UNIVERSITY COLLEGE LONDON

Department of Cognitive, Perceptual and Brain Sciences

MSc in Cognitive and Decision Sciences (2010-2011)

Thesis

Mechanisms of Active Causal Learning

(11,441 words)

Supervisor: Dr David Lagnado

Candidate #: CDS4

Abstract

Much existing causal learning research focuses on highly constrained problems, usually framed within familiar, generically causal, situations. This thesis describes an experiment examining how people break down realistically complex causal learning problems; how they learn without the support of contextual cues; and how these cues themselves are learned. Participants had to identify the causal structure underlying probabilistic patterns of node activations through freely selecting multiple interventions. Their intervention choices and resulting structural endorsements were measured against those of an efficient Bayesian active learning benchmark model. Successful participants were relatively systematic and efficient in their interventional patterns, and two underlying simple strategies were identified. These strategies were independently analysed, and participants were assessed for their coherence with them. The thesis concludes that causal-structural learning is likely to be based in simple action-selection and causal attribution mechanisms, and suggests that individual differences in these mechanisms may be responsible for deep differences in individuals' ideologies.

Contents

1. Introduction

How do we go about establishing causal explanations for the changes we perceive in our surroundings? This question is subordinate to more general questions of what our idea of causality is and what purpose it serves in cognition. For this reason it is necessary to begin this thesis by introducing some concepts central to the current understanding of the structure and function of causal beliefs in psychology, so as to motivate the questions investigated in the main experiment. **§1.1** argues that powerful causal learning and representational abilities are fundamental to both our ability to organise reality and to make inferences based on this organisation. **§1.2** explains why the causal Bayes net formalism has proved itself to be an invaluable graphical analogue, capturing the essential properties of causal beliefs. **§1.3** introduces intervention, the active element of causal learning. **§1.4** describes the 'rational' active learning framework used in later analyses, **§1.5** justifies this approach as a part of a broader hierarchical rational framework for inference from data. Finally, in **§1.6,** research questions are posed, and in **§1.7**, extensions to previous work are described before moving onto the experiment in **§2.**.

1.1 The Role of Causality in Cognition

Psychologically speaking, causal knowledge is the part of our conceptual organisation of reality which allows us to predict, control and explain the changes which we observe in our surroundings. Ascribing an underlying causal structure to observed patterns of events, and to relationships between objects, is fundamental to our ability to form expectations about the consequences of our interactions with the world, and to reason about the possible antecedents to observed data. If we did not make the assumption that we exist within an environment governed by an underlying (causal) structure, we would be unable to make inferences about, or form a conception of, a reality beyond whatever data we observe.

However, we manifestly *do* interpret our observations as constituted by objects and events, taking place in a structured reality. This is true for humans in a phenomenological sense, as well as the pragmatic sense applicable to a broader range of organisms, that our actions make sense in the light of causal-structural beliefs. Much of the recent causal representation literature has demonstrated that our causal representations are undeniably model-based, as opposed to merely associative (see for example: Cheng, 1997; Lagnado et al., 2007; Shanks, 2007; Sloman, 2005; Waldmann, Hagmayer & Blaisdell, 2006; Woodward, 2003). It follows from this that we must have a robust concept of causal structure, and be adept at identifying these causal structures in data. Investigating how we go about doing this is the subject matter of this thesis.

1.2 Causal Bayes Nets and Causal Representation

Causal Bayes nets (hereafter CBNs), initially adopted from engineering and other applied sciences (e.g. Pearl, 2000/2009), have become the dominant way of formalising causal information. This is because they easily capture the probabilistic, inference guiding nature central to our mental representations of causal structure. CBNs are probabilistically parameterised directed acyclic graphs made up of 'nodes' which represent variables or propositions and 'directed edges' which represent causal links between these variables (see Figure 1).

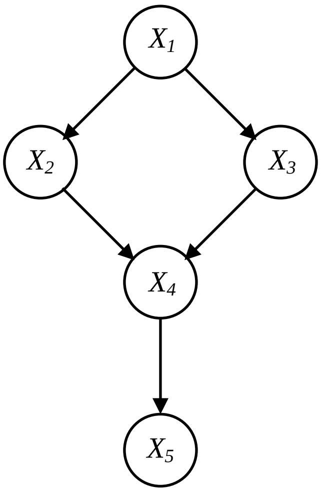


Figure 1. The directed acyclic graph for a simple 4 variable Bayes net. Adapted from (Pearl, 2009/2000)

By their definition, CBNs assert that all variables are independent of their non-descendants (other variables which are *not* direct or indirect effects of that variable) given their parents (any variable(s) which *are* direct causes of that variable). This is known as the 'parental Markov condition'

(Pearl 2000/2009). It captures mathematically what is intuitively the *role* of causal structure in conceptual organisation: if you know the state of all the direct causes of an event you don't need to know about anything else to predict if that event occurs. Rational inferences about the probability of states of variables, given the states of others, are easily computed from the joint probability distributions implied by a CBN:

$$P(X_1 \dots X_n) = \prod (P(X_i | \, parents(X_i))$$

Intuitively, the way in which multiple causes combine in producing effects is often not simple, and may depend heavily on specific knowledge of the *domain* in question. For example, taking two types of medication might actually *lower* your chances of recovery, whilst having an extra player on your side in a tug-of war might *disproportionately* increase your chance of winning. That said, the default intuition is that multiple causes normally have independent chances to produce their effects (Griffiths & Tenenbaum, 2009; Holyoak & Cheng, 2011). This is achieved mathematically with the 'noisy-OR' function, where the probability of the occurrence of an effect given its base rate of occurrence (B) and 'n' active causes, each with an independent probability of producing their effect (C_i), is given by:

$$P(E|B, C_1 \dots C_n) = 1 - (1 - B) \prod_i (1 - C_i)$$

Their probabilistic parameterisations allow CBNs to capture the subjectively *uncertain* nature (Laplace, 1814; Pearl, 2000/2009) of most of our causal beliefs. Causes in a CBN can work with only a certain probability, and variables can have a base rate of occurrence without being caused from within the system (i.e. due to exogenous causes beyond the scope or granularity of the model). This makes CBNs well suited to capturing approximate, partial, but robust, structural interpretations of the kind of complex, noisy data constitutive of sensory experience (e.g. Glymour, 2001; Gopnik et al, 2004). For these reasons, the use of CBNs to model causal representation garners the strengths of the

Bayesian approach to modelling cognitive processes, while capturing what is important about causal knowledge: its role in *simplifying* and facilitating reasoning from data (Krynski & Tenenbaum, 2007).

1.3 Intervention

The process of learning the causal structure underlying an encountered situation or domain is commonly divided into two stages. Passively observing the behaviour of a system allows learning about what patterns of dependency the system exhibits, while *actively intervening* to isolate and test subparts of the system, is required to isolate its unique and counterfactually predictive causal structure. It can be shown that observation alone is insufficient to identify a single causal-structural interpretation of data (see Steyvers et al, 2003), at least without the help of additional cues. In a mathematical sense, this is because there are always multiple ways of linking up and parameterising Bayes nets over sets of variables that are equally consistent with the same data.

In the simplest case, consider the covariation of two variables, A and B. Without any other cues to direction, this could be due to A's causing B or equally B's causing A. This could be true regardless of the exact frequencies of their respective activations and their degree of covariation. Two different structural interpretations would just yield correspondingly different parameters of causal strength and base activation rates. Similarly, three variables (A, B and C) could exhibit the same patterns of dependence and independence if linked in a chain (A→B→C) or with B as a common cause (A←B→C). This is because, in both cases, each variable is probabilistically dependent on both of the other two, but A and C become independent of each other conditional on B. These observationally equivalent structures form 'Markov equivalence classes' (see Figure 2) which can only be distinguished through additional information.

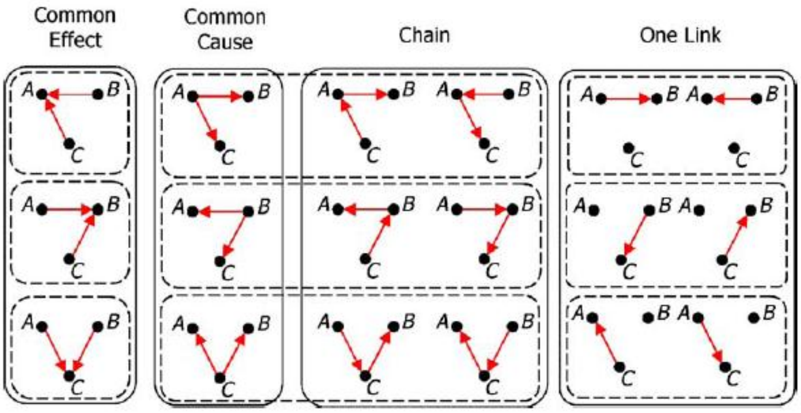


Figure 2. Taken from Steyvers et al, 2003 dashed boxes indicate structures are part of the same equivalence class.

For the majority of practical cases, we have strong cues to causal direction. Most commonly: temporal ordering (Lagnado & Sloman, 2006), knowledge of the causal roles of objects (Kemp, Goodman & Tenenbaum, 2010), or the implications of their spatial configuration, make certain causal interpretations of situations much more natural than others. But how are the validities of these cues established in the first place? And by what criterion? It is circular to define causation by its hallmarks, as these hallmarks will need, in turn, to be shown to apply only to actual causal structures to avoid trivialising the concept of causation.

Imagine trying to convince someone that it is the flagpole that causes its shadow and not visa-versa (Bromberger, 1966). Intuitively, the way to do this is to demonstrate that changing the flagpole – e.g. shaking it, or taking it down – affects its shadow, while changing the shadow - e.g. shining a light to blank it out - does *not* affect the flagpole. This is the essence of active learning about causation. Reaching into a system and fixing parts of it and/or holding parts of it constant, is the basic exploratory move for any embodied agent who needs to establish what depends on what.

In the literature this is known as an 'intervention'. In terms of the CBN formalisation, it can be represented by 'link breaking' graph surgery (see Figure 3) whereby the links to an intervened on variable are severed and its value is fixed exogenously. It has been shown that people are much better at identifying causal structures when they are allowed to intervene on them (e.g. Lagnado & Sloman, 2002, 2004; Hagmayer et al, 2007). Further, it has been shown that people are *sensitive* to the differences in the implications of natural patterns of activations versus those resulting from active interventions (Sloman & Lagnado, 2005). This sensitivity is, arguably, a necessary part of reasoning and inference (Pearl, 2000/2009) as *imagining* the effects of interventions on our mental causal graphs allows for counterfactual thinking.

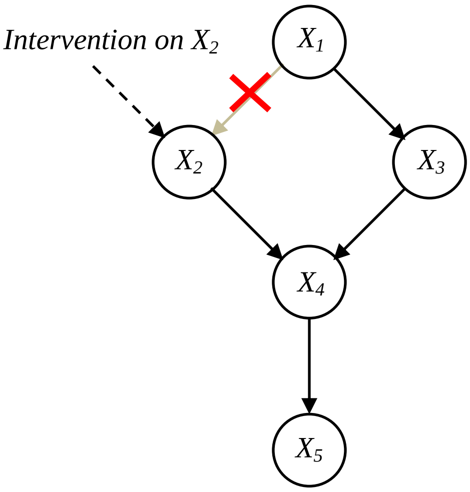


Figure 3. A depiction of an intervention on the DAG from Figure 1. The link from X_1 to X_2 is broken and X_2's value is fixed by the agent from outside the domain of the graph as indicated by the dashed arrow.

The interventions described above are known as 'ideal' or 'link breaking'. Meder et al. (2010) identify four other weaker intervention types which exert only probabilistic influences on their targets, under the hypothesis that this is more typical of many of the interventions we perform in everyday life. However, even weak interventions are sufficient to break the deadlock between Markov equivalent causal-structural interpretations of data, while 'ideal' interventions are easier to model and provide clearer experimental results and so will be the subject matter of this thesis.

1.4 The Bayesian 'Benchmark' for Active Causal-Structural Learning

Different interventions on a system vary in their effectiveness in distinguishing between possible structures. If we define an optimal intervention to be one which has the leads to the largest *expected reduction in uncertainty* about the true underlying causal structure given data and what is already known, it is possible to calculate optimally efficient interventions for solving simple problems by using Bayes theorem and information theory (e.g. Tong & Koller, 2001). These optima can then be used to assess the performance of different learning strategies and the actions of human learners.

Using information entropy (Shannon, 1951), we can quantify our uncertainty (*U*) about the true causal structure for a problem as

$$U = -\sum_{i=1}^{n} P(H_i) log_2 P(H_i)$$

where $H_{1\ldots n}$ is a hypothesis space consisting of *n* distinct causal structural hypotheses consistent with the number of variables in the current problem. *U* is a measure of how much we *do not* know about the underlying system. The more peaked a probability distribution across hypotheses, the smaller the value of *U,* and the more certain one can be that the structure with the maximum posterior probability is the actual structure.

