Staying afloat on Neurath's boat – Heuristics for sequential causal learning

Neil R. Bramley[1] (neil.bramley@ucl.ac.uk), **Peter Dayan**[2] (dayan@gatsby.ucl.ac.uk),
David A. Lagnado[1] (d.lagnado@ucl.ac.uk)
[1]Department of Experimental Psychology, UCL, 26 Bedford Way, London, WC1H 0DS, UK
[2]Gatsby Computational Neuroscience Unit, UCL, Alexandra House, 17 Queen Square, WC1N 3AR, UK

Abstract

Causal models are key to flexible and efficient exploitation of the environment. However, learning causal structure is hard, with massive spaces of possible models, hard-to-compute marginals and the need to integrate diverse evidence over many instances. We report on two experiments in which participants learnt about probabilistic causal systems involving three and four variables from sequences of interventions. Participants were broadly successful, albeit exhibiting sequential dependence and floundering under high background noise. We capture their behavior with a simple model, based on the "Neurath's ship" metaphor for scientific progress, that neither maintains a probability distribution, nor computes exact likelihoods.

"We are like sailors who on the open sea must reconstruct their ship but are never able to start afresh from the bottom. Where a beam is taken away a new one must at once be put there, and for this the rest of the ship is used as support." (Quine, 1969, p3)

Introduction

It is tremendously hard to learn causal models. Even in apparently simple circumstances, it is necessary to cope with a huge diversity of complex, noisy and probabilistic interactions, and thus to integrate, often painfully, over extended experience. Optimal reasoning with distributional causal beliefs places substantial demands on inference and storage. Nevertheless, in several studies (Bramley, Lagnado, & Speekenbrink, 2014; Coenen, Rehder, & Gureckis, 2014; Lagnado & Sloman, 2004, 2006; Steyvers, 2003) it has been shown that people can learn successfully from interventional data in probabilistic scenarios. Existing experiments have largely been confined to small structures, small data and semi-determinism, thus limiting the computational demands and the need for heuristics or approximations. Here, we report on two experiments designed to tax learning more severely, with a broad range of structures, long sequences of data points, and substantial noise (Experiment 1) whose level and nature participants have to infer as they learn (Experiment 2). We thereby examine how people deviate from rational norms, and explore what this can tell us about their psychological processes.

Representing causal structure

We adopt a ubiquitous framework for formalizing models of causal structure – the parametrized directed acyclic graph (Pearl, 2000). Arrows represent causal connections; and parameters encode the influence of parents (the source of an arrow) on children (the arrow's target). Such graphs can represent continuous variables and any forms of causal relationship; here we focus on binary $\{0,1\}$ variables and generative connections. We adopt Cheng's power PC (1997) parametrization for which the probability that a variable takes the value 1 is a noisy-OR combination of the *power* or strength S of any active causes in the model, together with an omnipresent background cause B that is exogenous to the model. S and B are assumed to be the same for all connections and components, and there is no other latent variable (although see Buchanan, Tenenbaum, & Sobel, 2010).

Optimal structure learning

The likelihood of a datum (a complete observation, or the outcome of an intervention) $\mathbf{d}$ given a noisy-or parametrized causal model m over variables X, with strength and background parameters S and B is

$$P(\mathbf{d}|m,S,B) = \prod_{x\in X} P(d_x|\mathbf{d}_{pa(x)},S,B) \quad (1)$$

$$P(d_x=1|\mathbf{d}_{pa(x)},S,B) = 1-(1-B)(1-S)^{\sum_{y\in pa(x)} d_y} \quad (2)$$

where $pa(x)$ denotes the parents of variable x in the causal model. We can thus compute the posterior probability of model $m \in M$ over a set of models $\mathcal{M}$ given a prior $P(M)$ and observations D. We can condition on S and B if known:

$$P(m|D,S,B) = \frac{P(D|m,S,B)P(m|S,B)}{\sum_{m'\in M} P(D|m',S,B)P(m'|S,B)} \quad (3)$$

or else marginalize over their possible values

$$P(m|D) = \frac{\int_{S,B} P(D|m,S,B)p(S,B)P(m)\ \mathrm{d}S\ \mathrm{d}B}{\sum_{m'\in M}\int_{S,B} P(D|m',S,B)p(S,B)P(m')\ \mathrm{d}S\ \mathrm{d}B} \quad (4)$$

If data arrive sequentially, we can either integrate them at the end, or update our beliefs sequentially, taking the current posterior as the new prior $P(M)$ for the next datum[1].

