Title

A region in human left prefrontal cortex selectively engaged in causal reasoning

Permalink

https://escholarship.org/uc/item/6ms537c4

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Pramod, RT
Chomik, Jessica
Schulz, Laura
et al.

Publication Date

2024

Peer reviewed

A Region in Human Left Prefrontal Cortex Selectively Engaged in Causal Reasoning

RT Pramod*, Jessica Chomik-Morales*, Laura E. Schulz, Nancy Kanwisher
{pramodrt, jchomik, lschulz, ngk}@mit.edu
Department of Brain and Cognitive Sciences, MIT
*denotes equal authorship

Abstract

Causal reasoning enables us to explain the past, predict the future, and intervene in the present. Does the brain allocate specialized cortical regions to causal reasoning? And if so, are they involved in reasoning about both physical and social causal relationships, or are they domain-specific? In a pre-registered experiment (Exp 1) we scanned adults using fMRI while they matched physical and social causes to effects (e.g., 'The car swerved to avoid a crash' -> 'Coffee spilled all over the car seat'; 'He was late for work' -> 'Tom was scolded by his boss') or physical and social descriptions of the same entity matched for difficulty and linguistic variables to the causal conditions (e.g., 'The brightest object in the sky'-> 'The closest star to earth'; 'She works at a hotel' -> 'She brings in guests' luggage'). A region in the left lateral prefrontal cortex (LPFC) responded significantly more strongly to causal than descriptive conditions in most subjects individually. Responses in this region in held-out data were high for both social and physical causal conditions, yet no greater than baseline for the two descriptive (non-causal) conditions. In a follow-up exploratory experiment (Exp 2), we tested a different task (answering causal versus non-causal questions about physical and social narratives, matched for linguistic variables). Again, we found that both the physical and social causal stimuli selectively engaged the LPFC region. Finally, in both experiments, we found that brain regions previously implicated in intuitive physical reasoning responded more to the physical causal than the physical non-causal stimuli. Collectively, these results suggest that a) a region in the LPFC is selectively engaged in causal reasoning independent of content domain and b) the hypothesized physics network (hPN) is selectively involved in physical causal reasoning across modalities (visual vs. linguistic).

Keywords: causal reasoning; fmri; intuitive physics

Introduction

The idea that intervening on some events changes the probability of others underlies scientific inquiry and commonsense reasoning alike (Pearl, 2000; Woodward, 2003); little wonder that David Hume called causality "the cement of the universe" (1740). But although causal reasoning has long been a topic of investigation in philosophy and cognitive science, little is known about its neural basis. Despite some recent advances (Operskalski & Barbey, 2017), many questions remain unanswered. In particular, does causal reasoning engage specialized cortical regions, or is it implemented by more general networks involved in a range of cognitive processes? If there are specialized networks for causal reasoning, are they involved in both physical and social causal reasoning, or are there separate domain-specific regions for causal reasoning about physical and social phenomena?

The question of whether causal reasoning engages domain-specific or domain-general machinery has been a topic of extensive debate (cf: Hirschfeld & Gelman, 1994). On the one hand, considerable evidence supports the existence of some domain-specific causal representations, starting early in infancy. Babies as young as three-months recognize that inanimate objects respond to contact causality whereas animate agents can interact at a distance and respond to internal motivations like preferences and goals (e.g., Baillargeon, 1994; Carey & Spelke, 1994; Spelke & Kinzler, 2007; Liu, Brooks, & Spelke, 2019). Children's causal inferences become increasingly sophisticated over development to the point that many researchers have argued that children's early understanding of the physical and psychological world takes the form of domain-specific intuitive theories (e.g., Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1992; Rhodes, 2014; Wellman & Gelman, 1992). Consistent with this idea, neuroimaging studies have found distinct cortical regions engaged in perceiving and understanding inanimate objects versus social agents, suggesting that causal reasoning about agents and objects may differentially recruit these two systems (Kanwisher, 2010; Fischer et al., 2016).