For the causal learning problems given to participants in the experiment described in this thesis, a model was devised which used Bayes theorem to calculate the change in the probability distribution across possible structures, given *every* possible intervention (I_j) and *every* subsequent pattern of activation A_k of the unfixed variables given that intervention. To do this, the *likelihood* of each 'intervention-activation' pattern was calculated, using the known parameters of the underlying CBN then summed over hypotheses to provide the denominator:

$$\sum_{i=1}^{n} P\left(A_k \middle| I_j, H_i\right)$$

Finally, the model calculated the posterior probability distribution over hypotheses using Bayes theorem:

$$P\left(H_i \middle| I_j, A_k\right) = \frac{P\left(A_k \middle| I_j, H_i\right) P(H_i)}{\sum_{i=1}^{n} P(A_k | I_j, H_i)}$$

. It is possible to estimate CBN parameters contemporaneously with learning structure (Nyberg & Korb, 2006). But, because the focus of this thesis is structure learning, this was not necessary. Instead, consistent parameters were used which participants were made aware of from the start.

Flat priors were used at the beginning of each problem. A minimum description length prior (Lu et al., 2009, Meder et al, 2010) was considered to reflect the common assumption (e.g. Sprites, Glymour & Scheines, 1992/2000) that we are biased to endorse sparse but powerful causal structures[1]. Ultimately though, it was decided that participants in a causal learning experiment would expect a variety of link numbers and structural patterns, making a flat prior a relatively rational choice.

Each intervention resulted in the observation of one of a subset of possible activation-intervention patterns, each with an informational value for how much they would change the model's uncertainty ($\mathit{\Delta U}$), about the true structure and a likelihood for observing exactly those activations given the intervention ($P(A|I)$). From this, it is possible to calculate the *expected* reduction in uncertainty (*EU*) of an intervention, using an analogue of the normal expected value formula:

$$EU_j = \sum_{k=1}^{n} \mathit{\Delta U}_k P\left(A_k | I_j\right)$$

and the definition of uncertainty given above.

Thus, on each trial, the benchmark model returns a vector of expected information gains (one for each intervention) and simply picks the highest one. If more than one intervention has the same value, it picks from them at random. The model is myopic, in the sense that it always opts to learn as much as possible with each trial, as opposed to planning series of interventions tactically, like a chess player. However, the model learns very quickly by being information greedy, and at least in the case of the experiment discussed in this thesis, the probability space shifts so much on each observation that it is intuitively unlikely that looking ahead would often lead to a different interventional decision.

Even with myopia, such a model is extremely computationally intensive. It must compute, and weigh up, the informational implications of every possible outcome before it picks a single intervention. Therefore, it should be stressed that this model is intended only to provide a pseudo-rational, or normative, *benchmark* for assessing the performance of participants in the experiment, rather than a seriously considered model of actual active learning.

[1] Such a bias makes intuitive sense when you consider that structuring data causally is all about simplifying

1.5 A Hierarchical Framework for Rational Causal Representation

It was noted in **§1.3** that adult causal judgements are typically strongly influenced by prior knowledge of temporal and spatial cues to causal structure. The lack of an account of the relationship between our use of these cues and of covarational and interventional information has long been a weakness of formal models of causal induction (Lagnado, 2011). Recently though, a general purpose framework for rational inductive learning and representation has been proposed (Goodman, Ullman & Tenenbaum, 2011; Kemp, Goodman & Tenenbaum, 2010; Griffiths & Tenenabum, 2009), which allows a hierarchical extension of the rational framework of **§1.4** to potentially *encompass* the use of these other elements.

The framework assumes that our causal knowledge consists of beliefs about causal structure, spanning many (interacting) levels of abstraction. Thus we not only have *specific* beliefs about relationships between *token* objects; but also: beliefs about the *roles* of object *types* in causal structures; beliefs about the *types* of structures which typically underlie *broad domains* or classes of events; and more abstracted beliefs, right up to a '*universal theory'* made up of our beliefs about the nature of causality itself (Figure 4). The hierarchical application of Bayes theorem to each of these levels provides a rational framework, showing how knowledge about causation, at increasingly abstract levels can be learned inductively from observed similarities and patterns across multiple causal learning instances. This high level causal knowledge then allows us to use the perceptual or spatiotemporal properties of situations as cues to their likely causal structures: Essentially, we learn to let higher level beliefs guide lower level causal inferences, by constraining the space of possible or likely causal structures underlying observed data.

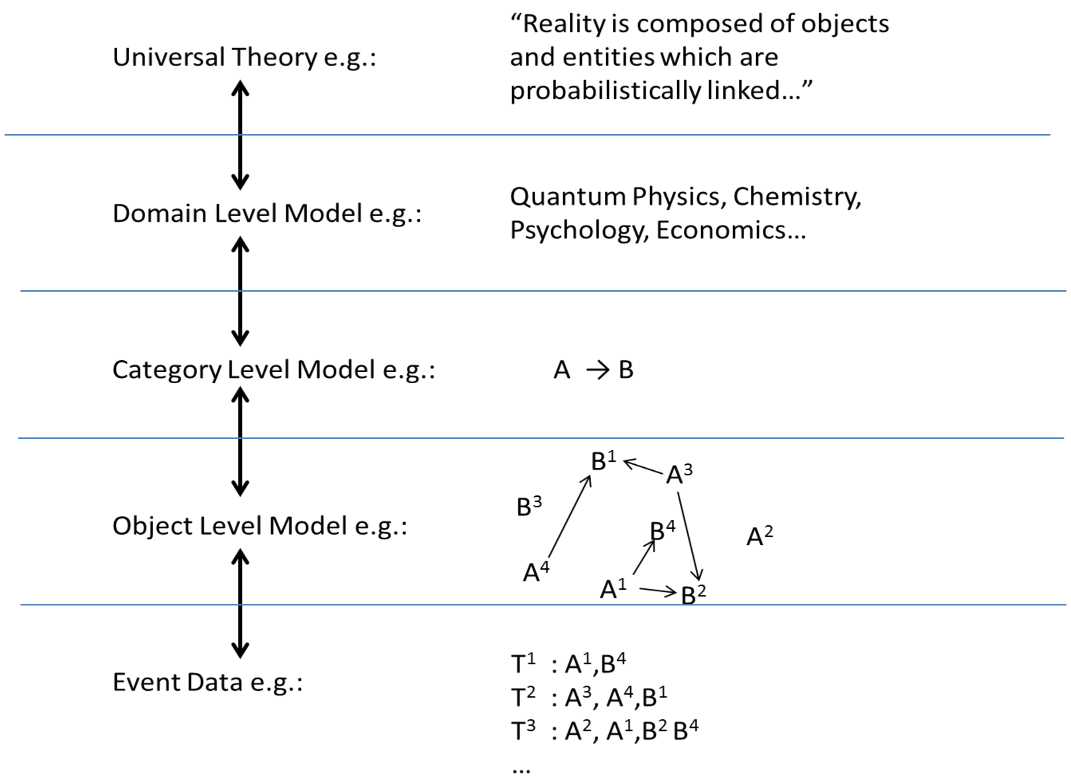

Universal Theory e.g.:
"Reality is composed of objects and entities which are probabilistically linked…"
Domain Level Model e.g.:
Quantum Physics, Chemistry, Psychology, Economics…
Category Level Model e.g.:
A → B
Object Level Model e.g.:
B1
A3
B3
B4
A2
A4
A1
B2
Event Data e.g.:
T1 : A1,B4
T2 : A3, A4,B1
T3 : A2, A1,B2 B4
…

Figure 4. A schematic of proposed hierarchy of our causal knowledge. Parameter estimates and generative probabilities can be included at the different levels but are left out here for simplicity. Note the 'category level' generalisation that As cause Bs. It is clear that such knowledge could *guide* object level inferences about subsequently encountered As and Bs. This is the essence of the framework.

This approach is one wing of a broad project reconceptualising of the mind as domain general, hierarchical probabilistic inference engine (Tenenbaum, et al, 2011); showing that it is *possible* for a cognitive agent to grow its entire system of beliefs about the world organically, by gradually discovering patterns in observed data at increasing levels of abstraction. The framework is undeniably powerful as it provides an intuitive solution to perennial problems of a priori knowledge in philosophy (such as Kantian a priori belief that all events have causes (Kant, 1781/1965)) and native knowledge in developmental psychology (e.g. Chomskian universal grammar (Chomsky, 1975)). However, the Bayesian learning algorithms used in the hierarchical framework are only intended to provide a *computational level* solution to the problems faced by cognitive agents (Marr, 1982) as they are intractable for all but very constrained learning problems (Jones & Love, 2011).

The parallel architecture of the brain lends support to the idea that many of our inferences take place over *some kind of* layered, distributed, and probabilistic representations, in the way that this framework suggests (Tenenbaum, Griffiths & Kemp, 2006). That said, it does *not* follow that we update our complex networks of beliefs given new data, in anything like a Bayes-optimal way. To do so would require that we effectively reweighted every neural connection on the basis of every new datum. While it has been shown that we are good at picking up on basic statistical information, like frequencies of covariation, *implicitly* (Lewicki, Hill & Czyzewska 1992), it has already been argued that, by being model based, causal knowledge requires something more than this. Potentially, we develop our sophisticated semantic networks by building up and amending our beliefs *locally* with simple but robust active mechanisms for or intervening on, or targeting attention to, stimuli we deem likely to resolve simple questions which we might have about the nature of our environment. Our implicit sensitivity to *basic* statistical information probably feeds this process but, arguably, some more explicit, focused organisational mechanism is also required. Do we learn at multiple of these hierarchical levels simultaneous, or one at a time, or do we abstract post hoc our higher level beliefs on further reflection? These dynamics are largely unknown. The experiment in this thesis is intended to examine what people actually *do* do when learning about causal structure so as to get a clearer picture the mechanisms involved.

1.6 Summary and Research Questions

So far this thesis has argued that:

1. We have causal beliefs in order to predict, control and explain our surroundings.

2. This necessitates that causal knowledge for humans is model-based, structural and probabilistic in form.

3. Causal Bayes nets seem to capture essential properties of causal knowledge, and by extension, data generated by a causal Bayes net has a natural causal interpretation.

4. Causal learning is fundamentally an *active* and embodied process of targeted interventions on encountered phenomena in order to encounter their counterfactually reliable covarational patterns.

5. Mechanisms of causal structural induction can be assessed against a ‘rational’ benchmark active-learning model, based in information theory and Bayes theorem.

6. The learning, and use of, non-statistical cues to causation can potentially captured by extending such the Bayesian framework hierarchically, but that the *basic* mechanisms of causal

learning must be identified by carefully controlling the available evidence at different hierarchical levels.

In the light of the above, the experiment in this thesis was designed to:

1. *Investigate the basic mechanisms of active causal learning used by humans when spatiotemporal cues are absent*.

2. *Investigate how learning a simple causal cue interacts with structure learning.*

1.7 Extensions to Previous Work

In order to investigate the above questions, a formal and abstract causal learning task was formulated, based loosely on the 'recover the graph' paradigm used by Lagnado and Sloman (2006), Kushnir et al (2009) and the rational framework of Steyvers et al (2003). This paradigm essentially involves asking participants to identify the structure of a CBN based on observing its natural and interventional patterns of activation. The introduction of a minimal cue follows from a previous study by Kemp, Goodman and Tenenbaum (2010). The experiment represents extensions of these paradigms in the following respects.

Kushnir et al. (2009) and Lagnado and Sloman (2006) looked only at learning of three and four node causal structures, and in the former case only by way of observational information. Their learning problems were framed within realistic causal scenarios (a hidden mechanism linking sticks in a box and the transmission of viruses around a computer network) meaning that prior knowledge was likely to play a large part in inferences. Participants were asked to complete the causal structures by drawing links between these objects on a paper diagram in the case of Kushnir et al, and by answering multiple specific questions in the case of Lagnado & Sloman.

In this experiment, problems ranged from simple to complex with up to six nodes and five causal links to examine how learning scaled up. Causal learning problems were deliberately posed *without* the provision of realistic scenarios to avoid, as far as possible, the influence of priors. Participants *were* able to perform interventions, and the endorsement of causal links between nodes took place *in* the interactive computer environment during the task, rather than afterwards, to minimise memory effects.

The decision to assess intervention choice using information theory follows from the work of Steyvers et al (2003) who measured the relative informativeness of a single simple intervention choice (i.e. fixing one node one time only), made after a series of forced observations. Their experiments also presented the causal learning tasks framed with a cover story (mind reading aliens).

This experiment also extends on this work by allowing participants a free rein to perform simple *and* complex interventions; to do so over as many trials as desired, and to do so without a fixed observational period at the start. This was done to investigate what *patterns* of contemporaneous and sequential interventions people might use to uncover a causal structure, how much they would try to achieve on each trial and to what degree they would break the problems down into simpler parts. This extension also allows more sophisticated measures such as interventional efficiency over time and the cumulative results of multiple interventions.

Finally, Kemp, Goodman and Tenenbaum (2010) have demonstrated that people can learn about abstract or domain level properties of data, and subsequently use this knowledge in inferences. What has not been investigated is how having a domain level cue available affects active learning, how attention is divided between cue learning and structure learning, and how knowing a cue affects intervention choices and subsequent causal ascriptions. The addition of a minimal domain-level cue in the second half of this experiment provides an opportunity to investigate some of these questions.

2. The Experiment

This section describes the causal inference experiment. Full details of the procedure are in **§2.1.4** followed by model predictions in **§2.2** and results and analyses in **§2.3**.

2.1 Methods

2.1.1 Participants

24 members of the UCL psychology subject pool (12 male) took part in the experiment. They were paid between £4 and £6 based on their performance. The mean age was 26.7 (SD = 6.6) and the range was 19 to 47. Four participants were excluded from the analysis because it was clear from their behaviour in the task that they had not fully understood the instructions.

