to both causal conditions. The same prefrontal regions also exhibited increased activity when participants were slower to respond to the task. Thus, this 'reverse causality effect' likely reflects processing demands rather than causal inference per se. An alternative interpretation of elevated prefrontal activity observed for one of the noncausal conditions is that it reflects the effortful search for a causal connection between sentences when such a connection is difficult to find. This interpretation would suggest that domain-general executive mechanisms become engaged when causal inferences are effortful and explicit. By contrast, semantic systems are engaged when we implicitly infer a known causal relationship.

Causal inferences are a highly varied class, and domain-general systems likely play an important role in many causal inferences not tested in the current study. 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., 2009; Meder & Mayrhofer, 2017). The discovery of novel causal relationships (e.g., 'blicket detectors'; Gopnik et al., 2001) and the identification of complex causes, even in the case of illness, may depend in part on domain-general neural mechanisms. The present results suggest, however, that causal knowledge is

embedded within higher-order semantic systems, and that biological causal knowledge is embedded with a semantic system relevant to animacy.

Acknowledgements

We thank the F.M. Kirby Research Center for Functional Brain Imaging at the Kennedy Krieger Institute for their assistance with data collection, and the participants for making this research possible. This work was supported by a grant from the National Science Foundation (BCS-2318685 to M.B.).

Figures

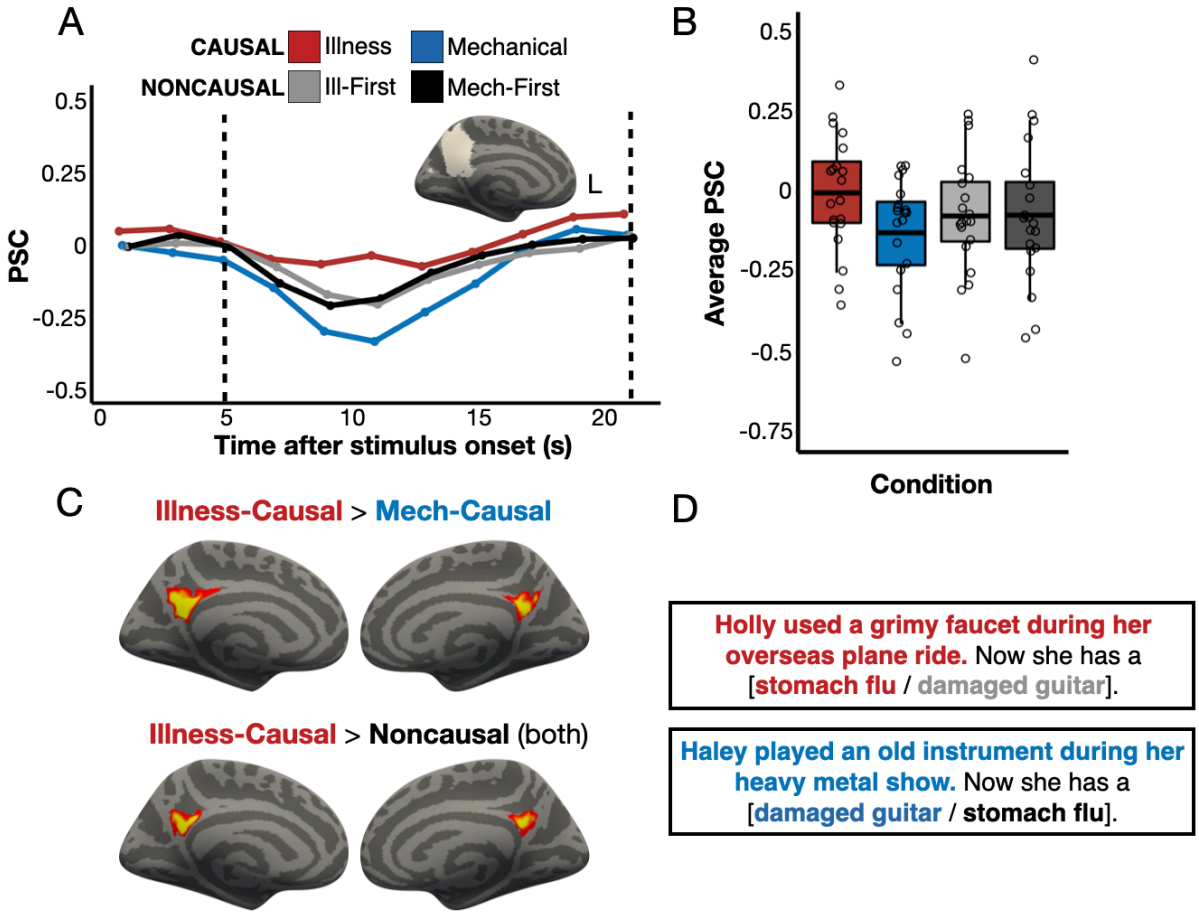


Figure 1. Responses to illness inferences in the precuneus (PC). Panel A: Percent signal change (PSC) for each condition among the top 5% *Illness-Causal* > *Mechanical-Causal* vertices in a left PC search space (Dufour et al., 2013) in individual participants, established via a leave-one-run-out analysis. Panel B: Average PSC in the critical window (marked by dotted lines in Panel A) across participants. The horizontal line within each boxplot indicates the overall mean. Panel C: Whole-cortex results (one-tailed) for *Illness-Causal* > *Mechanical-Causal* and *Illness-Causal* > *Noncausal* (both versions of noncausal 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.

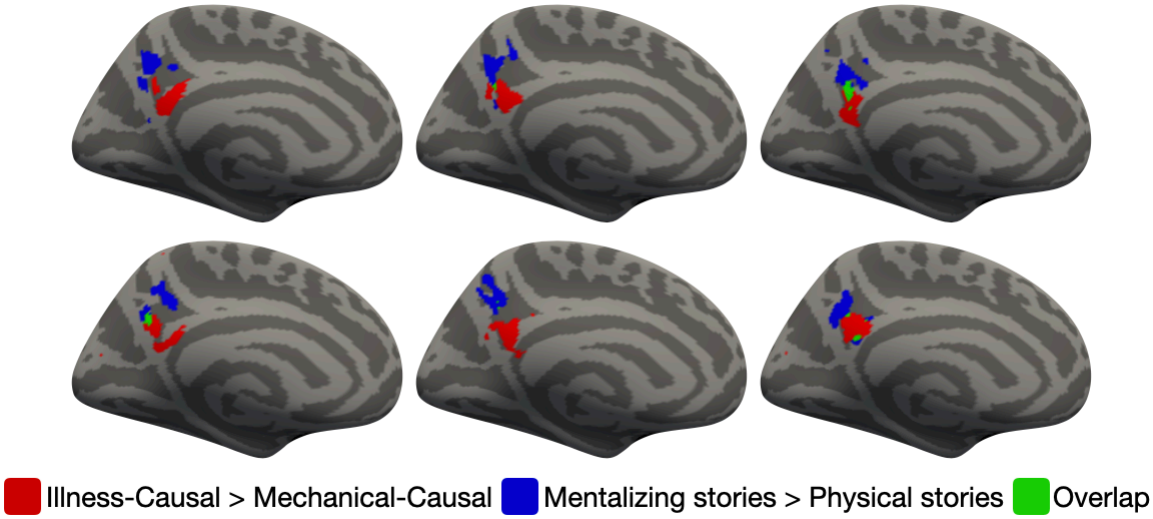


Figure 2. Spatial dissociation between univariate responses to illness inferences and mental state inferences in the precuneus (PC). The left medial surface of 6 individual participants were selected for visualization purposes. The locations of the top 10% most responsive vertices to *Illness-Causal > Mechanical-Causal* in a PC search space (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 search space are shown in blue. Overlapping vertices are shown in green.

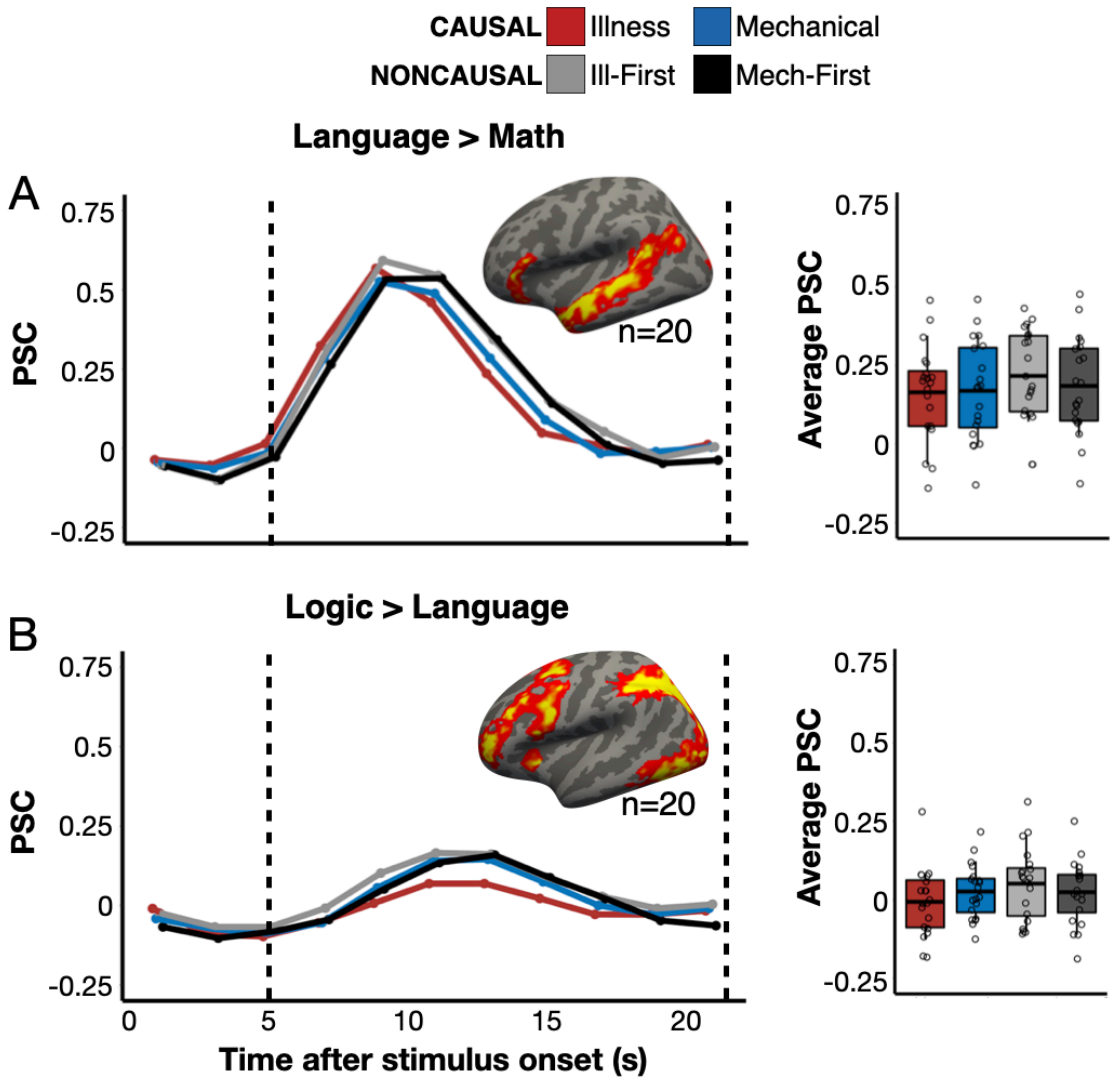


Figure 3. Individual-subject analysis of language- and logic-responsive vertices. Panel A: percent signal change (PSC) for each condition among the top 5% most language-responsive vertices (*language* > *math*) in a temporal language network search space (Fedorenko et al., 2010). Results from a frontal language search space (Fedorenko et al., 2010) can be found in Supplementary Figure 8. Panel B: PSC among the top 5% most logic-responsive vertices (*logic* > *language*) in a logic network search space (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. Boxplots display average PSC in the critical window (marked by dotted lines) across participants. The horizontal line within each boxplot indicates the overall mean.

