Causal inferencing relies on domain-specific systems: Evidence from illness causality

Miriam Hauptman (mhauptm1@jhu.edu)
Marina Bedny (marina.bedny@jhu.edu)
Department of Psychological and Brain Sciences, Johns Hopkins University
Baltimore, MD, USA

Abstract

Our remarkable ability to infer complex cause-effect relationships is thought to distinguish humans from all other species. Despite that causal inferencing pervades human cognition, it remains unclear whether this fundamental cognitive ability is supported by a unified, domain-general mechanism or multiple domain-specific mechanisms. Both the language and logical reasoning systems have been described as possible unified substrates of causal inferencing. The current study uses neuroimaging to offer insight into this debate. We specifically focus on the culturally universal and highly motivationally relevant case of inferring illness causes. Participants read causal and noncausal vignettes about illness and mechanical failure while undergoing fMRI. We find that inferring the causes of illness selectively activates the brain's 'animacy network,' particularly the precuneus. By contrast, a domain-general (i.e., 'content-invariant') preference for causal inferencing did not emerge, including in the language and logical reasoning networks. Together, this evidence suggests that domain-specific mechanisms enable causal inferencing.

Keywords: cognitive neuroscience; causal reasoning; language understanding; concepts and categories; fMRI

Introduction

A distinguishing feature of human cognition is our ability to reason about complex cause-effect relationships, particularly when causes are hidden (Tooby & DeVore, 1987, Lagnado et al., 2007; Rottman, Ahn, & Luhmann, 2011; Muentener & Schulz, 2014; Sloman & Lagnado, 2015; Goddu & Gopnik, 2024). Reasoning about the causes of illness provides a classic example (Keil et al., 1999; Waldmann, 2000; Meder & Mayrhofer, 2017; Legare & Shtulman, 2018). When reading something like, *Lucy attended a busy conference last week. Now she has COVID*, we naturally infer a causal relationship between crowded spaces and the (invisible) transmission of infectious disease. What cognitive and neural mechanisms support this kind of inferencing? A key point of debate concerns whether causal reasoning is carried out by a single, unified mechanism that operates regardless of domain (e.g., biology, physics, social cognition) or whether it is embedded separately within different domain-specific cognitive systems (Boyer, 1995; Gopnik et al., 2004; Tenenbaum, Griffiths, & Niyogi, 2007; Carey, 2011; Bender, Beller, & Medin, 2017).

The most extreme interpretation of the domain-general account reduces causal reasoning to general learning processes responsible for tracking covariation in events (e.g., Hume, 1739/1978; Kelley, 1973; Cheng & Novick, 1992; cf. Ahn et al., 1995; Lagnado et al., 2007). Such proposals view causal reasoning as a purely domain-general process.

A more recent set of theories posit the existence of a dedicated *cause* representation that is agnostic to semantic content, but in some cases interacts extensively with domain-specific knowledge (e.g., Gopnik et al., 2004; Tenenbaum et al., 2007; Carey, 2011). These 'dedicated causal engine' accounts are inspired by two main lines of evidence. First, causal learning in children and adults can be captured by content-invariant computational models, such as Bayes nets (Pearl, 2000; Gopnik et al., 2001; Schulz & Gopnik, 2004; Rehder & Burnett, 2005). Second, causal inferences often transcend domain-specific systems, such as when we infer that an agent was the cause of a flying inanimate object (Saxe, Tenenbaum, & Carey, 2005; Saxe, Tzelnic, & Carey, 2007; Muentener & Carey, 2010; Carey, 2011; see also Legare & Shtulman, 2018). These findings have inspired the related claims that abstract 'causal maps' guide causal learning across domains (Gopnik et al., 2004; Gopnik & Wellman, 2012), and that a unified representation of causality is situated within a 'central workspace,' where it interfaces with perceptual input (e.g., Michotte, 1963) and domain-specific conceptual knowledge (Saxe & Carey, 2006; Carey, 2011).

An alternative, but not mutually exclusive, possibility is that causal representations are built into domain-specific semantic systems dedicated to processing specific content (e.g., intuitive physics, biology, psychology) (Boyer, 1995; Gelman, 1990; Wellman & Gelman, 1992; Gerstenberg & Tenenbaum, 2017). Even young children have distinct 'intuitive theories' that express specific causal relationships (Callanan & Oakes, 1992; Wellman & Gelman, 1992; Gopnik & Wellman, 1992; Schult & Wellman, 1997; Keil, 2003). For instance, infants view human actions as driven by goals and desires, whereas they attribute the movement of inanimate objects to physical properties such as continuity and gravity (Woodward, 1998; Onishi & Baillargeon, 2005; Saxe et al., 2005; Baillargeon, 1995; Spelke et al., 1994). It has been proposed that domain-specific intuitive theories provide 'grammars of causal inference' that specify abstract causal laws (e.g., illness causes symptoms, but symptoms don't cause illness; Tenenbaum et al., 2007; Gerstenberg & Tenenbaum, 2017). Thus, causal representations may be embedded within domain-specific systems.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