2.1.2 Instruments and Materials

The main experiment was programmed in MATLAB Version R2010a, and the interface utilised MATLAB's Psychophysics Toolbox Version 3. This was followed by a paper-based questionnaire.

2.1.3 Design

Participants all saw the same four blocks of five problems. Problems were randomised within blocks but the order of the blocks was the same for all participants.

2.1.4 Procedure

Instructions were included within the main program (see Appendix **§7.1** for task instructions). Participants were instructed to read these instructions in their own time before beginning the task. There was no practice round due to the importance of recording participants' actions when facing the task for the first time. However, participants were tested on their understanding of the instructions by the experimenter before commencing the first block. Instructions which had not been fully understood would then be reinforced by the experimenter if necessary.

The aim of the task was to establish and endorse the causal structure underlying patterns of activation a group of 'nodes' which appeared as filled coloured circles on the screen. Each problem had a different underlying causal structure. The program determined which of these 'nodes' would activate on each trial according to a causal Bayes net with a structure corresponding to the block and randomised problem number (see Figure 5 for example some problems and Appendix **§7.2** for the complete list). The parameters of these causal Bayes nets were always the same: each node activated at random with a probability of 0.1 unless it was 'fixed' by the participant (see below). Where there was a causal link in the underlying structure, an active cause node would have a 0.9 probability of *causing* the activation of its effect node. Participants were told these parameters as percentages in the instructions. Causes were always generative (i.e. they always raised the probability of their effects). Multiple causes were integrated by the generic noisy-OR equation. The underlying structures were always acyclic and each trial was independent of previous trials. Block 1 problems all had three nodes, block 2 problems had four and blocks 3 and 4 had six nodes each.

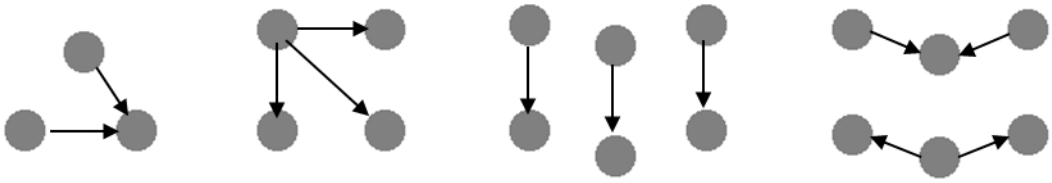


Figure 5. The hidden structures of problem 2 for blocks 1-4 (L to R)

2.1.4.1 Performing Trials

In each problem, participants were initially presented with a white screen on which between 3 and 6 grey filled circles would be scattered, representing the nodes (Figure 6). These circles were positioned randomly, to control for any directional bias toward endorsing causal links running in the directions in which they are typically depicted in diagrams, (i.e. left to right or top to bottom). In each new problem, the circles would be drawn in new random positions. This was done to make it clear to participants that they were distinct from the ones in the previous problem.

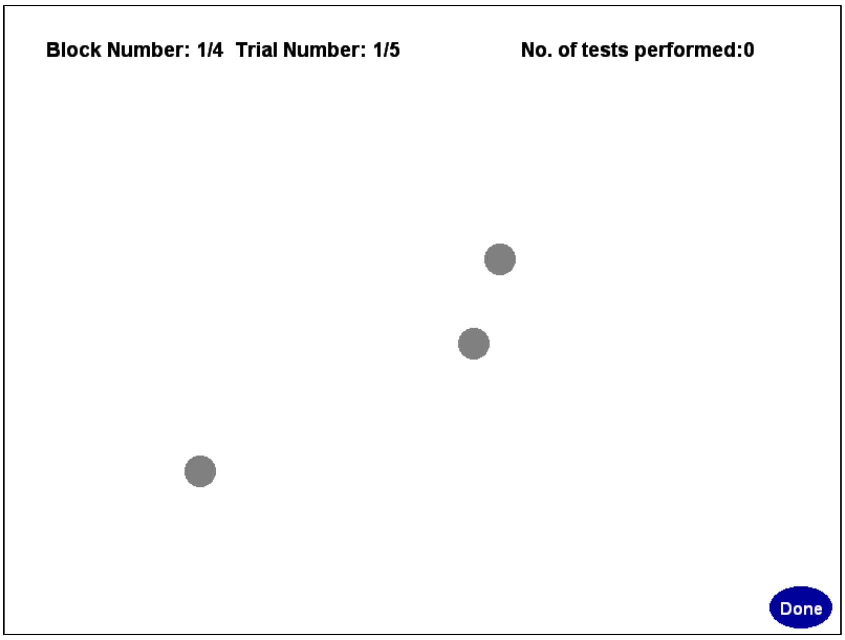


Figure 6. User interface for a problem in block 1.

Participants tested the network with the spacebar. Each press constituted one *trial*. The result of each trial was a 1.5 second period in which 'activated' nodes would turn red while the others would remain grey. After each trial, the program returned to the previous screen where all the nodes were grey.

Between trials, participants were able to 'fix' nodes. Each node could be fixed either to be 'definitely active' or 'definitely inactive' on subsequent trials. It was indicated on screen that a node was fixed, by the addition of a thick black circle surrounding it and 'fixed active' nodes were permanently red, while 'fixed inactive' nodes were permanently grey (Figure 7). Initially, all the nodes were 'unfixed'. Each time one was clicked it would cycle one step through its three states. Nodes did not reset to 'unfixed' after each trial but stayed in their selected state until they were clicked on again. This was to avoid biasing participants to perform more interventions with few or no nodes fixed.

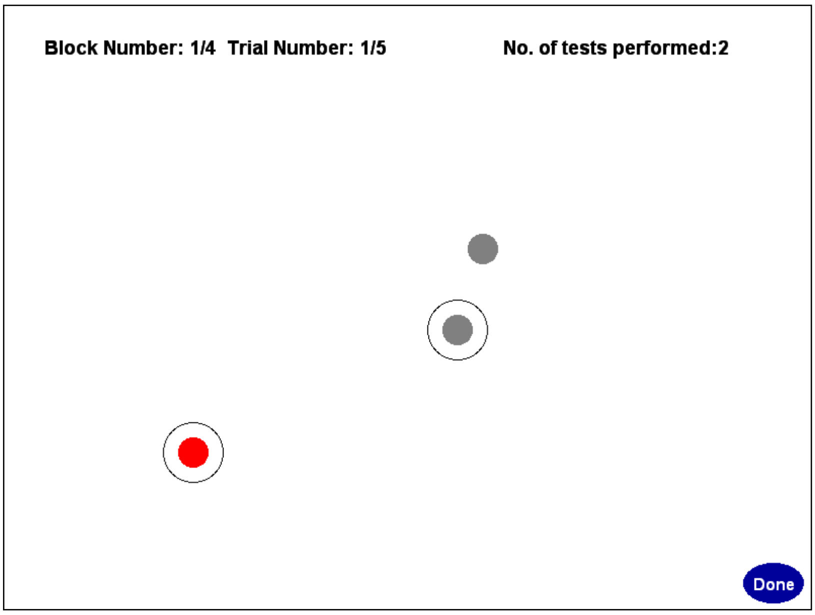


Figure 7. User interface showing one node fixed ‘active’ (left); one node fixed ‘inactive’ (middle) and one ‘free’ (right)

Fixed nodes behaved as ‘ideal’, or ‘link breaking’, interventions. The program can therefore be thought of as computing the activations of only the unfixed nodes on each trial, based on an underlying causal Bayes net, altered by removing links into fixed nodes and changing their base rates to 1 for ‘fixed active’ or 0 for ‘fixed inactive’ respectively.

After each test of the network, the program would loop back to the screen in which participants were able to fix and unfix nodes as they liked. This process would continue until participants indicated that they thought they had identified the causal structure underlying the behaviour of the nodes by pressing a button labelled ‘Done’ in the corner of the screen.

2.1.4.2 Endorsing a Structure

Participants were instructed to endorse the causal structure which they thought was correct by clicking in the spaces between pairs of nodes on the screen (Figure 8) to add directed arrows as they went along. In this way, participants would effectively ‘draw’ the structure which they thought was responsible for the data, link by link. One click between nodes made an arrow appear, a second click reversed its direction, and a third removed it again. The initial direction of the arrow was randomised to control for participants to select the first direction to appear. Where the spaces between pairs of arrows overlapped, clicks in the overlapping space would cycle randomly through the different pairs.

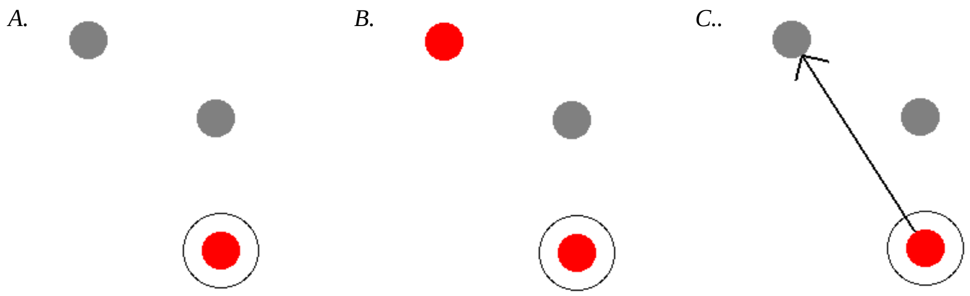


Figure 8. A. The bottom node is fixed to 'active'. B. The spacebar is pressed and the top node activates. C. The user decides to endorse a link from the bottom to the top node.

2.1.4.3 Feedback

After each problem there was a feedback screen which displayed the true causal structure overlaid with participants' endorsed arrows. The causal links that the participants correctly endorsed would go green, arrows which they incorrectly endorsed would go red and arrows which they missed would now appear in grey (Figure 9).

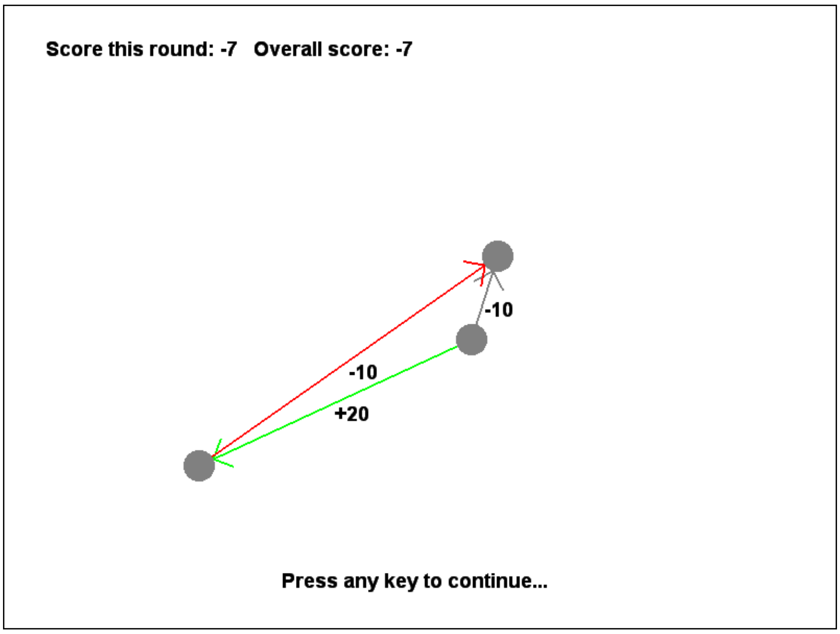
Participants also received a score for each problem and were able to see their aggregate score over the experiment so far. Their score for each problem/ was computed as follows:

+20 points for each correctly endorsed arrow

-10 points for each wrongly endorsed arrow

-10 points for each missed arrow

-1 point for each trial taken



Score this round: -7 Overall score: -7

Press any key to continue...

Figure 9. A typical feedback screen from the experiment.

The scoring system was designed to incentive participants to try to be as accurate and efficient. At the end they were paid £4 plus an additional penny for every three points scored, capped at £6 for scores over 600.

