Formalizing Neurath’s Ship: Approximate Algorithms for Online Causal Learning

Neil R. Bramley\textsuperscript{1,}\textsuperscript{*}
Peter Dayan\textsuperscript{2}
Thomas L. Griffiths\textsuperscript{3}
David A. Lagnado\textsuperscript{1}

\textsuperscript{1}Experimental Psychology, UCL, London, WC1H 0DS, UK
\textsuperscript{2}Gatsby Computational Neuroscience Unit, UCL, London, W1T 4JG
\textsuperscript{3}Department of Psychology, UC Berkeley, CA 94720-1650

\textsuperscript{*}Corresponding author
February 11, 2016

Address for correspondence:
Room 201, 26 Bedford Way
University College London
London, UK
WC1H 0DS
Tel: +44 7914419386
Email: neil.bramley@ucl.ac.uk
Abstract

Higher-level cognition depends on the ability to learn models of the world. We can characterize this at the computational level as a structure-learning problem with the goal of best identifying the prevailing causal relationships among a set of relata. However, the computational cost of performing exact Bayesian inference over causal models scales extremely poorly, becoming infeasible for more than a few relata. This implies that the cognitive processes underlying causal learning must be substantially approximate. This need for approximation is captured by the Neurath’s ship metaphor in philosophy of science, where theory change is cast as a stochastic and gradual process shaped as much by people’s limited willingness to abandon their current theory when considering alternatives as by the ground truth they hope to approach. Inspired by this metaphor and by algorithms for approximating Bayesian inference in machine learning, we propose an algorithmic-level model of causal structure learning under which learners represent only a single global hypothesis that they update locally as they gather evidence. We propose a related scheme for understanding how, under these limitations, learners choose informative interventions that manipulate the causal system to help elucidate its workings. We find support for our approach through two behavioral experiments.

KEYWORDS: active learning; causal learning; theory change; resource rationality; intervention
By adulthood, a normal person will have developed a sophisticated and structured understanding of the world. The “blooming buzzing confusion” (James, 2004, p.462) of moment-to-moment sensory experience will have given way to a more coherent dance of objects and forces, relata and causal relationships. Such representations enable humans to exploit their physical and social environments in flexible and inherently model-based ways (Dolan & Dayan, 2013; Griffiths & Tenenbaum, 2007; Sloman, 2005). An important question therefore, is how people learn appropriate causal relationships from the data they gather by observing and manipulating the world. Much recent work on causal learning has used Pearl’s (2000) causal Bayesian network framework to demonstrate that people make broadly normative causal inferences based on cues like observed contingencies or the outcomes of interventions, which are manipulations or tests of the system (e.g. Griffiths & Tenenbaum, 2009; Holyoak & Cheng, 2011; Lagnado & Sloman, 2004, 2006; Lagnado, Waldmann, Hagmayer, & Sloman, 2007). Related work has begun to explore how people engage in “active learning” – selecting interventions on variables in systems of interest in order to be effective at reducing uncertainty about the true causal model (Bramley, Lagnado, & Speekenbrink, 2015; Coenen, Rehder, & Gureckis, 2015; Sobel & Kushnir, 2006; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003).

Models of human causal learning based on Bayesian networks have tended to focus on what Marr (1982) called the computational level. This means that they consider the abstract computational problem being solved and its ideal solution rather than the actual cognitive processes involved in reaching that solution – Marr’s algorithmic level. In practice the demands of computing and storing the quantities required for exactly solving the problem of causal learning are intractable for any non-trivial world and plausibly-bounded learner. Even a small number of potential relata permit massive numbers of patterns of causal relationships. Moreover, real learning contexts involve noisy (unreliable) relationships and the threat of exogenous interference, further compounding the complexity of normative inference. Navigating this space of possibilities optimally would require maintaining probability distributions across many models and updating all these probabilities whenever integrating new evidence. This evidence might in turn be gathered piecemeal over a lifetime of experience. Doing so efficiently would require choosing maximally informative interventions, a task which poses even greater computational challenges: consideration and weighting of all possible outcomes, under all possible models for all possible interventions (Murphy, 2001; Nyberg & Korb, 2006).

In order to understand better the cognitive processes involved in learning causal relationships, we present a detailed exploration of how people, with their limited processing resources,
represent and reason about causal structure. In particular, we draw on the literature on algorithms for approximating probabilistic inference in computer science and statistics. Previous experiments on human causal learning have focused on situations with small numbers of possible structures, semi-deterministic relationships and limited choices or opportunities to intervene. These constraints have limited the computational demands, and thus the need for heuristics or approximations. Therefore, to explore how people learn in real contexts, we ran two experiments designed to tax learning more severely, with a broader range of structures, multiple opportunities to gather evidence, and substantial noise (Experiment 1) whose level and nature participants have to infer as they learn (Experiment 2).

As a framework for understanding these results, we consider a class of algorithms for causal structure learning inspired by an old idea about theory change. The idea can illustrated by a metaphor, originally attributed to Otto Van Neurath (1932/1983) but popularized by Quine, who writes:

“We [the theorist] 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. In this way, by using the old beams and driftwood the ship can be shaped entirely anew, but only by gradual reconstruction.” (1969, p3)

The Neurath’s ship metaphor describes the piecemeal growth and evolution of scientific theories over the course of history. In the metaphor, the theorist (sailor) is cast as relying on their existing theory (ship) to stay afloat, without the privilege of a dry-dock in which to make major improvements. Unable to step back and consider all possible alternatives, the theorist is limited to building on the existing theory, making a series of small changes with the goal of improving the fit.

We argue that people are in a similar position when it comes to their beliefs about the causal structure of the world. We propose that people normally maintain only a single hypothesis about the global causal model, rather than a distribution over all possibilities. They update their hypothesis by making local changes (e.g. adding, removing and reversing individual connections, nodes or subgraphs) while leaning on the rest of the model for support. We show that by doing this, a learner could end up with a relatively accurate causal model without ever representing the whole hypothesis space or storing all the old evidence, but that their causal beliefs will exhibit a particular pattern of sequential dependence. We provide a related account
of bounded intervention selection, based on the idea that learners adapt to their own learning
limitations when choosing what evidence to gather next, attempting to resolve local rather than
global uncertainty. Together, our Neurath’s ship model and local-uncertainty-based schema for
intervention selection provide a step towards an explanation of how people might achieve a
resource rational (Griffiths, Lieder, & Goodman, 2014; Simon, 1982) trade-off between accuracy
and the cognitive costs of choosing interventions and updating beliefs.

The paper is organized as follows. We first formalize causal model inference at the com-
putational level. We then motivate and present our Neurath’s ship framework for belief change
and active learning as an approximate algorithmic-level solution to this problem. We then
present two behavioral experiments and show that participants’ overall patterns of judgments
and intervention choices are in line with the predictions of our framework. Finally, we fit
our models to the experimental data and show that they provide a better fit than baseline or
computational-level competitors. We provide additional details about the formal framework
and model specification in Appendix A and additional details about the modeling and model
variants in Appendix B. Also, where indicated, additional results and figures are provided in
Supplementary materials at [http://www.ucl.ac.uk/lagnado-lab/el/nsm](http://www.ucl.ac.uk/lagnado-lab/el/nsm).

A computational-level framework for active structure learning

Before presenting our theoretical framework, we lay out our computational-level analysis
of the problem of structure learning. This can be broken down into three interrelated elements:
(1) the representation of a causal model (2) performing inference over possible models given
evidence (observations and the outcomes of interventions), and (3) selecting interventions to
gather more evidence and support this inference. We introduce the three elements here, pro-
viding more detail where indicated in Appendix A.

Representation

For representing causal models we adopt a standard approach – the parameterized directed
acyclic graph (Pearl, 2000, see Figure 1a). Nodes represent variables (i.e. the component parts
of a causal system); arrows represent causal connections; and parameters encode the combined
influence of parents (the source of an arrow) on children (the arrow’s target)\(^1\). Such graphs
can represent continuous variables and any form of causal relationship; but here we focus on

\(^1\)Following standard graph nomenclature, we will often refer to the space between a pair of nodes in a model as an
“edge”, so that an acyclic causal model defines each edge as taking one of three states: forward \(\rightarrow\), backward \(\leftarrow\), or
inactive \(\emptyset\).
systems of binary \{0 = \text{absent}, 1 = \text{present}\} variables and and assume generative connections – meaning we assume that the presence of a cause will always raise the probability that the effect is also present.

We also adopt Cheng’s Power PC (1997) convention for parameterization, which provides a simple way to capture how probabilistic causal influences combine. This assumes that causes have independent chances of producing their effects, meaning the probability that a variable takes the value 1 is a noisy-OR combination of the power or strength \(w_S\) of any active causes of it in the model, together with that of an omnipresent background strength \(w_B\) encapsulating the influence of any causes exogenous to the model. We write \(w = \{w_S, w_B\}\). The probability that variable \(x\) takes the value 1 is thus

\[
P(x = 1|pa(x), w) = 1 - (1 - w_B)(1 - w_S)^{\sum_{y \in pa(x)} y} ,
\]

where \(pa(x)\) denotes the parents of variable \(x\) in the causal model (see Figure 1a for an example). In the current work, we assume \(w\) is the same for all connections and components².

### Inference

In this framework, each causal model \(m\) over variables \(X\) with strength and background parameters \(w\), assigns a probability to each datum \(d = \{x \ldots z\}\). In turn, this probability depends on the collection of variables whose values are determined by propagation in the model; and those that are fixed, being determined through intervention \(c\) (see Figure 1b). The space of all possible interventions \(C\) is made up of all possible combinations of fixed and unfixed variables. We use Pearl’s Do\([\_]\) operator (Pearl, 2000) to denote what is fixed on a given test. For instance, Do\([x = 1, y = 0]\) means a variable \(x\) has been fixed “on” and variable \(y\) has been fixed “off”, with all other variables free to vary³. Interventions allow a learner to override the normal flow of causal influence in a system, initiating activity at some components and blocking potential influences between others. This means they can provide information about the presence and direction of influences between variables that is typically unavailable from purely observational data (see Bramley, Lagnado, & Speekenbrink, 2015, for a more detailed introduction), without additional cues such as temporal information (Bramley, Gerstenberg, & Lagnado, 2014). For instance, in Figure 1b, we fix \(y\) to 1 and leave \(x\) and \(z\) free (Do\([y = 1]\)).

²We also restrict ourselves to situations with no latent variables, although we note that imputing the presence of hidden variables is another important and computationally-challenging component of causal inference (Buchanan, Tenenbaum, & Sobel, 2010; Kushnir, Gopnik, Lucas, & Schulz, 2010).

³We include the pure observation Do\([\emptyset]\) in \(C\).
Under the $x \rightarrow y \rightarrow z$ model we would then expect $x$ to activate with probability $w_B$ and $z$ with a probability of $1 - (1 - w_B)(1 - w_S)$.

In total, the probability of datum $d$, given intervention $c$, is just the product of the probability of each variable that was not intervened upon given the states of its parents in the model

$$P(d|m, w, c) = \prod_{x \in X} P(x|\{d, c\}_{pa(x)}, w).$$

(2)

To perform inference optimally, we treat the true model as a random variable $M$. Our prior belief $P(M)$ is then an assignment of probabilities, adding up to 1 across possible models $m \in M$ in the set of models $M$. When we observe some data $D = \{d^i\}$, associated with interventions $C = \{c^i\}$, we can update these beliefs with Bayes theorem by multiplying our prior by the probability of the observed data under each model and dividing by the average probability of those data across all the possible models:

$$P(m|D, w; C) = \frac{P(D|m, w; C)P(m)}{\sum_{m' \in M} P(D|m', w; C)P(m')}.$$  

(3)

We will typically treat the data as being independent and identically distributed, so $P(D|m, w; C) = \prod_i P(d^i|m, w; C)$. If the data arrive sequentially (as $D^t = \{d^1, \ldots, d^t\}$; and similarly for the interventions), we can either store them and update at the end, or update our beliefs sequentially, taking the posterior $P(M|D^{t-1}, w; C^{t-1})$ at timestep $t - 1$ as the new “prior” for datum $d^t$. If we are also unsure about the parameters of the true model (i.e. $w_B$ and $w_S$) we have to treat them as random variables too and average over our uncertainty about them to compute a marginal posterior over models $M$ (see Appendix A.1).

![Causal model representation. a) An example causal Bayesian network, parametrized with strength $w_S$ and base rate $w_B$. The tables give the probability of each variable taking the value 1 conditional on its parents in the model and the omnipresent background noise rate $w_B$. b) Visualisation of intervention Do[y = 1]. Setting $y$ to 1 renders it independent of its normal causes as indicated by the scissors symbols.](image-url)
Choosing interventions

It is clear that different interventions yield different outcomes, which in turn have different probabilities under different models. This means that which interventions are valuable for identifying the true model depends strongly on the hypothesis space and prior. For instance, fixing $x$ to 1 and $y$ to 0 ($Do[x = 1, y = 0]$) is (probabilistically) diagnostic if you are primarily unsure whether $x$ causes $z$ because $p(z | Do[x = 1, y = 0])$ differs depending whether $pa(z)$ includes $x$. However, it is not diagnostic if you are primarily unsure whether $x$ causes $y$ because $p(y | Do[x = 1, y = 0])$ is the same regardless of whether $pa(y)$ includes $x$.