Scope for approximation

Learning is hard because the number of possible graphs grows rapidly with the number of components (3, 4 and 5-variable problems have 25, 543, 29281 respectively) and there is no known closed form update for densities over S and B in noisy-OR models. To understand how people might mitigate this computational explosion, we take inspiration from machine learning.

Approximating with a few hypotheses One common approximation is based on a manageable number of individual hypotheses, or particles (Liu & Chen, 1998), with weights

[1]For the present, we ignore the related question of active learning – i.e., the efficient selection of interventions. See the discussion.

corresponding to their relative likelihoods. Sophisticated reweighting and resampling schemes allow particle filters impressive fidelity.

In rodent learning (Courville & Daw, 2007), and human categorisation (Sanborn, Griffiths, & Navarro, 2010) and binary decision making (Vul, Goodman, Griffiths, & Tenenbaum, 2009), it has been proposed that people's beliefs actually behave more like a single particle, capturing why individuals often exhibit fluctuating and sub-optimal judgements, whereas group-level posteriors are smooth.

Local search A related simplification is to edit these particle hypotheses only locally – for instance adding, subtracting and reversing individual connections to one's current causal structure in searching for changes that make the model more likely (Cooper & Herskovits, 1992). This is approximate since the complex dependencies between the connections imply that one cannot guarantee to be able to learn each one separately (although see Fernbach & Sloman, 2009).

Prior assumptions People might also exploit simplifying priors, for instance, expecting causal connections to be strong (high *S*trength) and sparse (low *B*ackground noise) (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008), and structures to be "well designed" (Bramley, Gerstenberg, & Lagnado, 2014), lacking redundant connections, or unconnected components. These would be sensible, since causal models simplify inference only to the extent that their structure reduces the number of relata per variable. Mayrhofer and Waldmann (2011) suggest that people might favor deterministic causal structures, accommodating noisy data by assuming that causal connections are occasionally "broken". Their study assumed an absence of background noise; but one could imagine an equivalent accommodation treating inexplicable events as being 'miraculous'. This suggests the heuristic proxy for likelihood judgments for a model as a simple count of the number of variables lacking explanation.

A class of simple structural change models

The resulting picture of a heuristic causal learner is reminiscent of Neurath and Quine's (1969) metaphor for theory change in science. Here, the theorist is cast as relying on their theory to stay afloat, without the privilege of a drydock to make major improvements. At most local changes to patch leaks and to improve the theory are possible, without the whole space of possibilities ever being considered.

Similarly, we propose that causal learners might: (1) maintain only a single causal model (a single particle) b_{t-1} at time $t-1$; (2) search for local improvements (adding, subtracting, reorienting edges) in order to (3) (approximately) maximize the number of aspects of the new data $\mathbf{d}_t$ for which their model can account (Figure 1). Iterating this procedure leads to reasonable, though sub-optimal, causal structure judgments without either representing more than one causal model or remembering old evidence.

We parametrized a whole class of such models via two constructs: the dissimilarity between b_{t-1} and a potential new b_t, and the suitability of that b_t for capturing $\mathbf{d}_t$. We quantified dissimilarity in two ways. One is simple difference $E^*_{b_t b_{t-1}}$, which is 1 iff b_t is non-identical to b_{t-1} and 0 otherwise. The second is the *Edit distance* $E_{b_t b_{t-1}}$, which counts the number of edits (additions, subtractions, reversals of links) going from b_{t-1} to b_t (ranging from 0 to 6 for a 4 variable problem).

We quantified the suitabilities via two approximate likelihoods. One, $L_{b_t}(\mathbf{d}_t)$, is the correct noisy-OR likelihood under a prospective new belief b_t. The second, *explanatory inAdequacy* $A_{b_t}(\mathbf{d}_t)$, just counts the number of component states that the prospective model fails to explain.