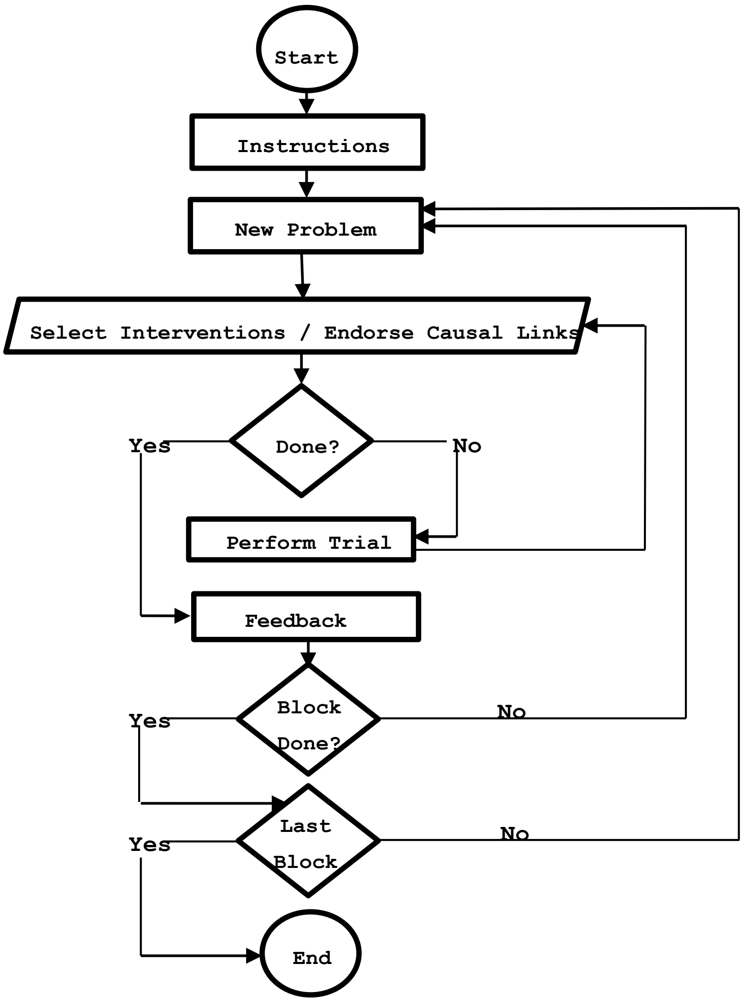
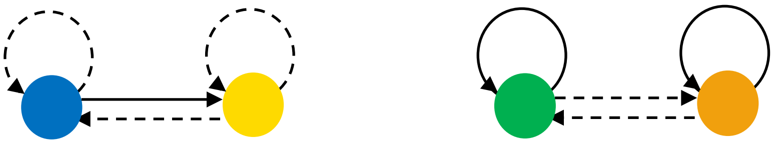


Figure 10. A flowchart of the program used in the experiment

2.1.4.4 Colour Cues

Blocks 3 and 4 also differed slightly from blocks 1 and 2. Instead of the inactive nodes being grey, they were one of two colours (see Figures 11 and 12 and Appendix **§7.2**). In block 3 there were always 3 blue and 3 yellow nodes, and in block 4 there were always 3 orange and 3 green nodes. Regardless of colour all nodes still turned red when activated. Participants were told after block 2 that their task would be the same as before but that: "[they] may find the colours helpful and may be asked about them afterwards."



Figures 11 & 12. The two colours of node in blocks three and four. Solid arrows indicate the types of causal link that were present while the dashed arrows indicate those which were absent.

The colours actually corresponded to the underlying structures in such a way as to provide a potentially helpful category level cue (see Figure 3 and **§1.5**). In block 3, causal links always originated from blue nodes and acted on yellow nodes; there were no causal links between blue nodes, between yellow nodes or from yellow nodes to blue nodes. In block 4, causal links always ran between nodes of the same colour, so there were no causal links from green to orange or from orange to green.

2.1.4.5 Coding the Questionnaire Responses

At the end of block 4, participants were thanked for participating and the program ended. They were then given a very short questionnaire (Appendix **§7.4**) to complete which asked them to describe how they thought they had solved the task; to describe any specific strategies they had used (including a diagram if possible); and to describe the roles of the colours in blocks 3 and 4 if identified. Responses were also coded in the following way:

By 'Strategy'

0 = Participants who reported having no specific strategy, or left the strategy question blank.

1 = Participants who reported having and described a strategy for solving the task.

Those who were coded as '1' were later recoded based on the emergence of two distinct strategies one or other of which was described by every participant coded as a '1'. These are described in detail in the results, but ultimately '1' was used for the simpler strategy and '2' for the more complex one.

By Colour Cues:

The other two questions on the questionnaire asked whether participants found the colours helpful in blocks 3 and 4. If they answered 'yes', they were asked to describe the roles of the yellow and blue nodes in block three, then green and orange nodes in block four. Participants were coded based on whether they correctly identified the limiting roles of the colours (links from blue to yellow only in block 3 and links within colours in block 4). The codes used were:

0 = Unable to describe the role of either colour/Answered ‘no’ to finding the colours helpful.

1 = Able to describe the roles of the colours in block 3 only.

2 = Able to describe the roles of the colours in block 4 only.

3 = Able to describe the roles of the colours in both block 3 and 4.

2.2 Benchmark model predictions

The number of possible structures increases super exponentially with the number of variables involved with: 24, 542 and 3.78 million permissibly directed acyclic graphs with 3, 4 and 6 ‘nodes’ respectively. There were 9, 65 and 665 possible intervention patterns in blocks 1, 2and 3/4 respectively, each with 2^n possible activation patterns where ‘n’ is the number of unfixed nodes in that intervention As an example, an intervention which fixes one node to ‘on’ in block 1, leaves the other two nodes unfixed, and so makes a set of four possible ‘activation-intervention’ patterns (Figure 13), each with a distinct likelihood under each hypothesis. Interventions in which all of the nodes were fixed were not included because they always and trivially had an informational value of zero. The structure in which none of the nodes were causally connected was also not included in the hypothesis space because the task instructions implied that there would be at least one causal connection in each problem/structure. The superexponential[2] increase in the calculations required to select interventions, meant that it was not possible to fully model blocks 3 and 4. After one week of processing on a high-end PC (8-core, 12 Gb RAM, Linux Desktop) the model had only just selected its *first* intervention for the for problem 1 in block 3. Formal analysis using the benchmark model is therefore mainly restricted to blocks 1 and 2.

[2] Approaches exponentiality from above, i.e. $f(0) = 1, f(g)f(h) \leq f(g.h)\ \forall g, h \geq 0$

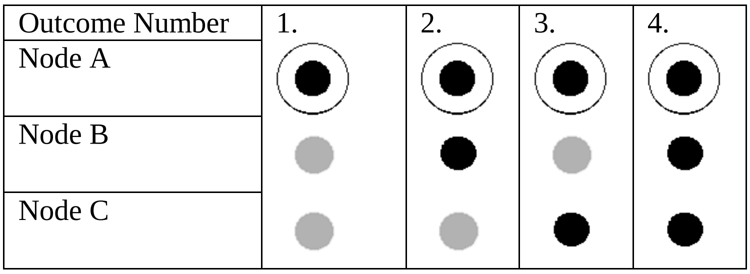


Figure 13. Each column depicts schematically a possible outcome of an intervention which fixes node A to be definitely active. In a general sense seeing nodes B or C activate is evidence for the existence of causal links from A to one or both of them.

To establish the 'informativeness' of different interventions for learning about the structure of a causal system, the benchmark learner model was tested 1000 times on the three and four node problems. This allowed the frequency with which different interventions were selected to be measured, as well as the typical patterns of interventions and corresponding shifts in the probability distribution. In these tests, the model was parameterised to keep performing trials until the posterior probability of the likeliest structure was greater than 0.9 and then to endorse that structure. Uncertainty typically decreased sharply in the last few trials so the model is robust to small changes in this value. The results of these tests are summarised in Table 1.

2.2.1 What Interventions did the Model Select?

For three-node learning, *fixing one active node* at a time was the most frequent choice of the benchmark learner model at (70.58%; Table 1). The only other type of intervention frequently chosen was one with *one node fixed on and another fixed off* (28.77%). For four node problems, the modal intervention type was to fix *one node on and one off* (45.58%, Table 2) followed by fixing just *one on* (36.89%). However, also selected were: '*two on, one off*' (12.35%), *'one on two off'* (4.51%) and '*two on'* (0.67%). While it was not possible to do this for six node problems, the vector of expected values (Table 3) that was computed for the first intervention gives an idea. Fixing '*one on'* was the most informative move from a flat prior but '*one on, one off*', '*one on, two off*' and *'two on, one off'* followed close behind suggesting that they were also likely to come into play as the probability distribution evolved.

Possible Intervention Types For Structures With Three Nodes	No. Possible Outcomes	Expected Value For First Trial (Bits)	Frequency Chosen By Bayesian Learner (%)
All three nodes free	8	0.2016	0
One node fixed on, two nodes free	4	1.0165*	70.58*
One node switched off, two nodes free	4	0.0677	0
One node on, one node off, one free	2	0.4872*	28.77*
Two nodes off, one free	2	0.0025	0

Table 1. The expected value column shows how informative these interventions are for updating from a flat prior (so on the first trial for each problem). As the probability distribution becomes more peaked, these values shift around, which is why the model does not always select the same intervention.

Contrary to the common claim (e.g. Steyvers et al, 2003), that starting with passive observations is a rational way to initially narrow down the hypothesis space of possible structures, in this task, the 'nothing fixed' intervention was *never* selected. This was due to the high probability of nothing activating by chance (.73 block 1, .65 in block 2 and .53 in blocks 3 and 4). The chance of nothing activating, if everything is unfixed, is given by the product of the probability that each node fails to activate at random, (e.g. 0.9^n where *n* is the number of nodes). When nothing activates, nothing can be learned about the underlying structure, making passive observation a relatively uninformative move, at least when intervening carries no additional cost.

Intervention Types *Chosen* By Bayesian Learner For Structures With Four Nodes	No. Of Possible Outcomes	Expected Value For First Trial (Bits)	Frequency Chosen (%)
One node on, three free	8	1.3094*	36.89%*
One node on, one off, two free	4	0.7749	45.58%*
Two nodes on, two free	4	0.9550	0.67%
Two nodes on, one off, one free	2	0.3825	12.35%*
One node on, two off, one free	2	0.2741	4.51%

Table 2. To save space, this tables only shows the interventions which were ever selected by the model for four node problems. Five other, generally less informative, interventions are excluded.

2.2.2 When Were the Different Interventions Used and What Role Did they Play?

To try to illustrate this dynamic, Figures 14 and 15 show two typical patterns of the model's intervention choices based on its observations. Two of the problems in block 1 were selected as exemplars of this inferential process. On average the model took 6.4 trials to reach 90% certainty about the three node networks, and 11.0 for four node networks.

2.2.3 Examples

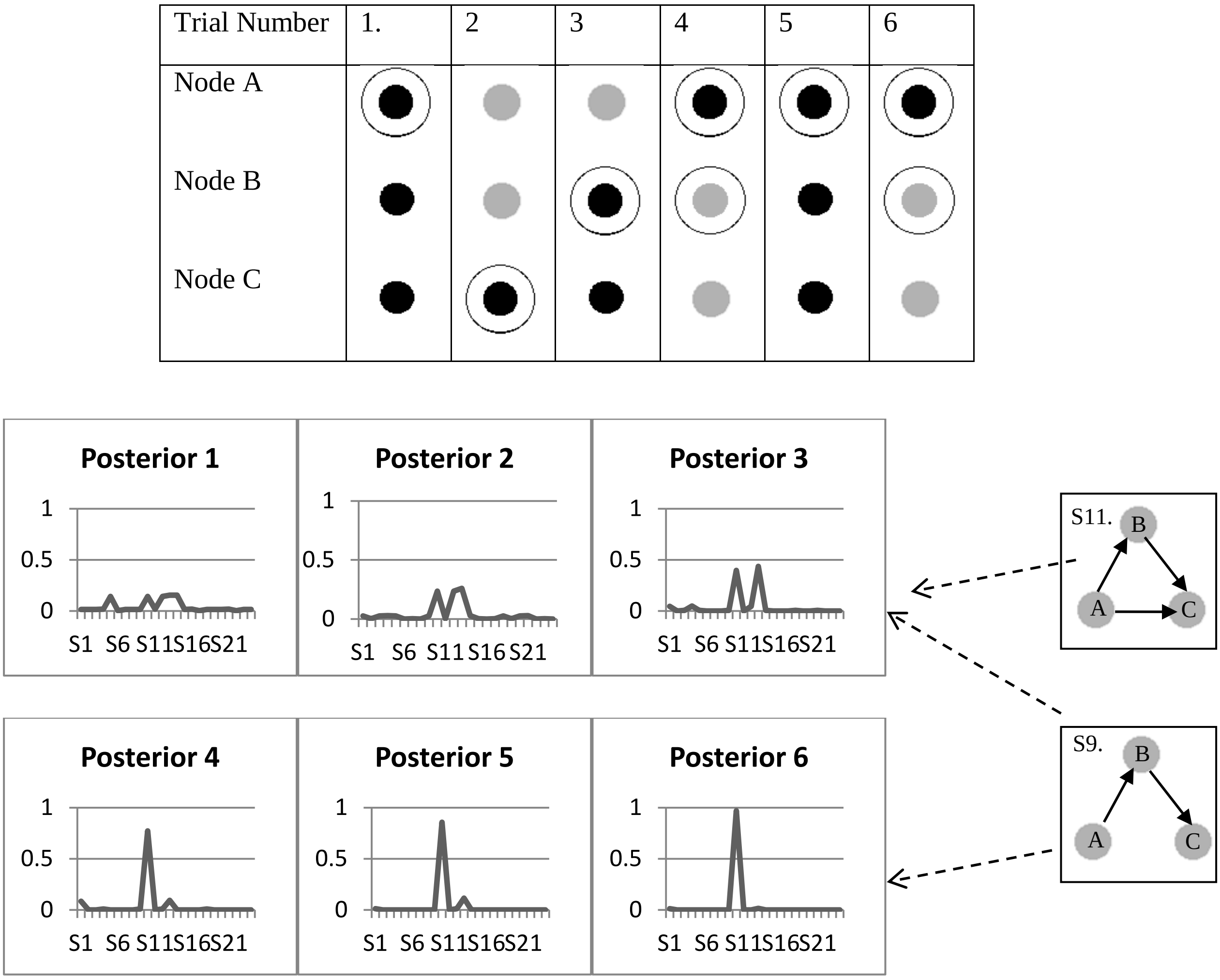


Figure 14. Schematic of interventions and resulting activations. Each column is one trial showing intervened on nodes (with black rings) and their outcomes.