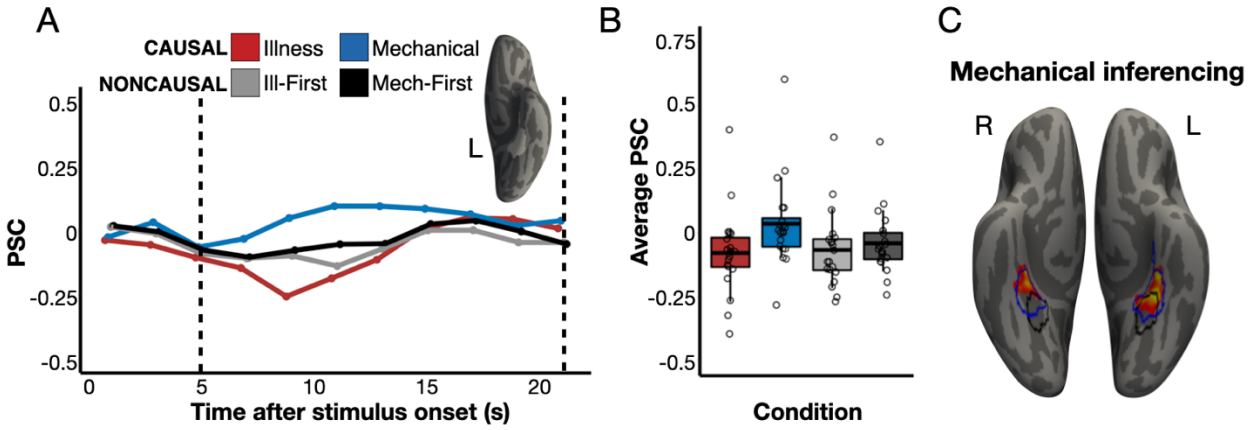
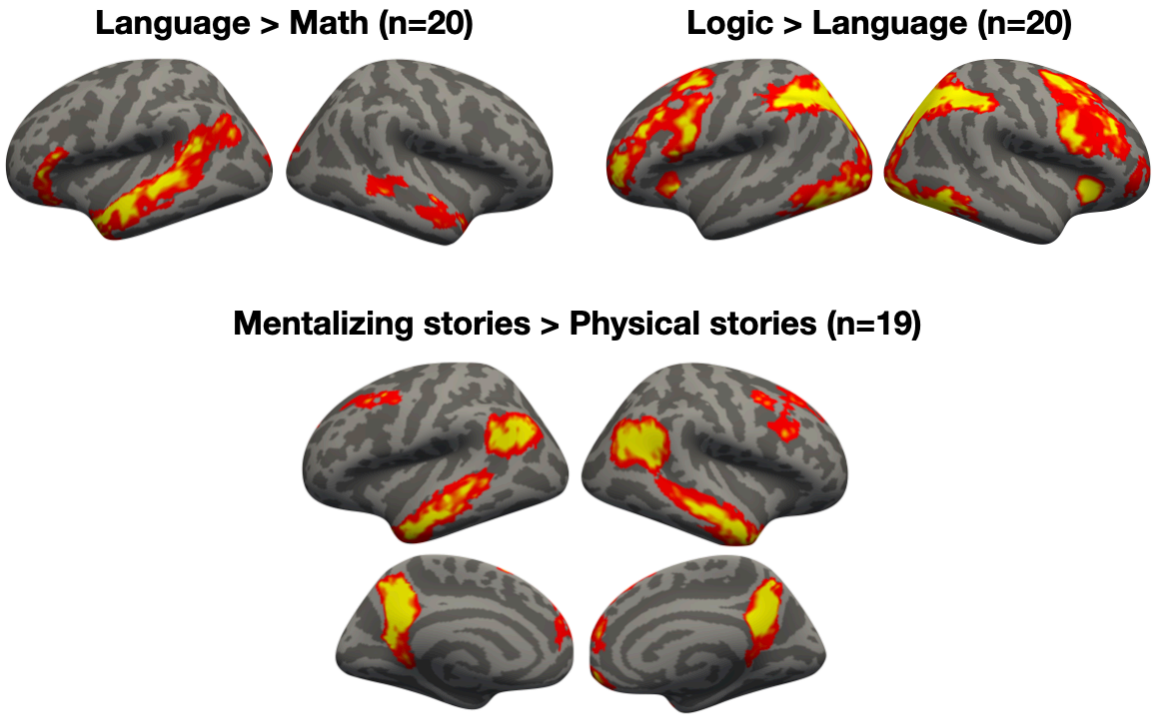
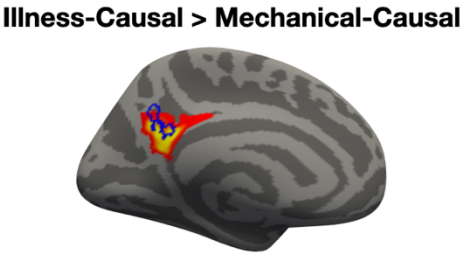


Figure 4. Responses to mechanical inferences in anterior parahippocampal regions (anterior PPA). Panel A: Percent signal change (PSC) for each condition among the top 5% *Mechanical-Causal* > *Illness-Causal* vertices in a left anterior PPA search space (Hauptman, Elli, et al., 2025) in individual participants, established via a leave-one-run-out analysis. Panel B: Average PSC in the critical window (marked by dotted lines in Panel A) across participants. The horizontal line within each boxplot indicates the overall mean. Panel C: The intersection of two whole-cortex contrasts (one-tailed), *Mechanical-Causal* > *Illness-Causal* and *Mechanical-Causal* > *Noncausal* that 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. Similar to PC responses to illness inferences, anterior PPA 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., 2018) is overlaid in black. The average PPA location from separate study involving verbal place stimuli (Hauptman, Elli, et al., 2025) is overlaid in blue.