The current study: inferring illness causality

The current study uses neuroimaging (fMRI) to offer insight into the 'one vs. many mechanisms' problem in causal reasoning. Cognitive neuroscience is ideally suited to test this question because each potentially relevant mechanism is associated with a functionally distinct brain network. Here, we investigate the automatic causal inferences that people make during language comprehension, with a specific focus on the universal yet culturally variable phenomenon of inferring the causes of illness (Ackerknecht, 1982; Foster, 1976; Legare & Gelman, 2008; Lock & Nguyen, 2010).

Illness inferencing offers a key, ecologically relevant test case for understanding the contribution of domain-specific vs. domain-general mechanisms to causal inference. Illness is itself a biological process and could depend on a domain-specific system, namely 'intuitive biology' (Inagaki & Hatano, 2006; Atran, 1998; Medin & Atran, 1999; Keil, 1992; Keil et al., 1999). Intuitive biology is thought to encompass knowledge about the structure and behavior of biological entities (e.g., animals, plants), which share certain core properties such as reproduction, growth, and heterogeneous internal structure (Keil, 1992). Evidence for the domain-specificity of the biological system comes from studies showing that infants and young children have distinct intuitions about the properties of living vs. nonliving things, e.g., living things have different insides and need energy to grow (Gelman, 1988; Simons & Keil, 1995; Setoh et al., 2013; Inagaki & Hatano, 2004). Young children additionally distinguish biological from psychological processes, for instance favoring biological over psychological/moral explanations for illness (Springer & Ruckel, 1992; Notaro, Gelman, & Zimmerman, 2001; Raman & Winer, 2004; Legare & Gelman, 2008).

Although illness exclusively affects biological entities, reasoning about illness causality can draw upon multiple semantic domains, including physical, social, and in some cases even mentalistic knowledge (e.g., smoking causes cancer; interacting with sick people causes a cold; evil eye of a neighbor causes fever; Legare et al., 2012; Legare & Shtulman, 2018). Some explanations of illness reference multiple domains simultaneously (e.g., negative emotions cause blocked arteries, which cause heart attack; Lynch & Medin, 2006; Legare & Gelman, 2008). Variability in the semantic domains referenced in illness explanations, both within and across cultures (e.g., Foster, 1976) suggests that a domain-general causal mechanism may be needed to pool together such disparate knowledge during causal inferencing.

Recent neuroimaging evidence has identified a potential neural substrate of biological knowledge in the precuneus (PC) (Fairhall & Caramazza, 2013a, 2013b; Fairhall et al., 2014; Peer et al., 2015; Deen & Freiwald, 2022; Hauptman, Elli, et al., 2023). We hypothesized that if inferring the causes of illness depends on a domain-specific system, it would selectively activate the animacy-preferring PC (see pre-registration). The PC was originally identified as part of the mentalizing network (Saxe & Kanwisher, 2003; Saxe et al., 2006). However, unlike some other parts of this network, the PC is not selective for mental state content (Saxe & Kanwisher, 2003; Saxe & Wexler, 2005), instead activating when people reason about humans and animals in ways that do not require mentalizing (e.g., judging semantic category membership; Fairhall & Caramazza, 2013b). These findings suggest that the PC encodes biological knowledge.

Different proposals have been offered regarding what cognitive systems might support a domain-general causal mechanism. One possibility is that causal inferencing is supported by general logical reasoning abilities (Khemlani, Barbey, & Johnson-Laird, 2014), which enable a variety of non-causal inferences (e.g., disjunctive syllogism: *P or Q, not P, therefore Q*; Lea, 1995; Mody & Carey, 2016; Cesana-Arlotti, Kovács, & Téglás, 2020). Prior work on logical fallacies made during causal inferencing tasks has motivated its portrayal as a form of both deductive and inductive logical reasoning (Goldvarg & Johnson-Laird, 2001; Johnson-Laird & Khemlani, 2017; see Waldmann & Hagmayer, 2013 for a review). Logical reasoning selectively activates a frontoparietal network (Reverberi et al., 2007; Monti, Parsons, & Osherson, 2009; Monti & Osherson, 2012), inspiring the claim that logic-responsive frontal cortex is the seat of causal thinking (Khemlani et al., 2014; Operskalski & Barbey, 2017).