On the other hand, causal reasoning supports prediction, intervention, explanation, and causal reasoning across domains, and we can construct the same kinds of structured representations of causal variables and their relationships regardless of the content the variables stand for (e.g., Griffiths & Tenenbaum, 2009; Pearl, 2000). Even young children can use patterns of covariation between interventions and outcomes to draw causal inferences, and they can do so whether they are reasoning about blocks that activate a machine, flowers that make a monkey sneeze, or animals that scare a puppet (Schulz & Gopnik, 2004). Causal relationships have distinctive syntactic markers across languages (Landau & Jackendoff, 1993; Naigles, 1990; Yuan & Fisher, 2009; Gleitman & Gilette, 1998; Slobin, 1982) and, in ordinary language, we express causal relationships with the same syntax and even some of the same words, independent of content domain ("She made the toy go"; "She made the girl smile"). These considerations suggest that a single system may underlie both physical and social causal reasoning.

In the current study, we used fMRI to investigate causal reasoning in the brain by scanning adults while they performed causal and non-causal reasoning tasks in both social and physical domains. In Experiment 1 we tested whether any brain regions are selectively activated during causal (versus non-causal) reasoning and if so, whether these brain regions are recruited equally for physical and social causal reasoning. We also test the separate, mutually non-exclusive hypothesis that brain regions previously implicated in domain-specific reasoning about intuitive physics and intuitive psychology will be selectively engaged in physical and social causal reasoning respectively (compared to both non-causal tasks in the target domain and causal tasks in the other domain). In an exploratory follow-up study, Experiment 2, we test the generality of our findings using a different causal, non-causal contrast (a narrative task) in both domains.

Methods

Experiment 1 ("Ice Cream" Task)

Participants We recruited 18 participants (Mean age = 25.4 y, range 22 - 31; 12 identifying as female; 6 identifying as male) from the Greater Boston area. Participants had normal or corrected-to-normal vision, and no MRI contraindications. All study procedures were approved by Massachusetts Institute of Technology (MIT) Committee on the Use of Humans as Experimental Subjects (#0403000096). Participants were asked to provide written informed consent prior to participation and were paid $75 for a 2-hr scan session. The methods and analyses of this experiment were pre-registered prior to data collection.

Experimental Design and Procedure We used a within-subjects, 2-by-2 blocked design with two orthogonal factors: Causality (causal vs descriptive) and Domain (physical vs social). Participants were asked to match sentences describing causes to sentences describing their effects, or to match sentences that describe the same entity. Each block consisted of stimuli from one of the four conditions (physical causal, social causal, physical descriptive, social descriptive). Each block showed four images of ice cream scoops (top) and four images of ice cream cones (bottom) that participants were to match, progressing from left to right (Figure 1A). Each stimulus block lasted 20 seconds; participants could match the items at their own pace within that time. Each of the four conditions occurred in two blocks within each run, with condition order organized palindromically within a run. Across the eight runs, each condition occurred equally often in each serial position. Each run consisted of 8 stimulus blocks and three 18-second fixation blocks (white cross in the center of the screen) appearing once at the beginning, middle, and end of each run. The complete stimulus set contained 64 pairs of sentences for each of the four conditions. Within each block, the spatial position of the scoop images was shuffled across participants.

Experiment 2 ("Short Stories" Task)

Participants We recruited 8 participants (Mean age = 24.6 y, range 20-34; 4 identifying as female, and 4 as male) from the Greater Boston area. Participants had normal or corrected-to-normal vision, and no MRI contraindications.