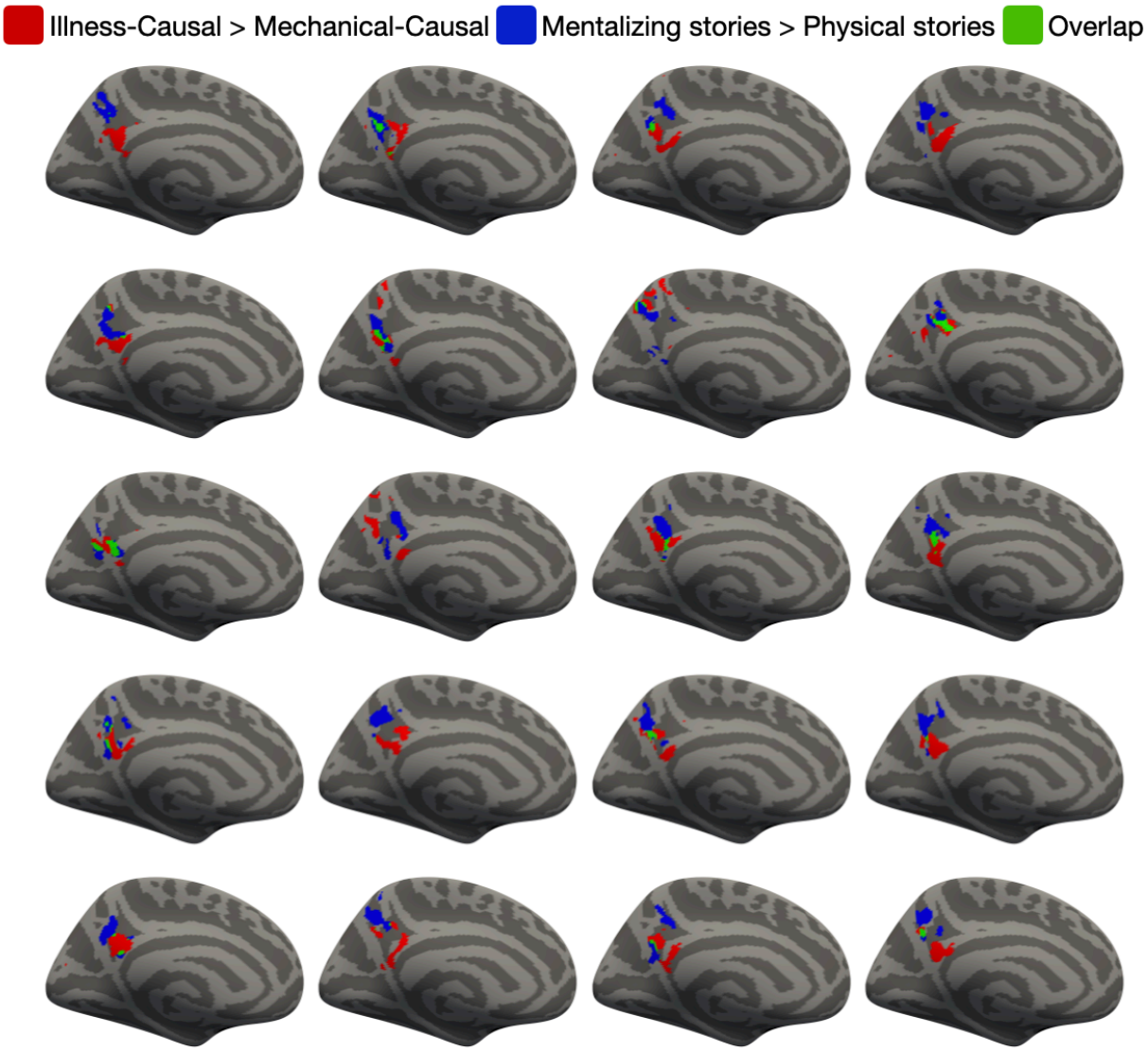
Supplementary Figures



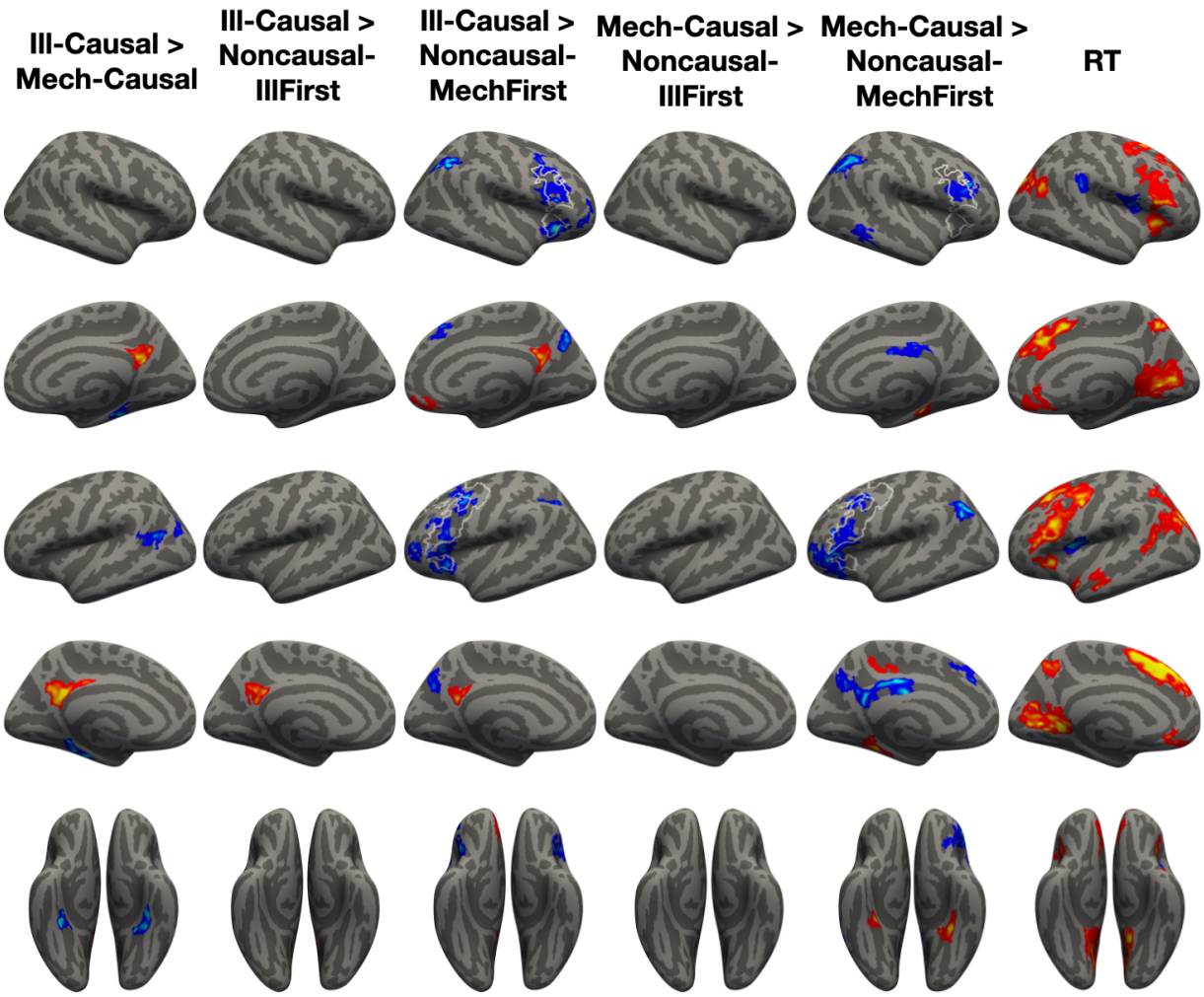
Supplementary Figure 1. Functional localization of language, logical reasoning, and mentalizing networks (see Monti et al., 2009; Fedorenko et al., 2010; Dodell-Feder et al., 2011; Liu et al., 2020). Group maps for each contrast of interest (one-tailed) are corrected for multiple comparisons ($p < .05$ FWER, cluster-forming threshold $p < .01$ uncorrected). Vertices are color coded on a scale from p=0.01 to p=0.00001.



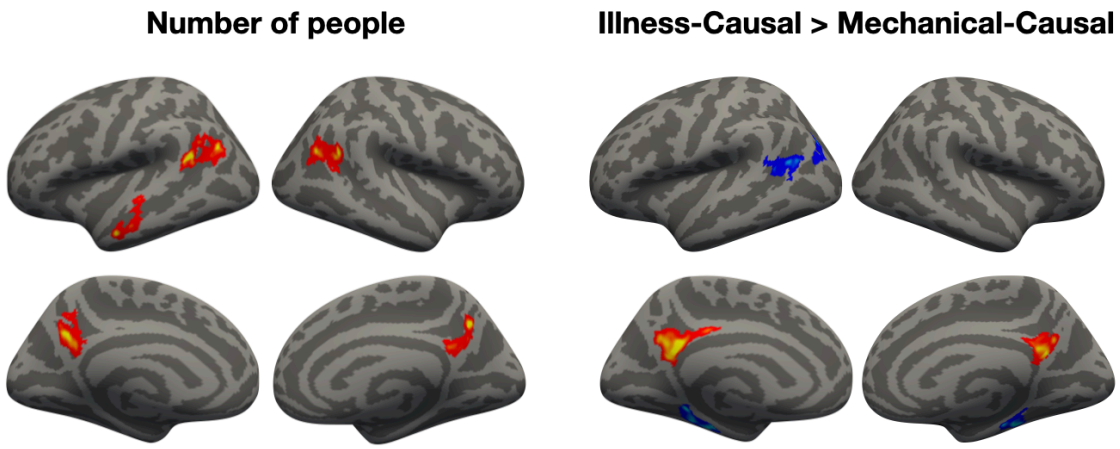
Supplementary Figure 2. Overlap between left precuneus (PC) responses to illness inferences in the current study and people-related stimuli in a separate study (Fairhall & Caramazza, 2013b). The average location from a separate study comparing people and place concepts (Fairhall & Caramazza, 2013b) is overlaid in blue on the response to illness inferences observed in the current study. The 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.



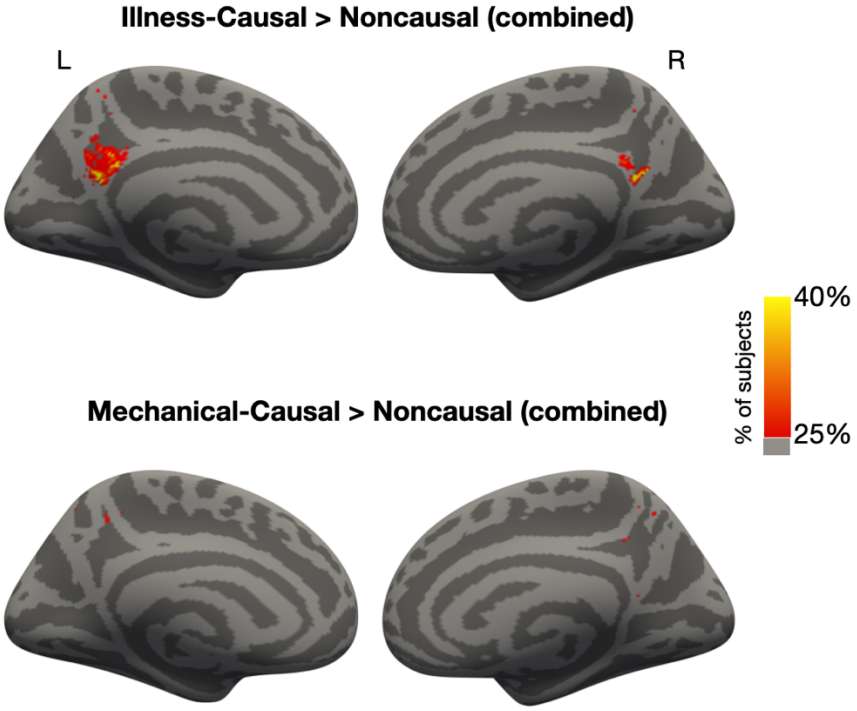
Supplementary Figure 3. Spatial dissociation between responses to illness inferences and mental state inferences in left precuneus (PC). The left medial surface of all participants (n=20) is shown. The locations of the top 10% most responsive vertices to *Illness-Causal > Mechanical-Causal* in a PC search space (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 search space are shown in blue. Overlapping vertices are shown in green.



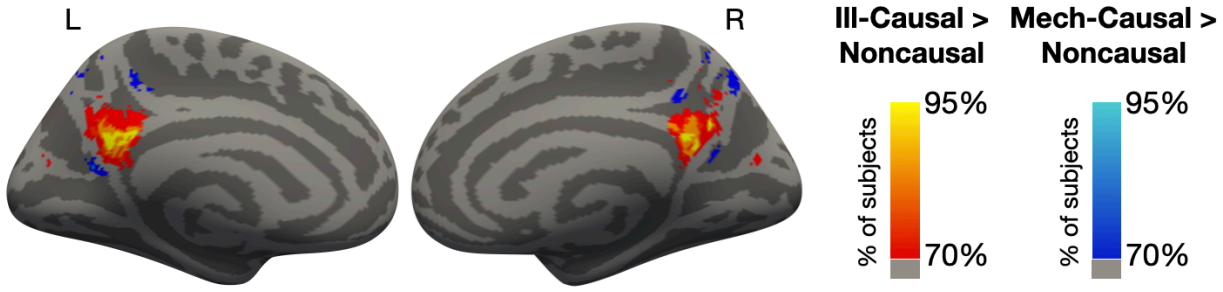
Supplementary Figure 4. Full whole-cortex univariate results. Regions whose activity scales with response time (RT) are displayed under "RT." Frontal RT regions are outlined in white on the lateral surface for other contrasts where frontal effects are observed. Group maps (two-tailed) are corrected for 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.



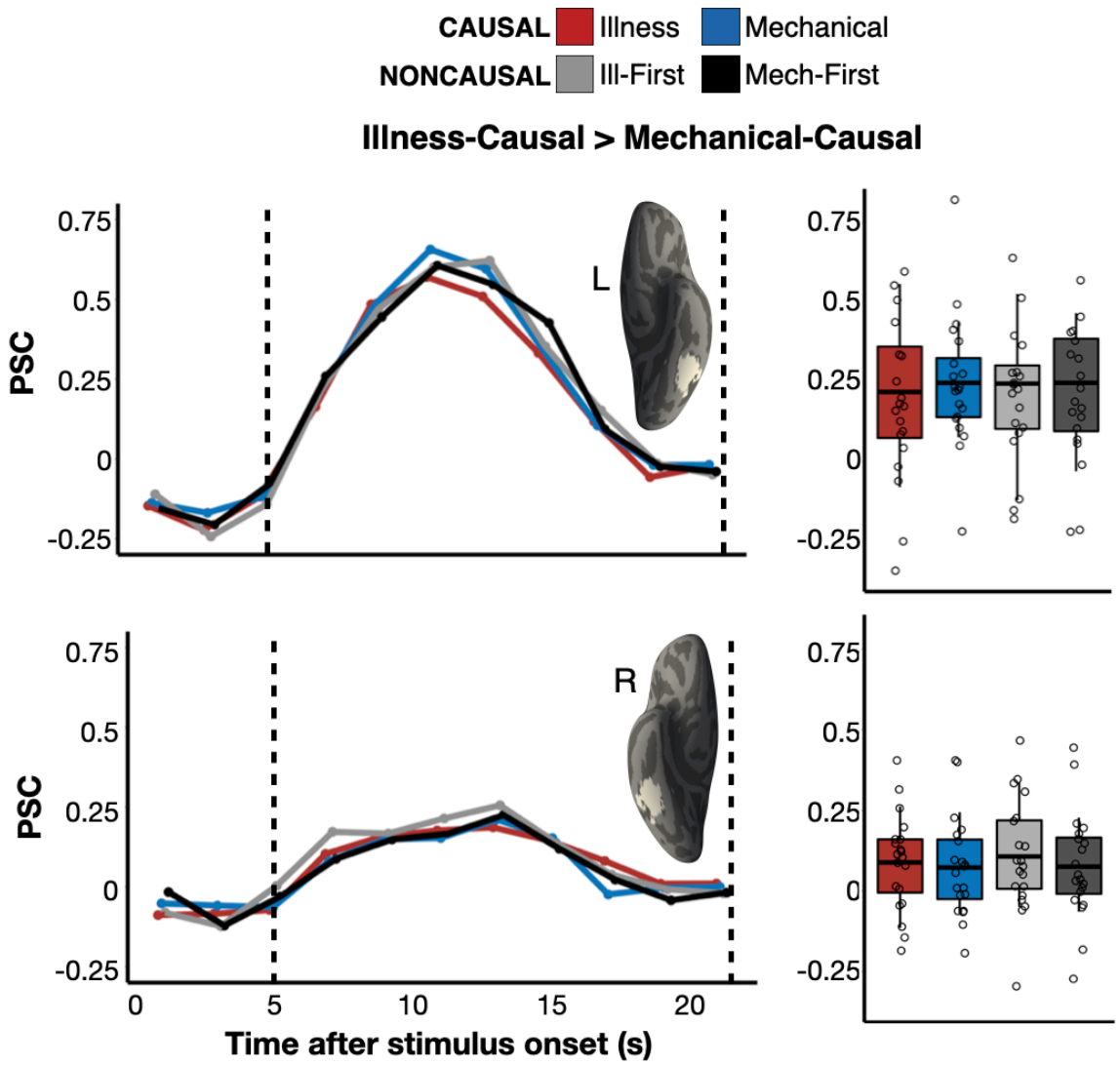
Supplementary Figure 5. Comparison of whole-cortex results for number of people in each vignette (left) and illness inferences (right) from the same GLM. Group maps (two-tailed) are corrected for multiple comparisons ($p < .05$ FWER, cluster-forming threshold $p < .01$ uncorrected). Vertices are color coded on a scale from p=0.01 to p=0.00001.