Language is another candidate system that could support causal inferencing (Kuperberg et al., 2006; Mason & Just, 2011; Prat et al., 2011). It has been suggested that language enables humans to combine information from otherwise encapsulated semantic domains (e.g., intuitive physics, social cognition; Spelke, 2003; 2022), a process that appears to underlie many instances of causal inference (Legare & Shtulman, 2018). Natural languages are efficient transmitters of causal information (Pinker, 2003; Tooby & DeVore, 1987; Solstad & Bott, 2017), especially culturally accumulated knowledge about imperceptible causes, such as in the case of illness (Harris & Koenig, 2006; Legare & Gelman, 2008; Legare et al., 2012). Importantly, language processing selectively activates a left-lateralized frontotemporal brain network (e.g., Fedorenko et al., 2010) that dissociates from the frontoparietal logic network (e.g., Monti et al., 2009), allowing us to assess the unique contributions of each system.

To investigate the neurocognitive mechanisms underlying illness inferencing, we showed participants two-sentence vignettes about human agents that either elicited causal inferences about illness (*Illness-Causal*), elicited causal inferences about a non-illness domain (i.e., mechanical failure; *Mechanical-Causal*), or contained illness-related language but were not causally connected (*Noncausal*). The same participants also performed a language and logic localizer task. If causal representations are built into domain-specific knowledge systems, inferring illness causality but not mechanical causality should activate PC. Alternatively, if a domain-general causal mechanism enables causal inferencing, we should observe a preference for both *Causal*

conditions compared to *Noncausal* stimuli in the language and logical reasoning networks. It is additionally possible that there is a distinct neural circuit dedicated to causal inferencing in frontal cortex, which is important for high-level reasoning (e.g., Collins & Koechlin, 2012; Donoso, Collins, & Koechlin, 2014). We test this last possibility using whole-cortex analysis.

Method

Open science practices

Our methods and analytical procedures were pre-registered prior to data collection (https://osf.io/cx9n2/).

Participants

Twenty adults (7 females, 13 males, 25-37 years old, $M = 29$ years ± 3 SD, all with or pursuing graduate degrees) participated in the study. Two additional participants were excluded from the final dataset due to head motion (>2 mm) and an image artifact. All participants were screened for cognitive/neurological disabilities (self-report). Participants gave written informed consent and were compensated $30 per hour. The study was approved by the Johns Hopkins Medicine Institutional Review Boards.

Causal inferencing experiment

Stimuli Participants read two-sentence vignettes in 4 conditions, 2 *Causal* and 2 *Noncausal* (Figure 1C). Each vignette focused on a human agent. The first sentence described something the agent did or experienced and served as the potential cause. The second sentence described the potential outcome (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 transmission of pathogens, exposure to environmental toxins, and genetic mutations. *Mechanical-Causal* vignettes elicited inferences about physical causes of mechanical damage to inanimate objects (e.g., houses, jewelry). 2 *Noncausal* conditions used the same sentences as in the *Illness-Causal* and *Mech-Causal* conditions but in a shuffled order: illness cause with mechanical outcome (*Noncausal-Illness First*) or mechanical cause with illness outcome (*Noncausal-Mech First*). Explicit causality judgments collected from a separate group of online participants (n=26) verified that both *Causal* conditions were more causal than *Noncausal* conditions, $t(25) = 36.97$, $p < .0001$. In addition, *Illness-Causal* and *Mech-Causal* items were equally causal, $t(25) = -0.64$, $p = 0.53$.

Illness-Causal and *Mech-Causal* vignettes were constructed in pairs, such that each member of a given pair shared parallel or near-parallel phrase structure. All 4 conditions were also matched (pairwise t-tests, all ps > 0.3, no statistical correction) on linguistic variables known to modulate activity in language regions (e.g., Pallier, Devauchelle, & Dehaene, 2011; Shain, Blank, et al., 2020): number of characters, number of words, average number of characters per word, average word frequency, average bigram surprisal (https://books.google.com/ngrams/), and average syntactic dependency length (Stanford Parser; de Marneffe, MacCartney, & Manning, 2006).

Procedure We used a magic detection task to encourage participants to process the meaning of the stimuli without making explicit causality judgments. Participants saw 'magical' catch trials that closely resembled the experimental trials but were fantastical (e.g., drinking lava). On each trial, participants indicated via button press whether 'something magical' occurred in the vignette (Yes/No). Both sentences in a vignette were presented simultaneously, one above the other (7 s), followed by an inter-trial interval (12 s). Each participant saw 38 trials per condition plus 36 'magical' catch trials (188 total trials) in one of two versions, counterbalanced across participants, such that individual participants did not see the same sentence in both *Causal* and *Noncausal* conditions. The experiment was divided into 6 10-minute runs containing a similar number of trials per condition per run presented in a pseudorandom order.