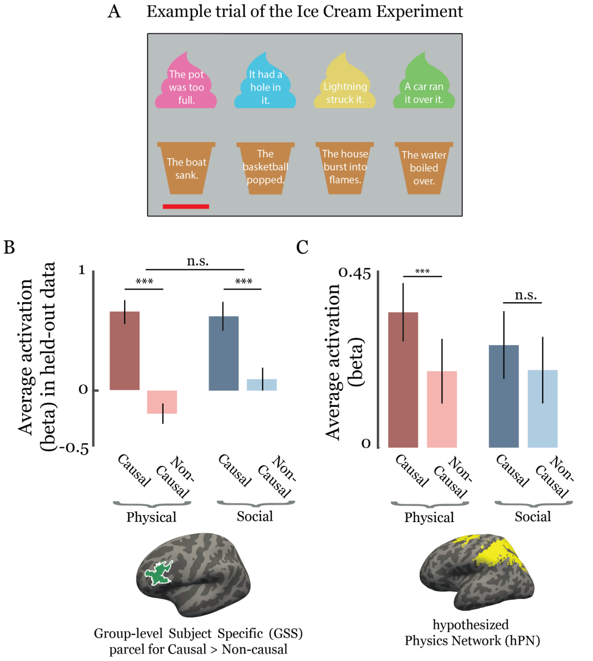


Figure 1. Experiment 1 (Ice Cream). (A) Example trial showing causes (scoops, top row) and effects (cones, bottom row). Red underbar indicates the current effect to be matched. Bar plots show the average fMRI response (N = 18) to causal and non-causal conditions for both physical and social domains in data independent from those used to define the fROI in (B) the left lateral prefrontal cortical (LPFC) fROI and (C) the hypothesized Physics Network fROI. *** indicates p < 0.0005; n.s. is not significant.

Experimental Design and Procedure Experiment 2 was a reading comprehension task in which subjects read 3-sentence narratives, each describing an event or situation that entailed a social cause or physical cause (SCn, PCn), or that described socially or physically non-causal (spatial or constitutive) relations (SSn, PSn), and were then asked multiple-choice questions (see Figure 3A). Using button boxes, participants chose the correct answers to multiple-choice questions after each short story to ensure that they were reading and understanding the sentences (see Figure 3A). Each stimulus block lasted 22 seconds, with the story presentation lasting for 13s followed by the multiple-choice question to which participants had 9 seconds to respond. The complete stimulus set contained 16 stories for each of the four conditions (PCn, SCn, PSn, SSn). Each condition was shown four times in each run, and each condition appeared equally often in each serial position across runs. Each run consisted of 16 stimulus blocks, five 14s fixation blocks (one at the beginning and one after every four runs), and two-second fixation periods between each block. The fixation blocks showed a white cross in the center of a gray screen.

Linguistic and Difficulty Controls Verbal stimuli for both Experiments 1 and 2 were matched across conditions for grade level and word count using online tools. In Experiment 1, we observed that functionally localized language regions in individual participants, as determined using parcels from: https://evlab.mit.edu/funcloc/, did not exhibit a greater response to causal compared to descriptive condition in either physical or social domains, indicating an approximate linguistic match of our causal and descriptive conditions. Stimuli were adjusted with behavioral piloting for both Experiment 1 and 2 before the fMRI experiment was run to match difficulty across causal and non-causal conditions. Behavioral data collected during the fMRI scans showed higher accuracy for the causal than descriptive conditions in the Experiment 1 (N=18, Average Accuracy = 91.5% and 84.9% for causal and non-causal conditions respectively; $p < 0.05$ for a Wilcoxon signed rank test). Reaction times were not significantly different (Average Reaction Time = 5.58 s and 5.6 s for causal and non-causal conditions respectively; $p > 0.05$ for a Wilcoxon signed rank test). Behavioral data collected inside the scanner for Experiment 2 showed low task accuracy across conditions and was highly variable across participants, likely due to button-assignment confusion in the scanner. Behavioral testing of the same paradigm outside the scanner showed no significant differences of accuracy and response time across conditions (N = 20).

Intuitive Physics Network Localizer Participants in both Experiments 1 and 2 were scanned on two runs of a "Physics localizer" to identify regions in the frontal and parietal hypothesized Physics Network (hPN), using the same stimuli and contrast reported by Fischer et al. (2016). Participants view short video clips depicting rotating block towers and are asked to determine which side the tower is most likely to fall towards (the physics condition) or if the tower contains more blue than yellow blocks (the control condition). Each run included 23 blocks (10 physics, 10 color, 3 rest) and lasted approximately 6.9 minutes.

Theory of Mind (ToM) Localizer Participants in Experiment 1 were scanned on two runs of a standard theory of mind localizer (Jacoby et al, 2016) to identify brain regions previously implicated in inferences about others' mental states (Saxe & Kanwisher, 2003). Participants answer a true/false question after reading a short story describing a false representation that is either mentally held by a person ('False Belief' condition) or physically present on a map/photo ('False Photo' condition). Each run contained 5 False Belief and 5 False Photo stories. Each story was presented for 10s after which a true/false question about either the true state of the world or the false representation was presented for 4s, with a 12s inter-stimulus interval. Each run lasted 4.5 minutes.