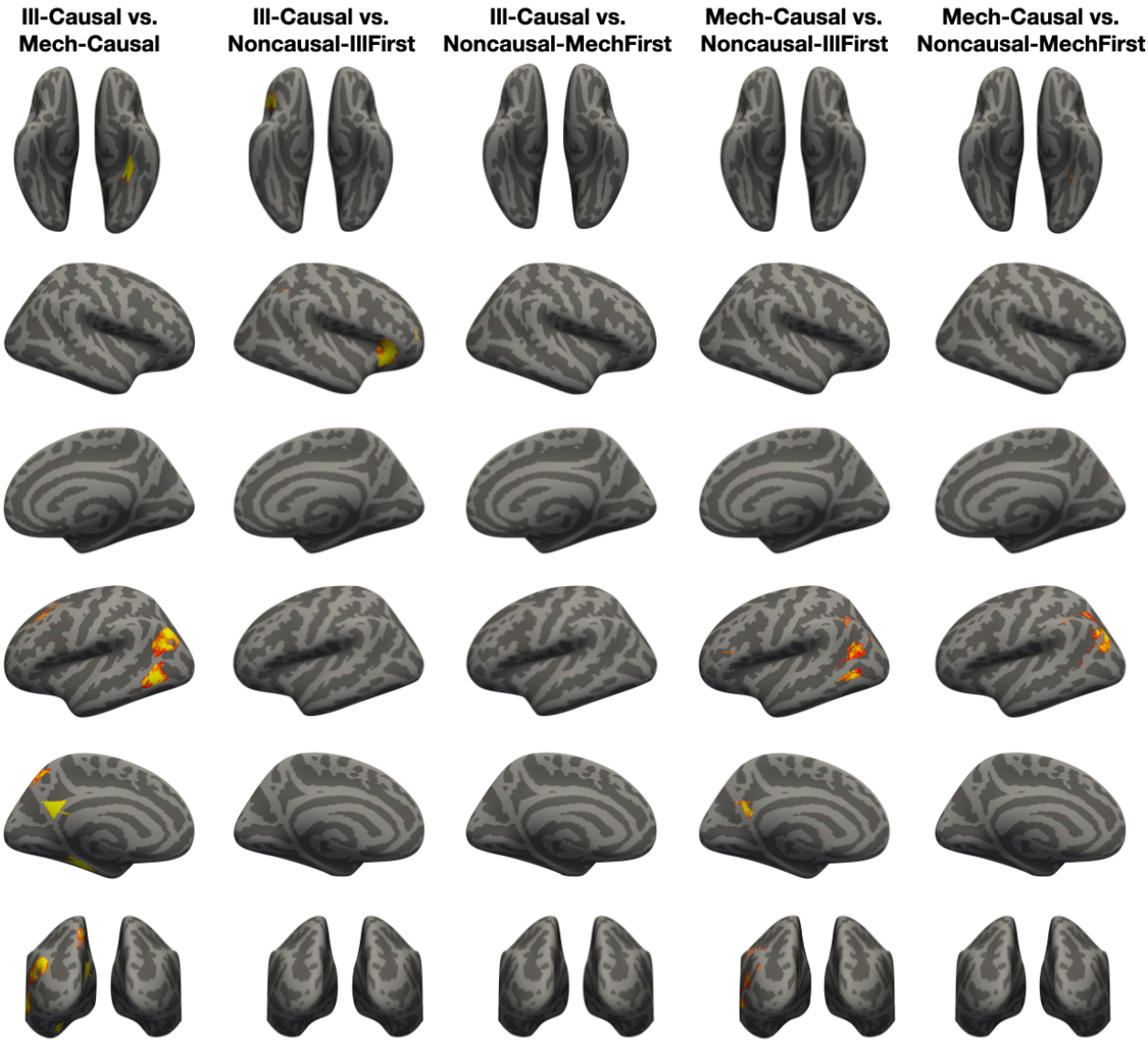
Supplementary Figure 6. Group overlap in univariate contrasts comparing causal (*Illness-Causal*, *Mechanical-Causal*) and noncausal conditions (*Noncausal-Illness First* + *Noncausal-Mechanical First*) in the PC. Each vertex in a PC search space (Dufour et al., 2013) was color-coded according to the proportion of participants who showed significant activation ($p < .05$ uncorrected) at that location.



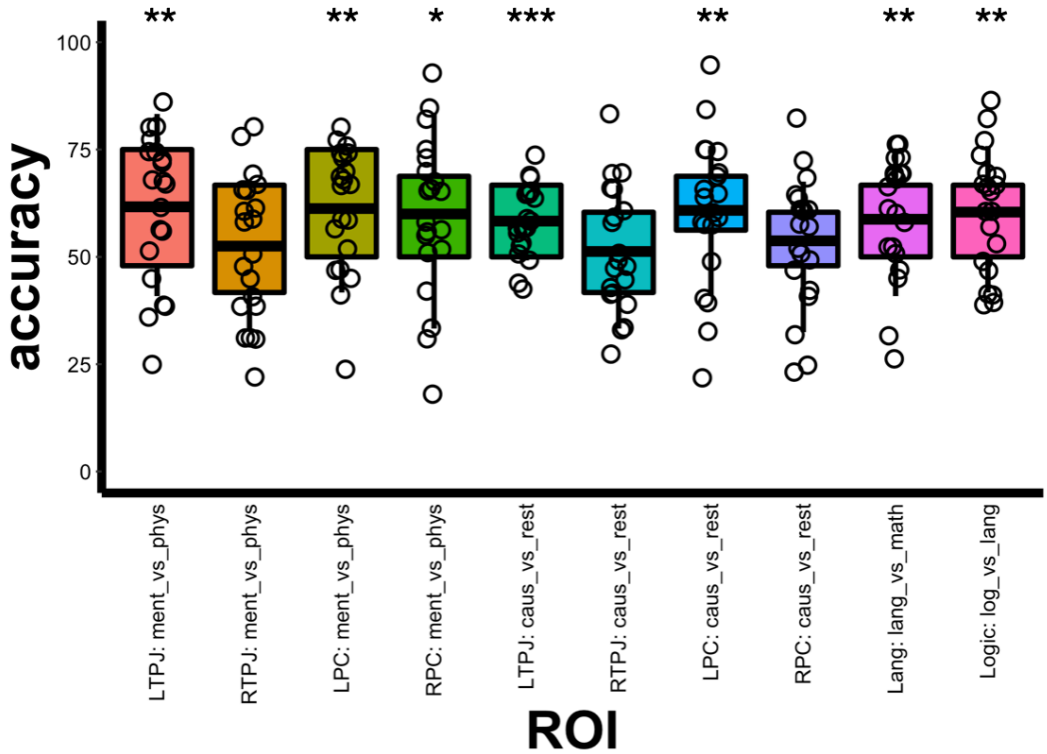
Supplementary Figure 7. Group overlap in univariate contrasts comparing causal (*Illness-Causal*, *Mechanical-Causal*) and noncausal conditions (*Noncausal-Illness First* + *Noncausal-Mechanical First*) in the PC, winner-take-all approach. Each vertex in a PC search space (Dufour et al., 2013) was color-coded according to the proportion of participants who showed a preference for *Illness-Causal* > *Noncausal* compared to *Mechanical-Causal* > *Noncausal* (red) and vice versa (blue) at that location.



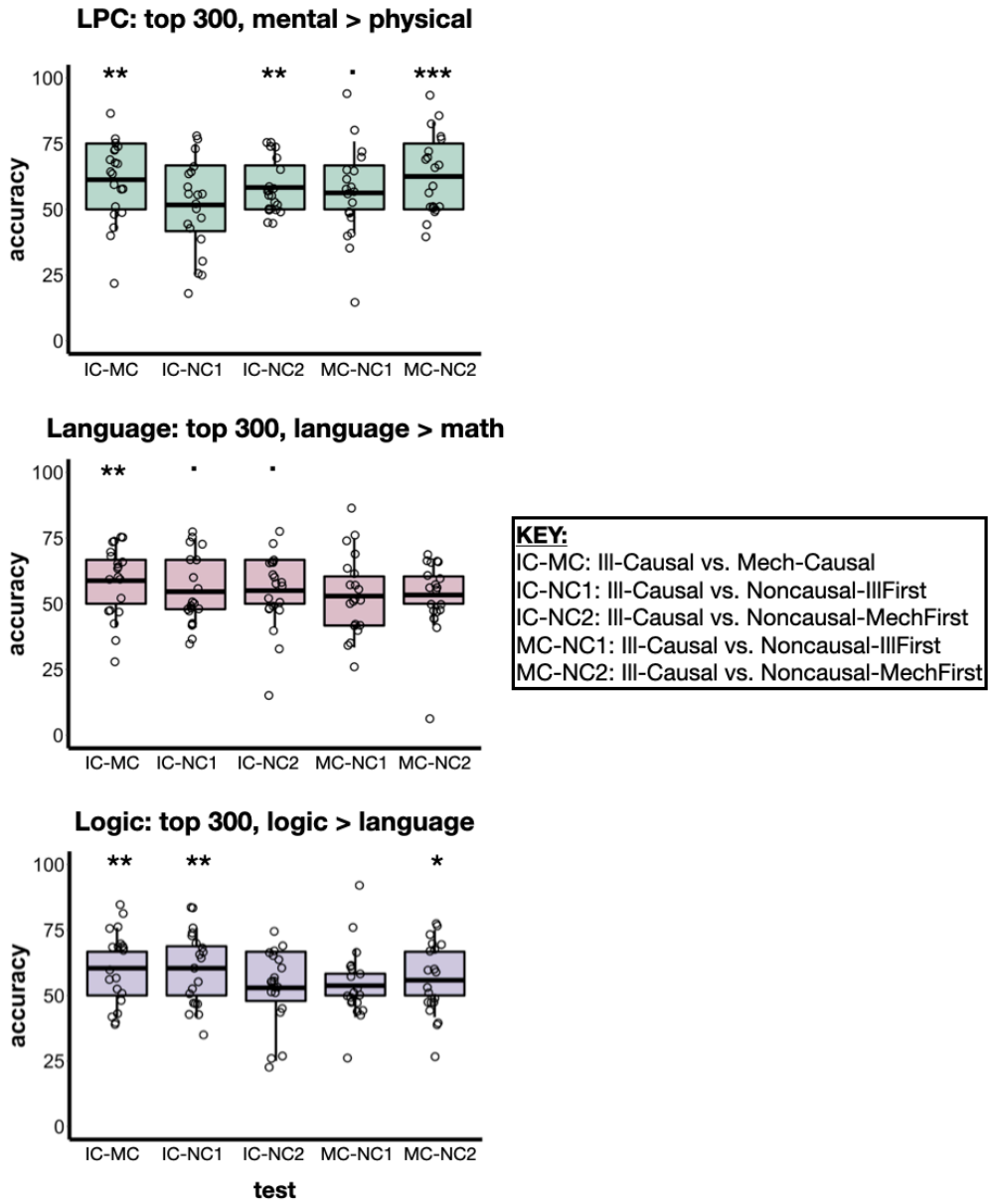
Supplementary Figure 8. Responses to illness inferences in the fusiform face area (FFA). Percent signal change (PSC) for each condition among the top 5% *Illness-Causal > Mechanical-Causal* vertices in left and right FFA search spaces (Julian et al., 2012) in individual participants, established via a leave-one-run-out analysis. Average PSC in the critical window (marked by dotted lines in Panel A) across participants is displayed via boxplot. The horizontal line within each boxplot indicates the overall mean.