We considered the eight viable combinations of these constructs (singletons labeled E, E^*, L, A; pairs labeled E^*L, E^*A, EL, EA). Each model can be taken to generate a likelihood for a subject's choices based on a softmax probability that the model assigns to a choice of b_t. For instance, for EA, this probability is

$$P(b_t) = \frac{\exp(E_{b_t b_{t-1}}\theta_1 + A_{b_t}(\mathbf{d}_t)\theta_2)}{\sum \exp(E_{b_t b_{t-1}}\theta_1 + A_{b_t}(\mathbf{d}_t)\theta_2)} \tag{5}$$

with parameters θ_1 and θ_2 that can be fit to maximize the likelihood. For the moment, we assume that subjects search over all possible edits; how they actually perform this search is an important question for the future.

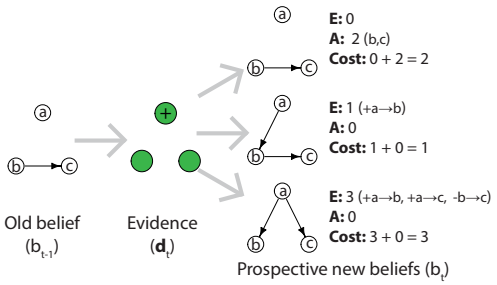

Figure 1: A simple structure change model. The learner encounters data that are not well explained by their model so they search for a local change that improves it. By balancing the edit-cost E against reduced inability to explain the latest outcome A, they opt to add a connection $a \rightarrow b$.

Experimental rationale

To explore these approximations, we considered sequential structure learning in appropriately difficult problems. If subjects really maintain only a single causal belief and make local edits, we expect sequential dependence, and a tendency to get stuck in local optima. If they forget old evidence, relying on the current structure itself, we expect to observe recency effects whereby participants may return to judgments previously rejected. Finally, if they rely on generic priors we expect to see better performance when the true causal structure is conformant.

We therefore designed two online studies based on the paradigm used in Bramley, Lagnado and Speekenbrink (2014) (demo at ucl.ac.uk/lagnado-lab/el/ns15a). Participants

interacted with a series of probabilistic causal systems involving 3-4 variables, repeatedly selecting interventions (or tests) to perform in which any number of the variables were either fixed on or off, while the remainder were left free to vary. The tests people chose, along with the true underlying causal model, S and B, jointly determined the data they saw. We systematically varied the number of connections between components in the problem set, along with S and B.

Experiment 1

In Experiment 1, we restricted ourselves to the effects of "expected" uncertainty (Yu & Dayan, 2003) by training subjects explicitly on the true prevailing values of B and S.

Methods

Participants We recruited 150 participants (85 male, mean$\pm$SD age 35 ± 10 from MTurk, split randomly between 9 conditions (group size 16.7 ± 3.4). They were paid \$1.50 and received a bonus of 10c per correctly identified connection on a randomly chosen test for each problem (max= \$6.00, mean$\pm$SD \$$3.68 \pm .75$). The task took an average of 41 ± 20 minutes.

Design We included five 3-variable and five 4-variable problems (see Figure 2). Within these, we varied the sparseness of the causal connections, ranging between a single connection (devices 1; 6) to fully connected structures (5; 10). We included problems exemplifying three key types of causal structure: forks (diverging connections), chains (sequential connections) and colliders (converging connections).

There were three different levels of causal strength $S \in [1, .85, .6]$ and three different levels of background noise $B \in [0, .15, .4]$ making $3 \times 3 = 9$ between-subjects conditions. For instance, in condition 1 ($S = 1; B = 0$) the causal systems were perfectly deterministic, with nothing activating without being intervened on, or caused by, an active parent, and connections never failing to cause their effects. Meanwhile, in condition 9, ($S = 0.6; B = 0.4$) the outcomes were very noisy, with probability .4 that a variable with no active parents would activate, compared to a probability $1 - (1 - .6)(1 - .4) = 0.76$ for a variable with one active parent.