The value of an intervention can be quantified relative to a notion of uncertainty. We can define the value of an intervention as the expected reduction in uncertainty about the true model after seeing its outcome\(^4\). To calculate this expectation, we must average, prospectively, over the different possible outcomes $d' \in D_e$ (where $D_e$ is the space of possible outcomes of intervention $c$) weighted by their marginal likelihoods under the prior. For a greedily optimal sequence of interventions $c_1, \ldots, c_t$, we take $P(M | D^{t-1}, w; C^{t-1})$ as our prior each time. The most valuable intervention $c_t$ at a given time point is then

\[
\arg \max_{c \in C} \mathbb{E}_{d' \in D_e} \left[ \Delta H(M | d', D^{t-1}, w; C^{t-1}, c) \right],
\]

where $\mathbb{E}_{d' \in D_e}$ denotes the expected value (i.e. average) over outcomes $d'$ and $\Delta H(.)$ denotes reduction in uncertainty. We use Shannon entropy (2001) to measure uncertainty (see Appendix A.2).

Algorithms for causal learning with limited resources

Unfortunately, both inference and choosing interventions scale so poorly in the number of variables, they are fundamentally intractable for any plausibly bounded learner (Cooper, 1990; van Rooij, Wright, Kwisthout, & Wareham, 2014). The number of possible graphs grows rapidly with the number of variables they relate (3-, 4- and 5-variable problems have 25, 543 and 29281 respectively). Additionally, evidence will often arrive gradually, requiring beliefs be updated frequently if they are to make use of the latest evidence. Active intervention selection adds extra complexity because there are many possible interventions (3-, 4- and 5-variable

---

\(^4\)Strictly this is greedy rather than optimal because planning several steps ahead can result in a different intervention being favored. However, planning ahead was shown to make little difference for the similar problems explored in Bramley et al (2015). Also, it has been shown (Nemhauser, Wolsey, & Fisher, 1978) that greedy strategies are guaranteed to be competitive with optimal strategies for some active learning problems, although to the authors’ knowledge this has not yet been shown for graph inference.
problems permit 27, 81 and 243 patterns of fixed “on”, fixed “off” and free components, each of which might yield many outcomes (up to 8, 16 and 32 respectively, depending how many variables are left free to vary). All combinations of potential model, intervention and outcome should be averaged over in order to select the most valuable intervention. This implies that people must find a considerably more economical way to approximate model inference while maintaining satisfactory accuracy.

Various techniques have been developed in recent years in machine learning and statistics that make approximate learning efficient in otherwise intractable circumstances. Additionally, research in these fields on active learning and optimal experiment design has identified a range of reasonable heuristics for selecting queries when the full expected information calculation of (Equation 4) is intractable. We will take inspiration from some of these ideas to give a formal basis to the intuitions behind the Neurath’s ship metaphor. We will then use this formal model to generate predictions that we will compare to participants’ behaviour in our experiments.

Approximating with a few hypotheses

One common approximation, for situations where a posterior cannot be evaluated in closed form, is to maintain a manageable number of individual hypotheses, or “particles” (Liu & Chen, 1998), with weights corresponding to their relative likelihoods. The ensemble of particles then acts as an approximation to the desired distribution. Sophisticated reweighting and resampling schemes can then filter the ensemble as data are observed, approximating Bayesian inference.

These “particle filtering” methods have been used to explain how humans and other animals might approximate the solutions to complex problems of probabilistic inference. In associative learning (Courville & Daw, 2007), categorization (Sanborn, Griffiths, & Navarro, 2010) and binary decision making (Vul, Goodman, Griffiths, & Tenenbaum, 2009), it has been proposed that people’s beliefs actually behave most like a single particle, capturing why individuals often exhibit fluctuating and sub-optimal judgment while maintaining a connection to Bayesian inference.

Causal learning carries a particular motivation for frugal representation. Since the hypothesized role of a causal model for a cognitive system is to represent accumulated knowledge compactly and make inference cheaper (Pearl, 2000), it is plausible that there is particular pressure to minimize how many models must be stored, weighted and averaged over in inference. In line with this, Bonawitz, Denison, Gopnik, and Griffiths (2014) proposed a “win-stay, lose-sample” scheme that can be applied in the context of causal learning, suggesting learn-
ers maintain a single structural hypothesis, resampling from the posterior only after observing data that strongly contradicts it. Unlike particle filtering, this scheme guarantees that the single hypothesis remains a sample from the posterior distribution at every point, providing an explanation for how aggregate behavior can resemble Bayesian inference even as individuals exhibit sub-optimal choices and statistical dependencies among those choices.

**Sequential local search**

The idea that people’s causal theories are like particles requires they also have some procedure for sampling or adapting these theories as evidence is observed. Another class of useful machine learning methods involves generating sequences of hypotheses, each linked to the next via a form of possibly stochastic transition mechanism. Two members of this class are particularly popular in the present context: Markov Chain Monte Carlo (MCMC) sampling, which asymptotically approximates the posterior distribution; and (stochastic) hill climbing, which merely tries to find hypotheses that have high posterior probabilities.

MCMC algorithms involve stochastic transitions with samples that are typically easy to generate. Under various conditions, this implies that the sequences of (dependent) sample hypotheses form a Markov chain with a stationary distribution that is the full, intended, posterior distribution (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). The samples will appear to “walk” randomly around space of possibilities, tending to visit more probable hypotheses more frequently. If samples are extracted from the sequence after a sufficiently long initial, so-called burn-in, period, and sufficiently far apart (to reduce the effect of dependence), they can provide a good approximation to the true posterior distribution. There are typically many different classes of Markov chain transitions that share the same stationary distribution, but differ in the properties of burn-in and subsampling.

The stochasticity inherent in MCMC algorithms implies that the sequence sometimes makes a transition from a more probable to a less probable hypothesis – this is necessary to sample multi-modal posterior distributions. A more radical heuristic is only to allow transitions to more probable hypotheses — this is called “hill-climbing”, attempting to find, and then stick at, the best hypothesis (Tsamardinos, Brown, & Aliferis, 2006). This is typically faster than a full MCMC algorithm, but is prone to become stuck in a local optimum, where the current hypothesis is more likely than all its neighbors, but less likely than some other more distant hypothesis.

Applied to causal structure inference, we might in either case consider transitions that change at most a single edge in the model (Cooper & Herskovits, 1992; Goudie & Mukherjee,
A simple case is Gibbs sampling (Geman & Geman, 1984), starting with some structural hypothesis and repeatedly selecting an edge (randomly or systematically) and re-sampling it (either adding, removing or reversing) conditional on state of the other edges. This means that a learner can search for a new hypothesis by making local changes to their current hypothesis, “rethinking” each of the edges in turn, conditioning on the state of the others without ever enumerating all the possibilities. By constructing a short chain of such “rethinks” a learner can easily update a singular hypothesis without starting from scratch. The longer the chain, the less dependent or “local” the new hypothesis will be to the starting point.

The idea that people might update their judgments by something like MCMC sampling is explored by Lieder, Griffiths and Goodman (2012; under review). They argue that under reasonable assumptions about the costs of resampling and need for accuracy, it can be rational to update one’s beliefs by constructing short chains where the the updated judgment retains some dependence on its starting state, arguing that this might explain anchoring effects (Kahneman, Slovic, & Tversky, 1982).

In addition to computational savings, updating beliefs by local search can be desirable for another reason. If the learner has forgotten some of the evidence they have seen, the location of their previous hypothesis acts like a very approximate version of a prior (as in a single-particle particle filter), carrying some of this forgotten information. This can make it advantageous to the learner to strike a good balance between editing their model to better account for the data they can remember, and staying close to their previous model to retain the connection to the data they have forgotten (Bramley, Lagnado, & Speekenbrink, 2015).

Neurath’s ship: An algorithmic-level model of sequential belief change

The previous section summarized two ideas derived from computer science and statistics that provide a potential solution to the computational challenges of causal learning: maintaining only a single hypothesis at a time, and exploring new hypotheses using local search based on sampling. In this section, we formalize these ideas to define a class of models of causal learning inspired by the metaphor of Neurath’s ship. We start by treating interventions as given, and only focus on inference. We then consider the nature of the interventions.

Concretely, we propose that causal learners make inferences by:

1. Maintaining only a single causal model (a single particle), i.e. $b^{t-1}$ at time $t - 1$.

2. After seeing evidence $d^t$, searching for local improvements by sequentially reconsidering
edges \( E_{ij} \in \{ 1 : i \rightarrow j, \; 0 : i \leftrightarrow j, \; -1 : i \leftarrow j \} \) (adding, subtracting or reorienting them) conditional on the current state of the edges in the rest of their model \( E_{\setminus ij} \) – e.g. with probability \( P(E_{ij} | E_{\setminus ij}, d^t, c^t, w) \).

3. After searching for \( k \) steps, then stopping and taking the latest version of their model as their new belief \( b^t \).

A detailed specification of this process is given in Appendix A.3.

Starting with any hypothesis and repeatedly resampling edges conditional on the others is a form of Gibbs sampling (Goudie & Mukherjee, 2011). Further, starting with the previous belief \( b^{t-1} \) allows the learner to make use of the data they have forgotten, since these data are represented in the location of \( b^{t-1} \). Resampling using the latest datum \( P(M | d^t, c^u, w) \) allows the learner to adjust their beliefs to encapsulate better the data they have just seen.

**Resampling, hill climbing or random change**

Following the procedure outlined above, the learner’s search steps would constitute dependent samples from the posterior of \( d^t \). However, it is also plausible that learners will try to hill-climb rather than sample, preferring to move to more probable local models more strongly than would be predicted by Gibbs sampling. In order to explore this, we will consider generalizations of the update equation allowing transitions to be governed by powers of the conditional edge probability (i.e. \( P^\omega(E_{ij} = e | E_{\setminus ij}, d^t, c^t, w) \)), yielding stronger or weaker preference for the most likely state of \( E_{ij} \) depending whether \( \omega > 1 \) or \( < 1 \). By setting \( \omega \) to zero, we would get a model that does not learn but just moves randomly between hypotheses, tending to remain local and by setting it to infinity we would get a model that always moved to the most likely state for the edge.

**Search length**

It is reasonable to assume that the number of search steps \( k \) that a learner performs will be variable, but that their capacity to search will be relatively stable. Therefore, we assume that for each update, the learner searches for \( k \) steps, where \( k \) is drawn from a Poisson distribution with mean \( \lambda \in [0, \infty] \).

The value of \( \lambda \) thus determines how sequentially dependent a learner’s sequences of beliefs are. A large \( \lambda \) codifies a tendency to move beliefs a long way to account for the latest data \( d^t \) at the expense of the older data \( d^1 \ldots d^{t-1} \) – retained only in the location of the previous belief \( b^{t-1} \) – while a moderate \( \lambda \) captures a reasonable trade-off between starting state and new evidence, and a small \( \lambda \) captures conservatism, i.e. failure to shift beliefs enough to account for
the latest data.

**Putting these together**

By representing the transition probabilities from model $i$ to model $j$, for a particular setting of hill climbing parameter $\omega$ and data $d^t$, with a transition matrix $R_{i}^{\omega}$ \(^5\), we can thus make probabilistic predictions about a learner’s new judgment $b^t \in B^t$. The probabilities depend on the previous judgment $b^{t-1}$ and their average search length $\lambda$. By averaging over different search lengths with their probability controlled by $\lambda$, and taking the requisite row of the resulting transition matrix we get the following equation

$$P(b^t = m | d^t, c^t, b^{t-1}, \omega, \lambda) = \sum_{0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} [((R_{i}^{\omega})^k)]_{b^t-1, m}$$

(5)

See Appendix A.3 for more details and Figure 2 for an example.

---

\(^5\)Note that transitions that would create a loop in the overall model get a probability of zero
Figure 2: An illustration of NS model of causal belief updating. a) An example search path: The learner starts out with a singly connected model at the top ($x \rightarrow y$ connection only). They update their beliefs by resampling one edge at a time $e \in \{\rightarrow, \leftarrow, \leftrightarrow\}$. Each entry $i, j$ in the matrices gives the probability of moving from model in the row $i$ to the model in the column $j$ when resampling the edge marked with the colored question mark. Lighter shades of the requisite color indicate low transition probability, darker shades indicate greater transition probability; yellow is used to indicate zero probabilities. Here the learner stops after resampling each edge once, moving from $b^{x^{-1}}$ of $[x \rightarrow y]$ to $b^y$ of $[x \rightarrow y, x \rightarrow z, y \rightarrow z]$. b) Assuming the edge to resample is chosen at random, we can average over the different possible edge choices to derive a 1-step Markov chain transition matrix $R^n_x$ encompassing all the possibilities. By raising this matrix to higher powers we get the probability of different end points for searches of that length. If the chain is short (small $k$) the final state depends heavily on the starting state (top) but for longer chains (large $k$), the starting state becomes less important, getting increasingly close to independent sampling from the desired distribution (bottom).

**Selecting interventions on Neurath’s ship: A local uncertainty schema**

In situations where a posterior is already hard to evaluate, calculating the globally most informative intervention – finding the intervention $e^i$ that maximizes Equation 4 – will almost always be infeasible. Therefore, a variety of heuristics have been developed that allow tests to be selected that are more useful than random selection, but do not require the full expected information gain be computed (Settles, 2012). These tend to rely on the learners’ current, rather
than expected, uncertainty (e.g. *uncertainty sampling* which chooses based on outcome uncertainty under the prior) or the predictions under just a few favored hypotheses (e.g. *query by committee*) as a substitute for the full expectancy calculation. The former relies on maintaining a complete prior distribution, making the latter a more natural partner to the *Neurath's ship* framework.

Having argued that learners are limited to consideration of a few local possibilities at a time we should expect them to face similar limitations in terms of which, and how many, alternatives they can explicitly try to distinguish with a chosen intervention. In line with this, we propose one way in which learners might select robustly informative interventions by attempting only to distinguish a few “local” possibilities at a time, requiring only “local” uncertainty estimates to target the possibilities on which to focus.