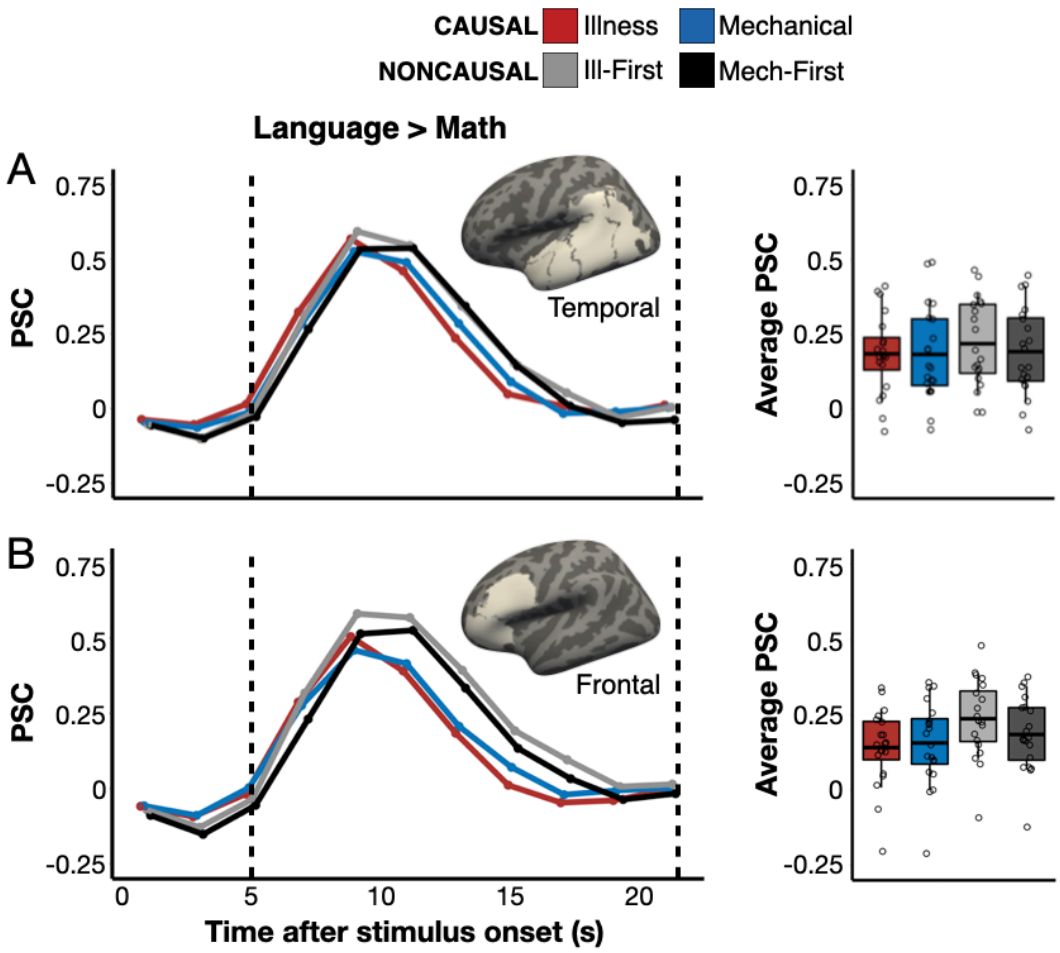
Supplementary Figure 9. Searchlight MVPA group maps. Whole-brain searchlight maps were thresholded using a vertex-wise threshold ($p < .001$ uncorrected) and a cluster size threshold (FWER $p < .05$, corrected for multiple comparisons across the entire cortical surface). Vertices are color coded on a scale from 55-65% decoding accuracy.



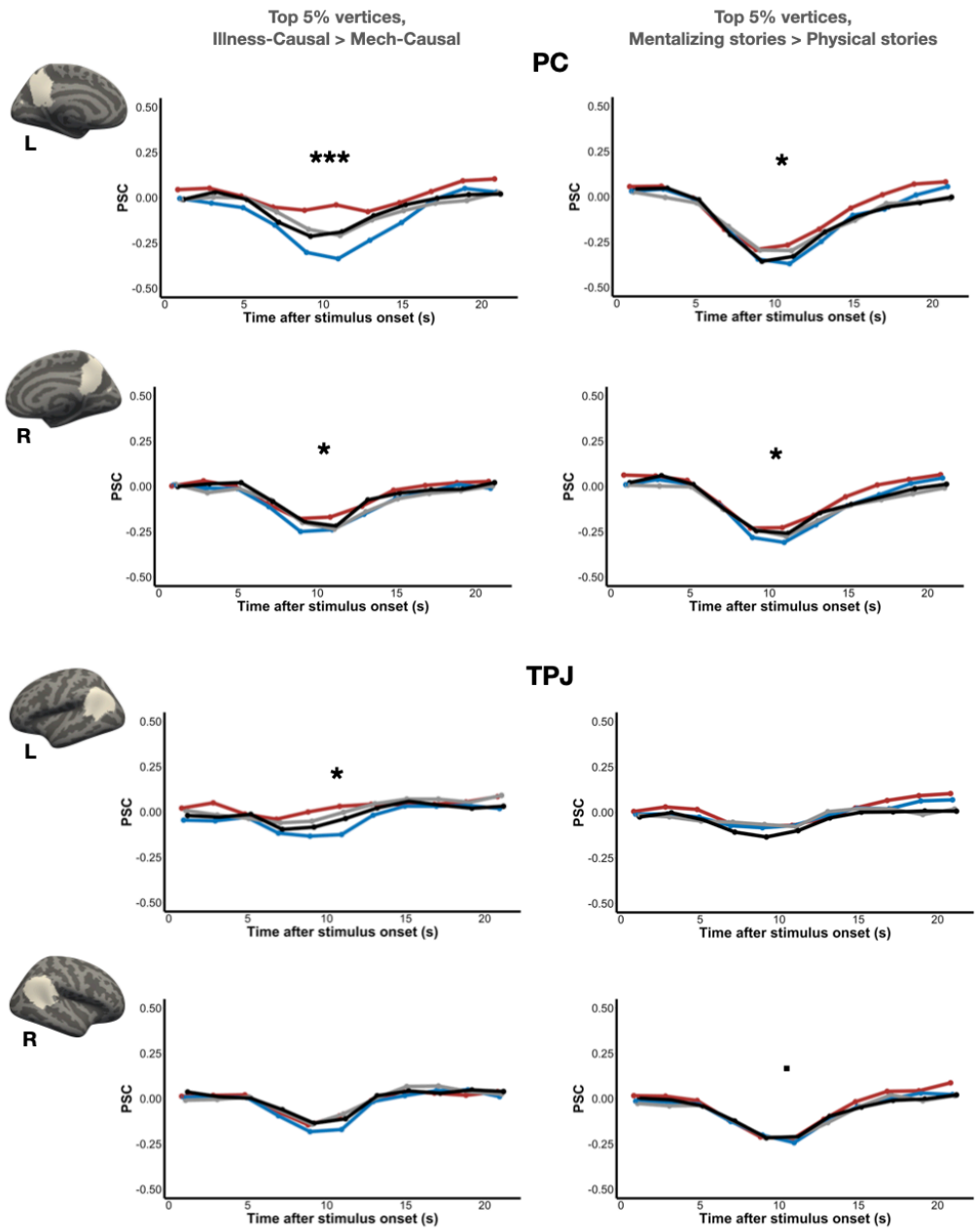
Supplementary Figure 10. Subject dispersion data for individual-subject MVPA performed in functional ROIs. We tested whether patterns of activity elicited during illness inferences vs. mechanical inferences could be decoded in each fROI: left and right PC, left and right TPJ, language network, logic network. In accordance with our preregistration, 2 types of fROIs were constructed using PC and TPJ search spaces: 1) top 300 most active vertices for mentalizing stories compared to physical stories in the mentalizing/animacy localizer (*Mentalizing stories* > *Physical stories*), and 2) top 300 most active vertices for both causal conditions compared to rest (*Illness-Causal* + *Mechanical-Causal* > *Rest*) Chance: 50%. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1. Full statistical results are included in Supplementary Table 2.



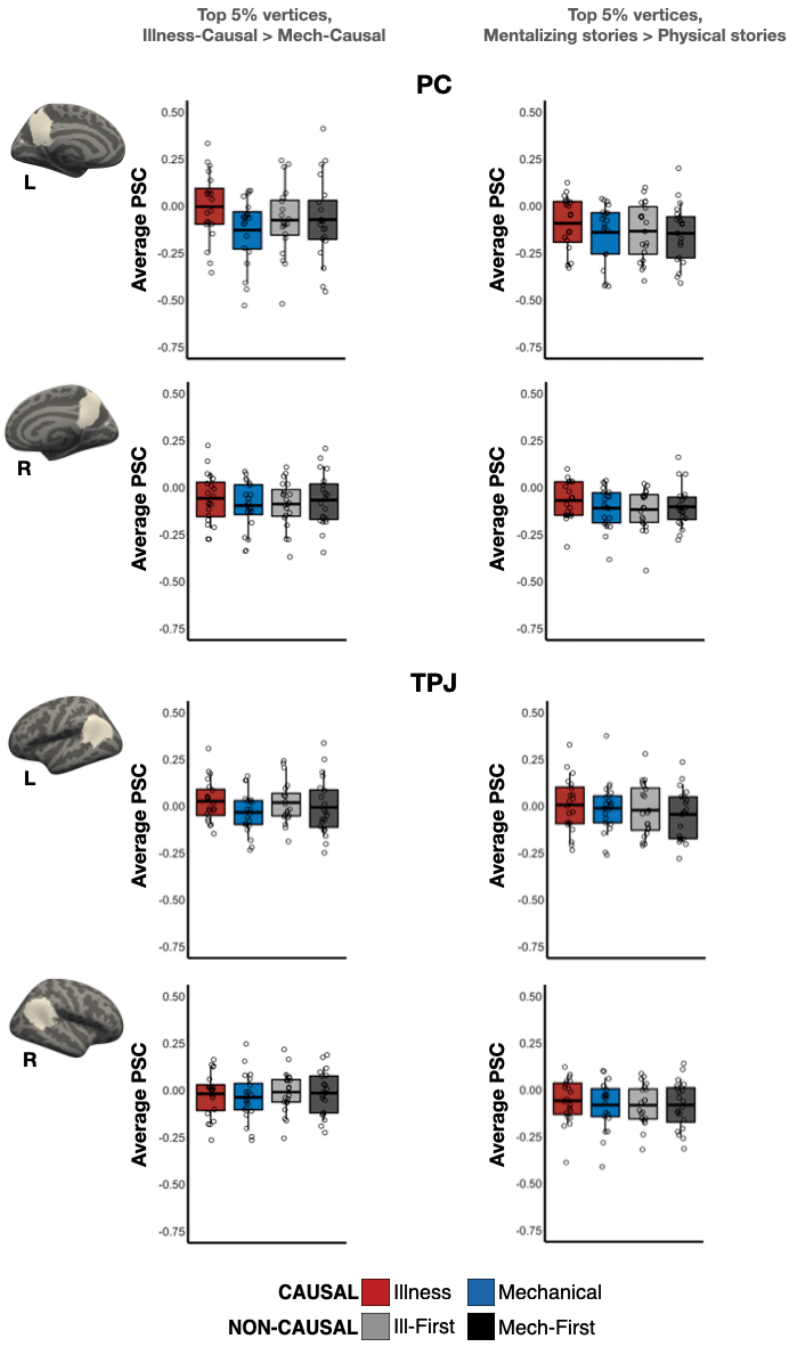
Supplementary Figure 11. Subject dispersion data for individual-subject MVPA performed in functional ROIs. 5 tests were performed in each fROI: left PC (LPC), language network, logic network. Chance: 50%. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1. Full statistical results are included in Supplementary Table 3.