Step 1.On the first trial the model first fixes a node as ‘definitely active’. It selects this node at random because all are a priori equal. Both other nodes activate. As a result those structures with direct or indirect links from A to B and or C become more probable (Posterior 1). Steps 2. – 3. do the same with nodes C and B. The activation of C given B makes a direct link from B to C more probable (Posterior 3) which is peaked at the two structures consistent with a direct link from A to B and from B to C (S9 and S11, right). Crucially, Step 4 effectively tests for the existence of a direct link from A to C which distinguishes between the two remaining contenders. Due to the probabilistic nature of all of the evidence in this task, it takes a repetition of steps 1 and 4 to achieve 90% certainty, and correctly endorse structure 9.

Trial No.	1	2	3	4	5	6	7	8	9
Node A	○	◉	●	◉	○	◎	◎	◎	◎
Node B	◉	●	●	●	○	●	○	○	◉
Node C	○	●	◉	●	◉	◉	◉	◉	○

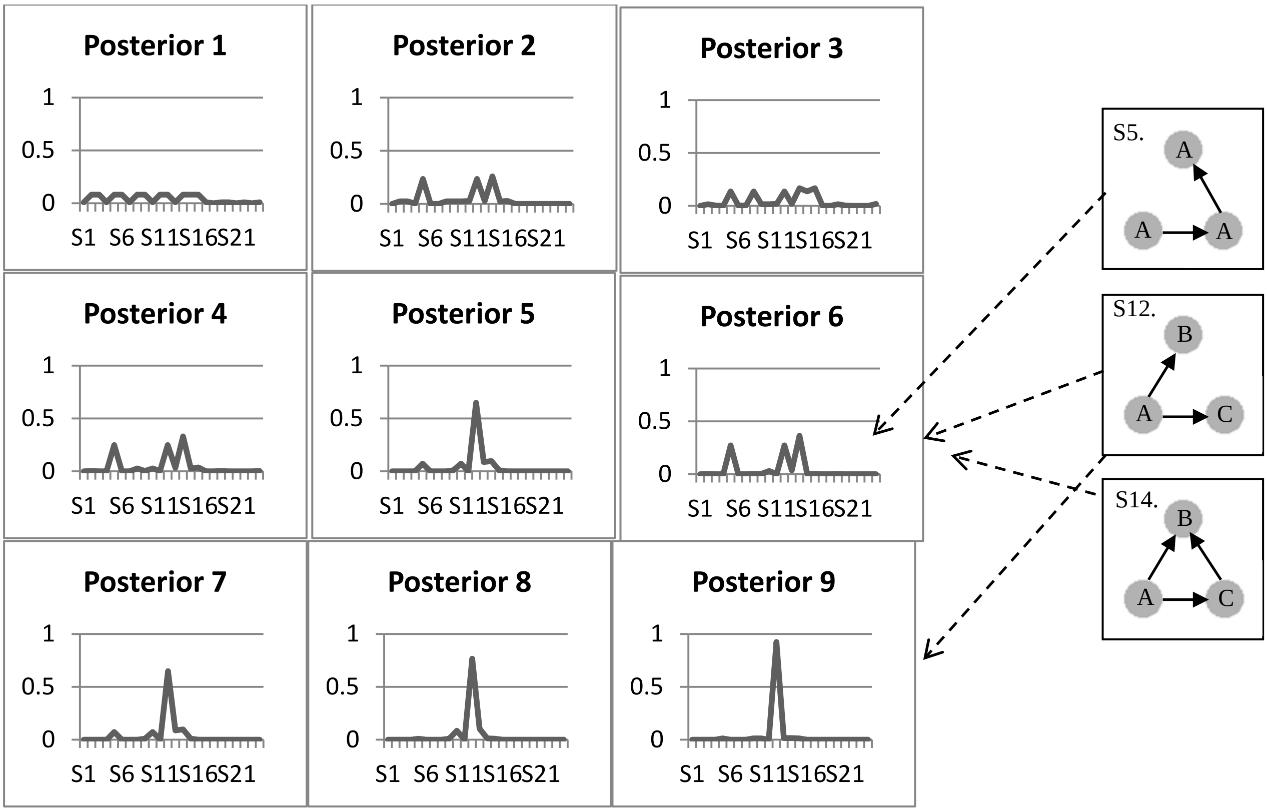


Figure 15. This is included to demonstrate the behaviour of the model when it receives conflicting information as frequently happens due to spurious random activations of the free nodes. Note in particular step 6. This observation actually causes uncertainty to increases significantly because, after reaching ~60% certainty that structure S12 is correct, it sees what looks like evidence for a link from C to B which is not present in S12. The model then repeats the same intervention multiple times until it sees node C fail to activate three times and it becomes very unlikely that there is a link here.

The model's choices, at least in block 1, lend themselves quite naturally to explanation in terms of the discovery of, and subsequent distinction between a few hypotheses. It is clear from this that the two common intervention patterns from table x serve distinct roles. The model always starts by performing trials with each node individually fixed to 'active', one at a time. Each corresponding activation leads to a big rise in the probability of structures that contain a link from the fixed node to those that activate. This narrows the model's search down to the few structures consistent with all intervention-activations so far. Outcomes where two free nodes activate are less discriminating. This is because two activations could be due to a structure with: two direct links (i.e. the fixed node is a 'common cause'), a 'chain' (i.e. where one of the free nodes is a direct effect and the other is an indirect one), or one with two direct links *and* an indirect one (as in the case of S11 and S14 in the examples. Thus, in cases where more than one activation occurs, the most informative interventions become those that effectively disambiguate between these options by testing for the presence of a direct link between the current fixed node and one of the free ones. This is achieved by double interventions that fix one of the two free nodes to be definitely inactive, thus testing for a direct connection between the 'fixed active' and the remaining unfixed node.

Something akin to this pattern continues when it comes to four node problems. Fixing single nodes to 'active' remains the most informative intervention initially. Similarly, double and triple activations necessitate follow-up interventions, which fix all but one of the resulting activated nodes to 'definitely inactive', in combination with the same 'fixed active' node. However, the hypothesis space is exponentially larger for even four nodes making it hard to represent some of the more complex effect of observations on the probability distribution. Certain interventions, which the model occasionally selects, e.g. where more than one thing node is fixed to 'active', change the probability distribution in a way that is too complex to lend itself to hypothetico-deductive explanation.

2.2.4 Colour-cue Consistent Interventions

As mentioned, it proved infeasible to compute the most efficient pattern of interventions for the six node networks in blocks three and four, for the full hypothesis space of possible structures (either with or without the hierarchical level for capturing rational cue use). It *was* possible, though, to test what an efficient learner should do *assuming* they have identified the role of the colours.

Of the ~3.8 million possible six node causal structures, only 570 were made up of only links from any three nodes to the other three, and only 511 were made up of only links *within* exclusive triples (see Figure 16). This meant that identification of the role of the colour cues in this task reduced the number of possibilities in blocks 3 and 4 by a factor of around 7000.

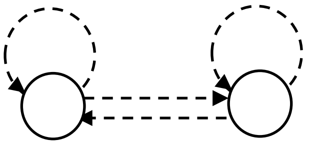


Figure 16. The dotted arrows indicate different possible causal links between two *classes* of object. Reflexive arrows indicate causation within members of the class.

For the blocks 3 and 4, the benchmark model averaged 9.6 and 9 trials respectively to endorse each structure, and the expected values of different interventions were markedly different from those computed from the full hypothesis space (see Table 4). Essentially, the model always fixed all three blue nodes in block 3, usually 'two on, one off' (56.25%) or 'one off, two on' (39.58%). Interventions which fixed yellow 'known effect' nodes were always worse than those that were the same but with the yellow nodes unfixed. Essentially, fixing a known effect *always* reduced the amount you could learn on that trial. In block 4 the model's behaviour is best understood as treating the problem as two simultaneous three-node problems, performing the same patterns of interventions as in block one simultaneously, *within* each colour.

Intervention type	**Outcomes**	**EU full H space flat prior.**	**EU if cue known Block3**	**Freq. Selected by model**	**EU if cues known Block 4**	**Freq. selected by model**
All Free	64	1.014	.435	0	0.398	0
One On	32	2.660**	1.589 / .289	0	1.215	0
One On, One Off	16	2.079*	1.607 / .144	0	0.685	0
Two On	16	2.358*	1.420 / .144	0	2.039 / 0.779	0.489*

Two On, One Off	8	1.772	1.460* / 0	0.396	1.504 / Var.	0.267*
One On, Two Off	8	1.513	1.624** / 0	0.563	0.550/Var.	0
Three On	8	1.761	1.064 / 0	0.021	1.600/Var.	0
Three On, One Off	4	1.191	Var. Low	0	1.066/ Var.	0.067
Two On, Two Off	4	1.179	Var. Low	0	0.972/Var.	0.156*
One On, Three Off	4	.972	Var. Low	0	.484/Var.	0

Table 4. Table shows the expected reductions in uncertainty (EU) for the top ten interventions that were eventually computable for the six node problems from a flat prior (column 3). Also the EU's and selection frequencies for the same problems were computed using the restricted hypothesis spaces for blocks 3 and 4 respectively as shown in columns 4-7. Interventions had different values depending on *which* colours were fixed. In block 3 any interventions on yellow nodes diminished the EU of an intervention, so the high figures in column 4 are those with only blue nodes fixed and the low have all yellow nodes fixed. Similarly.in block 4, column 6, the high values were for interventions fixes spread over both colours.

Testing the behaviour of a hierarchical extension of the model (with an additional layer for the hypothesis space of possible one-cue grammars) would be ideal. However, the model's behaviour when it is given the grammar at least gives us a criterion for grammar-consistent learning behaviour and a benchmark for grammar consistent performance. This allows some roughshod analyses of participants' abilities to alter their learning strategy to make use of the constraining role of grammar level knowledge.

Overall, the Bayesian active learning model averaged 443 points in the scoring metric of the task in blocks 1 and 2, and a further 657 points in blocks 3 and 4 when given the colour rules, putting an absolute theoretical ceiling on rational performance in the task at around 1100 points.

2.3 Results

The performance of the 20 participants included in the final analysis was highly varied and fell into three clusters which were used to guide further analysis (Figure 17):

1. Ten participants performed very well, forming a natural cluster with a mean score of 519.50 (SD 164.10), finding 57.0/66 (SD 8.67) causal links on average while wrongly endorsing a further 15.7 (SD 8.672).

2. A further four participants had mediocre performances, although were still a long way above chance (see below), with a mean score of -175.00 (SD 164.11), finding 35.5/66 (SD 4.51) causal links on average but wrongly endorsing about the same number 33.0 (SD 18.57).

3. Finally, the other six participants were not significantly above chance performance with a mean at -711.11 and (SD 143.45) finding 21.5/66 (SD 10.635) causal links and wrongly endorsing a further 59.0 (SD 30.47).

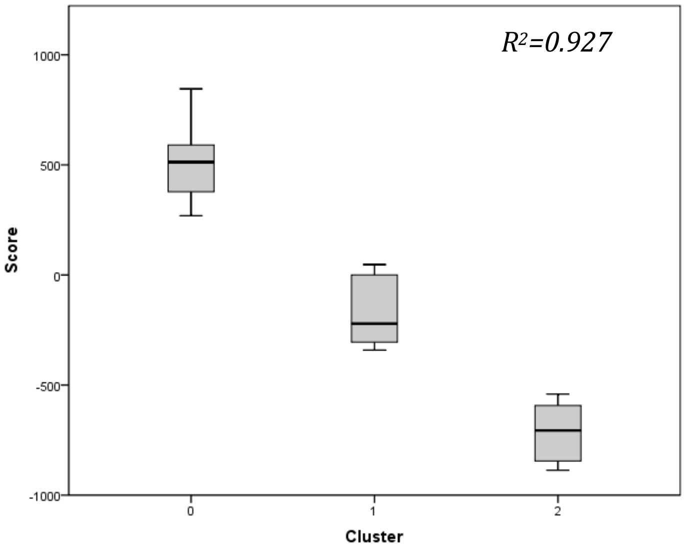


Figure 17.

Chance performance can be interpreted in several ways for an open-ended task like this. Never performing trials or adding links would leave a participant with $-10 \times MissedLinks = -660$ points, while a random performance, computed by a function which added the mean number of causal links per problem in each block, at random, without performing any trials, scored -909. By performing trials *and* still adding links at chance, one could do significantly worse than this. However the first, most conservative, measure of chance was used.

Taken as a whole, scores ranged from: 845 (near-optimal) to -887 (chance), mean 11.4, SD 576.6. No specific problem was found significantly easier or harder than any other, and individuals' performances were generally fairly consistent throughout the experiment. Participants endorsed 74.3 causal links on average over the Experiment that was not significantly higher than the true figure of 66.

While time spent on the task was not a significant factor, the number of trials performed *was* significantly positively correlated with performance (p .002 F(2,20)=13.89 R^2=0.44), suggesting either that having more patience or perseverance in the task improved performance, or equally that people who were generally more patient were generally better at this kind of task.

2.3.1 Measures of Interventional Efficiency and Sensitivity to Resultant Data

For all participants, the average *maximum posterior probability* of the s*ubjectively most likely structure* was calculated for blocks 1 and 2, using the benchmark model, based on the actual interventions chosen and activations seen by participants. From this is possible to say that, on average, people endorsed models when they had seen enough data to be normatively justified in being 41.5% (SD.27.7%) sure about which was the correct underlying structure. The 10 very successful participants were significantly better at choosing interventions that distinguished between possible structures (p<0.001), achieving an average of 63.02% posterior probability of the subjectively most likely hypothesis, than the 10 mediocre and chance performers. The latter groups achieved 35.8% and 9.4% respectively. Given their lack of the necessary information, it is not surprising that those in the lower performing groups were unable to endorse the correct hypotheses in the majority of cases. Additionally, participants in the high scoring cluster were significantly more likely (p<0.001) to select the subjectively most likely hypothesis when they had finished performing trials. They did so 60.0% of the time while the mediocre performers did so only 22.5% of the time and the chance performers just 8.3%. This second measure is highly dependent on the first ($R^2 = 0.67, F = 36.33,\ p < 0.001$) partly due to the fact that, presumably, the more starkly favoured the most likely structure is, the easier it is identify.