Procedure The causal systems were represented as grey circles on a white background. Participants were told that the circles were components of a causal system of binary variables, but were not given any further cover story. Initially, all components were inactive and no connection was marked between them. Participants performed tests by clicking on the components, setting them at one of three states "fixed on", "fixed off" and "free-to-vary", then clicking "test" and observing what happened to the "free to vary" components as a result. The observations were of temporary activity (graphically activated components would turn green and wobble). After each test, participants registered their best guess about the underlying structure. They did this by clicking between the components to select either no connection, or a clockwise or anti-clockwise connection, (represented as black arrows). Participants were incentivized to report their best guess about the structure, through receipt of a 10¢ bonus for each causal relation (or non-relation) correctly registered at randomly selected time points throughout the task.

Participants completed instructions familiarizing them with the task interface; the interpretation of arrows as (probabilistic) causal connections; the incentives for judgment accuracy; and the level of S and B in their condition. To train participants on S and B, they were shown first 10 unconnected components and forced to test them 5 times. The frequency with which the components activated reflected the true background noise level. Then, they were shown a set of two-component causal systems where component "A" was a cause of "B", and were forced to test these systems 5 times by fixing component "A" on. This indicated that the frequency with which "B" activated reflected the level of S combined with the background noise they had already learned (e.g. 76% of the time in condition 9).

After completing the instructions and correctly answering comprehension checks, participants solved a practice problem drawn from the five three-variable problems. They then faced the 10 test problems in random order, with randomly orientated unlabeled components. They were given six tests per three variable problem and eight tests per four variable problem. After the final test for each problem they received feedback telling them the true connections.

Results

Performance by condition We expected the quality of participants' judgments to be bracketed by those of a random ($\frac{1}{3}$ per link, given the three possibilities) and a Bayes-optimal observer. For the latter, we calculated the posterior distributions over the task using Bayesian integration based on the outcomes the participants actually observed, calculating the likelihoods using the true causal strength S and background noise B, assuming a uniform prior at the start of each problem. By reporting the MAP structure (guessing in the event of ties) participants could have achieved accuracies ranging between $.84 \pm 0.14$ in condition 2 and $.55 \pm 0.09$) in the nosiest condition, 9 (see Figure 3, blue circles). Optimal learning predicts differences by condition, with a considerable reduction in accuracy going from no to high background noise, and a more moderate reduction going from perfectly strong to highly unreliable causal connections.

Participants significantly outperformed chance in all nine conditions (all p values $< .05$ for t-tests comparing to $\frac{1}{3}$). However they underperformed the Bayes-optimal observer (t-test p values $< .05$) in all conditions bar condition 2 $S = 0.85, B = 0$ (p=0.07). Like the optimal observer, participants became less accurate as noise increased, with a main effect of *B*ackground noise $F(2,147) = 6.34, \eta^2 = 0.07, p = 0.002$ with lower performances for $B = 0.1, t(147) = -2.23, p = 0.03$ and $B = 0.4, t(147) = -3.5 p < .001$ compared to $B = 0$, but no main effect of *S*trength $F(2,147) = 1.2, p = 0.3$.

Participants marked more causal connections per problem than the optimal learner, mean$\pm$*SD* estimates 2.93 ± 1.4 com-

pared to 2.75 ± 1.4, $t(2998) = 3.5, p = 0.0005$. The true proportion was 2.6. The number of connections participants marked on average was affected by both B and S, going from 2.78 ± 1.5 for $B = 0$ to 3.14 ± 1.4 for $B = 0.4$, and 2.77 (SD=1.4) for $S = 0.6$ to 3.01 (SD=1.4) for $S = 1$.

Performance by problem Average accuracy on three variable problems was fractionally higher than on four variable problems $.55 \pm 0.34$ compared to $.52 \pm 0.29$, $t(1463) = 2.0, p = 0.04$, and tests were completed marginally quicker with medians 12.3s and 14.6s. Due to the unrestricted timing of the study, test times were highly positively skewed. Therefore, we tested for a difference between medians by permutation test (Higgins, 2004), finding it significant $p < .0001$. However, there was no main effect of the number of connections on judgement accuracy $F(1,1498) - 2.1, \eta^2 = 0.001, p = 0.14$.