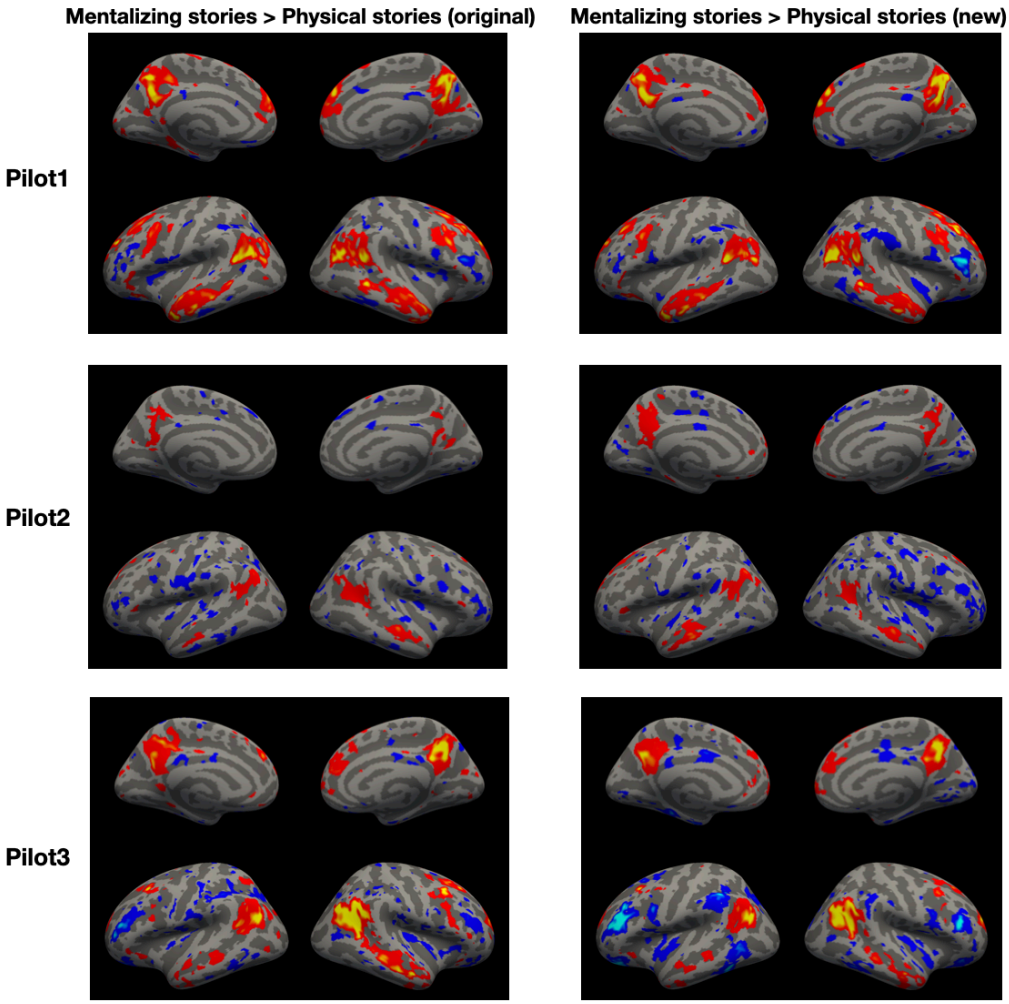
Supplementary Figure 12. Responses to causal inference in the language network. Panel A: Percent signal change (PSC) for each condition among the top 5% most language-responsive vertices (*language* > *math*) in a temporal language network search space (Fedorenko et al., 2010). Panel B: The same results in a frontal language search space (Fedorenko et al., 2010). Boxplots display average PSC in the critical window (marked by dotted lines) across participants. The horizontal line within each boxplot indicates the overall mean.



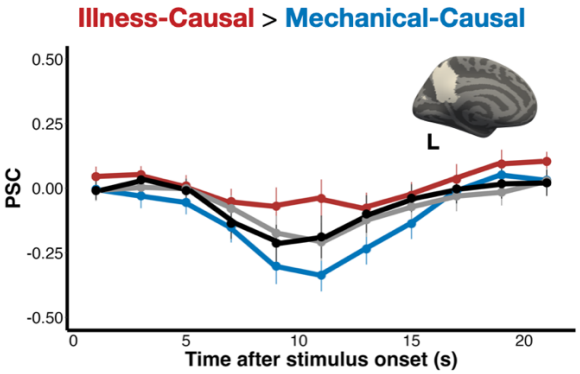
Supplementary Figure 13. Responses to illness inferences in bilateral PC and TPJ. Percent signal change (PSC) for each condition among the top 5% *Illness-Causal* > *Mechanical-Causal* vertices in bilateral PC and TPJ search spaces (Dufour et al., 2013) in individual participants, established via a leave-one-run-out analysis, is shown. We hypothesized that the PC and TPJ would exhibit a preference for illness inferences and report all data for completeness (see preregistration https://osf.io/6pnqg). Significance codes for *Illness-Causal* > *Mechanical-Causal* comparison (paired samples t-tests): 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1. Subject dispersion data are shown in Supplementary Figure 14.



Supplementary Figure 14. Subject dispersion data for responses to illness inferences in bilateral PC and TPJ (see Supplementary Figure 13). We hypothesized that the PC and TPJ would exhibit a preference for illness inferences and report all data for completeness (see preregistration https://osf.io/6pnqg). Boxplots display average PSC in the critical window (5-21 s) across participants. The horizontal line within each boxplot indicates the overall mean.



Supplementary Figure 15. 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., 2011) but contained new physical stories that included more vivid physical description and did not refer to animate agents. Individual-subject maps are shown at $p < .01$ uncorrected.



Supplementary Figure 16. Percent signal change (PSC) for each condition among the top 5% Illness-Causal > Mechanical-Causal vertices in a left PC search space (Dufour et al., 2013) in individual participants, established via a leave-one-run-out analysis. Error bars indicate the standard error of the mean for each condition at each timepoint.

Supplementary Materials

Supplementary Table 1: Illness types present in the stimulus set.

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 2: Results of preregistered MVPA for *Illness-Causal* vs. *Mechanical-Causal* in individual-subject functional ROIs. Each fROI was created by selecting the top 300 vertices for each contrast (see 'Contrast') in each search space. Accuracy refers to classifier performance against chance (50%) for *Illness-Causal* vs. *Mechanical-Causal*. Permuted and Bonferroni-corrected (across fROIs) 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). Visualizations of these results are displayed in Supplementary Figure 10.

Search space	**Contrast**	**Accuracy**	**t**	**Permuted p**	**Bonferroni adj. p**
LTPJ	ment_vs_phys	61.70%	2.96	0.0053	0.032
RTPJ	ment_vs_phys	52.50%	0.71	0.2573	1
LPC	ment_vs_phys	61.30%	3.44	0.003	0.0112
RPC	ment_vs_phys	60%	2.4	0.0176	0.108
LTPJ	caus_vs_rest	58.30%	4.16	0.0004	0.0024
RTPJ	caus_vs_rest	51.30%	0.39	0.3305	1
LPC	caus_vs_rest	60.80%	2.94	0.0055	0.0336
RPC	caus_vs_rest	53.70%	1.1	0.1516	1
Logic	logic_vs_lang	60.40%	3.46	0.0017	0.0026
Language	lang_vs_math	58.80%	2.76	0.0069	0.0124

Supplementary Table 3: MVPA results for all tests in select individual-subject functional ROIs. Each fROI was created by selecting the top 300 vertices for each contrast in each search space: left PC (LPC) = top *main experimental conditions* > *rest,* language = top *language* > *math* (language/logic localizer), logic = top *logic* > *language* (language/logic localizer). Accuracy refers to classifier performance against chance (50%) for each test. Permuted and Bonferroni-corrected (across fROIs) p-values are reported. Visualizations of these results are displayed in Supplementary Figure 11.

Search space	**Test**	**Accuracy**	**t**	**Permuted p**	**Bonferroni adj. p**
LPC	ill-causal vs. mech-causal	61.30%	3.44	0.0018	0.007
LPC	ill-causal vs. noncausal1	51.70%	0.43	0.3345	1
LPC	ill-causal vs. noncausal2	58.30%	3.45	0.0018	0.007
LPC	mech-causal vs. noncausal1	56.20%	1.66	0.0542	0.284
LPC	mech-causal vs. noncausal2	62.50%	3.81	0.0005	0.003
Language	ill-causal vs. mech-causal	58.80%	2.76	0.0097	0.093
Language	ill-causal vs. noncausal1	54.60%	1.5	0.083	1
Language	ill-causal vs. noncausal2	55%	1.61	0.0675	0.93
Language	mech-causal vs. noncausal1	52.90%	0.85	0.1957	1
Language	mech-causal vs. noncausal2	53.30%	1.09	0.1609	1
Logic	ill-causal vs. mech-causal	60.40%	3.46	0.0029	0.0195
Logic	ill-causal vs. noncausal1	60.40%	3.27	0.0029	0.03
Logic	ill-causal vs. noncausal2	52.90%	0.88	0.1928	1
Logic	mech-causal vs. noncausal1	53.80%	1.23	0.1121	1
Logic	mech-causal vs. noncausal2	55.80%	1.97	0.0425	0.4785

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 noncausal conditions, $t(25) = 36.97$, $p < .001$ (causal: $M = 3.51 \pm 0.78$ SD, noncausal: $M = 1.10 \pm 0.45$ SD). In addition, *Illness-Causal* and *Mechanical-Causal* items received equally high causality ratings, $t(25) = -0.64$, $p = .53$ (*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: Full behavioral results.

Accuracy on the magic detection task was at ceiling ($M = 97.9\% \pm 2.2$ SD). There were no significant differences across the 4 main experimental conditions (*Illness-Causal*, *Mechanical-Causal*, *Noncausal-Illness First, Noncausal-Mechanical First*), but participants were more

accurate on *Illness-Causal* trials compared to 'magical' catch trials ($F_{(4,76)} = 2.81$, $p = .03$; *Illness-Causal:* $M = 98.8\% \pm 2.2$ SD; 'magical' catch trials*:* $M = 96.4\% \pm 3.8$ SD).