A potentially more fine-grained measure of participants' ability to select informative interventions is the *average informativeness* of each intervention, as a percentage of the informativeness of the intervention that the benchmark model *would have selected* in the same situation. This can be thought of as a measure of participants' 'interventional efficiency'. This measure effectively controls for the fact that observing a long string of varied and randomly selected interventions would still, in the limit, provide enough data to allow a high degree of posterior certainty about the correct structure, despite there being no method behind the pattern of interventions chosen. Unlike the average maximum posterior probability, this measure does *not* differentiate between the clusters at all, with efficiency scores with similar means: 48.4% (SD 14.4%), 39.6% (SD 9.4%) and 38.0% (SD 12.0%) respectively. Between cluster contrasts were not significant (with p-values of: 1:2 0.12, 2:3 0.85, 1:3 0.14). This might be due to lower performing participants being less thorough, i.e. not repeating tests on the same node. For this reason, the frequency with which intervention choices were 'optimal' according to the benchmark model was also tested. Results were 25.0% (SD 16.0%), 19.1% (SD 9.2%) and 17.0% (SD 12.6%) respectively. While there was a trend, once again, there were no significant differences between clusters (p-values: 1:2 = 0.29, 2:3=0.82, 1:3=0.49). These findings suggest that being sensitive to the implications of the outcomes of your interventions may constitute a larger part of successful active learning than the choices of the interventions themselves.

2.3.2 Strategies

In order to try to establish what was behind the large individual differences in performances between clusters, the questionnaire at the end of the experiment probed whether participants thought they had used a strategy, and if so, whether they were able to describe what they had done. 9 of the 10 participants in the high performing cluster claimed to have used a strategy to solve the task while only 2 of the 4 mediocre performers and 3 of the 6 at-chance performers did. Overall, reporting having a strategy was correlated with performance ($p = 0.005$).

As mentioned in the **§2.1.4**, the 14 participants who reported having a strategy were coded post-hoc according to how they described this strategy. This was because, with no exceptions, these descriptions came in one of two general forms:

1. Simple link endorsement (n=4/14):

"I activated nodes one by one to see which other nodes they were activating" (Participant 3)

"Seeing which nodes light up when I tested an individual one. Based on that principle I made a judgment that they should be linked together" (7)

"It's always one fixed node each time, I test all of them" (20)

"Simply link the arrow to the one making the other ones active" (22)

This strategy was formalised as: Fix one node on and test it one or several times. Add a causal link from the fixed node to any other node which activated on the majority of trials.
Repeat with each node (Figure 18).

2. Simple link endorsement + additional disambiguation step (n=10):

The first step like with simple link endorsement with the additional step as described below:

"I activated one node and if this activated, for example, two other nodes, I fixed one of the other nodes to test whether it was a model like this [diagram of common cause structure] or like this [diagram of chain structure]. I fixed both activated nodes to test for that."(1)

"If there w[as] more than one responsive node, using fixed inactive node to see how they were interrelated"(2)

"If two or more light up > deactivate some to see if the activated causes a specific knot or if one of the other knots that are caused by the activated [one] are causing other knots to light up > repeat" (6)

"Fixed deactivated nodes help pinpoint the 'chain' reactions and the successive linking" (18)

"Switched each 'node' on individually, saw which ones light up. When there was more than only 1, inactivated one of the nodes to see which one was causing the activation" (24)

This was formalised as an additional step which was performed whenever more than one unfixed node activated: Perform one or several trials with the same fixed ‘active’ node and all but one of the previously activated nodes fixed to ‘inactive’. If the unfixed node does not activate, remove the causal link added in step one. Repeat this procedure with each of the activated nodes (Figure 18).

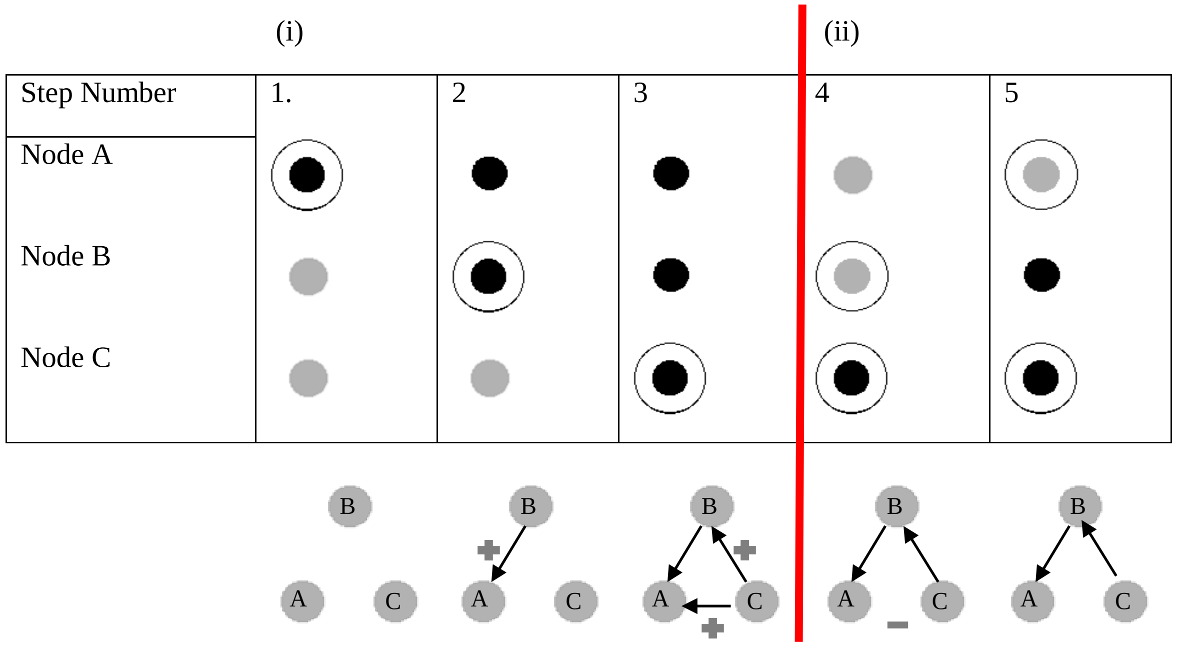


Figure 18. (i) Illustrates ‘simple link endorsement’, each step can be repeated multiple times (see below). (ii) further illustrates a disambiguation step, resulting from the double activation in trial three. Below are the causal links endorsed by these strategies as a result of each step. For the ‘simple link endorsement’ strategy, the model would stop at the red bar while for the ‘disambiguation’ strategy it would also perform the additional step and therefore remove the link from C→A resulting in a chain structure.

2.3.3 Assessing the Strategies

The two strategies were formalised as models which simply applied the rules specified above for each problem in the experiment and endorsed the resulting structure. Several participants reported performing each trial several times: 2 times (participants 6, 12, 16), 5 times (17, 20), ‘several times’(21) and the mean number of trials performed (209.25 and 355.05 for ‘simple endorsers’ and ‘disambiguators’ respectively) suggest 2 to 3 trials per step (see below). Therefore, in order to assess the strategies independently, the models were tested on the problems in the experiment 100 times, in each in of three conditions (allowing the efficiency and sensitivity of the two strategies to be measured as well as their frequencies of intervention choices and overall performance):

1. Performing 1 trial on each link.
2. Performing 3 trials on each link and rounding the resulting activations to 1 or 0.
3. Performing 5 trials on each link and again rounding the result.

Performing 2 or 4 trials were also tested for the disambiguation strategy but the result was that, in cases where 1/2 or 2/4 trials resulted in activations, a decision had to be made whether to round up or down. Both options were tested, but performance was significantly worse in both conditions than those with an odd number of trials per node. Therefore, only the 1, 3 and 5 trial conditions are included here.

Before assessing the performance of the strategies themselves, it is important to ascertain whether the strategies which participants *reported* using actually did describe what they did in the task. A rough measure of this is to test whether there is a match between the *frequencies* of different types of interventions chosen by participants, and the *frequencies* of the different types of interventions chosen by the strategies. For example, to be consistent with ‘simple link endorsement’, you should *always* select interventions with single nodes fixed to ‘active’ and all others left unfixed. However, the 4 participants who reported only using this first step were a fair distance from only using this intervention. While the single node active intervention was the modal choice in all blocks, it was selected only 46% of the time in block 1, 70% in block 2 and 70% in blocks 3 and 4. Other selected interventions were spread across the other categories (Table 5). However, the extremely low sample size (n=4) for this group makes it unwise to draw any strong conclusions from this.

Much more confirmatory, were the frequencies for the 10 people that described themselves as using the disambiguation strategy. Most interventions should also be of the single-activated-node type but, with diminishing likelihood, interventions with one node fixed active *and* up to n-2 nodes fixed to be inactive should be selected in cases where multiple activations are observed. The full results of this comparison are shown in Table 5, but in general, there *was* remarkably close correspondence between human and model intervention frequencies for the mixed strategy, with a mean difference of only 5.5%.

In terms of performance in the task, the simpler strategy actually performed marginally *better*. This was due to its savings in efficiency. It took 285 trials to achieving 697.2 points on average in the three trials per node condition compared to 476 trials to achieve 657.7 for the model with the disambiguation step added (see Tables 6 and 7). The simple link endorsement strategy was also more efficient, and picked the optimal intervention more frequently, but endorsed structures while the maximum posterior probability was lower, and selected the most probable structure given the data less frequently (Table 8).

Block	Intervention Type	Purported Simple Endorsers	Simple Endorsement Model	Purported Disambiguators	Disambiguating Model
1	**All free**	.321	0	.114	0
	One on	.464*	1.0	.590	.695
	One on, one off	.0622	0	.1921	.305
2	**All free**	.0574	0	.0403	0
	One on	.703**	1.0	.466	.523
	One on, one off	.145	0	.281	.225
	One on, two off	.0304	0	.136	.251
3&4	**One on**	.699**	1.0	.609	.672
	One on, one off	.154	0	.153	.220
	One on, two off	.040	0	.0636 (.62 / .38)	.0841
	One on, three off	.0176 (0/1.0)	0	.0426	.0219
	Two on	.0141	0	.0172 (.19 / .81)	.0024

Table 5. – Match between purported strategy intervention frequencies and actual intervention frequencies. In most cases patterns in block 3 and 4 were similar enough to be collapsed together . The splits in brackets in the last three rows indicate the few notable divergences.

This analysis shows that the main *advantage* of performing the second ‘disambiguation’ step was that it significantly lowered the probability that extra over-determining causal links would be endorsed by mistake. It did this by effectively pruning out most links which were initially endorsed due to their *indirect* causal influence on another node. This means that, had the payoffs been weighted more towards accuracy than frugality, the more complex strategy would have paid out higher. For example, if the cost of wrongly endorsing an arrow was as high as the payoff for correctly identifying one, scores would be the other way around (400.50 / 539.06).

No of times each test performed:	Mean Score	Splits by Block:				Tests Req:	Causal Links		
							Right,	Wrong,	Missed
1	514.4	129.40	143.60	120.20	121.20	95.00	60.44	**54.38**	5.56
3	**697.2 ****	131.80	158.60	200.40	206.40	285.00	64.63	**29.67**	1.37
5	623.2	110.60	127.90	189.10	195.60	475.00	65.65	**21.13**	0.35

Table 6. Performance of the ‘simple link endorsement’ strategy.

No of times each test performed:	Mean Score	Splits by Block:				Tests Req:	Causal Links		
							Right,	Wrong,	Missed
1	642.0	129.45	166.91	172.04	173.59	180.82	56.79	**22.08**	9.21
3	**657.7 ****	122.06	153.65	187.55	194.43	475.87	63.74	**11.86**	2.26
5	468.1	89.86	88.63	144.43	145.17	753.58	65.41	**8.07**	0.59

Table 7. Performance of the ‘disambiguation’ strategy.

The performance of participants who reported doing ‘simple link endorsement’ was not significantly different to those who reported using the disambiguation step as well, but there was a dramatic difference (p=0.026) in the number of wrong links endorsed, with an average of 53.25 (SD 33.925) for the simple link endorsers compared to 19.20 (SD 13.863) for the disambiguators. When this is compared to how many of the 66 *correct* links were also identified (39.5 and 54.1 respectively) this amounts to a massive false positive rate for those who did not perform the disambiguation step. It must be noted again though, that while these findings are serendipitous, the pure ‘simple link endorsement’ group was too small to draw strong conclusions based on comparisons with them.