There was a significant main effect of device type $F(5,1444) = 2.91, \eta^2 = 0.007, p = 0.02$ (see Figure 2). Accuracy was lowest for chains (devices 3; 8) 0.49 ± 0.28, and highest for colliders 0.57 ± 0.30 (4; 9). Taking the chain as treatment group, the main effect of device was driven by higher accuracy on colliders (4; 9) $t(1497) = 3.2; p = 0.001$, and marginally higher performance on singly- (1; 6) and fully-connected (5; 10) structures.

Changing judgements Comparing participants' sequences of structure judgments indicates that they shift markedly less frequently than the optimal observer, changing an average of 0.94 ± 1.3 connections after each test compared with 1.78 ± 1.5, $\chi^2(6) = 1920, p < .0001$ (see Figure 3b)

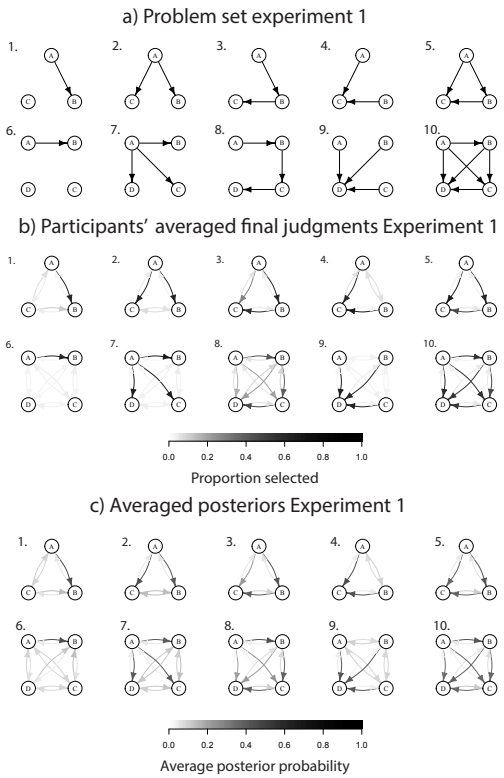

Figure 2: a) The problems faced by participants. b) Weighted average final judgments by participants. Darker arrows indicate that a larger proportion of participants marked this link in their final model. c) Bayes-optimal marginal probability of each edge in $P(M|\mathbf{d}_{1:T}, S, B)$, averaged over participants.

Discussion

Participants identified causal connections above chance even in the most complex and noisy situations we tested. Nevertheless, they were systematically less accurate than they could have been. This is hardly surprising given the considerable complexity of the inferences, and invites comparison with the heuristics discussed earlier. That response times do not increase greatly going from three- to four-variable problems argues against explicit Bayesian-like calculations, as these grow at least $O(2^N)$ with increasing number of variables N. Nevertheless, that the ensemble behavior across all participants resembles the (averaged) posteriors (Figure 2) is in line with the idea that individuals' judgments can be plausibly thought of as individual particles. The strong sequential dependence in judgments argues firmly against their representing the whole distribution. Finally, systematic overconnecting, especially for high B, fits with subjects' failing to compute the exact likelihoods even when they know the parameters, but rather relying on more generic or heuristic approximations.

As a hint that the heuristic models discussed above might therefore offer a better model of the subjects' behavior, the green dots in figure 3 show the case of EA with $\theta_1 = \theta_2 \to \infty$ (so that the MAP structure is chosen at each iteration), and with ties broken randomly. This matches more closely the subjects' performances per condition, and also their patterns of sequential judgment edits.

Modeling

To test the models more formally, we fit the likelihoods of the various combinations, as in the example of equation 5, to the judgments $b_{t=1:T}$ of all participants, for all problems. We expect the resulting θ parameters to be such that lower dissimilarities and fewer explanatory inadequacies lead to more probable selection. Judgments at $t = 0$ were assumed to be an unconnected causal model, but starting evaluation at $t = 1$, when a judgment was already in place, produces comparable results.