Language/logic localizer experiment

A localizer task was used to identify the language and logic networks in each participant. The task had three conditions: language, formal logic, and math. Participants judged whether two visually presented sentences, one in active and one in passive voice, shared the same meaning (language), whether two logical statements were consistent (logic; e.g., *If either not Z or not Y then X* vs. *If not X then both Z and Y*), or whether a variable had the same value across two equations (math; for details see Liu et al., 2020). Each trial lasted 20 s. Following prior studies, the language network was identified by contrasting *language* > *math* and the logic network by contrasting *logic* > *language* (Liu et al., 2020; Kanjlia et al., 2016; Monti et al., 2009). The use of functional localizers represents an improvement upon past fMRI studies of causal inferencing, which rely on anatomical landmarks to make inferences about relevant cognitive processes.

fMRI methods

Acquisition and preprocessing Whole-brain fMRI data was acquired on a 3T Phillips Achieva Multix X-Series scanner at F.M. Kirby Research Center. T1-weighted structural images were collected in 150 axial slices with 1 mm isotropic voxels using the magnetization-prepared rapid gradient-echo (MP-RAGE) sequence. T2*-weighted functional BOLD scans were collected in 36 axial slices (2.4 2.43 mm voxels, TR=2 s). 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.

Whole-cortex analysis For the main causal inferencing experiment, the GLM modeled the four main conditions (*Illness-Causal*, *Mech-Causal*, *Noncausal-Illness First, Noncausal-Mech First*) and the magic catch trials during the 7 s display of the vignettes after convolving with a canonical hemodynamic response function and its first temporal derivative. For the language/logic localizer experiment, a separate predictor was included for each condition (language, logic, math) during the 20 s display of the stimuli.

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

Individual-subject ROI analysis (univariate) We defined individual-subject functional ROIs in the animacy (PC), language (frontal and temporal) and logic (frontoparietal) networks. Domain-specific animacy ROIs were created in a left PC mask previously shown to respond to social information (Dufour et al., 2013). We used an iterated leave-one-run-out procedure, which allowed us to perform sensitive individual-subjects analysis while avoiding circular analysis (Vul & Kanwisher, 2011). In each participant, we identified the most illness inferencing-responsive voxels in the PC mask (top 5% of voxels, *Illness-Causal* > *Mech-Causal*) in 5 of the 6 runs and extracted percent signal change (PSC) for each condition compared to rest in the held-out run (*Illness-Causal*, *Mech-Causal*, *Noncausal-Illness First, Noncausal-Mech First*). For all analyses, PSC was extracted and averaged over the entire duration of the trial (17 s total), allowing 4 s to account for the hemodynamic lag.

Language ROIs were identified by taking the most language-responsive voxels (top 5%) in left frontal and temporal language areas (group search space: Fedorenko et al., 2010) using the *language* > *math* contrast. A logic-responsive ROI was identified by taking the most logic-responsive voxels (top 5%) in the left frontoparietal network (group search space: Liu et al., 2020) using the *logic* > *language* contrast. We then extracted the PSC for each of the four main conditions.

MVPA We performed MVPA (Hanke et al., 2009) to test whether patterns of activity in the PC distinguished illness inferencing from mechanical inferencing. In each participant, we identified the top 300 voxels most responsive to causal inferencing across domains (i.e., both *Illness-Causal* + *Mech-Causal* > Rest) in a left PC mask (Dufour et al., 2013). For each voxel in each participant's causal PC ROI, we obtained one observation per condition per run (z-scored beta parameter estimate of the GLM). A linear support vector machine was then trained and tested on the data using cross-validation. We compared classifier performance to chance (50%, one-tailed test) using an empirical null distribution generated via a permutation and bootstrap approach (Schreiber & Krekelberg, 2013; Stelzer et al., 2013).

Results

Behavioral results

Accuracy on the magic detection task was at ceiling (M = 97.9% ± 2.2 SD) and there were no significant differences across the 4 main experimental conditions (*Illness-Causal*, *Mech-Causal*, *Noncausal-Illness First, Noncausal-Mech First*), $F(3,57) = 2.39$, $p = .08$. A one-way repeated measures ANOVA evaluating response time revealed a main effect of condition, $F(3,57) = 32.63$, $p < .0001$, whereby participants were faster on *Illness-Causal* trials (M = 4.73 ± 0.81 SD) compared to *Noncausal-Illness First* (M = 5.33 s ± 0.85 SD) and *Noncausal-Mech First* (M = 5.27 s ± 0.89 SD). There were no differences in response time between *Mech-Causal* (M = 5.15 s ± 0.88 SD) and any other conditions. Accuracy on the localizer task was above chance for all conditions and all participants and was highest in the language task, followed by math and logic (language: M = 98.1% ± 5.8 SD, math: M = 93.8% ± 6.4 SD, logic: M = 67.5% ± 14.0 SD).