The idea that learners will focus on distinguishing only a few alternatives at a time requires specifying how they choose which of the many possible subsets of the full hypothesis space to target with a particular test. In the current work, we will consider three possible varieties of focus, one motivated by the *Neurath's ship* framework (*edge focus*) and two inspired by existing ideas about bounded search and discovery in the literature (*effects focus* and *confirmation focus*). While these are by no means exhaustive they represent a principled starting point. Furthermore, by licensing quite different intervention preferences, they allow us to diagnose individual and trial-by-trial differences in focus preference.

**Representing uncertainty**

Using just the latest evidence $d^t$ to update beliefs is very frugal, as the learner does not have to store any of the older evidence. However, the use of a single hypothesis $b^{t-1}$ is an extremely impoverished substitute for a full distribution, primarily because it does not encode anything about learner’s current uncertainty. Neither overall confidence in the current hypothesis, nor relative confidence in its subparts need be represented to update beliefs in the way we proposed in the previous section. However if learners really have no conception of their own uncertainty, there are severe limits on the quality of their inference and their ability to select interventions. Crucially, any ability to choose interventions dynamically must depend on the learner having some notion of their current uncertainty. Thus, a more moderate possibility is that learners will have approximate ways of estimating their confidence in their current hypothesis and its subcomponents. For instance, learners might update “confidence” weights on the edges in the current model via various statistics encountered as they update their beliefs. In the next section we will explore the idea that learners use such local uncertainty estimates.
to target their interventions dynamically during learning, although, as we discuss below, we remain agnostic about exactly how these are calculated.

**Active learning based on local uncertainty**

The test that maximizes expected global information is often that whose outcome is most uncertain (Settles, 2012, see uncertainty sampling). However, research suggests people often have a preference for some of the more certain queries in the option set (Markant, Settles, & Gureckis, 2015; Parpart, Schulz, Speekenbrink, & Love, 2015). This apparently inefficient behavior led Markant et al. (2015) to speculate that people might be choosing tests targeted to minimizing “local” areas of uncertainty (e.g. between a subset of hypotheses or relative to a rough grained division of the hypothesis space), rather than the global uncertainty over all the possibilities. Queries that optimally reduce expected uncertainty about one local aspect of a problem are liable to differ from those that promise high global uncertainty reduction. For example, Figure 3b shows two trials taken from our experiments, and shows that the expected values of each of a range of different intervention choices (shown in Figure 3a) are very different depending whether the learner is focused on resolving global uncertainty all at once, or on resolving some specific “local” aspect of it. This illustrates the idea that a learner might choose a test that is optimally informative with respect to a modest range of options that they have in mind at the time (e.g. models that differ just in terms of the state of $E_{xz}$) yet appear sporadically inefficient from the perspective the goal of greedy global uncertainty reduction.

We have proposed a model of structure inference under which learners are only able to consider a small set of alternatives at a time, and only able to generate alternatives that are “local” in some dimension. Locally driven intervention selection is a natural partner to this for at least two reasons: (1) Under the constraints of the Neurath’s ship framework, learners would not be able to work with the prospective distributions required to estimate global expected informativeness, but could potentially estimate expected informativeness with respect to a sufficiently narrow sets of alternatives. (2) Evidence optimized to distinguishing local possibilities (focused on one edge at a time for instance) might better support sequential local belief updates (of the kind emphasized in our framework) than the globally most informative evidence (Patil, Zhu, Kopeć, & Love, 2014).

**The two stages of the schema**

The idea that learners focus on resolving local rather than global uncertainty results in a metaproblem of choosing what to focus on next, making intervention choice a two stage pro-
cess. We write $L$ for the set of all possible focuses $l$, and $L \subseteq L$ for the subset of possibilities that the learner will consider at a time, such as the state of a particular edge or the effects of a particular variable. The procedure is:

**Stage 1** Selecting a local focus $l^t \in L$

**Stage 2** Selecting an informative test $c^t$ with respect to the chosen focus $l^t$

Different learners might differ in the types of questions they consider, meaning that $L$ might contain different varieties and combinations of local focuses. We first formalize the two stages of the schema, and then propose three varieties of local focus that learners might consider in their option set $L$ that differ in terms of which and how many alternatives they include.

As mentioned above, we assume that the learner has some way of estimating their current local confidence. We will assume confidence here is approximately the inverse of uncertainty, so assume for simplicity that learners know their current uncertainty in the form $H(l|D^{t-1}, E^{t-1}, w)$ for all $l \in L$ (the assumption we examine in the discussion). They then choose (Stage 1) the locale where they are currently least certain

$$l^t = \arg\max_{l \in L} H(l|D^{t-1}, E^{t-1}, w; C^{t-1})$$  \hspace{1cm} (6)

However, in carrying out Stage 2 we make the radical assumption that learners do not know $P(l^t|D^{t-1}, E^{t-1}, w; C^{t-1})$, but rather, consistent with the method of inference itself, only consider the potential next datum $d'$. This means that the intervention $c^t$ itself is chosen to maximize the expected information with respect to a uniform prior over the “local” alternatives under consideration. Specifically, we assume that $c^t$ is chosen as

$$c^t = \arg\max_{c \in C} \mathbb{E}_{d \in D_e} [\Delta H(l^t|d, w, b^{t-1}; c)]$$  \hspace{1cm} (7)

where we detail the term in the expectation below for the three types of focuses.

Assuming real learners will exhibit some decision noise, we can model both choice of focus and choice of intervention relative to a focus as soft (Luce, 1959) rather than strict maximization giving focus probabilities

$$P(l^t|D^{t-1}, E^{t-1}, w; C^{t-1}) = \frac{\exp(H(l^t|D^{t-1}, E^{t-1}, w; C^{t-1}) \rho)}{\sum_{l \in L} \exp(H(l|D^{t-1}, E^{t-1}, w; C^{t-1}) \rho)}$$  \hspace{1cm} (8)
governed by some inverse temperature parameter \( \rho \), and choice probabilities

\[
P(c^l | l, w, b^{l-1}) = \frac{\exp(\mathbb{E}_{d' \in D_c} \Delta H(l|d', w, b^{l-1}; c^l) \eta)}{\sum_{c \in C} \exp(\mathbb{E}_{d' \in D_c} \Delta H(l|d', w, b^{l-1}; c) \eta)}
\]  

(9)

governed by inverse temperature \( \eta \).

**Three varieties of local focus**

**Edges**

An obvious choice, given the *Neurath’s ship* framework, would be for learners to try to distinguish alternatives that differ in terms of a single edge (Figure 3a), i.e. those they would consider during a single update step.

For a chosen edge \( E_{xy} \) we can then consider a learner’s goal to be

\[
\arg \max_{c \in C} \mathbb{E}_{d' \in D_c} \Delta H(E_{xy}|E_{xy}', d, w, b^{l-1}; c)
\]

(10)

(see Appendix A.4 for the full local entropy equations). This goal results in a preference for fixing one of the nodes of the target edge “on”, leaving the other free, and depending on the other connections in \( b^{l-1} \), either favors fixing the other variables “off” or is indifferent about whether they are “on”, “off” or “free” (Figure 3b).

**Effects**

A commonly proposed heuristic for efficient search in the deterministic domains is to ask about the dimension that best divides the hypothesis space, eliminating the greatest possible number of options on average. This is variously known as “constraint-seeking” (Ruggeri & Lombrozo, 2014) or “the split half heuristic” (Nelson, Divjak, Gudmundsdottir, Martignon, & Meder, 2014). In the case of identifying the true deterministic \( w_S = 1 \) and \( w_B = 0 \) causal model on \( N \) variables through interventions it turns out that the best split is achieved by querying the effects of a randomly chosen variable, essentially asking: “What does \( x \) do?” (Figure 3a)\(^6\). Formally we might think of this question as asking: which other variables (if any) are descendants of variable \( x \) in the true model? This a broader focus than querying the state of a single edge, but considerably simpler question than the global “which is the right causal model?” because the possibilities just include the different combinations of the other variables as effects (e.g. neither, either or both of \( y \) and \( z \) are descendants of \( x \) in a 3-variable model) rather

---

\(^6\)This is also the most globally informative type of test relative to a uniform prior in all of the noise conditions we consider in the current paper
than the superexponential number of model possibilities\textsuperscript{7}.

Relative to a chosen variable $x$, we can write an effect focus goal as

$$\arg \max_{c \in C} \mathbb{E}_{d \in D_e} \left[ \Delta H(De(x)|d, w, b^{t-1}; c) \right]$$

(11)

where $De(x)$ is is the set of $x$’s direct or indirect descendants. This goal results in a preference for fixing the target node “on” (e.g. $Do[x = 1]$) and leaving the rest of the variables free to vary (Figure 3b).

**Confirmation**

Another form of local test is to seek evidence that would confirm or refute the current hypothesis, against a single alternative “null” hypothesis. Confirmatory evidence gathering is a ubiquitous psychological phenomenon (Klayman & Ha, 1989; Nickerson, 1998). Although confirmation seeking is widely touted as a bias, it can also be shown to be optimal, e.g. under deterministic or sparse hypotheses spaces or peaked priors (Austerweil & Griffiths, 2011; Navarro & Perfors, 2011).

Accordingly, Coenen et al. (2015) propose that causal learners adopt a “positive test strategy” when distinguishing causal models. They define this as a preference to “turn on” a parent component of one’s hypothesis – observing whether the activity propagates to the other variables in the way that this hypothesis predicts. They find that people often intervene on suspected parent components, even when this is uninformative, and do so more often under time pressure. In Coenen et al’s tasks, the goal was always to distinguish between two hypotheses, so their model assumed people would sum over the number of descendants each variable had under each hypotheses and turn on the component that had the most descendants on average. However, this does not generalize to the current, unrestricted, context where all variables have the same number of descendants if you average over the whole hypothesis space. However, Steyvers et al (2003) propose a related rational test model that selects interventions with a goal of distinguishing a single current hypothesis from a null hypothesis that there is no causal connection.

Following Steyvers et al. (2003), for a confirmatory focus we consider interventions expected to best reduce uncertainty between their current hypothesis $b^{t-1}$ and a null $b^0$ in which there

\textsuperscript{7}The number of directed acyclic graphs on $N$ nodes, $|\mathcal{M}|_N$, can be computed with the recurrence relation $|\mathcal{M}|_N = \sum_{k \in N} (-1)^{k-1} \binom{N}{2} 2^k (N-k)|\mathcal{M}|_{N-1}$ (see Robinson, 1977)
are no connections (Figure 3a).

$$\arg \max_{c \in C} \mathbb{E}_{d \in D_c} [\Delta H\left(\{b^t, b^{t-1}\}; d, w, b^{t-1}; c\right)]$$

This goal results in a preference for fixing on the root node(s) of the target hypothesis (Figure 3b). The effectiveness of confirmatory focused testing depends on the level of noise and the prior, becoming increasingly useful later once the model being tested has sufficiently high prior probability.

**Implications of the schema**

The local uncertainty schema implies that intervention choice depends on two separable stages. Thus it accommodates the idea that a learner might be poor at choosing what to focus on but good at selecting an informative intervention relative to their chosen focus. It also allows that we might understand differences in learners’ intervention choices as consequences of the types of local focus they are inclined or able to focus on. Learners cognizant of the limitations in their ability to incorporate new evidence might choose to focus their intervention on narrower questions (i.e. learning about a single edge at a time) while others might focus too broadly and fail to learn effectively. In the current work we will fit behavior assuming that learners choose between these local focuses, using their patterns to diagnose which local focuses they include in their option set $L$, which of these they choose on a given test $l^t$ and finally how these choices relate to their final performance.
Figure 3: An illustrative example of local focused uncertainty minimization a) Three possible “local” focuses b) Expected value of 19 different interventions at the start of learning (i.) and after several tests have been performed (ii.) assuming: global expected information gain from the true prior (green squares, and shaded), effects of z focus (red circles), the relationship between x and y (blue triangles) and confirming b—1 (yellow diamonds), assuming a uniform prior over the requisite possibilities and a known wS and wB of .85 and .15. c) The value of these choices of focus according to their current uncertainty Equation 6. Note that confirmation is undefined at the start of learning where both current and null hypothesis are that there are no connections in the model.

Experimental rationale

To explore whether people use these or other approximations when updating causal beliefs and choosing interventions, it is essential to design appropriately difficult tests. The Neurath’s ship framework we have introduced has two distinct signatures. Making only local edits from
a single hypothesis, would lead to sequential dependence, and a tendency to get stuck in local optima, even if all the data are remembered. Conversely, forgetting old evidence, particularly using the current hypothesis as a proxy, would lead to recency effects whereby participants may make judgments inconsistent with evidence gathered earlier during learning.

In terms of interventions, if participants are locally focused we expect their hypotheses to deviate from optimal predictions in ways that can be accommodated by our local uncertainty schema, i.e. selecting interventions that are more likely to be targeted toward local rather than global uncertainty. If learners do not maintain confidence estimates we expect their intervention distributions to be relatively insensitive to what evidence has already been seen, while still providing local information. If people disproportionately focus on identifying effects, we expect to see relatively unconstrained interventions with one variable fixed “on” at a time. If people focus on individual edges we expect more constraining interventions with more variables fixed “off”. If confirmatory tests are employed, we expect to see more interventions on putative parents than on child nodes.

We therefore designed two studies based on the paradigm used in Bramley, Lagnado and Speekenbrink (2015). 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 parameters $w$ of the true underlying causal model, jointly determined the data they saw. We systematically varied the number of connections between components in the problem set, along with $w$ (and their knowledge thereof).

Experiment 1: Causal learning with known strengths

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 $w = \{w_S, w_B\}$.