A one-way repeated measures ANOVA evaluating response time revealed a main effect of condition, $F_{(4,76)} = 8.17$, $p < .001$, whereby participants were faster on *Illness-Causal* trials ($M = 4.73 \pm 0.81$ SD) compared to *Noncausal-Illness First* ($M = 5.33$ s ± 0.85 SD), *Noncausal-Mechanical First* ($M = 5.27$ s ± 0.89 SD) trials, and 'magical' catch trials ($M = 5.34$ s ± 0.89 SD). There were no differences in response time between *Mechanical-Causal* ($M = 5.15$ s ± 0.88 SD) and any other conditions.

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; $F_{(2,38)} = 44.28$, $p < .001$).

Accuracy on the mentalizing localizer task was not different across the mentalizing stories and physical stories conditions (mentalizing: $83.50\% \pm 15.7$ SD, physical: $90.50\% \pm 12.3$ 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; $F_{(1,19)} = 16.59$, $p < .001$).

Appendix 4: Individual-subject univariate fROI analysis in the fusiform face area (FFA).

In an exploratory analysis, we defined individual-subject fROIs in the fusiform face area (FFA). Illness inference fROIs were created in left and right FFA search spaces from a previous study on responses to images of faces in the ventral stream (Julian et al., 2012) using an iterated leave-one-run-out procedure. In each participant, we identified the most illness inference-responsive vertices in left and right FFA search spaces in 5 of the 6 runs (top 5% of vertices, *Illness-Causal* > *Mechanical-Causal*). We then extracted PSC for each condition compared to rest in the held-out run (*Illness-Causal*, *Mechanical-Causal*, *Noncausal-Illness First, Noncausal-Mechanical First*), averaging the results across all iterations.

In contrast to the PC, the FFA did not show a preference for illness inferences compared to mechanical inferences (leave-one-run-out individual-subject fROI analysis; repeated measures ANOVA, condition (*Illness-Causal*, *Mechanical-Causal*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 0.04$, $p = .84$, main effect of hemisphere, $F_{(1,19)} = 9.46$, $p = .006$, condition x hemisphere interaction, $F_{(1,19)} = 1.34$, $p = .26$; Supplementary Figure 8). Additionally, the FFA did not show a preference for illness inferences compared to noncausal vignettes, which contained illness-related language but were not causally connected (repeated measures ANOVA, condition (*Illness-Causal*, *Noncausal-Illness First*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 0.94$, $p = .34$, main effect of hemisphere, $F_{(1,19)} = 4.47$, $p = .05$, condition x hemisphere interaction, $F_{(1,19)} = 0.06$, $p = .82$; repeated measures ANOVA, condition (*Illness-Causal*, *Noncausal-Mechanical First*) x hemisphere (left, right): main effect of condition, $F_{(1,19)} = 0.07$, $p = .8$; main effect of hemisphere, $F_{(1,19)} = 7.59$, $p = .01$; condition x hemisphere interaction, $F_{(1,19)} = 2.72$, $p = .12$; Supplementary Figure 8). Thus, although the FFA exhibits a preference for images of animates (e.g., Kanwisher et al., 1997), the current evidence suggests

that this region is not sensitive to abstract causal knowledge about animacy-specific processes (i.e., illness).

References

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

Aglinskas, A., & Fairhall, S. L. (2023). Similar representation of names and faces in the network for person perception. *NeuroImage*, *274*, 120100. https://doi.org/10.1016/j.neuroimage.2023.120100

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, *21*(4), 547–569. https://doi.org/10.1017/S0140525X98001277

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

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

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

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, *7*(2), 84-91.

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

Bi, Y. (2021). Dual coding of knowledge in the human brain. *Trends in Cognitive Sciences*, *25*(10), 883-895.

Bi, Y., Wang, X., & Caramazza, A. (2016). Object Domain and Modality in the Ventral Visual Pathway. *Trends in Cognitive Sciences*, *20*(4), 282–290. https://doi.org/10.1016/j.tics.2016.02.002

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in cognitive sciences*, *15*(11), 527-536.

Black, J. B., & Bern, H. (1981). Causal coherence and memory for events in narratives. *Journal of Verbal Learning and Verbal Behavior*, *20*(3), 267–275. https://doi.org/10.1016/S0022-5371(81)90417-5

Blank, I. A., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows. *NeuroImage*, *219*, 116925. https://doi.org/10.1016/j.neuroimage.2020.116925

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

Callanan, M. A., & Oakes, L. M. (1992). Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive Development*, *7*(2), 213–233. https://doi.org/10.1016/0885-2014(92)90012-G

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

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

Carey, S. (1988). Conceptual differences between children and adults. *Mind and Language*, *3*(3),

167-181.
Carey, S. (2011). *The origin of concepts*. Oxford University Press.
Chang, C. H. C., Nastase, S. A., & Hasson, U. (2022). Information flow across the cortical timescale hierarchy during narrative construction. *Proceedings of the National Academy of Sciences*, *119*(51), e2209307119.
Chen, J., & Bornstein, A. M. (2024). The causal structure and computational value of narratives. *Trends in Cognitive Sciences*.
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
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*(2), 940–954. https://doi.org/10.1016/j.neuroimage.2007.11.044
Coley, J. D., Solomon, G. E. A., & Shafto, P. (2002). The development of folkbiology: A cognitive science perspective on children's understanding of the biological world. In P. Kahn & S. Kellert (Eds.), *Children and nature: Psychological, sociocultural and evolutionary investigations* (pp. 65-91). Cambridge, MA: MIT Press.
Connolly, A. C., Sha, L., Guntupalli, J. S., Oosterhof, N., Halchenko, Y. O., Nastase, S. A., Visconti di Oleggio Castello, M., Abdi, H., Jobst, B., Gobbini, M. I., & Haxby, J. V. (2016). How the human brain represents perceived dangerousness or "predacity" of animals. *Journal of neuroscience*, *36*(19), 5373-5384.
Cosmides, L., & Tooby, J. (2013). Unraveling the enigma of human intelligence: Evolutionary psychology and the multimodular mind. In *The evolution of intelligence* (pp. 145-198). Psychology Press.
Coutanche, M. N. (2013). Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us?. *Cognitive, Affective, & Behavioral Neuroscience*, *13*, 667-673.
Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395
Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., & Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *Neuroimage*, *97*, 271-283.
Davis, Z. J., & Rehder, B. (2020). A Process Model of Causal Reasoning. *Cognitive Science*, *44*(5), e12839. https://doi.org/10.1111/cogs.12839
Deen, B., & Freiwald, W. A. (2022). *Parallel systems for social and spatial reasoning within the cortical apex* (p. 2021.09.23.461550). bioRxiv. https://doi.org/10.1101/2021.09.23.461550
Dehaene-Lambertz, G., Monzalvo, K., & Dehaene, S. (2018). The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLoS biology*, *16*(3), e2004103.
DeJesus, J. M., Venkatesh, S., & Kinzler, K. D. (2021). Young children's ability to make predictions about novel illnesses. *Child Development*, *92*(5), e817-e831.
Devlin, J. T., Russell, R. P., Davis, M. H., Price, C. J., Moss, H. E., Fadili, M. J., & Tyler, L. K. (2002). Is there an anatomical basis for category-specificity? Semantic memory studies in PET and fMRI. *Neuropsychologia*, *40*(1), 54–75. https://doi.org/10.1016/S0028-3932(01)00066-5

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*(3), 1144–1179. https://doi.org/10.1152/jn.00529.2019

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage*, *55*(2), 705–712. https://doi.org/10.1016/j.neuroimage.2010.12.040

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*(1), 27–42. https://doi.org/10.1016/0749-596X(90)90008-N

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*(9), e75468. https://doi.org/10.1371/journal.pone.0075468

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*(7), 2017–2032. https://doi.org/10.1002/hbm.24350

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*(28), 7900–7905. https://doi.org/10.1073/pnas.1602413113

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Fairhall, S. L., & Caramazza, A. (2013a). Brain Regions That Represent Amodal Conceptual Knowledge. *Journal of Neuroscience*, *33*(25), 10552–10558. https://doi.org/10.1523/JNEUROSCI.0051-13.2013

Fairhall, S. L., & Caramazza, A. (2013b). Category-selective neural substrates for person- and place-related concepts. *Cortex*, *49*(10), 2748–2757. https://doi.org/10.1016/j.cortex.2013.05.010

Fairhall, S. L., Anzellotti, S., Ubaldi, S., & Caramazza, A. (2014). Person- and Place-Selective Neural Substrates for Entity-Specific Semantic Access. *Cerebral Cortex*, *24*(7), 1687–1696. https://doi.org/10.1093/cercor/bht039

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*(3–6), 401–408. https://doi.org/10.1080/02643290244000293

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*(1), 132–153. https://doi.org/10.1111/nyas.13046

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*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

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*(10), 2151–2163. https://doi.org/10.1162/jocn.2009.21387

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*(3), 325–340. https://doi.org/10.1016/S0926-6410(01)00007-6

Foster, G. M. (1976). Disease Etiologies in Non-Western Medical Systems. *American Anthropologist*, *78*(4), 773–782. https://doi.org/10.1525/aa.1976.78.4.02a00030

Fugelsang, J. A., & Dunbar, K. N. (2005). Brain-based mechanisms underlying complex causal thinking. *Neuropsychologia*, *43*(8), 1204–1213. https://doi.org/10.1016/j.neuropsychologia.2004.10.012

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*(3), 371–395. https://doi.org/10.1037/0033-295X.101.3.371

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition*, *38*(3), 213-244.

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

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.

Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face perception, not generic within-category identification. *Nature Neuroscience*, *7*(5), 555–562. https://doi.org/10.1038/nn1224

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

Goldvarg, E., & Johnson-Laird, P. n. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*(4), 565–610. https://doi.org/10.1207/s15516709cog2504_3

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

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*(1), 3–32. https://doi.org/10.1037/0033-295X.111.1.3

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*(5), 620–629. https://doi.org/10.1037/0012-1649.37.5.620

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, *138*(6), 1085.

Gutheil, G., Vera, A., & Keil, F. C. (1998). Do houseflies think? Patterns of induction and biological beliefs in development. *Cognition*, *66*(1), 33–49. https://doi.org/10.1016/S0010-0277(97)00049-8

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A Python Toolbox for Multivariate Pattern Analysis of fMRI Data. *Neuroinformatics*, *7*(1), 37–53. https://doi.org/10.1007/s12021-008-9041-y

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*(10), 2539–2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008

Hatano, G., & Inagaki, K. (1994). Young children's naive theory of biology. *Cognition*, *50*(1), 171–188. https://doi.org/10.1016/0010-0277(94)90027-2

Hauptman, M., Elli, G., Pant, R., & Bedny, M. (2025). Neural specialization for 'visual' concepts emerges in the absence of vision. *Cognition*, *257*, 106058.

Häusler, C. O., Eickhoff, S. B., & Hanke, M. (2022). Processing of visual and non-visual naturalistic spatial information in the. *Scientific Data*, *9*(1), 147. https://doi.org/10.1038/s41597-022-01250-4

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, *180*, 4-18.