Domain-specific responses to illness inferencing in precuneus (PC)

We find that animacy-responsive PC responds selectively to causal inferences about illness. Inferring illness causes (*Illness-Causal*) activated the PC more than inferring physical causes of mechanical failure (*Mech-Causal*) (one-way repeated measures ANOVA, $F(1,19) = 28.69$, $p < .0001$). Illness inferencing additionally activated the PC more than illness-related language that was not causally connected (both *Noncausal* conditions) (one-way repeated measures ANOVA, $F(1,19) = 13.23$, $p = .002$) (Figure 1A). MVPA likewise revealed that responses to *Illness-Causal* and *Mech-Causal* vignettes produced spatially distinguishable neural patterns in left PC, $t(19) = 3.50$, $p < .001$.

In whole-cortex analysis, the PC was the only cortical region to show a preference for causal inferencing about illness. Illness inferencing activated the PC more than causal inferencing about a non-illness domain (*Mech-Causal* condition) and more than both *Noncausal* conditions ($p < .05$, corrected for multiple comparisons) (Figure 1B). Responses to illness inferencing in PC overlap with previously reported responses to people- and animal-related concepts (e.g., Fairhall & Caramazza, 2013a; Hauptman, Elli, et al., 2023).

No evidence for domain-general responses to causal inferencing in language or logic networks

We failed to find increased activity for causal inferencing in either the language or logical reasoning networks.

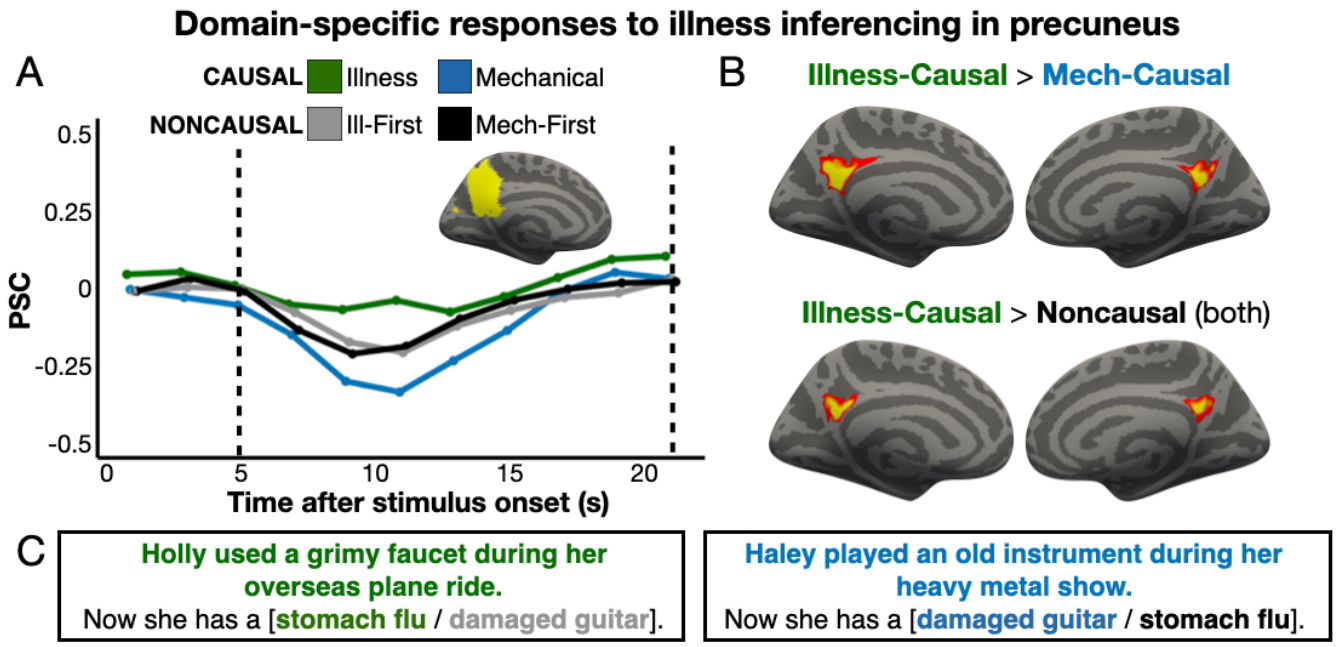

Figure 1: Domain-specific responses to illness inferencing in the precuneus (PC). Panel A: Percent signal change (PSC) for each condition among the top *Illness-Causal* > *Mech-Causal* voxels in a left PC mask (Dufour et al., 2013), established via an individual-subjects leave-one-run-out analysis. Panel B: Whole-cortex results for *Illness-Causal* > *Mech-Causal* and *Illness-Causal* > *Noncausal* (both versions) contrasts, corrected for multiple comparisons ($p < .05$ FWER, cluster-forming threshold $p < .01$ uncorrected). 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.