Methods

Participants

We recruited 150 participants (85 male, mean±SD age 35 ± 10) from Amazon Mechanical Turk\(^8\), split randomly between 9 conditions (group size 16.7 ± 3.4). They were paid $1.50 and

---

\(^8\)Mechanical Turk (http://www.mturk.com/) is a web based platform for crowd-sourcing short tasks widely used in psychology research. It offers an well validated (Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013; Gosling, Vazire, Srivastava, & John, 2004; Hauser & Schwarz, 2015; Mason & Suri, 2012) subject pool diverse in age and background, suitable for high-level cognition tasks where tight control of experimental conditions are not critical.
received a bonus of 10c per correctly identified connection on a randomly chosen test for each problem (max = $6.00, mean±SD $3.68 ± 0.75). The task took an average of 41 ± 20 minutes.

Design

We included five 3-variable and five 4-variable problems (see Figure 4a). 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 $w_S \in [1, 0.85, 0.6]$ and three different levels of background noise $w_B \in [0, 0.15, 0.4]$ making $3 \times 3 = 9$ between-subjects conditions. For instance, in condition 1 ($w_S = 1; w_B = 0$) 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, ($w_S = 0.6; w_B = 0.4$) the outcomes were very noisy, with probability $0.4$ that a variable with no active parent would activate, compared to a probability $1 - (1 - 0.6)(1 - 0.4) = 0.76$ for a variable with one active parent.

Procedure

The causal systems were represented as gray 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 (see Figure 5). 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 forward or backward connection (represented as black arrows). Participants were incentivised to report their best guess about the structure, through receipt of a 10c 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 value of w in their condition. To train participants on w, they were first shown 10 unconnected components and forced to test them 5 times. The frequency with which the components activated reflected the true background noise level. They were then shown a set of two-component causal systems in which component “x” was a cause of “y”, and were forced to test these systems 5 times with component x fixed on. This indicated that the frequency with which y activated reflected the level of wS 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 oriented 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. The task can be tried out at http://www.ucl.ac.uk/lagnado-lab/el/ns15a.
Figure 4: The true models and aggregated results. a) The problems faced by participants. b) Weighted average final judgments by participants. Darker arrows indicate that a larger proportion of participants marked this edge in their final model. c) Bayes-optimal final marginal probability of each edge in $P(M|D^T, E^T w)$, averaged over participants’ data.
Results

We analysed participants’ sequences of judgments and then their intervention choices, dividing both sections in subsections focusing on differences by condition, device and trial. To assess the differences we made comparisons between subgroups and relative to efficient learner simulations that were both qualitative, in terms of the patterns of judgments and interventions, and quantitative, covering judgment accuracy and intervention quality.

Judgments

By condition  We expected the quality of participants’ judgments to be bracketed by those of a random responder (\( \frac{1}{3} \) per edge, 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 \( w_S \) and background noise \( w_B \), assuming a uniform prior over models at the start of each problem. By reporting the most probable structure in the posterior (guessing in the event of ties) participants could have achieved accuracies ranging between \( 0.84 \pm 0.14 \) in condition 2 and \( 0.55 \pm 0.09 \) in the noisiest condition, 9 (see Figure 6a, 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...
Running head: FORMALIZING NEURATH’S SHIP

to highly unreliable causal connections.

Participants significantly outperformed chance in all nine conditions (all \( p \) values < 0.05 for t-tests comparing to \( \frac{1}{3} \)). However they underperformed the Bayes-optimal observer (t-test \( p \) values < 0.05) in all conditions bar condition 2, where there was a still a trend \( w_S = 0.85, w_B = 0, (p = 0.07) \). Like the optimal observer, participants became less accurate as noise increased, with a main effect of background noise \( w_B \): \( F(2, 147) = 6.34, \eta^2 = 0.07, p = 0.02 \) with lower performances for \( w_B = 0.1 \), \( t(147) = -2.23, p = 0.03 \) and \( w_B = 0.4 \), \( t(147) = -3.5, p < 0.001 \) compared to \( w_B = 0 \), but were not affected significantly by strength \( w_S \): \( F(2, 147) = 1.2, p = 0.3 \).

Participants marked more causal connections per problem than the optimal learner, mean±SD estimates 2.93 ± 1.4 compared to 2.75 ± 1.4, \( t(2998) = 3.5, p < 0.001 \). The true proportion was 2.6. The number of connections participants marked on average was affected by both \( w_S \) and \( w_B \), going from 2.77 ± 1.4 for \( w_S = 0.6 \) to 3.01 ± 1.4 for \( w_S = 1 \), and from 2.78 ± 1.5 for \( w_B = 0 \) to 3.14 ± 1.4 for \( w_B = 0.4 \).

Under the assumption that participants soft-maximized over the posterior we inferred best fitting subjective \( w_S^* \) and \( w_B^* \) parameters by maximum likelihood estimation (see Appendix B), finding average estimates that were considerably less diverse across conditions and closer to \( w_S = 0 \) and \( w_B = 1 \) than the true prevailing values (Figure 6c).

By device  
Average accuracy on three variable problems was fractionally higher than on four variable problems 0.55 ± 0.34 compared to 0.52 ± 0.29, \( t(1463) = 2.0, p = 0.04 \), and the smaller problems were completed marginally more quickly with medians 12.3s and 14.6s respectively. 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 \( p < 0.0001 \). There was no relationship between the number of connections in a device and judgment accuracy \( F(1, 1498) = 2.1, \eta^2 = 0.001, p = 0.14 \). However there was a significant main effect of device type \( F(5, 1444) = 2.91, \eta^2 = 0.007, p = 0.02 \) (see Figure 4). Accuracy was lowest for chains (devices 3; 8) 0.49 ± 0.28, and highest for colliders 0.57 ± 0.30 (4; 9). The only significant pairwise difference was a higher level of accuracy identifying colliders (4; 9) \( t(1497) = 3.2; p = 0.001 \) than chains. The most common judgment errors were to mistake the chain for a fully connected structure (30/150 participants on 3-variable problems) or the reverse (20/150 participants on 3-variable problems, Table 1). Judgments on the four-variable chain were particularly diverse with the most frequent final judgment only shared by 7 of the 150 participants.
Table 1: The true underlying structure for each problem and the three most frequent final judgments. In all cases but the 3-variable chain the most frequent final judgment was the correct model. ∅ indicates a final judgment that there were no connections.

<table>
<thead>
<tr>
<th>N variables</th>
<th>True model</th>
<th>Most frequent</th>
<th>N</th>
<th>2nd</th>
<th>N</th>
<th>3rd</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Single $x \rightarrow y$</td>
<td>$x \rightarrow y$</td>
<td>40</td>
<td>$x \rightarrow y, z \rightarrow y$</td>
<td>15</td>
<td>$z \rightarrow x, z \rightarrow y$</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>Common cause $x \rightarrow y, x \rightarrow z$</td>
<td>$x \rightarrow y, x \rightarrow z$</td>
<td>43</td>
<td>$x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>14</td>
<td>$x \rightarrow y, z \rightarrow y, x \rightarrow z$</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>Chain $x \rightarrow y, y \rightarrow z$</td>
<td>$x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>30</td>
<td>$x \rightarrow y, y \rightarrow z$</td>
<td>28</td>
<td>$x \rightarrow y, x \rightarrow z$</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Common effect $x \rightarrow z, y \rightarrow z$</td>
<td>$x \rightarrow z, y \rightarrow z$</td>
<td>41</td>
<td>$x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>15</td>
<td>$y \rightarrow x, z \rightarrow x, y \rightarrow z$</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>Fully connected $x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>$x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>39</td>
<td>$x \rightarrow y, x \rightarrow z$</td>
<td>20</td>
<td>$x \rightarrow z, y \rightarrow z$</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>Single $w \rightarrow x$</td>
<td>$w \rightarrow x$</td>
<td>26</td>
<td>$w \rightarrow y, y \rightarrow x, y \rightarrow z$</td>
<td>3</td>
<td>$w \rightarrow x, y \rightarrow x, z \rightarrow x, w \rightarrow y, z \rightarrow y, w \rightarrow z$</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Common cause $w \rightarrow x, w \rightarrow y, w \rightarrow z$</td>
<td>$w \rightarrow x, w \rightarrow y, w \rightarrow z$</td>
<td>26</td>
<td>$w \rightarrow x, y \rightarrow x, y \rightarrow z$</td>
<td>3</td>
<td>$w \rightarrow x, y \rightarrow x, z \rightarrow x, w \rightarrow y, z \rightarrow y, w \rightarrow z$</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Chain $w \rightarrow x, x \rightarrow y, y \rightarrow z$</td>
<td>$w \rightarrow x, x \rightarrow y, y \rightarrow z$</td>
<td>7</td>
<td>$w \rightarrow x, y \rightarrow x, y \rightarrow z$</td>
<td>6</td>
<td>$w \rightarrow x, w \rightarrow y, x \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Common effect $w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>$w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>25</td>
<td>$w \rightarrow z, w \rightarrow y, x \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>3</td>
<td>$y \rightarrow z$</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Fully connected $w \rightarrow x, w \rightarrow y, x \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>$w \rightarrow x, w \rightarrow y, x \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>18</td>
<td>$w \rightarrow x, w \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>7</td>
<td>$w \rightarrow x, w \rightarrow y, x \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>4</td>
</tr>
</tbody>
</table>

**By trial** Comparing participants’ sequences of structure judgments indicates that they make markedly fewer changes to their model between trials than the optimal observer, changing an average of $0.94 \pm 1.3$ connections after each test compared with $1.78 \pm 1.5$, $\chi^2(6) = 1920, p < 0.0001$ (see Figure 6b). Furthermore, participants’ final judgments displayed clear recency and hints of primacy. We assessed this by looking at the likelihood of $d^1 \ldots d^T$ under the final model judgment $b^T$ (see Figure 6c). There was a significant relationship between trial and data likelihood under $b^T F(5, 4494) = 3.3, \eta^2 = 0.004, p < 0.005$ for the three variable problems. Polynomial contrasts revealed a significant positive linear effect $\beta = 0.04, t = 3.2p = 0.001$ (recency) and a positive quadratic trend $\beta = 0.02, t = 1.8, p = 0.06$ (associated with the hint of primacy of the first trial Figure 6c). For four variable problems there was a marginal main effect of trial $F(7, 5992) = 1.864, \eta^2 = 0.002, p = 0.07$ again with a positive linear effect again revealed by polynomial coding $\beta = 0.04, t = 3.0, p = 0.003$. To ensure this was not an artifact of participants’ self-selection of interventions, we checked the likelihood of the data under the final maximum a posteriori model $\max P(M|D^T, w; C^T)$ finding no relationship with trial for either three $F(5, 4994) = 0.7, \eta^2 = 0.001, p = 0.6$ or four $F(7, 5992) = 1.1, \eta^2 = 0.001, p = 0.32$ variable problems.
Figure 6: Experiment 1 results. 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 with mean search length of 1. Red squares: random guessing. b) Bars show average number of edits (additions, subtractions or reversals of connections) between all \( t - 1 \) and \( t \) judgments, as compared to Bayesian, Neurath’s ship and random choice simulations. c) Probability of data \( d^t \) given final model judgement \( b^T \) on three and four variable problems, ± standard errors. Increase in likelihood for more recent data points (e.g. those closer to \( T \)), is indicative of recency. d) Boxplot of best fitting \( w_S \) and \( w_B \) parameters assuming learners soft-maximised over \( P(M|D^t, w^*;C^t) \).
Interventions

Participants’ intervention selections, from the 27 legal interventions in 3-variable problems and 81 legal interventions in 4-variable problems, were clearly non-random. The distribution of selections differed from a uniform distribution over the legal choices for both three-variable $\chi^2(26) = 9752, p < 0.0001$ and four-variable $\chi^2(80) = 25206, p < 0.0001$ problems (see Figure 7, and additional plots in the supplementary materials available at http://www.ucl.ac.uk/lagnado-lab/el/nsm).

Participants’ intervention choices over all problems and time points were also quite different from those of an efficient learner. In order to compare the two, we simulated the task 150 times with the same number of simulations per condition as we had participants (Figures 7), stochastically generating the outcomes of the simulations’ intervention choices according to the true model and true $w$, as had been done for participants. Simulated efficient active learners would perfectly track the posterior and always select the best intervention $c^t = \arg \max_{c \in C} \mathbb{E}_{d \in D} [\Delta H(M|d, D^{t-1}, w; C^{t-1}, c)].$ Participants’ overall distribution of interventions differed substantially from the efficient learner simulations for both 3-variable $\chi^2 = 1538, p < 0.0005$ and 4-variable problems $\chi^2 = 1538, p < 0.0005$.

Participants’ most frequent choices were “one-on” interventions (Do$x = 1$, Do$y = 1$ etc) accounting for 52% of three-variable tests and 42% of four-variable tests. This was followed by the uninformative intervention, where all variables were fixed “on”, with Do$x = 1, y = 1, z = 1$ selected 10.6% of the time for 3-variable problems and Do$[w = 1, x = 1, y = 1, z = 1]$ selected 8.9% of the time for four-variable interventions. Observations with no variable fixed accounted for 7.4% of all tests across three- and four-variable problems, followed by interventions fixing two variables “on” and those fixing one “on” and one “off”. Over all tests, variables were fixed “on” 39% of the time, fixed “off” 8% of the time and left free to vary 53% of the time. In comparison, the efficient active learner simulations predominantly selected “one-on” interventions 74% and 66% for 3- and 4-variable problems, “one-on, one-off” interventions and “one-on, two-off” interventions, fixing 29% of variables “on”, 9% “off” and leaving 62% free.

As with causal judgments, we expected the overall amount of information gathered through participants’ sequences to be bracketed by that which would be achieved by intervening at ran-

---

9In semi-deterministic conditions 1, 2, 3, 4, and 7, simulated learners occasionally identified the true model before the final trial. On subsequent trials they would select interventions uniformly from the set of possible choices. This is why the optimal learner appears sometimes to make an uninformative intervention, setting all the variables to a given state.