Data acquisition Both anatomical and functional data were acquired from a Siemens 3T MAGNETOM Prisma scanner, using a 32-channel head coil (scan parameters are the same as in Pramod et al., 2022). Participants viewed stimuli through a mirror projected to a 12x16" screen behind the scanner, at a visual angle of approximately 14°x19°.

fMRI data preprocessing and analysis

Preprocessing Neuroimaging data were preprocessed using FreeSurfer and additional analyses were performed using custom scripts written in MATLAB (similar as in Pramod et al., 2022). All analyses were performed in the native volume space of each subject. The general linear model (GLM) used to estimate the voxel-wise activations (beta) for our experimental conditions in both experiments included one regressor per stimulus condition (a 'boxcar' function that was set to 1 for the duration of the block and 0 otherwise) and 6 nuisance regressors based on the motion estimates (x, y, and z translation; roll, pitch, and yaw of rotation).

Group-level parcel The left lateral prefrontal cortical (LPFC) parcel was derived from Experiment 1 data using the Group-constrained Subject-Specific (GSS) method (Fedorenko et al, 2010; Julian et al 2012), as follows. Participants' individual's binary contrast maps (causal > non-causal, aggregated across both physical and social domains, thresholded with $p < 0.05$) were overlaid in MNI (fsaverage) space to create an overlap map. This map was then spatially smoothed with a gaussian filter of FWHM = 6 mm and then thresholded to contain only those voxels with at least 15% overlap across subjects. Subsequently, a watershed image segmentation algorithm was applied to divide the thresholded subject overlap map into non-overlapping parcels. Finally, a subset of these parcels in which at least 80% of the subjects show some activated voxels were selected. We identified one of these parcels in the left lateral prefrontal cortex that responded more to causal compared with non-causal conditions in both physical and social domains. In addition to generating the LPFC parcel from all runs of Experiment 1 for analyzing Experiment 2, we also generated parcels from even and odd runs of Experiment 1 separately such that we could use one half of the data to define the functional ROI and the other half to compute activations for the stimulus conditions. To define fROIs for the hPN we used parcels from a previous study (Pramod et al 2022). ToM parcels (thresholded maps for rTPJ, rSTS, vmPFC, dmPFC and vmPFC) were from https://saxelab.mit.edu/use-our-theory-mind-group-maps/.

Defining functional Regions of Interest (fROI) Functional ROIs were defined separately in each participant as the intersection of relevant subject-specific functional contrast map thresholded at $p < 0.001$ uncorrected (Physics > Color task for the Physics Localizer, False Belief > False Photo for Theory of Mind Localizer, or Causal > Non-causal for our two experiments) with the relevant group-level anatomical parcel (GSS parcels for hPN and LPFC causal region, and thresholded subject-averaged contrast maps for ToM regions). If we failed to find at least 50 voxels within the fROI, we reduced the contrast map threshold to $p < 0.05$. Thus, individual subject fROI contained only those voxels that showed strong functional contrast and fell within the group-level parcel. This allowed the fROI locations and sizes to vary across subjects yet restricted them to a specific region across subjects. Response magnitudes derived from each fROI are based on data independent of those used to define the fROI.