If anything, the opposite pattern emerged: individual-subject ROI analysis revealed higher activity for both *Noncausal* conditions (*Noncausal-Illness First* + *Noncausal-Mech First*) compared to both *Causal* conditions (*Illness-Causal* + *Mech-Causal*) in both the frontal and temporal language network (temporal: one-way repeated measures ANOVA, $F(1,19) = 4.85$, $p = .04$) (Figure 2A). The same effect was marginally significant in the logic network, $F(1,19) = 3.88$, $p = .06$ (Figure 2B). These effects likely the reflect the greater difficulty associated with integrating unrelated sentences.

In whole-cortex analysis, we similarly did not observe any regions that were active in both the *Illness-Causal* > *Noncausal* and *Mech-Causal* > *Noncausal* contrasts.

Discussion

Using illness inferencing as a case study, we find that the domain-specific, animacy-preferring precuneus (PC) supports causal inferences about illness. Causal inferences about illness elicited increased activity in the PC compared to both i) causal inferencing about a non-illness domain (i.e., mechanical failure) and ii) closely matched but causally unconnected sentences. These results suggest that causal inferences made during language comprehension rely on domain-specific causal mechanisms that are recruited depending on the semantic domain in focus.

Our results fail to provide clear evidence for a domain-general mechanism for causal inferencing during language comprehension. Using a sensitive functional localization approach, we found that neither the language nor the logical reasoning networks exhibited a preference for causal inferencing. We also found no evidence for a distinct general-purpose causal inference mechanism outside these networks, i.e., a brain region that exhibited a robust preference for causal inferencing across both illness and mechanical stimuli.

With respect to the language network, our findings are consistent with prior evidence showing that language areas are most sensitive to linguistic input at the level of individual clauses/sentences (e.g., Jacoby & Fedorenko, 2020; Blank & Fedorenko, 2020). However, our findings contradict a small number of past experiments suggesting that the language network enables causal inferencing during comprehension (Kuperberg et al., 2006; Mason & Just, 2011; Prat et al., 2011). Unlike in this past work, the causal and noncausal stimuli used in our experiment were completely controlled for linguistic form: both conditions used the same sentences. We hypothesize that the language network responds more to causal or non-causal passages depending on which are more difficult to process, whether due to increased linguistic complexity (e.g., longer sentences) or the behavioral pressures imposed by explicit causal judgment tasks (see Kuperberg et al., 2006). This pattern does not support the hypothesis that the language network is the 'engine' for causal inferencing during language processing.

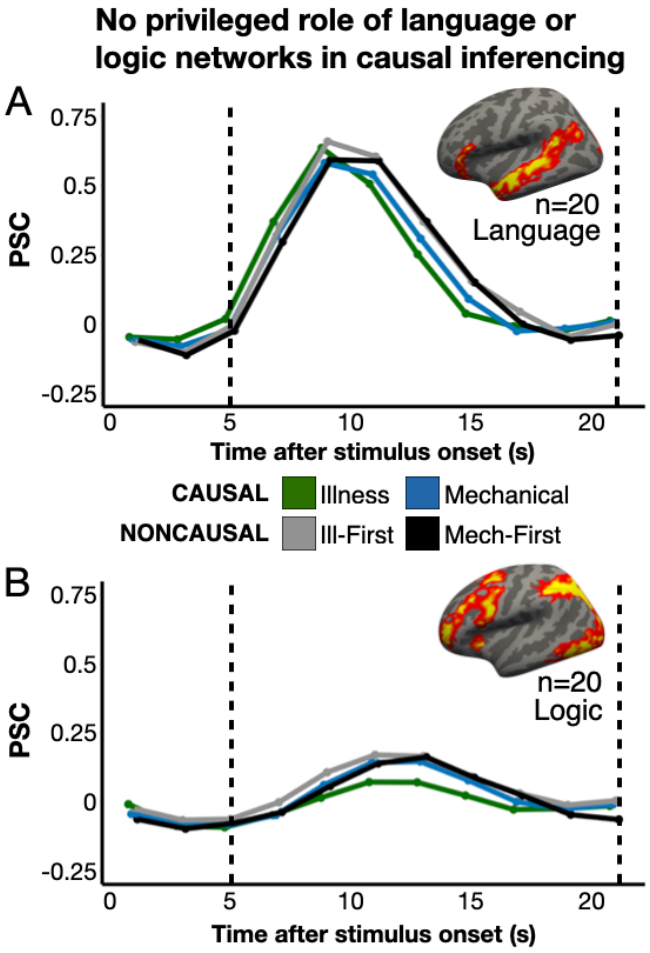

Figure 2: Individual-subjects analysis of language- and logic-responsive voxels. Panel A: percent signal change (PSC) for each condition among the top 5% most language-responsive voxels (*language* > *math*) in a temporal language network mask (Fedorenko et al., 2010). Panel B: PSC among the top 5% most logic-responsive voxels (*logic* > *language*) in a logic network mask (Liu et al., 2020). Group maps for each contrast of interest are corrected for multiple comparisons.