10Because the counts for some interventions were very small, these were calculated using the $\chi^2$ test with Monte-Carlo simulated p-values as implemented by R’s chisq.test function. Simulated p-values were used wherever the test is reported without a bracketed degrees of freedom (e.g. $\chi^2$ rather than $\chi(,)^2$).
dom and by intervening to maximize expected information. A simple way to assess the quality of a sequence of intervention choices is by calculating the maximum of the posterior given the interventions and resulting outcomes codifying the probability that a Bayesian learner would correctly guess the true structure given this evidence\textsuperscript{11}. Participants’ sequences of interventions were considerably more useful on average than the uniformly sampled interventions, but also significantly less useful than those of the information maximizing simulations (Figure 7d). For example, there is a large difference between the chance of a simulated ideal learner guessing the correct model after the final trial given participants’ interventions $0.29 \pm 0.30$ compared to uniformly selected $0.02 \pm 0.04$, $t(2998) = 35, p < 0.0001$ or expected information maximizing interventions $0.60 \pm 0.33$, $t(2998) = -26, p < 0.0001$.

**By condition** Participants’ interventions also differed significantly by condition (Figure 7a) with $\chi^2 = 668, p < 0.0005$ for both three-variable and $\chi^2 = 1570, p < 0.0005$ for four variable problems. “One-on” interventions became marginally more common when connections were more unreliable, going from 50\% when $S = 1$ to 54\% when $S = 0.6$ for 3-variable problems $\chi^2(2) = 5.6, p = 0.06$, and 39\% when $S = 1$ to 54\% when $S = 0.6$ for 4-variable problems $\chi^2(2) = 35, p < 0.0001$. This is broadly in line with the behavior of the efficient learner simulations, that almost always selected “one-on” interventions when background noise $w_B$ was high and edge strength $w_S$ was low. However, the condition effect is much weaker for participants than for the efficient learner simulations.

**By trial** Intervention choices also varied across trials (Figure 7b) 3-variable $\chi^2 = 233, p < 0.0005$ and 4-variable $\chi^2 = 670, p < 0.0005$. Specifically, the proportion of single-variable (e.g. $\text{Do}[x=1]$) interventions decreases over trials ($\chi^2(5) = 24.3, p = 0.0002$ for 3-variable problems and $\chi^2(7) = 34, p < 0.0001$ for 4-variable problems, while the proportion of a range of other more constrained interventions (two “on”, three “on”, one “on” one “off” and one “on” two “off”) increase. Again, while this shift toward more constrained interventions is in line with the behaviour of the efficient learner simulations, the shift is much less pronounced.

**By device** Intervention choices also differed significantly between the different test devices for both 3-variable $\chi^2 = 136, p = 0.014$ and 4-variable $\chi^2 = 382, p = 0.0015$ problems although no clear pattern emerged.

Intervention selection on the chain (3;8) and fully-connected (5;10) structures is of particular

\textsuperscript{11}One can also use posterior entropy or expected score yielding very similar results.
interest because these devices can only be reliably identified if a specific intervention \((Do|x = 1, y = 0)\) for 3-variable problems and \(Do[w = 1, x = 0, y = 0]\) for 4-variable problems. Participants were more likely to perform these interventions on these problems, doing so on 19% compared to 13% on other problems in the 3-variable case, and on 6.7% compared to only 3.6% on other problems in the 4-variable case. For comparison, the ideal active learner simulation selected \(Do[x = 1, y = 0]\) on 54% of 3-variable chains and fully-connected problems compared to on 12% of other 3-variable problems and 18% chains and fully-connected compared to 7.4% of other 4-variable problems. The probability of a random responder selecting these interventions by chance during a 3-variable problem was \(1 - (1 - 1/27)^6 = 0.20\) and during 4-variable problem was \(1 - (1 - 1/81)^8 = 0.09\) meaning that participants’ probability of selection of this intervention was still below chance overall.

**By current hypothesis** Finally, intervention choices also differed depending on participants’ most recently reported belief \(b^{t−1}\), \(\chi^2 = 2363, p < 0.0005\) for 3-variable problems and \(\chi^2 = 59098, p < 0.0005\) for 4-variable problems. Components with at least one child according to the participants’ latest hypothesis \(b^{t−1}\) were more likely to be fixed “on”, 46% of the time compared to 32% for components without children \(t(37498) = 28.1, p < 0.0001\).

**Comparing overall patterns to simulated local learners** We also compared participants’ judgments to several other simulated learners’, each restricted to one of the types of local focus introduced in Section 4. We simulated 150 learners that selected interventions by minimizing the “local” uncertainty, either about the true state of a randomly chosen edge (Equation 10); the effects of a randomly selected node (Equation 11); or confirming a single hypothesis against a null of no connectivity\(^{12}\) (Equation 12). When one of the simulated learners did not generate a unique best intervention, it would sample uniformly from the joint-best interventions according to that criterion. The results of the simulations for the three variable problems are visualized in Figure 7c and for the four variables problems in the supplementary materials. Participants’ intervention sequences were significantly more globally informative than those that were driven by reducing individual edge uncertainty \(t(2998) = 7.1, p < 0.0001\) or by confirmatory testing \(t(2998) = 3.5, p = 0.0003\); see figure 7c. However, they were less informative than ones driven by identifying effects \(t(2998) = −7.8, p < 0.0001\). The overall distribution of judgments was most similar to the effect focused simulated learner with the lowest \(\chi^2\) statistics.

\(^{12}\) We used the latest most probable judgment \(\max p(M|D^{t−1}, w)\) in place of a current hypothesis \(b^{t−1}\) for edge focused and confirmatory testing.
(1974 and 3312) for 3- and 4-variable problems.
Figure 7: Intervention choices on three variable problems in Experiment 1 compared with matched active learning simulations. a) Across the three strength ($w_S$) and three background noise ($w_B$) conditions, b) Problems by trial. c) Comparing the overall distributions and global information of participants selections to simulations using the three proposed local focuses.
Discussion

Participants were clearly able to generate plausible causal models but also did so suboptimally. Averaged across participants, final model judgments resembled the posterior over models (Figure 4c), but individuals’ trajectories exhibited strong sequential dependence. This is consistent with our hypothesis that individuals normally maintain a single hypothesis and update it locally. As found in previous research (Bramley, Lagnado, & Speekenbrink, 2015; Fernbach & Sloman, 2009; McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016), participants were worst at separating the direct and indirect causes in the chain (3; 8) and fully-connected (5; 10) models. A closer look at participants’ intervention choices revealed that this was due to a common failure to generate the constrained interventions necessary to disambiguate these options.

Participants’ overall distributions of intervention selections most closely resembled effect focused local testing. Concretely, this means that they favored the class of interventions that fixed a single variable “on” at a time, leaving the rest free to vary, but their distributions of choices were relatively invariant across conditions and trials while the efficient learners’ were much more dynamic. Comparison with the final global information gathered (7d) revealed that they did not select which variables to target particularly efficiently, leading to a considerable discrepancy between the total information gathered by participants compared to an ideal active learner – Figure 6a (white compared to blue dots) and Figure 7d (green inverted triangles compared to black lines). However, participants also displayed hints of adaptation of strategy over the trials: with a preference for confirmatory testing, being more likely to fix variables “on” when they had children according to their latest hypothesis $b^{-1}$, and displaying a modest shift toward more constrained interventions in later trials. Averaged across participants though, the differences between behavior on early and late trials and high and low noise levels were relatively small compared to ideal behavior.

Experiment 2: Causal learning with unknown strengths

In Experiment 1 we focused on the effects of expected uncertainty. However, in general the reliability of connections $w_S$ and background noise $w_B$ will be unknown, meaning that a learner should really take their uncertainty about these sources of noise into account during inference (see Appendix A for the computational level details of how to incorporate uncertainty over $w$ in model inference and intervention choice). Therefore, in Experiment 2, we focused on cases where participants were not pretrained on $w$, but had to take this additional uncertainty
into account in their inferences.

Additionally, in Experiment 1 we could not rule out the possibility that participants’ judgments might appear consistent with local search because of lazy reporting (e.g. participants sometimes not bothering to update their model between trials). We addressed this in Experiment 2 by having two judgment elicitation conditions, differing according to whether the previous judgment remained as the default from one trial to the next, or if disappeared and had to be re-entered.

In addition, we proposed in Section 4 that learners would need to track their local confidences in order to choose where to focus subsequent tests. Thus, in Experiment 2, we elicit confidences about the edges in each judgment. If participants track local uncertainties we should expect these to correlate with the true uncertainties. Specifically, given the Neurath’s ship representation, we might also expect them to better reflect conditional uncertainty in the edge \( H(E_{ij}|E_{\setminus ij}, D^t; C^t) \) than the marginal \( H(E_{ij}|D^t; C^t) \). We also elicited predictions about the outcome of each chosen test before the outcome was revealed. If participants maintain only a single hypothesis, we expect this to be reflected in ability to predict these outcomes. Specifically, we would expect a Neurath’s ship learner’s outcome predictions to predominantly reflect the predictive distribution under their current hypothesis rather than averaged across all the model possibilities, and intervention usefulness judgments to bear little relation to the global information gained.

Finally, in Experiment 1, participants’ intervention selections showed hints of being motivated by a mixture of local aspects of the overall uncertainty, with some participants appearing most consistent with a single type of local uncertainty (i.e. always effect finding, or always focusing on individual edges), but the majority assuming to focus on a mixture of different local aspects of uncertainty. To test this idea more thoroughly, in Experiment 2 we explicitly probed participants’ beliefs about their intervention choices and elicited judgments of the “usefulness” of the outcomes they observed.

**Methods**

**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 incentivised to be accurate based on randomly selected trials as before, but this time with the opportunity to win Amazon vouchers rather than money directly. Participants were split randomly into 8 groups of mean size 13.8 ± 3.4, each of which
was presented with a different condition in terms of the value of $w$ and the way that they had to register their responses.

**Design and procedure**

![Experiment 2 additional measures](image)

Figure 8: Experiment 2 additional measures - a) Outcome expectation sliders b) Outcome informativeness elicitation c) Edge confidence sliders.
Figure 9: 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 \( \int_w P(M|D^T; C^T) p(w) \, dw \) averaged over participants’ data assuming a uniform independent prior over \( w \in [0, 1]\).

Experiment 2 used the same task interface as Experiment 1, but focused just on the three variable problems (devices 1-5) and an additional device (6’) in which none of the components was connected (Figure 9). There were two causal strength conditions \( w_S \in [0.9, 0.75] \) and two background noise conditions \( w_B \in [0.1, 0.25] \). 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 whether lazy reporting drove the results in Experiment 1, 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 judgment 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 make a choice for every connection after each test.

In addition to the structure judgments and interventions, we also elicited several additional measures from participants:

1. After selecting a test, but before seeing the outcome, participants were asked to predict what would happen to the variables they had left free. To do this they would set a slider for each variable they had left free to vary. The left pole of the slider was labeled “Sure off”, the right pole “Sure on” and the middle setting indicated maximal uncertainty (Figure 8a).

2. After observing the actual outcome of their chosen test, participants were asked how helpful they found that outcome for identifying the connections, setting a slider between “Much less certain” and “Much more certain”, with the center of the slider indicating that they were no more or less certain (Figure 8b).

3. After drawing their best guess about the causal model by setting each edge between the variables, participants were asked how sure they were about each edge. Again they would respond by setting a slider, this time between “Guess” on the left indicating maximal uncertainty over the edge’s three possible states \( \left[ \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right] \) such that the selected edge state was not thought to be more likely than either of the other possibilities, and “Sure” on the right indicating high confidence that that edge judgment was correct \( [1, 0, 0] \) (Figure 8c).

To minimise anchoring effects on these responses, the sliders’ selectors would not appear until participants clicked on the slider. They could not move on until they had set all the relevant sliders at each stage. Participants were trained and tested on interpretation of the slider extremes and midpoints in an additional interactive page during the instructions. All sliders were mapped to a 21 point scale. Participants were also incentivised to provide accurate slider judgments through chance to win an additional Amazon voucher. Concretely, participants were told: “There will be an additional £20 Amazon voucher for the participant whose slider judgments are the most accurate/self-consistent.” and participants self-consistency was measured by their Brier score (1950) measuring the average divergence between their slider outcome predictions and edge confidences and their actual probability of a correct prediction or judgment.

Participants faced the six devices in random order, with six tests per device followed by feedback as in Experiment 1. Then they faced one additional test problem. On this problem,
the true structure was always a chain (Figure 9, device 7'). On this final problem, participants did not have to set sliders. Instead, after they selected each test, but before seeing its outcome, they were asked why they had selected that intervention. Labels would appear on the nodes and participants were invited to “Explain why you chose this combination of fixed and unfixed components. Use labels ‘A’ ‘B’, ‘C’ to talk about particular components or connections” in a text box that would appear below the device. Responses were constrained to be at least 5 characters long. The chain (device 3) was chosen for this problem because in Experiment 1 and in Bramley et al (2015) participants often did not select the crucial $\text{Do}[x = 1, y = 0]$ intervention that would allow them to distinguish a chain from a fully connected model (device 5) making this an interesting case for exploring divergence between participants’ behavior and ideal active learning.

Finally, at the end of the experiment participants were asked to estimate the reliability $w_S$ of the true connections: “In your opinion, how reliable were the devices? i.e. How frequently would fixing a cause component ON make the effect component turn ON too?” and the level of background noise $w_B$: “In your opinion, how frequently did components activate by themselves (when they were not fixed by you, or caused by any of the device’s other components)?” by setting sliders between “0% (never)” and “100% (always)”.