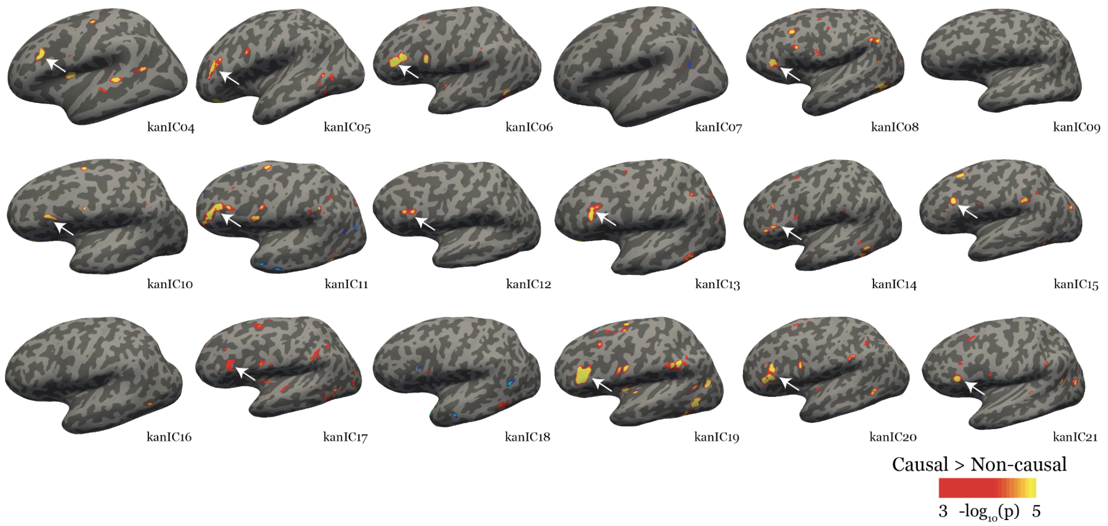


Figure 2: Whole-brain causal (PC+SC) > non-causal (PD+SD) contrast map for each participant in Experiment 1

Results

A brain region selectively involved in causal reasoning

Do any brain regions exist that are selectively involved in causal reasoning? To find out, we first conducted whole-brain analyses of Experiment 1 in each participant individually. We found a region in the left lateral prefrontal cortex (LPFC) that responded significantly more strongly to causal than descriptive conditions in 14 out of the 18 participants individually (Figure 2). This region did not reach significance in a group random effects analysis, presumably because its exact anatomical location varied across participants. We therefore conducted a group-level subject-specific (GSS) analysis to test for regions that respond significantly more to causal than descriptive conditions across both physical and social domains (see Methods and Fedorenko et al., 2010). To avoid double-dipping, we used one half of the data from the Experiment 1 to identify both the GSS ROI and functionally selective voxels therein (i.e., defining the functional ROI), and the other half of the data to measure average activations for the four stimulus conditions within the identified fROI.

The GSS analysis revealed a region in the left lateral prefrontal cortex (LPFC) that showed significantly higher responses for causal compared with non-causal conditions in the Experiment 1. Data independent of those used to define the LPFC fROI showed that causal conditions evoked greater responses than non-causal conditions in this region in both physical (average response ± standard error of mean: 0.66 ± 0.1 and -0.19 ± 0.09 for causal and descriptive conditions respectively; $p < 0.0005$ for a Wilcoxon signed rank test) and social domains (average response ± standard error of mean: 0.62 ± 0.12 and 0.096 ± 0.096 for causal and descriptive conditions respectively; $p < 0.0005$ for a Wilcoxon signed rank test). There was no significant difference between responses to physical and social conditions ($p = 0.22$), and no significant interaction of Causality x Domain ($p = 0.12$) in a two-way ANOVA with causality and domain as factors, implying the domain-general nature of causal reasoning in the LPFC fROI. Notably, the descriptive (non-causal) conditions did not respond significantly differently from rest/fixation in this region ($p > 0.1$), despite being matched for difficulty and linguistic variables.

To further test the domain-general nature of the LPFC fROI concerning causality, we conducted a Multivariate Voxel Analysis (MVPA). We found that the activity pattern correlation within causality (i.e., correlation between PC and SC, and between PD and SD), was significantly higher ($p < 0.0005$) than activity pattern correlations between causality (i.e, correlation between PC and SD, and between SC and PD). This finding indicates that the LPFC fROI contains domain-general information that distinguishes between causal and descriptive conditions. We also found that the pattern of response across voxels was more correlated ($p < 0.005$) within domain (i.e, correlation between PC and PD, and between SC and SD) than between domains (i.e, correlation between PC and SD, and SC and PD) indicating the presence of domain information within the fROI.

Thus, we found striking evidence for a highly selective response in the LPFC fROI to causal reasoning for both physical and social causes, compared to non-causal reasoning (Figure 1B). To test whether this finding generalizes to a different task, we measured responses to the causal and spatial conditions in Experiment 2 in the same region. To define the individual-subject fROIs for this analysis, we aligned the LPFC parcel obtained from Experiment 1 to the individual subject's brain in Experiment 2 and selected voxels within that parcel that showed higher responses to causal compared with spatial conditions ($p < 0.001$ uncorrected within subjects which was reduced to $p < 0.05$ if fewer than 50 voxels) in even runs of Experiment 2. We then computed the average response of the selected voxels to the four stimulus conditions using data from the odd runs. We then conducted the opposite analysis, using odd runs to select voxels and even runs to compute responses, and then averaged the responses for the four experimental conditions across the two splits. We found that average activations to causal conditions were higher than non-causal (spatial or constitutive) conditions in both physical (average response ± standard error of mean: 0.29 ± 0.21 and 0.04 ± 0.22 for causal and spatial conditions respectively, p = 0.1 for a Wilcoxon signed rank test) and social (average response ± standard error of mean: 0.34 ± 0.31 and -0.08 ± 0.24 for causal and spatial conditions respectively, p = 0.04 for a Wilcoxon signed rank test) domains, but reached significance only in the social domain. A power analysis suggests that the causal > spatial effect in the physical domain should be detectable at 80% power and $p < 0.05$ in a Wilcoxon signed rank test with N = 18 participants, a prediction we are currently testing.