With respect to the logical reasoning network, our results suggest that causal inferences elicited during language comprehension are supported by different neural mechanisms than formal logical reasoning (e.g., *If X then Y = If not Y then X*?). This finding coheres with prior evidence that the logic network is specialized for symbolic, 'content-free' stimuli, such as statements relating variables X and Y as opposed to concrete nouns (Monti et al., 2009; Feng et al., 2021).

Together, the current results suggest that causal inferences elicited during language comprehension depend on domain-specific semantic systems. In the case of illness causality, we rely on neural circuits that support biological knowledge (e.g., Atran, 1998; Keil, 1992). Whereas prior neuroscience work on this system has focused on simple binary judgments about words or word pairs (e.g., Fairhall & Caramazza, 2013a, 2013b; Fairhall et al., 2014; Deen & Freiwald, 2022), the current results suggest that this system also encodes rich causal and relational information.

Importantly, multiple domain-specific systems might contribute to a single inference. The vignettes used in the current study were designed to elicit biological inferences familiar to a lay audience. In reality, when faced with imperceptible, life-threatening causal processes, people across cultures commonly combine knowledge from multiple domains to explain the emergence of illness (e.g., contact with blood plus witchcraft leads to AIDS; Lynch & Medin, 2006; Legare & Gelman, 2008; Legare & Shtulman, 2018). We hypothesize that different domain-specific systems (e.g., biology, theory of mind) become engaged depending on the inference. For instance, both the PC and right temporoparietal junction (Saxe & Kanwisher, 2003) may become engaged when making the above inference. Relatedly, given vast differences in ideas about illness causality across cultures (e.g., germ theory, divine retribution, vitalism), future work can address how cultural input (e.g., medical education) shapes domain-specific responses to illness inferencing.

Our results do not rule out the possibility that domain-general mechanisms enable causal inferences under some circumstances, even inferences about the causes of illness. The vignettes used in the current study stipulate the cause in the first sentence, allowing participants to reason from causes to effects. By contrast, illness reasoning performed by medical experts proceeds from effects to causes and involves identifying potential illness causes within highly complicated and interconnected causal systems (Schmidt, Norman, & Boshuizen, 1990; Norman et al., 2006; Meder & Mayrhofer, 2017). Future studies should examine whether such complex inferences rely on domain-general reasoning systems.

Another open question not addressed in the current study concerns what mechanisms support learning novel causal relationships, such as when learning unfamiliar causal powers of objects (e.g., 'blicket detectors'; Gopnik et al., 2001) or forming new scientific theories about complex causal processes (e.g., illness transmission; Lock & Nguyen, 2010). It is possible that domain-general reasoning mechanisms support the discovery of novel causes. However, we hypothesize that forming new causal connections may also draw upon domain-specific causal knowledge in the domain that bears the most resemblance to the novel process. Future neuroimaging work can help test these possibilities.

Conclusion

The current results suggest that causal inferences during language comprehension rely on domain-specific neural machinery. In the case of illness inferencing, a domain-specific system for biological knowledge becomes engaged. By contrast, we find no evidence for a neural mechanism that is domain-general and operates regardless of semantic domain, either in the language system itself, the logical reasoning system, or anywhere else in the brain. Our results support the hypothesis that, at least for simple cases, causal inferences are supported by neural substrates that represent domain-specific semantic information.

Acknowledgements

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

References

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

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*(3), 299–352.

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

Baillargeon, R. (1995). Physical reasoning in infancy. In M. Gazzaniga (Ed.), *The cognitive neurosciences*. MIT Press.

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.

Blank, I. A., & Fedorenko, E. (2020). No evidence for differences among language regions in their temporal receptive windows. *NeuroImage, 219*, 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.

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

Cesana-Arlotti, N., Kovács, Á. M., & Téglás, E. (2020). Infants recruit logic to learn about the social world. *Nature Communications, 11*(1), 5999.

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

Collins, A., & Koechlin, E. (2012). Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology, 10*(3), e1001293.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage, 9(*2), 179–194.

Deen, B., & Freiwald, W. A. (2022). Parallel systems for social and spatial reasoning within the cortical apex. *bioRxiv.*

Donoso, M., Collins, A. G. E., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science, 344*(6191), 1481–1486.

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.

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.

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.

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.

Fairhall, S. L., & Caramazza, A. (2013a). Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience, 33*(25), 10552–10558.

Fairhall, S. L., & Caramazza, A. (2013b). Category-selective neural substrates for person- and place-related concepts. *Cortex, 49*(10), 2748–2757.

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.

Feng, W., Wang, W., Liu, J., Wang, Z., Tian, L., & Fan, L. (2021). Neural correlates of causal inferences in discourse understanding and logical problem-solving: a meta-analysis study. *Frontiers in Human Neuroscience*, *15*, 666179.