We also considered two baseline models. One is a a parameter-free model that assumes each judgment is a random draw from all possible causal models $p(b_t = m) = Unif(M)$ (leading to a probability $\frac{1}{3}$ for each link). The other model is a variant of the Bayes-optimal model that allows decision noise to corrupt choices from the true posterior at t, $P(M|D,S,B)_t$. For this, we considered

$$P(b_t|D) = \frac{\exp(P(M|D,S,B)_t \theta_1)}{\sum_{m \in M} \exp(P(m|D,S,B)_t \theta_1)} \quad (6)$$

controlled again by an inverse temperature parameter θ_1.

Separately, we estimated maximum likelihood S^* and B^*

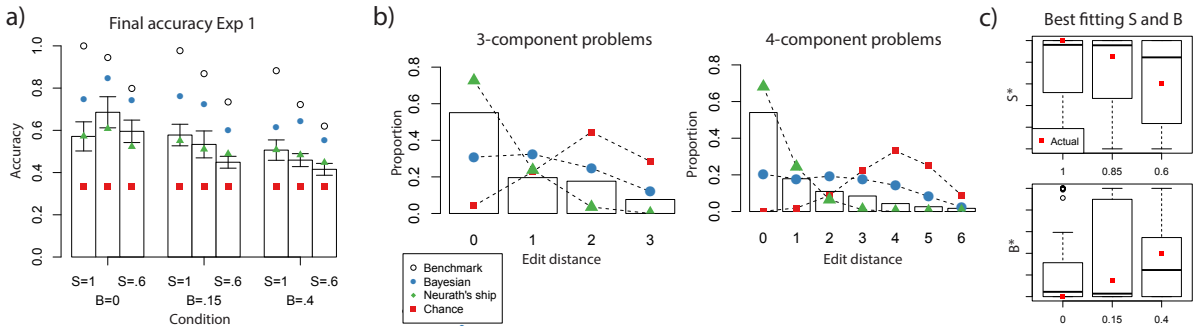

Figure 3: a) Mean final accuracy with standard errors. White circle: benchmark (greedy expected information gain maximizing) Bayesian learner. Blue circles: Bayesian learner that maximizes over the posterior after seeing participants' interventions. Green triangles: "Neurath's ship" simulation simply minimizing number of edits E and failures to explain A. Red squares: random guessing. b) Bars show average number of edits (additions, subtractions or reversals of connections) between all t and t+1 judgments, as compared to Bayesian, "Neurath's ship" and random choice simulations. c) Boxplot of best fitting S^* and B^* parameters assuming learners soft-maximised over $P(M|D,S^*,B^*)$.

parameters for participants assuming Equation 6.

Fitting the models Altogether we fit 9 different (fixed effects) models separately to each of the 150 participants. Models were fit using maximum likelihood as implemented by R's `optim` function, and compared using their BIC scores to accommodate their different numbers of parameters. Results are detailed in Table 1.

Table 1: Experiment 1 - Models fitted to individuals' judgments by maximum likelihood. McFadden's pseudo-R^2 is reported, alongside BIC, median soft-maximization weighting parameter estimates θs. N best fit according to BIC, and average final judgment accuracy for those best fit.

Model	BIC	Rsq	θ_1	θ_2	N fit	Accuracy
Baseline	104535	0			0	
$P(M\|D,S,B)$	91629	0.13	8		0	
L	97532	0.07		2.9	0	
A	98152	0.07		-1	1	.33
E^*	80406	0.24	-4.3		1	.33
E	58892	0.44	-2.2		35	.33
E^* L	73047	0.31	-4.6	3	1	.49
E^* A	74202	0.3	-4.4	-1.1	1	.31
E L	**51146**	**0.52**	**-2.4**	**4**	**60**	**.63**
E A	51665	0.52	-2.4	-1.3	51	.56

Model fit results and discussion

These results show that the large majority of participants are best described by variants of the structural change model of causal judgment that simply balances judgment inertia against a desire to accommodate the latest evidence. Participants were split fairly evenly between being better captured by the true likelihoods L compared to the simple explanatory inadequacy A proxy. Furthermore, estimated S^* and B^* values were less variable over conditions than the true values and stronger and sparser on average (Figure 3c), in line with the idea that participants relied on simplifying assumptions over trained likelihoods. No participant was best described by soft-maximising over the Bayesian posterior. Participants with average accuracy levels at chance were predominantly best captured by the E only model, indicating that their judgments were sequentially dependent but did not meaningfully reflect the data. The better fit for models using edit distance E rather than E^* suggests that participants do not just stick with the same model, but rather tend to make local, rather than drastic, changes.