Thus, we find evidence for a region in the left lateral prefrontal cortex (LPFC) that responds more to causal compared with non-causal conditions in both physical and social domains, in both Experiment 1 and Experiment 2.

Evidence for physical (but not social) causal reasoning in the hypothesized Physics Network

Here we tested the hypothesis that reasoning about physical and social causes will engage cortical mechanisms previously implicated in domain-specific processing of physical and social information. For instance, the hypothesized Physics Network (hPN) – previously shown to be engaged during physical reasoning tasks (Fischer et al 2016) and to carry invariant information about object mass (Schwettmann, Tenenbaum & Kanwisher, 2020) and physical stability (Pramod et al, 2022) – could be engaged more during physical than social causal reasoning. To test this hypothesis, we identified the hPN in each participant individually using an independent localizer (see Methods) and computed responses to causal and non-causal conditions within the fROI for both Physical and Social domains.

In Experiment 1, the average activation within the hPN fROI for the causal condition was significantly higher than the descriptive (non-causal) condition only in the physical but not the social domain (average activation ± standard error of mean: 0.33 ± 0.086 and 0.22 ± 0.096 for physical causal and physical non-causal conditions respectively, $p < 0.0005$ for a Wilcoxon signed rank test; average activation ± standard error of mean: 0.25 ± 0.098 and 0.21 ± 0.099 for social causal and social non-causal conditions respectively, p = 0.13 for a Wilcoxon signed rank test). Further, the effect of causality was stronger in physical than social domain (p = 0.037 for the interaction effect between causality and domain in an ANOVA).

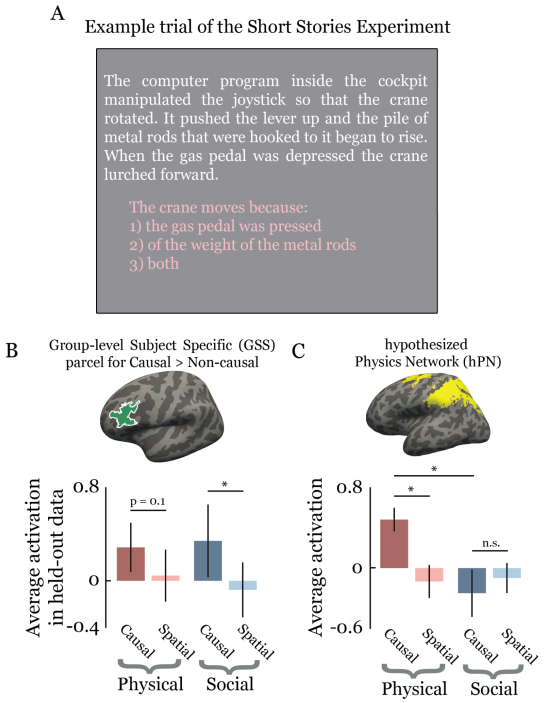
In Experiment 2 as well, the hPN responded more strongly in the causal than spatial conditions, but only for the physical conditions (average activation ± standard error of mean: 0.48 ± 0.12 and -0.13 ± 0.16 for physical causal and physical spatial conditions respectively, p = 0.017 for a Wilcoxon signed rank test; average activation ± standard error of mean: -0.25 ± 0.24 and -0.099 ± 0.15 for social causal and social spatial conditions respectively, p = 0.67 for a Wilcoxon signed rank test). As before, the average activation for physical causal condition was significantly higher than the social causal condition (p = 0.027 for a one-tailed Wilcoxon signed rank test).