Foster, G. M. (1976). Disease etiologies in non-western medical systems. *American Anthropologist, 78*(4), 773–782.

Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate-inanimate distinction as examples. *Cognitive Science, 14*(1), 79–106.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology, 20*(1), 65–95.

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

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

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

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.

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.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language, 7*(1–2), 145–171.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*(6), 1085–1108.

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.

Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development, 77*(3), 505–524.

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

Hume, D. (1978). *A treatise of human nature.* Oxford, England: Oxford University Press. (Original work published 1739).

Inagaki, K., & Hatano, G. (2004). Vitalistic causality in young children's naive biology. *Trends in Cognitive Sciences, 8*(8), 356–362.

Inagaki, K., & Hatano, G. (2006). Young children's conception of the biological world*. Current Directions in Psychological Science, 15*(4), 177–181.

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.

Johnson-Laird, P. N., & Khemlani, S. (2017). Mental models and causation. In M. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning.* Oxford University Press.

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.

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., 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*, MIT Press.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 7*(8), 368–373.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist, 28*(2), 107–128.

Khemlani, S. S., Barbey, A. K., & Johnson-Laird, P. N. (2014). Causal reasoning with mental models. *Frontiers in Human Neuroscience, 8*, 849.

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., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. In A. Gopnik, L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

Lea, R. B. (1995). On-line evidence for elaborative logical inferences in text. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1469.

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.

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.

Legare, C. H., & Shtulman, A. (2018). Explanatory pluralism across cultures and development. In J. Proust, M. Fortier (Eds.), *Metacognitive diversity: An interdisciplinary approach*. Oxford University Press.

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.

Lock, M. M., & Nguyen, V. K. (2018). *An anthropology of biomedicine.* John Wiley & Sons.

Lynch, E., & Medin, D. (2006). Explanatory models of illness: A study of within-culture variation. *Cognitive Psychology, 53*(4), 285–309.

De Marneffe, M.-C., MacCartney, B., & Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *Lrec* (Vol. 6, pp. 449-454).

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.

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. (1999). Introduction. In D. L. Medin, S. Atran (Eds.), *Folkbiology*. MIT Press.

Michotte, A. (1963). *The Perception of causality*. Basic Books.

Mody, S., & Carey, S. (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition, 154*, 40–48.

Monti, M. M., & Osherson, D. N. (2012). Logic, language and the brain. *Brain Research, 1428*, 33–42.

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.

Muentener, P., & Carey, S. (2010). Infants' causal representations of state change events. Cognitive Psychology, 61(2), 63–86.

Muentener, P., & Schulz, L. (2014). Toddlers infer unobserved causes for spontaneous events. *Frontiers in Psychology, 5*, 1496.

Norman, G. R., Grierson, L. E. M., Sherbino, J., Hamstra, S. J., Schmidt, H. G., & Mamede, S. (2018). 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.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255–258.

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.

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.

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.

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.

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.

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.

Reverberi, C., Cherubini, P., Rapisarda, A., Rigamonti, E., Caltagirone, C., Frackowiak, R. S. J., Macaluso, E., & Paulesu, E. (2007). Neural basis of generation of conclusions in elementary deduction. *NeuroImage, 38*(4), 752–762.

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.

Saxe, R., & Carey, S. (2006). The perception of causality in infancy. *Acta Psychologica, 123*(1), 144–165.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage, 19*(4), 1835–1842.

Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science, 16*(12), 995–1001.

Saxe, R., Tzelnic, T., & Carey, S. (2007). Knowing who dunnit: Infants identify the causal agent in an unseen causal interaction. *Developmental Psychology, 43*(1), 149–158.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43*(10), 1391–1399.

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.

Schult, C. A., & Wellman, H. M. (1997). Explaining human movements and actions: Children's understanding of the limits of psychological explanation. *Cognition, 62*(3), 291–324.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*(2), 162–176.

Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences, 110*(40), 15937–15942.

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.

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.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology, 66*(1), 223–247.

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.

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.

Spelke, E. S., Katz, G., Purcell, S. E., Ehrlich, S. M., & Breinlinger, K. (1994). Early knowledge of object motion: Continuity and inertia. *Cognition, 51*(2), 131–176.

Springer, K., & Ruckel, J. (1992). Early beliefs about the cause of illness: Evidence against immanent justice. *Cognitive Development, 7*(4), 429–443.
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.
Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz (Eds.), *Causal Learning*. Oxford University Press: New York.
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.
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.
Waldmann, M. R. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 53–76.
Waldmann, M. R., & Hagmayer, Y. (2013). Causal reasoning. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology.* Oxford University Press.
Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*(1), 337–375.
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.
Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*(1), 1–34.