Strategy	Freq. Opt. Intervention Selected	Av. Intervention Efficiency	Av. Max. Post .Prob.	Freq. Selected Max Posterior Structure	Score (Blocks 1&2)
Simple Endorsement	0.238	0.560	0.605	0.3	275
Disambiguation	0.130	0.4106	0.846	0.6	293

Table 8. ‘ Benchmark’ measures of strategy performance as described in 2.3.1

2.3.4 Colour Cues

Questionnaire responses were coded according to the ability of participants to describe the role of the colour cues in blocks 3 and 4 after completing the task. Based on this coding scheme, colour cues appeared to help participants to solve the problems in blocks 3 and 4, if they were able to recognise their value. Ten participants were able to describe one or both colour cues. Of these participants, the three who were able to describe both, averaged 317.7 (SD 165.5) points in blocks 3 and 4, the four who could only describe the block 3 cue averaged 57.0 (SD 335.4), the three that noticed only the block 4 cue averaged 133.0 (SD 61.0). The other 10 participants that could describe neither averaged -261.4 (SD 432.1). There was a significant effect of noticing either or both colour cues compared to not (F=7.198 p=0.015). Follow up studies with additional control conditions are necessary to really explicate this effect though due to the interaction between and the two potential causes:

1. The boost in performance due to making use of the colour cues.

2. Noticing the colour cues used the same skills that made some participants better performers in the task in general.

As mentioned in Benchmark Model Predictions section, it was not possible to model a benchmark hierarchical Bayesian solution to the optimal way to learn and integrate a simple cue but it *was* possible to compute efficient interventions *given* the grammar. This makes it possible to investigate whether the people markedly changed their interventional strategy if they realised the role of the cue which would offer some for the first interpretation.

Intervention choices in the *last problem faced* in blocks 3 and 4 were investigated for those participants who had been able to describe the roles of the colour cues for those blocks. For block 3, even participants savvy to the role of the colour cue were still fixed yellow ‘effect’ nodes on 50.0% of their interventions, and the efficient interventions: ‘two on, one off’ and ‘one on, two off’ on the blue nodes only, were only selected 19.12% (and all by one participant). For the last-seen problem in block four, 79.8% of interventions were performed within one colour only, leaving the other colour free. This accords with the intuition that participants would tend to focus their attention on one colour at a time once they understood the within-colour-causation-only cue representing a paradigm of divergence of serial human learning from the parallel learning of the Bayesian active learning model. Based on this limited data it is not really possible to confirm or disconfirm the role of the cue in participants inferences but only to guide future work (see **§3.4**).

3. Discussion

Overall, the analyses of the results of this experiment were broadly confirmatory of this researcher's intuitions as well as being elucidatory of active causal learning mechanisms and suggesting various directions for future work.

3.1 Simple Mechanisms for Active Causal Learning

The patterns of intervention choice and resulting changes in the probability distribution of the benchmark model for the 3-variable cases (**§2.2.3** Examples; Figures12 & 13); and the intervention choices and ascriptions of the 'disambiguation' strategy (**§2.3.3** Assessing Strategies; Figure 18), and the actions many of the participants (Appendix **§7.3**), were *fundamentally very similar*. The benchmark model tended to save its disambiguating steps until after testing all the nodes individually, while people and the disambiguating model tended to do this as they went along. However, the actual order of a pattern of interventions is logically irrelevant to its informational outcome, and the order of intervention choices chosen by the Bayesian benchmark model was an incidental by-product of its myopia. The number of times each node was tested was also flexible for the Bayesian model, while fixed for the 'disambiguation' model. The number of trials the Bayesian model would perform on a node would vary, depending on whether, and how much, conflicting information was observed. However, although this was not investigated formally, most purported strategy users appeared to exhibit a similar, or even exaggerated, flexibility in their deviations from the strategy models, performing more tests on nodes when they received evidence which conflicted with the links they had already identified, or when the already identified links 'explained away' (Hagmayer & Lagnado, forthcoming) the activations.

With the above qualifications in mind, the fit between the 3-node Bayesian benchmark, the disambiguation strategy, and participant's actions is striking. It suggests strongly that a mechanism which is: 1. Achievable within working memory limitations. 2. Involves no quantitative calculations. 3. Appears to be used by real learners - can result in, more or less, the same patterns of intervention, and similar causal conclusions as the maximally intensive Bayesian benchmark model (at least in some circumstances, see below).

3.2 Ecological Validity

It remains to be tested whether the strategies identified in this thesis are robust to application to environments or domains with different causal properties. Higher or lower base rates, higher or lower typical causal strengths, sparser or more causally dense environments, continuous variables, weaker interventions, degrees to which the Markov assumption is violated, or different typical integration properties all might necessitate learning mechanisms which are more or less constrained/rigorous. However, because these strategies essentially break structural induction down into a process of asking simple, targeted, questions of the data it seems likely that they would continue to work across a broad range of situations. Essentially they ask of objects: "What does this thing *do*?" and when faced with several answers, ask: "Are these direct or indirect effects?" By intervening positively, activating the target 'cause' nodes, the precise parameters of a network can largely be ignored, or lumped together as unspecified noise. This avoids the need for exact knowledge about (or at least explicit estimation of) of a causal domain's underlying parameters. While this is something required for the Bayesian model, it is presumably a rare commodity in real-world causal learning.

As the structures in the experiment got more complex, the simple strategies stayed simple, yielding similar performance with only linear increases in required interventions. Meanwhile, the Bayesian solution diverged, selecting more complex interventions as the hypothesis space grew and fast becoming intractable. These interventions had increasingly obscure effects on the expanding probability distributions of possible structures, and thus required inordinately more computation. So, in order to make modest savings in the number of interventions required, the Bayesian model eats up superexponentially *more* processing power. This computational inefficiency makes Bayesian-intervention-selection a poor choice for balancing the costs and benefits of active causal learning in real environments *regardless* of what one assumes the precise computational architecture of the brain *or* the precise properties of real environments to be.

Given that we are not completely ignorant about the properties of real environments, it is possible to make this point even more strongly. It is relatively uncontroversial that: 1. The actual environment has a complex structure, 2. Processing frugality is at an evolutionary premium in organisms with limited energetic resources, and 3. Active exploration of (at least the local and physical) structure of our environment is the meat of being an embodied cognitive agent living in the world (Simon, 1956; Clark, 1998). When structure is complex, mechanisms which scale up linearly are better than those that suffer from combinatorial processing explosions. And, when active intervention is relatively cheap, maximising your interventional efficiency becomes secondary to having robust and efficient mechanisms for active causal structure induction.

A slight limitation of these results are the unexpectedly large individual differences in performance in the task, including a significant numbers of participants who performed at chance, or were excluded for failing to understand the task instructions (10/24). This high variability is attributed, by the author, to weakness in the interface of the task environment, as would be implausible to claim that there is such massive variance in real world causal learning ability.

Lack of fluency in computer interaction, or with low-tech computer games, or a general dislike of tasks perceived as 'mathematical' may underlie the failures of some of these participants to fully engage with the task. Accordingly, it seems likely that modification of the task to make the learning environment more naturalistic and intuitive, while still steering clear of casting it in a specific causal domain, might well eliminate much of this variance. However, the very good performance of the other participants in what was really a very difficult task shows that we *do* have these causal learning abilities and *are* able to identify much more complex causal structures than we are given credit for in many existing publications.

To follow up from this study, it would be good to perform a thorough modelling study of "simple intervention + causal attribution" heuristics, including the ones identified in this thesis, and test them on a broad variety of CBN parameterisations as well as on real environments. The flexibility of human performances over the same tasks could also be tested. This could then guide conclusions about the extent to which we might have *multiple* simple mechanisms for active causal learning which we somehow select from dependent on the properties of the environment. Or conversely, the extent to which we might adjusting the parameters (i.e. how many simple tests to perform, how many, if any, disambiguation steps) of a *single* causal learning mechanism, so that it works well given the levels of noise, typical causal strengths, and costs and payoffs of particular environments. The sympathies of the author lie toward the latter interpretation. Perhaps we are able to use our implicit perceptual sensitivity to covarational frequencies and patterns (Lewicki, Hill & Czyzewska, 1992) to continually readjust the rigor of our action selection and causal ascription mechanisms as we attend to different aspects of our surroundings.

3.3 Causal Attribution Mechanisms and Ideological Differences

The comparison between the performance of the 'simple link endorsement' and the 'disambiguation' strategies (**§2.3.2**) suggests an interesting potential divergence between basic causal attribution mechanisms. Prima facie, much of the role of causal knowledge is pragmatic. Essentially, we often use our causal knowledge to get answers to questions like: "What do I need to *do* in order to make X occur?" Accordingly, we are often happy with approximate and simplified causal representations (**§1.2**) for situations which we know to be more complex when more closely investigated (Casini et al, 2011). In the light of these considerations, the 'simple link endorsement' strategy, which does not worry about the possibility of causes being indirect, seems to be *the* paradigmatically pragmatic approach to endorsing causal structure. However, it is often vital to establish, not just the pragmatic, but also the *counterfactually reliable* structure of a domain. For example, to establish whether a new drug really works, or if an expensive policy will lower crime, one must *control for* the possible indirect action of placebo or policy change effects. Essentially, in some situations, the additional 'disambiguation' step is very important.

A possibility, perhaps worthy of further investigation, is that individual differences in *propensity* to disambiguate between multiple potential causal paths from actions to events, might underlie deep ideological differences between people. For example, people differ markedly in their propensity to endorse what might be argued to be *overdetermined* explanations of worldly phenomena. Examples include believing that there are magical, supernatural or religious connections between actions and events, in addition to their known physical-causal explanations. As a potentially disingenuous example, in the children's film, Dumbo famously attributes his ability to fly to his 'magic' feather. Later, when he is forced to try and fly without it, Dumbo changes his causal model to one in which the feather had an *indirect* influence on ability fly, mediated by its effect on his confidence. Had he not been forced, Dumbo probably would have retained his magical belief for some time. In the same way, it might be that some individuals are more inclined to accept correlations between an actions and distal effects as legitimate cause-effect relations, in the same way that the 'simple link endorsement strategy' endorses direct links to everything that follows from an activation. Thus, individual differences in a basic active learning mechanism might determine how individuals develop their belief systems and populate them slightly different causal structures.

3.4 Cue Integration

Unfortunately it was not possible to perform a thorough analysis of participants' identification, and use of, the simple colour cues due to the many confounds and low sample size. But, the fact that some people *did* notice the colour cues, and that noticing the colour cues *was* correlated with performance, was encouraging. Furthermore, participants' questionnaire and oral responses provided some useful information to guide follow up studies. Anecdotally, it appears that the task put such a load on concentration and working memory that many participants barely noticed what colour the nodes they were testing were. A couple of participants reported using them as placeholders for remembering which nodes they had already tested suggesting the interpretation that we may only learn only at a single 'level' at a time. Potentially, noticing the role of the colours might have been dependent on whether participants focused their attention directly on them, temporarily ignoring the primary goal of the task, or learned the relations when 'resting' on the feedback screens between problems. A follow up experiment, with more trials and no feedback phase, is necessary to really get to the bottom of the interaction. Another observation, drawn from participants' verbal reports and the fact that more people noticed one or other colour cue (7) than noticed both (3), was that realising that blues cause yellows block 3 was likely to have a *blocking* effect on participants' chances of realising that greens cause greens and oranges cause oranges in block 4. This coheres with the hierarchical framework, in hindsight, considering that block 3 constituted all of the experience that participants would have had with the causal properties of 'coloured nodes'. Participants would be perfectly sensible in having an expectation at the start of block 4 that colours have something to do with the causal *roles* of nodes (e.g. as either cause or effect) making it unlikely that they would consider new colours to correspond to the, conceptually very different, *causal grouping* of block 4.

4. Conclusions

Overall, these results show that people can be very good at learning complex probabilistic causal structure, without domain knowledge, so long as they are given a free rein to intervene. In this task, people appeared to go about this by sequentially targeting interventions on single objects, finding out what they do, and building up a model on the basis of this, often performing an additional 'disambiguation' step when what they see could correspond to multiple linkages. The intervention choices of the simple strategies identified in this thesis correspond closely to the Bayes-'optimal' intervention choices for simple problems, proving that this is an efficient way to work on a local scale, but the strategies also scale up in a way that is not possible for Bayesian inferences.

The results are in line with the rational hierarchical framework, in that people show some signs of being able to identify and make use of a minimal 'grammars' from a simple binary cues, although follow-up studies are required to show this forcefully. Additionally, the *effort* required in completion of the task, and the difficulty many people found in solving it, testifies to the extent to which people normally rely on their existing knowledge in making causal inferences, only learning actively to the degree that a situation is unfamiliar.

Generally, these results are most in keeping with the view that while causal representation is complex the basic mechanisms through which we interact with and learn about the causal structure of the environment, are simple, active and heuristic.

5. Acknowledgements

Thanks to David Lagnado, Maarten Speekenbrink, Eric Schulz, Anne Hsu, Paula Parpart, and Christos Bechlivanidis.

6. References

Bromberger, S. (1966). *The Journal of Philosophy*, 63 (20), 597-606.

Casini, L, Illari, P. M., Russo, F & Williamson, J. (2011). Models for prediction, explanation and control: recursive Bayesian networks. *Theoria*, 26 (1), 5-33.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review,* 104, 367-405.

Chomsky, N (1975). *Reflections on Language.* New York: Pantheon.

Clark, A. (1998). *Being There.* Cambridge MA: MIT Press.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, 118 (1), 110-9.

Gopnik, A, Glymour, C, Sobel, D, Schulz, L, Kushnir, T & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 1-31.

Griffiths, T. L. & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review,* 116, 661–716.

Hagmayer, Y. & Lagnado, D. A. (Forthcoming). Causal models in judgment and decision making. In Dhami, M., Schlottman, A., Waldmann, M. (ed.*) Judgment and decision making as a skill: Learning, Development and Evolution*. Cambridge: CUP.