A demo of Experiment 2 can be viewed at http://www.ucl.ac.uk/lagnado-lab/el/ns15b.
Results

Figure 10: Experiment 2 results. a) Performance by condition in Experiment 2. White circles, crosses, plusses: benchmark (greedy expected information gain maximizing) Bayesian learner assuming a uniform and uniform (UU), strong and uniform (SU) and sparse and strong (SS) prior. Blue diamonds, crosses, plusses: Bayesian learner that maximizes over posterior after seeing participants’ interventions, assuming UU, SU and SS priors. Crosses: assuming a “strong” prior on \( w_S \) but a uniform prior on \( w_B \). Green triangles Neurath’s ship simulation with mean search length \( \lambda \) of 2 assuming UU prior (see Section 3). Red squares: random guessing. b) Participants final judgments of amount of background noise (\( w_B \)) and strength (\( w_S \)), rescaled from 100 point scale to 0-1, and best-fitting \( w_S \) and \( w_B \) estimates assuming ideal Bayesian updating, error bars \( \pm \) SD. c) Visualizations of Uniform-Uniform, Strong-Uniform and Strong-Sparse priors on \( w_S \) and \( w_B \). d) Plot of the likelihood of \( d^1 \ldots d^T \) under \( b^T \) \pm standard errors.


c) Priors

d) Primacy/Recency

Judgments

As in Experiment 1, participants significantly outperformed chance in all conditions (all 8 t.statistics > 6.1 all p values < 0.0001) and underperformed a Bayes optimal observer observing the same data as them. Because the noise was unspecified, we explored several reasonable
priors on w (always assuming that $w_S$ and $w_B$ were independent) when computing posteriors. Firstly we considered a uniform-uniform prior that made no assumptions about either $w_S$ or $w_B$ (UU) where $w \sim \text{Uniform}(0, 1)^2$. We also considered a strong-uniform (SU) variant, following (Yeung & Griffiths, 2011), expecting causes to be reliable – $w_S \sim \text{Beta}(2, 10)$, but making no assumptions about background noise – $w_B \sim \text{Uniform}(0, 1)$. Additionally, we considered a sparse-strong (SS) variant following Lu et al (2008), encoding an expectation of high edge reliability – $w_S \sim \text{Beta}(2, 10)$, and relatively little background noise – $w_B \sim \text{Beta}(10, 2)$. Choice of parameter prior made little difference to the Bayes optimal observer’s judgment accuracy (Figure 10). Thus participants significantly underperformed the Bayes optimal observer in all conditions regardless of the assumed prior, except for condition 2 ($w_S = 0.75; w_B = 0.1, \text{remain}$) under the SU prior, and 4 ($w_S = 0.75; w_B = 0.25, \text{remain}$) under all three considered priors.

Comparison with Experiment 1

Performance in Experiment 2 was comparable to the 3-variable problems in Experiment 1. For example, mean ± SD accuracy in Experiment 2, $[w_S = 0.75, w_B = 0.1]$ was 0.63 ± 0.27 and $[w_S = 0.75, w_B = 0.25]$ was 0.58 ± 0.31 while Experiment 1 condition 5 [$w_S = 0.85, w_B = 0.15$] was 0.60 ± 0.33. Participants were actually closer in accuracy to an ideal learner in Experiment 2 than in Experiment 1 with an average accuracy drop of only .12 compared to .15 for the 3-variable problems in Experiment 1. This suggests that participants were able make reasonable structure judgments without knowledge of the exact parameters. Supporting these conclusions, we found that participants’ final judgments of $w_S$ and $w_B$ and best fitting estimates assuming rational updating $w_S^*$ and $w_B^*$ suffered bias and variance: for $w_S = \{0.9, 0.75\}$, mean±SD judgments were \{0.74±0.21; 0.62±0.23\} and estimates were \{0.67±0.35; 0.52±0.38\}; for $w_B = \{0.1; 0.25\}$ the mean ± SD judgments were \{0.37±0.25; 0.48±.20\} and estimates were \{0.29 ± 0.39; 0.28 ± .033\} (Figure 10 b)\(^{14}\).

As with Experiment 1, participants were not affected by the reliability of the connections themselves $w_S \ t(106) = 0.88, p = 0.37$ but were affected by higher levels of background noise $w_B \ t(108) = 2.7, p = 0.008$. There was no difference in performance between the two judgment elicitation conditions $t(108) = 0.67, p = 0.50$. However, analysis of variance revealed an effect of condition on final judgment accuracies of $F(7, 103) = 2.87, \eta^2 = 0.16, p = 0.008$ with a significant interaction between $S$ and judgment type, with a 0.21 additional drop in accuracy

\(^{13}\)In all cases we numerically integrated over priors on w by drawing 1000 samples from their distributions and averaging across the resulting posteriors.

\(^{14}\)Fifty-eight participants’ final $w_B$ judgments were incorrectly stored, so the N for $w_B$ judgments was 53 rather than 111.
going from $S = 0.9$ to $S = 0.75$ in the disappear condition compared to the remain condition.

Accuracy was not significantly related to device in Experiment 2, $F(5, 771) = 0.3, \eta^2 = 0.002, p = 0.9$, nor did participants become more accurate on the final problem when identifying a chain structure for the second time (device 7') with accuracy of $0.62 \pm 0.30$ and $0.62 \pm 0.33$ respectively $t(220) = -0.07, p = 0.9$. The most frequent error once again was mistaking the chain structure for the fully connected structure, made by 17/111 participants, although this was reduced to 11/111 when facing the chain structure again on device 7', with only a single participant making the same error twice.

Average edit distance between sequential judgments about the same device was significantly increased by removing the record of previous judgments between trials, going from .85 in the remain condition to 1.1 in the disappear condition. Edit distances were still significantly lower than those predicted by random responding $\chi^2(3) = 2180, p < 0.0001$ and optimal $\chi^2(3) = 78, p < 0.0001$ judgments in the disappear condition but was only lower than random $\chi^2(3) = 1546, p < 0.0001$ in the remain condition but not significantly different from the ideal observer $\chi^2(3) = 4.8, p = 0.14$. Participants still appeared to display recency, with a relationship between likelihoods of $d^1 \ldots d^T$ under the final model choice $F(5, 4656) = 6.264, \eta^2 = 0.005, p < 0.0001$ driven by a significant positive linear contrast $\beta = 0.01, t = 4.9, p < 0.0001$.

While this time likelihoods under the model that is most likely according to a normative Bayesian estimate at $d^T$ did differ by trial $F(5, 4656) = 4.5, \eta^2 = 0.005, p = 0.0004$, the positive relationship between trial and data likelihood under participants’ final judgment persisted when controlling for this $\beta = 0.9, t = 3.2, p = 0.001$.

**Additional measures**

Participants’ edge confidence judgments increased significantly $F(1, 13984) = 1676, \eta^2 = 0.11, p < 0.0001$ over trials, going from $0.57 \pm 0.20$ on the first trial to $0.78 \pm 0.19$ by the final trial (Figure 10 d). The probability of changing an edge at the next time point was weakly inversely related to the learners’ reported confidence in it $t(11653) = -7.1, p < 0.0001$ with changed edges given confidence $0.67 \pm 0.21$ compared to unchanged edges $0.70 \pm 0.21$. Reported edge confidences were more consistent on average with the conditional probability of the edge states given the rest of the current model than the marginal probability of the edge-state in the full posterior under the $UU$ prior $r^{\text{cond}} = 0.20$ vs $r^{\text{mar}} = 0.17, p = 0.04$, although the difference was not significant for the $SU$ or $SS$ priors.

As predicted, reported outcome expectation probabilities were also more closely related to the predictive distribution under the participants’ latest structure judgment $b^{t-1}$. $r = 0.28, F(1, 14012) =$
1248, $\eta^2 = 0.08, p < 0.0001$ than marginalized over the full posterior $r = 0.22, F(1, 14012) = 743, \eta^2 = 0.05, p < 0.0001$. Participants judgments of the “helpfulness” of each outcome they observed correlated with the global information provided by that outcome $r = 0.16, F(1, 4660) = 134, \eta^2 = 0.04, p < 0.0001$.

**Interventions**

The overall distribution of intervention choices was broadly similar to Experiment 1, in that “one on” interventions were the most frequently chosen, making up 39% of selections, and the overall distribution differed substantially from uniform $\chi^2(26) = 5495, p < 0.0001$ and efficient information maximizing, assuming UU $\chi^2(26) = 2318, p < 0.0001$, SU $\chi^2(26) = 2269, p < 0.0001$ or SS $\chi^2(26) = 2414, p < 0.0001$ priors on $w_S$ and $w_B$. However, unlike Experiment 1, constrained “one-on one-off” interventions were almost as common as single “one-on” interventions, making up 38% of tests compared to 12% across 3-variable problems in Experiment 1. We discuss this difference in the Discussion. The intervention selections and informativeness of intervention sequences were not closely consistent with global expected information, nor any single type of local focus (See figures in Supplementary Materials), but could again be consistent with a mixture of local effect focused, edge focused and confirmation focused queries.

**Free explanations**

For device 7', participants gave free explanations for their intervention choices on each of their six tests. The overall distribution of intervention choices did not differ significantly from the original presentation of the chain (device 3) $\chi^2 = 31, p = 0.21$ suggesting that the different response format did not affect the intervention choices that participants made. In order to assess what the explanations tell us about participants’ intervention choices, we asked two independent coders to categorize the free responses into 8 categories. The categories were chosen in a partly data-driven, partly hypothesis-driven way: 1. An initial set of categories were selected, with the goal of distinguishing the approximations introduced in Section 4 from global strategies like uncertainty sampling or expected information maximization. 2. A subset of the data was then checked and the categories were refined to better delineate them responses with minimal membership ambiguity.

The eight resulting categories were:

1. The participant just wanted to learn about one specific connection. [Corresponding to *edge focused* testing]
2. The participant wanted to learn about two specific connections.
3. The participant wanted to learn about all three connections. [Corresponding to globally focused testing]
4. The participant wanted to learn what a particular component can affect but did not mention a specific pattern of connections. [Corresponding to effect focused testing]
5. The participant wanted to test / check / confirm their current hypothesis. [Corresponding to confirmatory testing]
6. The participant wanted to learn about the randomness in the system (as opposed to the location of the connections). [Corresponding to a focus on learning about noise rather than structure]
7. The participant chose randomly / by mistake / to use up unwanted tests / they say they did not understand what they are doing / it is clear they were not engaging with the task.
8. The participant’s explanation was complex / underspecified / did not seem to fall in any of the above categories.

A supplementary file (available at http://www.ucl.ac.uk/lagnado-lab/el/nsm) contains all the materials given to coders and the full set of participant responses. Coders were permitted to assign more than one category per response, but had to select a primary category. When the category referred to particular component label(s), the rater would record these, and when it referred to a specific connection they would record which direction (if specified) and the components involved. These details will be used to facilitate a quantitative comparison between participants’ explanations and our model fits in the next Section. Raters normally just selected one category per response, respectively selecting additional categories on only 8% of trials. Inter-rater agreement on the primary category was 0.73, and Cohen’s $\kappa = 0.64 \pm 0.04$, both higher than their respective heuristic criteria for adequacy of 0.7 and 0.6 (Krippendorff, 2012; Landis & Koch, 1977).

Figure 11 shows the proportion of responses in the different categories across the six trials. On the first trial participants were most likely to be categorized as 4. – focused on identifying what a particular variable could effect. On subsequent trials they most frequently categorized as 1. – focusing on learning about a specific connection. Toward the end, explanations became more diverse and were increasingly categorized as 5. – confirmatory testing or 6. learning about the noise in the system. Individuals almost always gave a range of different explanations across their six tests, falling under $3.0 \pm 0.99$ different categories on average, with only 5/111 participants providing explanations from the same category all six times (3 all-fours, 1 all-
Explanation type was predictive of performance $F(8, 657) = 13.75, \eta^2 = 0.14, p < 0.0001$. Taking category 7 – unprincipled or random intervening – as the reference category with low average performance of 10.2 points out of a possible 21, categories 1, 2, 4, 5, and 6 were all associated significantly higher final scores $[14.5, 12.9, 13.9, 13.9, 13.9]$ points, all $p's < 0.001$.

Figure 11: Free explanations for interventions agreed codes over the six tests in Experiment 2, problem 7'.

Discussion

In Experiment 2 we saw that people were able to identify causal structure effectively without specific parameter knowledge. Comparing a range of plausible prior assumptions about edge reliability $w_S$ and the level of background noise $w_B$ – strong and sparse, strong and uniform and uniform and uniform – yielded little difference in structure judgment or intervention choice predictions.
Participants’ overall judgment accuracy was not affected by the remain/disappear reporting condition, but this did affect sequential dependence and interact with the different noise conditions. Judgments exhibited less sequential dependence when the record of the previous judgment did not carry over into the next trial (disappear condition), although we note that by focusing on lower noise levels and only three variable problems, the expectation that a Neurath’s ship learner should exhibit dramatic sequential dependence is also reduced. The interaction between connection reliability $w_S$ and reporting condition can be understood as being due to better performance in the disappear than the remain condition, when noise was low but worse performance when noise was high. This is consistent with the idea that participants were more reliant on a record of their previous hypothesis when noise was substantial.

The Neurath’s ship idea was also supported by the additional measures elicited from participants during the task. Under Neurath’s ship, edge confidence judgments could only be made relative to the existing model (e.g. conditioning on, or “leaning on” the rest of the ship for support in the metaphor). Consistent with this, participants’ edge-wise confidence judgments were more similar to the conditional probabilities of those connections given their current model, than their marginal probabilities averaged over the full posterior. Likewise, with a single hypothesis rather than distributional beliefs, intervention outcome predictions could only be generated by the current hypothesis rather than averaged and weighted over all possible models. Again, we found that participants’ expectation judgments more closely in line with their current hypothesis than the marginal likelihoods.