Experiment 2

The fact that many participants are well captured by the model that relies on heuristic likelihoods suggests that people will still be able to learn causal models well even if they do not know S and B parameters explicitly. We therefore designed a second experiment (demo at ucl.ac.uk/lagnado-lab/el/ns15b) to test this effect. Furthermore, by asking subjects to re-register every link after every new test, we fixed a potential shortcoming of Experiment 1, in which the inertia in judgments might have arisen from subjects' response laziness (i.e., not being bothered to change links) rather than inferential heuristics.

Participants 111 UCL undergraduates (mean±SD age 18.7 ± 0.9, 22 male) took part in Experiment 2 as part of a course. They were incentivized as previously, but this time with the opportunity to win Amazon vouchers rather than money directly. They were split randomly into 8 conditions mean size 13.8 ± 3.4.

Design and procedure Experiment 2 used the same task interface as Experiment 1, but focused just on the three variable problems. There were two background noise conditions $B \in [.1, .25]$ and two causal strength conditions $S \in [.9, .75]$. However, unlike in Experiment 1, participants were not trained on these parameters, but only told that: "the connections do not always work", and "sometimes components can activate by chance".

To assess the influence of laziness, we examined two reporting conditions between subjects: *remain* and *disappear*. In the *remain* condition, judgments stayed on the screen into the next test, so participants did not have to change anything if they wanted to register the same judgement at t as at $t-1$. In the *disappear* condition, the previous judgment disappeared as soon as participants entered a new test. They then had explicitly to select what they wanted for every connection after each test.[2] At the end of the task, people were asked to estimate, in 100 tries how often: "components turn on by themselves?" (B) and "how often do the causal connections

[2]We also elicited additional judgments about expected outcomes of interventions, confidence in individual connections and 'helpfulness' of each outcome; however we do not report on these here for space reasons.

work?"(S).

Results and modeling

Performance in Experiment 2 was comparable to the 3-variable problems in Experiment 1. For example, mean±SD accuracy in Experiment 2, $[B = 0.1, S = 0.75]$ was $.63 \pm 0.27$ and $[B = 0.25, S = 0.75]$ was $.58 \pm 0.31$ while Experiment 1 condition 5 $[B = 0.15, S = 0.85]$ was $.60 \pm 0.33$. This suggests that people can make reasonable structure judgments without knowledge of exact parameters. Supporting these conclusions – participants' final judgments of S and B suffered bias and variance: for $B = \{.1; .25\}$ the mean±SD estimates were $\{.37 \pm .24; .48 \pm .20\}$ respectively; for $S = \{.9, .75\}$, mean±SD estimates were $\{.75 \pm .21; .64 \pm .23\}$.

As with Experiment 1, participants were affected by higher levels of background noise $B\ t(108) = 2.7, p = 0.008$, but not the reliability of the links themselves $S\ t(106) = 0.88, p = 0.37$, and there was no difference in performance between the two judgment elicitation conditions $t(108) = 0.67, p = 0.50$. Analysis of variance revealed an effect of condition on final judgment accuracy $F(7,103) = 2.87, \eta^2 = 0.16, p = 0.008$ with a significant interaction between S and judgment type, with a .21 additional drop in accuracy going from S=0.9 to S=0.75 in the disappear condition compared to the remain condition.

To check if the structure change model in Experiment 1 was driven by lazy reporting, we fit the models as before[3]. We found that once again the large majority of participants were fit by variants of the structural change model, both when judgments remain (47/53) and when they disappear (48/58), this time with a larger proportion better fit by EL than EA (32/47 for remain and 32/48 for disappear conditions), suggesting some sensitivity to the noisy-or aspect of the likelihoods at least for three variable problems. 5/53 and 10/53 in the remain and disappear conditions respectively were best fit by the model based on the Bayesian posterior $P(D|M)$.