Do the brain regions previously implicated in Theory of Mind (ToM) reasoning show social domain-specific effects for causal reasoning? To find out, we identified ToM fROIs in individual participants using an independent localizer (see Methods) and computed average activations to both physical and social, causal and non-causal conditions within each fROI. However, none of the fROIs showed higher activations for causal compared to non-causal conditions in the social (or physical) domain.

Thus, the hypothesized Physics Network (hPN) shows higher activations for causal compared to non-causal conditions only in the physical domain thereby providing evidence for a domain-specific mechanism for causal reasoning in the human brain. We did not find evidence for selective processing of social causal information in regions previously implicated in theory of mind.

Discussion

Our study investigated causal reasoning in the human brain by testing two hypotheses: 1) that one or more brain regions are selectively engaged during causal reasoning independent of content domain (social versus physical), and 2) that previously identified domain-specific regions for understanding the physical and social world are involved in causal reasoning only within those domains. We found evidence for both hypotheses: We identified a region in the left lateral prefrontal cortex (LPFC) that responded more to causal than non-causal conditions in both physical and social domains; and we found that the hypothesized Physics Network (hPN) – previously implicated in physical scene understanding – responds more during causal than non-causal conditions but only in the physical domain. We did not find evidence for social causal reasoning in regions previously implicated in Theory of Mind (rTPJ, rSTS, vmPFC, dmPFC or mmPFC), presumably because most of our social causal stimuli did not require reasoning about belief contents (e.g. 'She stayed up late last night' -> 'Karen overslept this morning'). The current results are unlikely to reflect generic task difficulty or linguistic effects as these were matched across conditions. Thus, our findings across two experiments suggest that both domain-general and domain-specific brain regions are involved in causal reasoning.



The computer program inside the cockpit manipulated the joystick so that the crane rotated. It pushed the lever up and the pile of metal rods that were hooked to it began to rise. When the gas pedal was depressed the crane lurched forward.

The crane moves because:
1) the gas pedal was pressed
2) of the weight of the metal rods
3) both

Figure 3: Experiment 2 (Short Stories). (A) Example trial from the Physical Causal condition showing the short narrative (white text) and the corresponding multiple choice question (pink text). Bar plot showing the average fMRI response (for N = 8 participants) to causal and spatial conditions in both physical and social domains within (B) the left lateral prefrontal (LPFC) fROI and (C) the independently defined hPN fROI. Response magnitudes are computed from data independent from those used to define fROIs.* is p < 0.05 and n.s. is not significant.

Our findings are loosely consistent with the few prior studies that have investigated causal reasoning with neuroimaging (Fugelsang et al, 2005; Operskalski 2017). Two of these have reported that causal reasoning engages brain regions in or near the LPFC region found in the current study. One study (Kuperberg et al, 2006) found activation in left DLPFC and other brain regions when participants judged the causal relationship of a third sentence to the first two sentences describing a scenario, although these effects could reflect generic difficulty confounds. Another study (Satpute et al. 2005) found that left DLPFC and right precuneus were more activated when participants judged the causal versus associative relationship of word pairs. However, because these studies only showed activations from group analyses, they could not see the striking evidence for selectivity revealed in our study: a region *that not only responded significantly more to causal than non-causal stimuli in 14 out of 18 subjects individually* (Figure 2), but *responds not at all (i.e., no more than during rest) to similar, linguistically and difficulty-matched but non-causal stimuli*.

Although these results are novel with respect to our understanding of the functional organization of the human brain, they are gratifyingly consistent with decades of work in cognitive, computational, and developmental science. As discussed above, abundant prior work provided evidence for both domain-specific representations of distinct kinds of causal relationships in the physical and social world and domain-general capacities for causal inference broadly (Hirschfeld & Gelman, 1994). The ability to formalize the domain-general ways in which people integrate new evidence and prior, domain-specific causal knowledge through hierarchical Bayesian inference has been one of the key insights in bridging the gap between these forms of reasoning (Griffiths & Tenenbaum, 2009). This suggests that investigating the neural pathways that might link regions involved in domain-general and domain-specific causal inference is an important area for future work.

