# Mechanism of learning and inference in the prefrontal cortex

**Shuyi Luo (shuyi.luo@psy.ox.ac.uk)**
Centre for Integrative Neuroimaging, Department of Experimental Psychology
University of Oxford, Oxford, UK

**Ali Mahmoodi (ali.mahmoodi1367@gmail.com)**
Centre for Integrative Neuroimaging, Department of Experimental Psychology
University of Oxford, Oxford, UK

**Matthew Rushworth (matthew.rushworth@psy.ox.ac.uk)**
Centre for Integrative Neuroimaging, Department of Experimental Psychology,
University of Oxford, Oxford, UK

## Abstract

**Despite causal relationships being inherently unobservable in a direct manner, people can infer these relationships with limited data. Using a one-shot causal inference task with fMRI, we investigated the inference process at both behavioural and neural levels. Our findings reveal that participants integrated observed causal evidence with their prior beliefs about the underlying causal structures that prevailed in the world to infer unobservable causal relationships. This process engaged a midbrain region linked to dopamine and learning but also a specific and circumscribed region of dorsolateral prefrontal cortex (dlPFC) in which activity was related to several aspects of the inference process.**



**Keywords:** causal inference; fMRI; dlPFC


## Introduction

Causal inference is a fundamental cognitive ability that enables individuals to learn about and predict the world. Humans are adept at causal reasoning even when given sparse, ambiguous data (e.g., Corlett et al., 2004). In this study, we hypothesised that the nature of inference is shaped by people's prior beliefs about underlying causal structures, which constrains the otherwise potentially unlimited ways of interpreting the observations (Tervo et al., 2016). To understand this process, we designed a one-shot causal inference task, in which participants inferred causal relationships based on ambiguous observations.

## Results

Participants (*n*=32) completed a one-shot causal inference task in a 3T scanner, involving four trial types (Figure 1A). Causal evidence was presented in a two-stage manner ("two dishes") with causal roles indicated by color frames learnt beforehand. Participants saw compound foods each comprising two elements (fruits) and were informed about whether they caused a symptom (an allergic reaction) or not. They then obtained more information about one of elements and then finally made an inference about the other element. Each trial included two rating phases, one before and one after the presentation of evidence about one of the elements (Figure 1C). Trials were structured to allow belief updates about causal structures (Figure 1B). Blocks consisted of 3–5 trials from a specific condition, with two randomly interspersed no-cause trials per block. To ensure one-shot learning, participants were informed that the same fruits would not reappear.

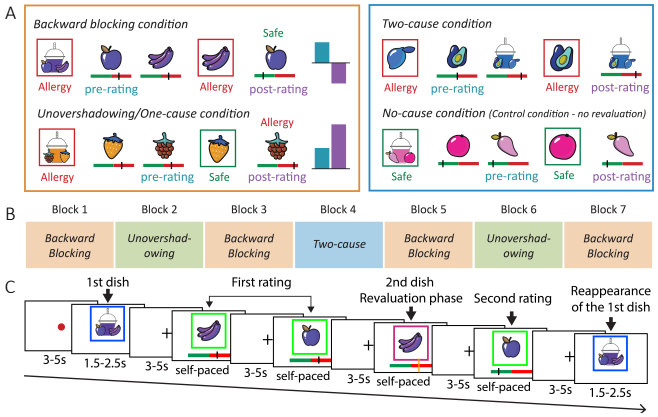


Figure 1: Task design

**Bayesian learner.** Causal induction can be considered a Bayesian inference process (Gopnik & Tenenbaum, 2007), wherein observed causal evidence is integrated with prior beliefs about underlying causal structures. We constructed a Bayesian model (Griffiths et al., 2011) to investigate how the task would be solved normatively. In this model, the learner holds prior beliefs about causal structures, denoted as $p(h)$, where $h \in H$. Additionally, we assume the updating of such beliefs. The simulation results illustrate how updating $p$(h) influences causal inference in the backward blocking condition, with the effect decreasing as the probability of a two-cause structure increases (Figure 2A).

**Behavioural results.** First, participants made revaluations of the causal roles of elements even while the elements were absent (Figure 2B). In the backward blocking trials, participants decreased their causal rating of the unexperienced fruit after learning that the other fruit from the compound alone caused symptoms. By contrast, in the unovershadowing condition, participants increased their causal rating of the unexperienced fruit after learning that the other fruit did not cause symptoms. Second, we compared backward blocking before and after two-cause trials to test whether these inferences shifted with participants' beliefs about causal structures. We found that, given the same causal evidence, participants' revaluation of the unexperienced fruit significantly decreased after two-cause trials (Figure 2C). The effect

was more salient for participants who initially held a stronger belief in one-cause structure (Figure 2D).