Hagmayer, Y., Sloman, S., Lagnado, D., & Waldmann, M. R. (2007). Causal Reasoning Through Intervention, 86-101.

Holyoak, K. J. & Cheng, P. W. (2011). Causal Learning and Inference as a Rational Process: The New Synthesis. *Annual Review of Psychology*, 62, 135–63.

Jones, M & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioural and Brain Sciences* (in press).

Kant, I. (1781/1965). *Critique of Pure Reason*. London: Macmillan.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to Learn Causal Models. *Cognitive Science*, (in press).

Krynski, T. R. & Tenenbaum, J. (2003). The role of causal models in reasoning under uncertainty. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Kushnir, T, Gopnik, A, Schulz, L. E. & Danks, D. (2003). Inferring hidden causes. In R. Alterman & D. Kirsh (Eds.) *Proceedings of the Twenty-Fifth Annual Meeting of the Cognitive Science Society*, (pp. 699–703). Mahwah, NJ: Erlbaum.

Lagnado, D. (2011). Causal thinking. In McKay-Illari, P., Russo, F., Williamson, J. (ed.) *Causality in the Sciences*. Oxford: OUP.

Lagnado, D. & Sloman, S. (2002). Learning causal structure. In W. Gray & C. D. Schunn (Eds.), *Proceedings of the twenty-fourth annual conference of the cognitive science society*, Mahwah, NJ: Erlbaum.

Lagnado, D. & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory & Cognition*, 30, 856-876.

Lagnado, D., & Sloman, S. (2006). Time as a Guide to Cause. *Cognition*, 32 (3), 451- 460.

Lagnado, D., Waldmann M., Hagmayer, Y & Sloman, S. (2007). Beyond covariation: cues to causal structure. In A. Gopnik, A. & L. Schulz (Eds.), *Causal Learning: Psychology, Philosophy, and Computation*, 154–72. London: OUP.

Laplace, P. L. (1814/2009) *Essai philosophique sur les probabilities.* New York: CUP

Lewicki, P., Hill, T., Czyzewska, M. (1992). Nonconscious acquisition of information. *American Psychologist*, 47, 6, 796-801.

Marr, D (1982). *Vision.* New York: Freeman & Co.

Meder, B., Gerstenberg, T., Hagmayer, Y., & Waldmann, M. R. (2010). Observing and intervening: rational and heuristic models of causal decision making. *Psychology*, 119-135.

Nyberg, E. & Korb, K. (2006). Informative Interventions. In F. Russo & J. Williamson (Eds.) *Causality and Probability in the Sciences.* College Publications: London.

Pearl, J. (2000/2009). Causality. New York: CUP (2nd edition).

Simon, H. (1956). Rational Choice and the Structure of the Environment. *Psychological review*, 63, 129-38.

Shanks, D. R. (2007). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section*, 37, 1, 1-21.

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal*, 30, 50-64.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives*. Oxford: OUP.

Sloman, S. & Lagnado, D. (2005). Do we "do"? *Cognitive Science*, 29 (1), 5-39.

Sprites, P., Glymour, C. & Scheines, R. (1993/2000). *Causation, Prediction, and Search (Springer Lecture Notes in Statistics)*. Cambridge, MA: MIT Press. (2nd edition).

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27, 453-489.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science, 10*, 309-318.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* 331, (6022), 1279-1285.

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the information given: Causal models of learning and reasoning. *Current Directions in Psychological Science*, 15, 307-311.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

7. Appendix

7.1 Task Instructions

Page 1:

“Your task will be to identify how a number of points, or "nodes", are causally related to one another.

In each trial you will see a number of these nodes placed randomly on the screen. They can either be ‘active’ (Red) or ‘inactive’ (Grey). Initially they are all inactive. Each time you hit the space bar some of them may activate and turn red for 1 second. Below you can see an example of: an inactive node and an active node.

Press any key when you are ready to continue…”

Page 2:

“Each time you press the space bar to test the nodes, there is a 10% chance of each node activating all by itself. Additionally, some of the nodes, when activated, will CAUSE other nodes to activate too. This will only happen if there is a causal link between them. Your task is to find these causal links.

In this experiment, causal links are not perfect. In fact, they work 90% of the time. This means that if a node is active and it has a causal link running to another node, there is a 90% chance of that second node activating.

Each hit of the space bar is equal to observing one test of all the nodes, and each test is independent of all your previous tests.

Press any key when you are ready to continue…”

Page 3:

“In order to test what causal links there are, you may FIX as many of the nodes as you like. You may either fix them so they are DEFINITELY ACTIVE or so they are DEFINITELY INACTIVE.

You can do this by clicking on the nodes. One click fixes a node to be active, a second click fixes it to be inactive and a third frees it again. A black circle around a node indicates that it is fixed. If it is RED it is fixed to DEFINITELY ACTIVE and if it is GREY it is fixed to DEFINITELY INACTIVE.

A fixed node can still influence other nodes but cannot be influenced itself. Remember that nodes will stay fixed until you release them.

Below you can see an example of: A free, a fixed active and a fixed inactive node.

Press any key when you are ready to continue…"

Page 4:

"When you think you know where a causal link exists, click in the space between the nodes. This will make an arrow appear. Clicking a second time changes the direction of the arrow. Clicking a third time removes the arrow again.

(If two or more arrow spaces overlap, keep clicking and you will cycle through the different arrows until you get the one you want.)

You add and remove arrows in the same phase as testing the network so you should not click "Done" until you have selected the arrows which you think are correct. Below is an example of an arrow between two nodes:

Press any key when you are ready to continue…"

Page 5:

"Once you are confident that you have selected the correct arrows click on the blue button labelled "Done", at the bottom right of the screen. Only do this once you have selected your arrows.

The real causal links for this trial will be revealed and compared to your choices. Arrows which you got right will turn GREEN. Arrows which you got wrong will turn RED and arrows that you missed will now appear in GREY.

For every arrow you get right, you win 20 points. For every arrow you get wrong you lose 10 points. Each time you test the nodes (every time you hit the space bar) you also lose 1 point.

The money you earn at the end depends on how many points you score over the whole experiment so it is important to try to identify the right causal links but also to do so with as few tests as possible.

If you are confident you understand what you are supposed to do, please press any key to continue…"

Page 6 (displayed between blocks 2 and 3):

"Well done on completing the first half of the experiment!

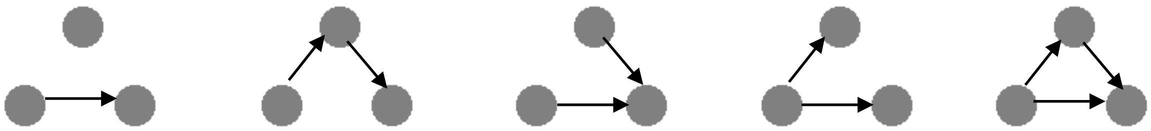
The second half is exactly the same as the first half but with one difference. There will now be two types of nodes instead of one: They will be Yellow and Blue in Block 3, and Green and Orange in Block 4.

Both types of nodes still go red when they are activated, and your task is the same as before. However you may find the colours helpful and you may be asked about them afterwards.

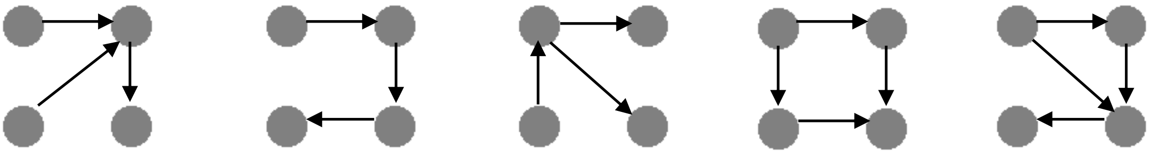
Please press any key when you are ready to continue..."

7.2 Problems Ordered by Block Number

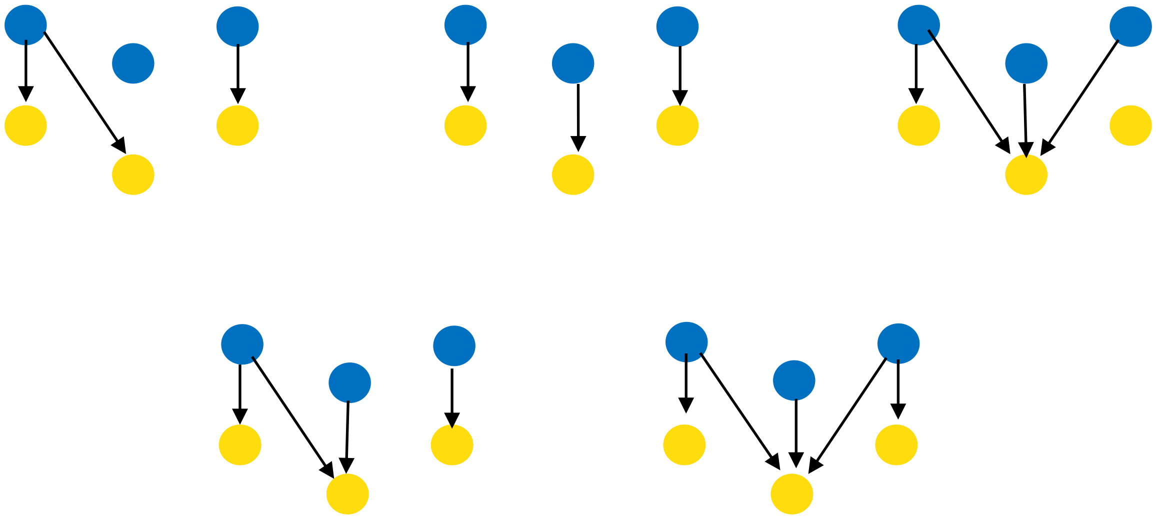
Block 1:



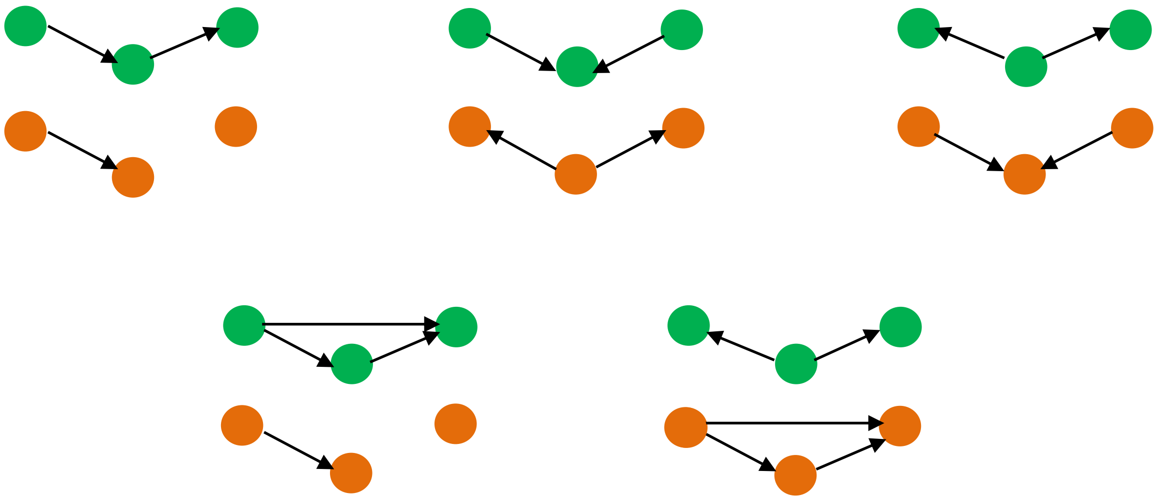
Block 2:



Block 3:



Block 4:



7.3 Individuals' Efficiency and Sensitivity Scores for Blocks 1 & 2

Participant	**Frequency Optimal**	**Efficiency**	**Sensitivity**	**Score (Blocks 1 and 2)**
1	0.131356	0.367732	0.8	284
2	0.666667	0.839952	0.7	241
3	0.318182	0.515791	0.4	-104
4	0.129542	0.329944	0.0	-195
5	0.183486	0.352745	0.1	-39
6	0.326733	0.352741	0.9	349
7	0.387097	0.528103	0.2	-61
8	0.165289	0.418202	0.1	-81
9	Excluded	-	-	-220
10	0.170732	0.485305	0.4	237
11	0.098765	0.297762	0.3	8
12	0.222222	0.498629	0.7	242
13	Excluded	-	-	0
14	0.027778	0.224759	0.0	-136
15	0.227273	0.455757	0.1	-132
16	0.315789	0.603234	0.5	134
17	Excluded	-	-	-102
18	0.190083	0.463002	0.4	209
19	Excluded	-	-	-90
20	0.152174	0.412963	0.4	222
21	0.166667	0.413094	0.5	200
22	0.084112	0.281331	0	-167
23	0.176471	0.462559	0.2	-174
24	0.164565	0.403001	0.7	225

7.4 Questionnaire

Questionnaire

Please describe, broadly, how you think you solved the task:

If you used a specific strategy or strategies (i.e. a particular combination of fixed and unfixed nodes) multiple times throughout the trials please describe what that strategy or those strategies were and how they helped you pin down the right causal links:

Did you find the colours of the nodes helpful in discovering the causal links in Blocks 3 and 4.

YES / NO

If so, how did the colours help in Block 3?

And, if so, how did the colours help in Block 4?

End