Several questions remain for future research. Causes always precede effects so all causal relationships are also temporal relationships. Future work must look at whether the LPFC region implicated here is specific for causal reasoning per se or whether it responds to temporal order and other abstract but non-causal relationships (e.g., part-whole relationships). Second, we tested two contexts for causal reasoning, matching causes and effects and reading causal narratives, however, there are many other ways to assess causal reasoning. It is not yet clear whether the LPFC region engaged here is involved in all forms of causal reasoning (counterfactual reasoning? inferring causal relationships from data?) or only a subset. And while we have preliminary evidence (i.e., responses to both the causal verbal tasks and the physics task) suggesting that this brain region responds to causal representations independent of modality, future work must determine how invariant these responses are. Third, the relationship between the LPFC regions activated by causal reasoning and the "multiple demand" (MD) system remains unclear. Although our causal and non-causal conditions were matched for difficulty, suggesting that this region does not overlap with the MD system, this question merits further investigation.

Overall, however, the current data suggest both that a region in human LPFC has a remarkably consistent, robust, and selective response to causal stimuli independent of domain, and that a region previously implicated in intuitive physical reasoning, the hPN, is engaged in domain-specific physical causal reasoning. Given the centrality of causal reasoning, it is perhaps unsurprising that there are regions selectively dedicated to it in the human cortex. However, humans' ability to use those brain regions to inquire about, experiment on, and explain the physical and social world, including the basis of causal reasoning itself, is a source of enduring wonder.

Acknowledgements

This research is funded by the MIT BCS Research Scholars Program and K. Lisa Yang (to JCM).

References

Baillargeon, R. (1994). Physical reasoning in young infants: Seeking explanations for impossible events. *British Journal of Developmental Psychology*, 12(1), 9–33.

Barrett, L. F., & Satpute, A. B. (2013). Large-scale brain networks in affective and social neuroscience: Towards an integrative functional architecture of the brain. In *Current Opinion in Neurobiology* (Vol. 23, Issue 3, pp. 361–372).

Blakemore, S. J., Fonlupt, P., Pachot-Clouard, M., Darmon, C., Boyer, P., Meltzoff, A. N., Segebarth, C., & Decety, J. (2001). How the brain perceives causality: an event-related fMRI study. *Neuroreport*, 12(17), 3741–3746.

Carey, S., & Spelke, E. (1994). Domain-specific knowledge and conceptual change. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 169–200). Cambridge University Press.

Fugelsang, J. A., Roser, M. E., Corballis, P. M., Gazzaniga, M. S., & Dunbar, K. N. (2005). Brain mechanisms underlying perceptual causality. Cognitive brain research, 24(1), 41-47.

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.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences of the United States of America*, 113(34), E5072–E5081.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, 73(2), 135–176.

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., & 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.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-Based Causal Induction. *Psychological Review*, 116(4), 661–716.

Hirschfeld, L. A., & Gelman, S. A. (Eds.). (1994). Mapping the mind: Domain specificity in cognition and culture. Cambridge University Press.

Hume, D. (1740). A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects.

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.

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.

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences of the United States of America,* 107(25), 11163–11170.

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.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences,* 16(2), 217–265.

Liu, S., Brooks, N. B., & Spelke, E. S. (2019). Origins of the concepts cause, cost, and goal in prereaching infants. *Proceedings of the National Academy of Sciences of the United States of America,* 116(36), 17747–17752.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language,* 17(2), 357–374.

Operskalski, J. T., & Barbey, A. K. (2017). Cognitive neuroscience of causal reasoning. In M. R. Waldmann (Ed.), The Oxford handbook of causal reasoning (pp. 217–242). Oxford University Press.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Spain: Cambridge University Press.

Pramod, R. T., Cohen, M. A., Tenenbaum, J. B., & Kanwisher, N. (2022). Invariant representation of physical stability in the human brain. *eLife,* 11, e71736.

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.

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

Schwettmann, S., Tenenbaum, J. B., & Kanwisher, N. (2019). Invariant representations of mass in the human brain. *eLife,* 8.

Slobin, D. I., & Bever, T. G. (1982). Children use canonical sentence schemas: A crosslinguistic study of word order and inflections. *Cognition*, 12(3), 229–265.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. In Developmental Science (Vol. 10, Issue 1, pp. 89–96).

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43, 337–375.

Woodward, J. (2005). Making Things Happen: A Theory of Causal Explanation. United Kingdom: Oxford University Press.

Yuan, S., & Fisher, C. (2009). "Really? She blicked the baby?": two-year-olds learn combinatorial facts about verbs by listening. *Psychological science,* 20(5), 619–626.