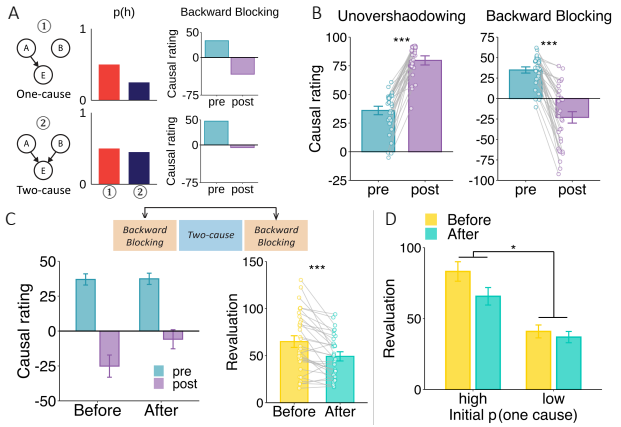


Figure 2: Behavioural results

**Inference during the revaluation phase.** We performed repetition suppression analysis to test whether participants made inferences about the unexperienced fruit while observing the outcome of the other item (i.e., revaluation phase). If an item's neural representation is active during the revaluation phase even when the item itself is absent then there should be suppressed activation upon its subsequent actual presentation. Consistent with this hypothesis, our results revealed significant repetition suppression effect in visual and visual association cortex regions when comparing post-rating to revaluation phases (Figure 3A). A representational similarity analysis (RSA) confirmed the existence of fruit-specific representations in both early visual cortex (EVC) and occipital fusiform gyrus (OFG; Figure 3B). We then performed an additional RSA, which revealed a stronger representation in OFG of the absent, unexperienced fruit during the revaluation phase (i.e., when showing the other fruit and associated outcome), when revaluation was more pronounced (Figure 3C).

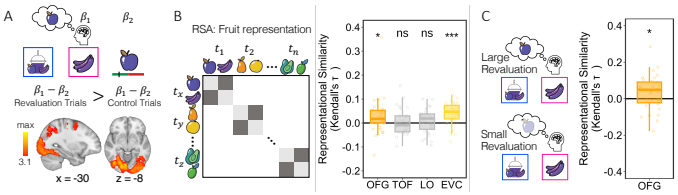


Figure 3

**Dissociative learning and inference in the brain.** During the revaluation phase, two distinct updates happened: an *experienced update* based on observed outcomes and an *inferred update* from reasoning about the unexperienced fruit (see illustration in Figure 4). We found that the experienced update is associated with the well-known prediction error related activity in the substantia nigra (SN) and ventral tegmental area (VTA; Figure 4A) and lateral orbitofrontal cortex (LOFC; Figure 4B), but these areas did not respond to the inferred update. Instead, the extent of the inferred update was linked to a distinct pattern of activity in dlPFC (MFG; Figure 4B). Moreover, dlPFC activation during revaluation was stronger in backward blocking trials before, rather than after, the two-cause condition (Figure 4C), consistent with the behavioural results (Figure 2C&D). More interestingly, our trial-wise analysis revealed that the stronger representation of the unexperienced fruit in OFG is linked to stronger activity in the dlPFC during revaluation phase (Figure 4D).

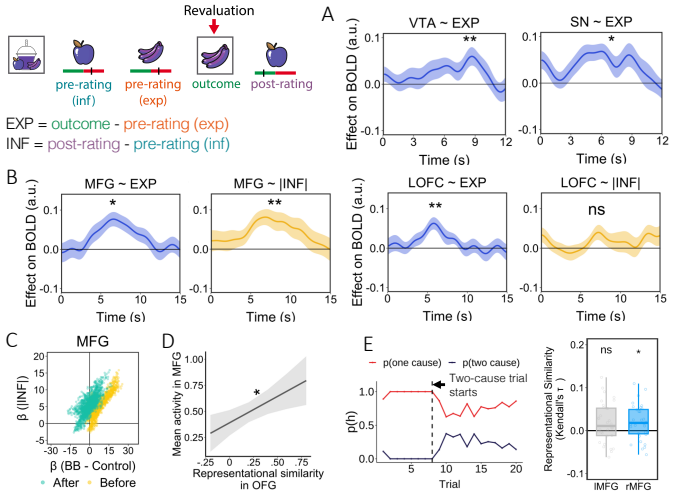


Figure 4: fMRI results

**dlPFC represents causal structure beliefs.** We are further interested in understanding how beliefs concerning causal structure are represented. We calculated participants' trial-by-trial causal structure beliefs ( $p(h),\ h \in H$ ) given their ratings of the unexperienced fruit. As well as many other aspects of inference, right dlPFC holds representations of causal structure beliefs (Figure 4E).

## Discussion

Participants inferred the causal role of elements in a compound even while they were absent. They did this by integrating the observation with their prior knowledge of the causal structures. Learning from observations is associated with prediction error-related activity in both midbrain and prefrontal regions. Inference from ambiguous observations is linked to dlPFC, which also carries information about underlying causal structure.

## References


Corlett, P. R., Aitken, M. R. F., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A. E., Robbins, T. W., Bullmore, E. T., & Fletcher, P. C. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron, 44*(5), 877–888. https://doi.org/10.1016/j.neuron.2004.11.022

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive development. *Developmental Science, 10*(3), 281–287. https://doi.org/10.1111/j.1467-7687.2007.00584.x

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and Blickets: Effects of Knowledge on Causal Induction in Children and Adults. *Cognitive Science, 35*(8), 1407–1455. https://doi.org/10.1111/j.1551-6709.2011.01203.x

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Current Opinion in Neurobiology, 37*, 99–105. https://doi.org/10.1016/j.conb.2016.01.014