Intervention patterns again were consistent with a mixture of locally motivated tests, but also globally much more informative compared to a random intervener’s. Once again, there was evidence of a move from an effect focus to a narrower edge or confirmatory focus on later trials. The largest difference from Experiment 1 was that constrained interventions (e.g. $\text{Do}[x = 1, y = 0]$) were chosen much more frequently. One explanation for this is that participants might have been forced to focus their attention more narrowly in Experiment 2, to compensate for their additional uncertainty about the noise by using more focused testing. Another possibility we cannot discount is that the different subject pools drove this difference. It is possible that mTurk’s markedly older and educationally diverse participants (Experiment 1) gathered evidence differently from the younger and scientifically trained UCL undergraduates (Experiment 2). This might have driven the tendency toward more tightly constrained tests in Experiment 2.

The idea that people relied on asking a mixture of different types of question (Section 4),
was borne out by our analysis of the coding of participants’ free explanations. Explanations almost always focused on one specific aspect of the problem, most frequently on a particular causal connection, or what a particular component can affect, but also sometimes on parameter uncertainty or confirming the current hypothesis. Furthermore, participants almost always referred to a mix of different local query types over the course of their six tests.

Subjective explanations are notoriously problematic (Ericsson & Simon, 1980, 1993; Russo, Johnson, & Stephens, 1989). Therefore, we must be careful in interpreting these results. One common issue is that eliciting responses concurrently with performing a task can change behavior, invalidating conclusions about the original behavior. We minimized this issue by eliciting explanations just after each intervention was chosen, before its outcome was revealed. Additionally, we did not find any difference in the distribution of interventions on the free response trials and those chosen the first time participants identified the chains structure.

A second issue is that there are limits on the kinds of processes people can describe effectively in natural language, with rule based explanations being typically easier to express than those involving more complex statistical weighting and averaging. That is, even if someone weighed several factors in coming to a decision, they might explain this by mentioning only the most significant, or recently considered of these factors, falsely appearing to have relied on a one-reason decision strategy. There is an active debate about this, including suggestions that people’s explanations for their choices are, in general, post-hoc rationalizations rather than genuine descriptions of process (Dennett, 1991; Johansson, Hall, Sikström, & Olsson, 2005), but also refutations of this interpretation (Newell & Shanks, 2014).

In sum, taken with appropriate caution, we suggest that this analysis does provides a valuable window on participants’ subjective sense of their intervention selection, with their relatively specific focus on one aspect of the uncertainty at a time consistent with the idea that they rely on a mixture of heuristic questions as introduced in Section 4In the next Section we compare participants’ explanations to our fitted models to test whether their explanations are borne out quantitatively in their behavior.

**Modeling**

In analyzing the two experiments, we established a strong qualitative correspondence, both between our Neurath’s ship framework and participants’ judgments, and between the two stage local intervention schema and participants interventions. In order to validate quantitatively that our framework offers a good description of participants’ behavior, we now fit the models
to the data and assess whether they go beyond the computational level framework in describing people’s behavior. By fitting the models separately to individual participants we can also assess individual differences in learning behavior and get a more fine grained picture of the computational processes.

**Judgments**

Over two experiments, we found evidence consistent with the thesis that people update their causal beliefs without enumerating all hypotheses, maintaining a distribution, or remembering all the evidence. Instead, we proposed a *Neurath’s ship* (NS) model (Section 3) whereby a learner maintains a single causal structural hypothesis during learning, and updates it bit by bit, by resampling edges, conditional on the rest of the model and the latest evidence.

**Models**

The fitted version of the NS model had 3 parameters:

1. An average search length parameter $\lambda$ controlling the probability of searching for different lengths $k$ on each belief update.

2. A search behaviour parameter $\omega$ controlling how strongly the learner would move toward the more likely state for an edge when updating it (recalling that $\omega = 1$ leads to probability matching, while $\omega = \infty$ leads to deterministic hill climbing and $\omega = 0$ to making random local edits).

3. A lapse parameter $\epsilon$ controlling the probability of making a mistake\(^{15}\).

Recalling how we constructed the *Neurath’s ship* predictions in Section 3 via a Markov matrix $R$ (Equation 5), this resulted in the following equation

$$P(b^t = m|d^t, e^t, w, b^{t-1}, \omega, \lambda) = (1 - \epsilon) \sum_0^\infty \frac{\lambda^ke^{-\lambda}}{k!} [(R_w^\omega)^k]_{b^{t-1}m} + \epsilon \text{Unif}(M)$$ (13)

We also considered three comparison models. *Baseline* is a parameter-free baseline that assumes each judgment is a random draw from all possible causal models

$$p(b^t = m) = \text{Unif}(M)$$ (14)

(leading to a probability of approximately $\frac{1}{3}$ for each edge). We also needed a stronger baseline model because even if judgments were not related to data $d^t$, they might still exhibit sequential dependence due to lazy responding, or disinclination to change judgments from trial to trial.

\(^{15}\)See Appendix B
Therefore we also considered a baseline variant of the NS model in which the search behavior parameter $\omega$ was fixed to 0 ($\text{NS}_{\omega=0}$), resulting in a model that moved randomly around the hypothesis space for $k$ steps on each update. For this model, small $k$ simply denotes more inertia. The third model, rational, was a variant of the Bayes-optimal observer (Section 2) that attempted to select the maximum a posteriori causal structure $\max P(M|D^t, w; C^t)$ with each judgment, with a soft maximization (Luce, 1959) governed by inverse temperature parameter $\theta$ and a lapse parameter $\epsilon$. For this, we considered

$$P(b^t = m|D^t, w) = (1 - \epsilon)\frac{\exp(P(M|D^t, w), \theta)}{\sum_{m' \in M} \exp(P(m'|D^t, w), \theta)} + \epsilon \text{Unif}(M)$$ (15)

Each of these belief models output a likelihood based on the probability that the model assigns to a belief of $b^t$, given the latest observed outcome $d^t$, or $D^t$ and the most recent judgment $b^{t-1}$. Because the choice of prior for Experiment 2 made negligible difference to our results, we only report models assuming uniform ($UU$) priors on $w$. For Experiment 2 we marginalized over the unknown values of $w$ rather than conditioning as in Experiment 1 as detailed in Appendix B.

**Evaluation**

To compare these models quantitatively, we used maximum likelihood optimization as implemented by R’s `optim` function\(^{16}\) to fit the model separately to each of the 150 participants in Experiment 1 and the 111 participants in Experiment 2. We used Bayesian Information Criterion (BIC, Schwarz, 1978) to compare the models while accommodating their differing numbers of parameters. Baseline acts as the null model for computing BICs and pseudo-$R^2$s (Dobson, 2010) for the other models. Recalculating the transition probabilities on the fly in the optimization of $\omega$ was infeasibly computationally intensive for the four-variable problems. So for Experiment 1 we first fit all three parameters to the three-variable problems only, then used the best fitting $\omega$ parameters from this fit when fitting the $\lambda$ and $\epsilon$ on the full data. In Experiment 2 we were able to fit all three parameters.

**Results and discussion**

As Tables 2 and 3, and Figure 12a show, the majority (164/261) of participants’ sequences of causal judgments were best fit by the NS model across both experiments. Only two participants were better described by the Baseline model in either experiment, while 55/261 across both

---

\(^{16}\)In Appendix B we give a detailed description of how the models were fit. We also report fits for several variants of the NS model relaxing the assumptions that learners’ starting belief is an unconnected graph
Table 2: Experiment 1 belief models. Columns: med = median estimated parameter (across participants), SD = standard deviation of parameter estimate across participants, N fit = number of participants best fit by each model (/150), M score = average score of those participants (/45), LogL = total log likelihood of model over all data, $R^2$ = McFadden’s pseudo-$R^2$, BIC = Bayesian information criterion. Best fitting model denoted with boldface.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda$ med</th>
<th>$\lambda$ SD</th>
<th>$\omega$ med</th>
<th>$\omega$ SD</th>
<th>$\theta$ med</th>
<th>$\theta$ SD</th>
<th>$\epsilon$ med</th>
<th>$\epsilon$ SD</th>
<th>N fit</th>
<th>M score</th>
<th>logL</th>
<th>$R^2$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0</td>
<td>52268</td>
<td>0</td>
<td>104535</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS $\omega=0$</td>
<td>0.7</td>
<td>49</td>
<td>0.15</td>
<td>0.18</td>
<td>41</td>
<td>15.0</td>
<td>-26674</td>
<td>0.49</td>
<td>54622</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS $\omega=1.0$</td>
<td>1.0</td>
<td>61</td>
<td>136</td>
<td>299</td>
<td>0.13</td>
<td>0.18</td>
<td>107</td>
<td>27.0</td>
<td>50278</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rational</td>
<td>17</td>
<td>728</td>
<td>0.16</td>
<td>0.38</td>
<td>2</td>
<td>42.0</td>
<td>-43702</td>
<td>0.16</td>
<td>88678</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Experiment 2 belief models. As in Table 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda$ med</th>
<th>$\lambda$ SD</th>
<th>$\omega$ med</th>
<th>$\omega$ SD</th>
<th>$\theta$ med</th>
<th>$\theta$ SD</th>
<th>$\epsilon$ med</th>
<th>$\epsilon$ SD</th>
<th>N rem</th>
<th>N diss</th>
<th>logL</th>
<th>$R^2$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>2</td>
<td>5.0</td>
<td>-15006</td>
<td>30013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS $\omega=0$</td>
<td>0.9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.10</td>
<td>0.29</td>
<td>9</td>
<td>5</td>
<td>10877</td>
<td>0.28</td>
<td>22585</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS $\omega=1.6$</td>
<td>1.6</td>
<td>82</td>
<td>20.2</td>
<td>277</td>
<td>0.04</td>
<td>0.20</td>
<td>28</td>
<td>29</td>
<td>7960</td>
<td>0.38</td>
<td>19963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rational</td>
<td>16.0</td>
<td>175</td>
<td>1.00</td>
<td>0.32</td>
<td>16</td>
<td>22</td>
<td>15.3</td>
<td>-10308</td>
<td>0.31</td>
<td>21447</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

experiments were best described by the stronger NS $\omega=0$ baseline model. These participants had average final scores consistent with chance performance (e.g. $15.0 = \frac{45}{3}$ in Experiment 1 and $8.4 \approx \frac{21}{3}$ in Experiment 2) suggesting they were confused or unmotivated by the task. Two participants in Experiment 1 and 38 in Experiment 2 were best fit by the Rational model. This is reflected in their performance, with these participants scoring near ceiling in Experiment 1 (42/45) and as high as the ideal learning simulations in Experiment 2 – 15.3/21 compared to an average of 15.5 for perfect Bayesian integration. This suggests that the design for Experiment 2 – being motivated primarily to look more closely at intervention choice – was not difficult enough either to force all the participants to update their beliefs locally or to distinguish local from global updating for all participants. Where performance is as good as the ideal learners’ it becomes impossible to distinguish our process account from the exact Bayesian account, and the Rational model with one less parameter will tend to be favoured by BIC.

The median fitted $\omega$ of 136 in Experiment 1 and 20.2 in Experiment 2 was suggestive of strong hill-climbing.

In line with predictions, participants’ average fitted search lengths ($\lambda$) were relatively small, suggesting that participants reconsidered an average of 1-2 edges per update. Because this parameter merely encodes a participant’s average search length this means that the same participant would sometimes not search at all, staying exactly where they are ($k = 0$), or might also sometimes search much longer (e.g. $k \gg \lambda$).

While the advantage of NS was stronger in the remain than the disappear condition in Exper-
iment 2, BIC still favoured the NS over Rational and Baseline models in the disappear condition, with a total BIC of 10782 compared to Baseline: 11498 NSω=0: and Rational: 12539.

Interventions

We also proposed that people focused on local aspects of the overall uncertainty when choosing interventions (Section 4), supporting their limited ability to update beliefs by gathering evidence well suited to updating beliefs locally. In this section we check this proposal quantitatively by fitting a class of locally focused intervention models, and making comparisons to a globally focused model and a baseline model.

Models

Each intervention model output a likelihood for an intervention choice of c′, depending on Dt−1, Ct−1 and b′t−1.

We compared the overall distribution of participants’ intervention selections and final performance with edge focused, effect focused and confirmation focused tests. We found that none of these models alone closely resembled participants patterns, but overall distributions were consistent with a mixture of different types of local tests. This was also supported by the the free-response coding in Experiment 2, showing that participants would typically report targeting a mixture of specific edges, effects of specific variables and confirming the current hypothesis. Therefore, we considered four locally driven intervention selection models, based on either an edge focus, an effect focus, a confirmation focus or a mixture of all three types of focus. For the edge model, the collection of possible focuses L contained the 3 (or 6) edges in the model. For the effect model, it contained the 3 components (or 4 in the 4-variable case). The confirmation model always had the same focus – comparing b′t to null b0 of no connectivity. The mixed model contained all 7 (or 11) focuses. As in Equations 8 and 9 in Section 4, each model would first compute a soft-max probability of choosing each possible focus l′ ∈ L. Within each chosen focus it would also calculate the soft-max probability of selecting each intervention, governed by another inverse temperature parameter η ∈ [0, ∞]. The total likelihood of the next intervention choice was thus a soft-maximization-weighted average of choice probabilities across possible focuses

\[ P(c′|η, \rho, D^{t-1}, b^{t-1}, w) = \sum_{l \in L} P(c|l, η, b^{t-1}, w) \frac{\exp(H(l|D^{t-1}, b^{t-1}, w; C^{t-1})ρ)}{\sum_{l' \in L} \exp(H(l'|D^{t-1}, b^{t-1}, w; C^{t-1})ρ)} \]  

(16)
where

\[
P(c|l, \eta, b^{t-1}, w) = \frac{\exp \left( \mathbb{E}_{d \in \mathcal{D}_c} [\Delta H(l|d, b^{t-1}, w; c)] \right)}{\sum_{c' \in C} \exp \left( \mathbb{E}_{d \in \mathcal{D}_c} [\Delta H(l|d, b^{t-1}, w; c')] \right)}
\]  

(17)

Positive values of \( \rho \) encode a preference for focusing on areas where the learner should be most uncertain, \( \rho = 0 \) encodes random selection of local focus, and negative \( \rho \) encodes a preference for focusing on areas where the learner should be most certain.