General Discussion

In sum, people were able to learn complex causal models, but exhibited strong sequential dependence and variability in their judgments. These patterns were well-captured by a heuristic model, inspired by "Neurath's ship", that maintains a single model, and attempts to account for incoming evidence by making local changes. However, we have not yet provided a plausible process model for the local search.

The model is still too simple in at least three respects. First, it assumes no memory of past evidence beyond the insufficient statistic of the current causal model. It is likely that subjects can remember some past experience, and combine it with the current datum when updating their beliefs. Of course, outside the lab setting, it is unlikely that our experience relevant to single causal models is adequately contiguous for this to be very useful in practice.

[3]For $P(D|M)$ we used importance sampling with 20,000 particles to marginalize over S and B, updating a density for each over the course of the 36 trials in the task.

Second, while participants' judgments showed high sequential dependence, they did occasionally change their model abruptly. The theory of unexpected uncertainty (Yu & Dayan, 2003), and substantial work on changepoint tasks (Speekenbrink & Shanks, 2010) are associated with the notion that people will sometimes "start over" if they are having consistently poor predictions from their existing model (Lakatos, 1976). Experiments in which the underlying structure changes over time would provide pointers.

Finally, we did not examine the selection of interventions, but only how to learn from them. Participants' interventions were far from perfectly efficient – in 100 simulations of the task, an active learning algorithm that selects interventions greedily to minimize its expected uncertainty over the space of possible structures, and updates beliefs optimally, achieves considerably higher final accuracy (mean 0.81, see white circles in Figure 3) compared with what could be achieved given the data participants actually saw (mean 0.69). This also raises further important questions.

Acknowledgements PD was supported by the Gatsby Charitable Foundation.

References

Bramley, N. R., Gerstenberg, T., & Lagnado, D. A. (2014). The order of things: Inferring causal structure from temporal patterns. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (p. 236-242). Austin, TX: Cognitive Science Society.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2014). Forgetful conservative scholars - how people learn causal structure through interventions. *Journal of Experimental Psychology: Learning, Memory and Cognition*.

Buchanan, D. W., Tenenbaum, J. B., & Sobel, D. M. (2010). Edge replacement and nonindependence in causation. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 919–924).

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

Coenen, A., Rehder, B., & Gureckis, T. (2014). Decisions to intervene on causal systems are adaptively selected. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 34th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, *9*(4), 309–347.

Courville, A. C., & Daw, N. D. (2007). The rat as particle filter. In *Advances in neural information processing systems* (pp. 369–376).

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of experimental psychology: Learning, memory, and cognition*, *35*(3), 678.

Higgins, J. J. (2004). *An introduction to modern nonparametric statistics*. Brooks/Cole Pacific Grove, CA.

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

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of experimental psychology. Learning, memory, and cognition*, *32*(3), 451–60.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can theories be refuted?* (pp. 205–259). Springer.

Liu, J. S., & Chen, R. (1998). Sequential monte carlo methods for dynamic systems. *Journal of the American statistical association*, *93*(443), 1032–1044.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological review*, *115*(4), 955.

Mayrhofer, R., & Waldmann, M. R. (2011). Heuristics in covariation-based induction of causal models: Sufficiency and necessity priors. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 3110–3115).

Pearl, J. (2000). *Causality*. New York: Cambridge University Press (2nd edition).

Quine, W. v. O. (1969). *Word and object*. MIT press.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*(4), 1144.

Speekenbrink, M., & Shanks, D. R. (2010). Learning in a changing environment. *Journal of Experimental Psychology: General*, *139*(2), 266.

Steyvers, M. (2003, June). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? optimal decisions from very few samples. In *Proceedings of the 31st annual conference of the cognitive science society* (Vol. 1, pp. 66–72).

Yu, A., & Dayan, P. (2003). Expected and unexpected uncertainty: ACh and NE in the neocortex. *Advances in neural information processing systems*, 173–180.