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: Evidence from everyday conversation. *Developmental Psychology*, *37*(5), 668–683. https://doi.org/10.1037/0012-1649.37.5.668

Hillis, A., & Caramazza, A. (1991). Category-specific naming and comprehension impairment: A double dissociation. *Brain: A Journal of Neurology*, *114 (Pt 5)*, 2081–2094. https://doi.org/10.1093/brain/114.5.2081

Inagaki, K., & Hatano, G. (1993). Young Children's Understanding of the Mind-Body Distinction. *Child Development*, *64*(5), 1534–1549. https://doi.org/10.2307/1131551

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

Inagaki, K., & Hatano, G. (2006). Young Children's Conception of the Biological World. *Current Directions in Psychological Science*, *15*(4), 177–181. https://doi.org/10.1111/j.1467-8721.2006.00431.x

Jacoby, N., Bruneau, E., Koster-Hale, J., & Saxe, R. (2016). Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *Neuroimage*, *126*, 39-48.

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*(6), 780–796. https://doi.org/10.1080/23273798.2018.1525494

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *Neuroimage*, *60*(4), 2357-2364.

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*(1), 79.

Kalish, C. W. (1996). Preschoolers' understanding of germs as invisible mechanisms. *Cognitive Development*, *11*(1), 83–106. https://doi.org/10.1016/S0885-2014(96)90029-5

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*(40), 11172–11177. https://doi.org/10.1073/pnas.1524982113

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, *17*(11), 4302-4311.

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*(2), 115–126. https://doi.org/10.1016/S0022-5371(84)90082-3

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

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

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

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

Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, *33*(25), 10235–10242. https://doi.org/10.1523/JNEUROSCI.0983-13.2013

Kragel, P. A., Carter, R. M., & Huettel, S. A. (2012). What makes a pattern? Matching decoding methods to data in multivariate pattern analysis. *Frontiers in neuroscience*, *6*, 162.

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), 1–16. https://doi.org/10.1162/jocn_a_00124

Kuperberg, G. R., Lakshmanan, B. M., Caplan, D. N., & Holcomb, P. J. (2006). Making sense of discourse: An fMRI study of causal inferencing across sentences. *Neuroimage*, *33*(1), 343-361.

Lagnado, D. A., Waldmann, M. R., Hagmaye, Y., & Sloman, S. A. (2007). Beyond Covariation: Cues to Causal Structure. In A. Gopnik & L. Schulz (Eds.), *Causal Learning* (1st ed., pp. 154–172). Oxford University Press: New York. https://doi.org/10.1093/acprof:oso/9780195176803.003.0011

Lee, H., & Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature Communications*, *13*(1), 4235. https://doi.org/10.1038/s41467-022-31965-2

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*(4), 607–642. https://doi.org/10.1080/03640210802066766

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*(3), 779–793. https://doi.org/10.1111/j.1467-8624.2012.01743.x

Legare, C. H., Wellman, H. M., & Gelman, S. A. (2009). Evidence for an explanation advantage in naïve biological reasoning. *Cognitive Psychology*, *58*(2), 177–194. https://doi.org/10.1016/j.cogpsych.2008.06.002

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

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*(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Lightner, A. D., Heckelsmiller, C., & Hagen, E. H. (2021). Ethnoscientific expertise and knowledge specialisation in 55 traditional cultures. *Evolutionary Human Sciences*, *3*, e37. https://doi.org/10.1017/ehs.2021.31

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*, e59340. https://doi.org/10.7554/eLife.59340

Lynch, E., & Medin, D. (2006). Explanatory models of illness: A study of within-culture variation. *Cognitive Psychology*, *53*(4), 285–309. https://doi.org/10.1016/j.cogpsych.2006.02.001

Mahon, B. Z., Anzellotti, S., Schwarzbach, J., Zampini, M., & Caramazza, A. (2009). Category-Specific Organization in the Human Brain Does Not Require Visual Experience. *Neuron*, *63*(3), 397–405. https://doi.org/10.1016/j.neuron.2009.07.012

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

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: structure and processes. *Current opinion in neurobiology*, *11*(2), 194-201.

Mason, R. A., & Just, M. A. (2011). Differentiable cortical networks for inferences concerning people's intentions versus physical causality. *Human Brain Mapping*, *32*(2), 313–329. https://doi.org/10.1002/hbm.21021

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

Medin, D. L., & Atran, S. (2004). The Native Mind: Biological Categorization and Reasoning in Development and Across Cultures. *Psychological Review*, *111*(4), 960–983. https://doi.org/10.1037/0033-295X.111.4.960

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*(3), 197–207. https://doi.org/10.1016/j.cogdev.2010.02.001

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*(30), 12554–12559. https://doi.org/10.1073/pnas.0902422106

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

Myers, J. L., Shinjo, M., & Duffy, S. A. (1987). Degree of causal relatedness and memory. *Journal of Memory and Language*, *26*(4), 453–465. https://doi.org/10.1016/0749-596X(87)90101-X

Noppeney, U., Price, C. J., Penny, W. D., & Friston, K. J. (2006). Two Distinct Neural Mechanisms for Category-selective Responses. *Cerebral Cortex*, *16*(3), 437–445. https://doi.org/10.1093/cercor/bhi123

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

Notaro, P. C., Gelman, S. A., & Zimmerman, M. A. (2001). Children's Understanding of Psychogenic Bodily Reactions. *Child Development*, *72*(2), 444–459. https://doi.org/10.1111/1467-8624.00289

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

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

Pakravan, M., Abbaszadeh, M., & Ghazizadeh, A. (2022). Coordinated multivoxel coding beyond univariate effects is not likely to be observable in fMRI data. *NeuroImage*, *247*, 118825.

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527. https://doi.org/10.1073/pnas.1018711108

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

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*(35), 11072–11077. https://doi.org/10.1073/pnas.1504242112

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*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

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

Pramod, R. T., Chomik, J., Schulz, L., & Kanwisher, N. (2023). A region in human left prefrontal cortex selectively engaged in causal reasoning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

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

Rabini, G., Ubaldi, S., & Fairhall, S. (2021). Combining Concepts Across Categorical Domains: A Linking Role of the Precuneus. *Neurobiology of Language*, *2*(3), 354–371. https://doi.org/10.1162/nol_a_00039

Raman, L., & Gelman, S. A. (2005). Children's Understanding of the Transmission of Genetic Disorders and Contagious Illnesses. *Developmental Psychology*, *41*(1), 171–182. https://doi.org/10.1037/0012-1649.41.1.171

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*(2), 255–274.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*(3), 264–314. https://doi.org/10.1016/j.cogpsych.2004.09.002

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., & Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature communications*, *9*(1), 1027.

Ritchey, M., & Cooper, R. A. (2020). Deconstructing the Posterior Medial Episodic Network. *Trends in Cognitive Sciences*, *24*(6), 451–465. https://doi.org/10.1016/j.tics.2020.03.006

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*(6), 1302–1320. https://doi.org/10.1111/j.1467-8624.1991.tb01607.x

Rottman, B. M., & Hastie, R. (2014). Reasoning about Causal Relationships: Inferences on Causal Networks. *Psychological Bulletin*, *140*(1), 109–139. https://doi.org/10.1037/a0031903

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

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*(5), 1233–1238. https://doi.org/10.1111/j.1460-9568.2005.04292.x

Saxe, R., & Carey, S. (2006). The perception of causality in infancy. *Acta Psychologica*, *123*(1), 144–165. https://doi.org/10.1016/j.actpsy.2006.05.005

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking peopleThe role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

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*(3), 229–234. https://doi.org/10.1093/scan/nsl034

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological science*, *17*(8), 692-699.

Schmidt, H., Norman, G., & Boshuizen, H. (1990). A Cognitive Perspective on Medical Expertise: Theory and Implications. *Academic Medicine*, *65*(10), 611–621.

Schreiber, K., & Krekelberg, B. (2013). The Statistical Analysis of Multi-Voxel Patterns in Functional Imaging. *PLoS ONE*, *8*(7), e69328. https://doi.org/10.1371/journal.pone.0069328

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162–176. https://doi.org/10.1037/0012-1649.40.2.162

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*, 107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307

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*(10), 6299–6319. https://doi.org/10.1093/cercor/bhac505

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*, e47391. https://doi.org/10.7554/eLife.47391

Simons, D. J., & Keil, F. C. (1995). An abstract to concrete shift in the development of biological thought: The insides story. *Cognition*, *56*(2), 129–163. https://doi.org/10.1016/0010-0277(94)00660-D

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, *7*(1), 12141.

Singer, M. (1994). Discourse inference processes. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 479–515). Academic Press.

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

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (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
Solstad, T., & Bott, O. (2017). Causality and causal reasoning in natural language. In M. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning*. Oxford University Press.
Spelke, E. S. (2003). What makes us smart? Core knowledge and natural language. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind*. Cambridge, MA: MIT Press.
Spelke, E. S. (2022). *What Babies Know: Core Knowledge and Composition Volume 1*. New York, NY: Oxford University Press.
Springer, K., & Keil, F. C. (1991). Early Differentiation of Causal Mechanisms Appropriate to Biological and Nonbiological Kinds. *Child Development*, *62*(4), 767–781. https://doi.org/10.2307/1131176
Springer, K., & Ruckel, J. (1992). Early beliefs about the cause of illness: Evidence against immanent justice. *Cognitive Development*, *7*(4), 429–443. https://doi.org/10.1016/0885-2014(92)80002-W
Steel, A., Billings, M. M., Silson, E. H., & Robertson, C. E. (2021). A network linking scene perception and spatial memory systems in posterior cerebral cortex. *Nature communications*, *12*(1), 2632.
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
Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489. https://doi.org/10.1207/s15516709cog2703_6
Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive Theories as Grammars for Causal Inference. In A. Gopnik & L. Schulz (Eds.), *Causal Learning* (1st ed., pp. 301–322). Oxford University PressNew York. https://doi.org/10.1093/acprof:oso/9780195176803.003.0020
Thompson-Schill, S. L. (2003). Neuroimaging studies of semantic memory: inferring "how" from "where". *Neuropsychologia*, *41*(3), 280-292.
Tooby, J., & DeVore, I. (1987). The reconstruction of hominid behavioral evolution through strategic modeling. In W. G. Kinzey (Ed.), *The evolution of human behavior: Primate models*. Albany, NY: SUNY Press.
Trabasso, T., & Sperry, L. L. (1985). Causal relatedness and importance of story events. *Journal of Memory and Language*, *24*(5), 595–611. https://doi.org/10.1016/0749-596X(85)90048-8
Varley, R. (2014). Reason without much language. *Language Sciences*, *46*, 232–244. https://doi.org/10.1016/j.langsci.2014.06.012
Varley, R., & Siegal, M. (2000). Evidence for cognition without grammar from causal reasoning and 'theory of mind' in an agrammatic aphasic patient. *Current Biology*, *10*(12), 723-726.
Vul, E. & Kanwisher, N. (2011). Begging the question: The non-independence error in fMRI data analysis. In S. J. Hanson, M. Bunzl (Eds.), *Foundational issues in human brain mapping*. MIT Press.
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
Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal

models: asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*(2), 222.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107*(3), 829-853.

Weiner, K. S., Barnett, M. A., Witthoft, N., Golarai, G., Stigliani, A., Kay, K. N., Gomez, J., Natu, V. S., Amunts, K., Zilles, K., & Grill-Spector, K. (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

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

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & van den Bosch, A. (2016). Prediction During Natural Language Comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

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

Woolgar, A., Golland, P., & Bode, S. (2014). Coping with confounds in multivoxel pattern analysis: What should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. *Neuroimage*, *98*, 506-512.

Yeshurun, Y., Nguyen, M., & Hasson, U. (2017). Amplification of local changes along the timescale processing hierarchy. *Proceedings of the National Academy of Sciences*, *114*(35), 9475-9480.