As with the belief modeling we also considered several alternatives. Baseline is a parameter-free model that assumes each intervention is a random draw from all possible interventions

\[
P(c^t) = \text{Unif}(C)
\]  

(18)

and, Global is a variant of the globally efficient intervention selection (Section 2) that attempts to select the globally most informative greedy test \( \arg \max_{c \in C} \mathbb{E}_{d \in \mathcal{D}_c} [\Delta H(M|d, D^{t-1}, w; C^{t-1}, c)] \). It has one inverse temperature parameter \( \theta \in [0, \infty) \) governing soft maximization (Luce, 1959) over the global expected information gains. For this, we considered

\[
P(c^t|D^{t-1}, w; C^{t-1}) = \frac{\exp(\mathbb{E}_{d \in \mathcal{D}_c} [\Delta H(M|d, D^{t-1}, w; c^t)])}{\sum_{c \in C} \exp(\mathbb{E}_{d \in \mathcal{D}_c} [\Delta H(M|d, D^{t-1}, w; c)])}
\]  

(19)

As with the belief modeling, for Experiment 2 we marginalized over the the unknown values of \( w \) rather than conditioning as in Experiment 1 as detailed in Appendix B.

Evaluation

All six models were fit to the data from both Experiment 1 and Experiment 2 in the same way as the belief models. The results are detailed in Tables 4 and 5.

Additionally, to compare model predictions of local focus choice \( l^t \) to participants’ self reports in problem 7' in Experiment 2, we computed the likelihood of each local focus prediction on each test. This was done by calculating \( P(c|l, \eta, b^{t-1}, w) \) for each of the local focuses we considered, using a fixed common \( \eta \) of 20. For each datapoint \( c^t \), we then calculated which \( l^t \) assigned the most probability to \( c^t \) the intervention actually chosen by the participant. Figure 13 plots the most likely focus of participants’ intervention choices in the final problem against the code assigned to their free responses.

Results and discussion

As Tables 4 and 5 and Figure 12b show, mixed was the best overall fitting model in both experiments, and the majority of participants 153/261 were fit by one of the local uncertainty driven models. Furthermore, Figure 13 shows that for effect and edge queries, there was a
Table 4: Experiment 1 intervention models. Columns: as in belief model Tables 2 and 3

<table>
<thead>
<tr>
<th>Model</th>
<th>η med</th>
<th>η sd</th>
<th>ρ med</th>
<th>ρ sd</th>
<th>θ med</th>
<th>θ sd</th>
<th>N fit</th>
<th>M score</th>
<th>logL</th>
<th>R²</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td>13.5</td>
<td>-41198</td>
<td></td>
<td>82396</td>
</tr>
<tr>
<td>Edges</td>
<td>6.7</td>
<td>13</td>
<td>11.4</td>
<td>680</td>
<td></td>
<td></td>
<td>5</td>
<td>23.8</td>
<td>-37292</td>
<td>0.09</td>
<td>75859</td>
</tr>
<tr>
<td>Effects</td>
<td>6.3</td>
<td>20</td>
<td>8.9</td>
<td>721</td>
<td></td>
<td></td>
<td>38</td>
<td>26.0</td>
<td>-33252</td>
<td>0.19</td>
<td>67779</td>
</tr>
<tr>
<td>Confirmatory</td>
<td>4.5</td>
<td>27</td>
<td></td>
<td></td>
<td>21</td>
<td>17.4</td>
<td>680</td>
<td>5</td>
<td>-38969</td>
<td>0.05</td>
<td>78575</td>
</tr>
<tr>
<td>Mixed</td>
<td>12.0</td>
<td>109</td>
<td>11.6</td>
<td>232</td>
<td>28</td>
<td>32.4</td>
<td>721</td>
<td>38</td>
<td>-33151</td>
<td>0.20</td>
<td>67577</td>
</tr>
<tr>
<td>Rational</td>
<td>3.2</td>
<td>6</td>
<td></td>
<td></td>
<td>33</td>
<td>26.2</td>
<td>114</td>
<td>109</td>
<td>-34128</td>
<td>0.17</td>
<td>68894</td>
</tr>
</tbody>
</table>

Table 5: Experiment 2 intervention models.

<table>
<thead>
<tr>
<th>Model</th>
<th>η med</th>
<th>η sd</th>
<th>ρ med</th>
<th>ρ sd</th>
<th>θ med</th>
<th>θ sd</th>
<th>N fit</th>
<th>M score</th>
<th>logL</th>
<th>R²</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td>7.5</td>
<td>-15365</td>
<td></td>
<td>30730</td>
</tr>
<tr>
<td>Edges</td>
<td>4.0</td>
<td>9</td>
<td>6.8</td>
<td>98</td>
<td></td>
<td></td>
<td>25</td>
<td>15.6</td>
<td>-12944</td>
<td>0.18</td>
<td>26718</td>
</tr>
<tr>
<td>Effects</td>
<td>3.2</td>
<td>9</td>
<td>15.4</td>
<td>410</td>
<td></td>
<td></td>
<td>8</td>
<td>9.2</td>
<td>-13000</td>
<td>0.13</td>
<td>26829</td>
</tr>
<tr>
<td>Confirmatory</td>
<td>2.5</td>
<td>14</td>
<td></td>
<td></td>
<td>10</td>
<td>9.7</td>
<td>680</td>
<td>5</td>
<td>-14180</td>
<td>0.04</td>
<td>28776</td>
</tr>
<tr>
<td>Mixed</td>
<td>3.8</td>
<td>9</td>
<td>6.4</td>
<td>705</td>
<td>18</td>
<td>14.7</td>
<td>721</td>
<td>38</td>
<td>-12418</td>
<td>0.19</td>
<td>25665</td>
</tr>
<tr>
<td>Global</td>
<td>3.0</td>
<td>8</td>
<td></td>
<td></td>
<td>36</td>
<td>14.4</td>
<td>114</td>
<td>109</td>
<td>-12850</td>
<td>0.14</td>
<td>26114</td>
</tr>
</tbody>
</table>

The intervention modeling suggests that some individuals relied on a single type of local focus, with 37 participants in Experiment 1, in particular described as relying on effect focused testing and 25 in Experiment 2 described as just using edge focused testing. As with the belief modeling, a moderate number of chance-level performing participants (37/261) were best described by the Baseline model. However, 69 participants across the two experiments were better described by the Globally efficient testing model than any local testing models. But, these were not the highest performing participants nor the most efficient gatherers of information, with lower average scores than those described by the mixed local focus model. This suggests that we do not yet have a good model of these participants’ choices. One possibility, consistent with the modest differences in overall choice distributions across trials (Figure 7, and Supplementary Materials), is that some of these participants might have established stereotyped patterns of intervention choices, e.g. reinforcing types of intervention they had discovered to be effective on past trials and problems. This would result in behavior that is not strongly dependent on the evidence seen or prior hypothesis on a particular trial, and does not pertain to a particular local focus, but which would tend to correlate with average global informativeness.
a) Model fits beliefs

i. Experiment 1

b) Model fits interventions

i. Experiment 1

ii. Experiment 2

Participant performance

Participant performance

iii. Experiment 2

iii. Experiment 2

Participant performance

Participant performance

c) Model parameters

i. Fitted $\lambda$

ii. Fitted $\rho$

Figure 12: Model fitting results. a) Experiment 1, i. Participants’ BIC values for the fitted belief change models (lower is better), best overall model indicated by rectangle, participant’s task performance coded by point colour (high = red, low = blue) b) Same plots for Experiment 2. c) Density estimates for mean chain lengths ($\lambda$) for belief update models and local focus rationality parameter ($\rho$) for intervention models.
Running head: FORMALIZING NEURATH’S SHIP

Figure 13: Model and free response correspondence. Each plot is for trials assigned a particular free response code, each bar is for the number of trials for which that local focus was most likely given the intervention choice. *Effect* and *edge* coded queries were also diagnosed as such by the model fitting while *confirmatory* coded queries were most likely to be diagnosed as querying the effects of the root node(s) in the true model which always was (or included) \( x \).

**General Discussion**

Actively learning causal models is key to higher-level cognition and yet radically intractable. We explored how people manage to identify causal models despite their limited computational resources. Over two experiments, we found that participants’ judgments reflected the true posterior and overall distributions of interventions reflected their average expected information, but that judgments were both sequentially dependent and exhibited recency in terms of consistency with the evidence seen so far. Intervention choices were generally informative but insufficiently reactive to the evidence that had already been observed.

We captured judgment patterns by assuming the learner maintained a single causal model rather than a full distribution, and tried to improve this model by making local changes to improve its ability to explain the latest data. We also captured participants’ interventions by assuming they focused stochastically on different local aspects of the overall uncertainty and tried resolve these, leading to behavior that was comparatively invariant to the prior or the environment, but robustly more informative than purely random intervening.

By casting our modeling in the language of machine learning, we were able to make strong connections between our Neurath’s ship model and established techniques for approximating distributions, sequential Monte-Carlo particle filtering and MCMC (specifically Gibbs) sampling. Likewise we were able to explicate intervention selections using the language of expected uncertainty reduction but relaxing the assumption that the goal was the global uncertainty in the full distribution. The combination of a single hypothesis (particle) and a Gibbs-

56
esque search, nicely reflects the Neurath’s ship intuition that theory change is necessarily piece-meal and that changes are evaluated against the backdrop of the rest of the existing theory.

Limitations of Neurath’s ship

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. This relates to the idea, in philosophy of science, of a “paradigm shift” (Lakatos, 1976). The current Neurath’s ship models do not naturally capture this but accommodate occasional large jumps by assuming a variable search length ($k$), meaning the search will sometimes be long enough to allow the learner to move to a radically different model in a single update. However we might also extend the Neurath’s ship framework to include a threshold on prediction accuracy below which a learner will start afresh e.g. by randomly sampling a model, or sampling from a hitherto unexplored part of the space. Experiments in which the underlying structure changes occasionally over time would provide pointers.

Our Neurath’s ship model assumed that participants’ judgment updates would always be based on only the latest datum $d^t$. This is very frugal, as the learner does not have to store any of the older evidence. However, in practice, learners might take a more moderate path, i.e. remembering some past evidence and reusing it when updating beliefs, particularly over shorter timescales like those in the current experiments. Assuming newer evidence is more likely to be remembered than old, one might explore using some halfway house between $D = d^t$ and $D^t = \{d^1, \ldots, d^t\}$ (also including the interventions $c^t$ and $E^t$) to drive the updating.

Another pragmatic limitation of the current modeling was the assumption of the noisy-OR functional form for the true underlying causal models. While we did take care to train participants on the sources of noise in both Experiments and the exact values in in Experiment 1, our own past work suggests that people may have simpler ways of evaluating likelihoods – for example, in Bramley, Dayan and Lagnado (2015), we found participants’ judgments could be captured by assuming they lumped sources of noise together and just counted the number of surprising outcomes under each model. Maximum likelihood values $w^*$ from participants in Experiment 1 had high variance but implied that subjects were assuming less noise than actually pertained (Figure 6) while in contrast, maximum likelihood estimates and elicited judgments at the end of Experiment 2 (Figure 10) suggested that the systems were considerably
more noisy than participants expected. We might explain this contrast with the idea that participants expected the causal systems to be highly reliable, and so reported their surprise upon experiencing substantial noise with high final noise estimates.

One possibility is that people actually formed likelihood estimates through simulation with an internal causal model. For instance, one might perform a mental intervention, activating a component of one's own internal causal model and keeping track of where the activation propagates. By simulating multiple times a learner could estimate the likelihood of different outcomes under their current model (Hamrick, Smith, Griffiths, & Vul, 2015), and by simulating under variations of the model, the learner could compare likelihoods generated on the fly. This simulation-based view provides a possible explanation for why participants more readily accommodated internal noise $w_s$ than background noise $w_B$. The former is “built in” to the inferred connections in their model, so more robust – i.e. $w_s$ will presumably remain constant for as long as the target causal system continues to function. In contrast, background noise comes by definition from sources outside the scope of the model and so might more reasonably change over time. Thus, it could be that participants attempt to estimate internal noise while remaining agnostic about the levels of background noise.

Future experiments and modeling might relax the assumption of noisy-OR likelihoods and allow induction of more diverse functional forms, or focus on well known domains where priors can be measured before the task. Another approach might be to render the noisy-OR formalization more transparent by visualizing the sources of exogenous noise alongside the target variables, for instance displaying varying numbers of nuisance background variables on screen for different background noise conditions.

**Alternative approximations and representations**

The choice of Gibbs sampling together with a single particle approximation is just one of numerous possible models of structure inference. For example we found (data not shown) that we could also achieve fairly good fits by assuming learners used a form of Metropolis-Hastings MCMC sampling – using an $M C^3$ proposal and acceptance distribution (Madigan & Raftery, 1994; Madigan, York, & Allard, 1995) – to generate local alternatives rather than Gibbs sampling. The two approaches make similar behavioral predictions but differ somewhat in their internal architecture – a Metropolis-Hastings sampler would first generate a similar alternative to their current belief, then make an accept-reject decision about whether to supplant their model with the new one, while the Gibbs sampler focuses on one subpart at a time and updates this conditional on the rest. Ultimately, the Gibbs sampler did a better job, as well aligning bet-
ter with the broader ideas of locality of inference implicit in the Neurath’s ship proposal.

A related application of MCMC to understanding cognition has been to design experiments where the trials form a Markov chain, uncovering participants’ or groups’ subjective functions (e.g. Sanborn, Griffiths, & Shiffrin, 2010). For instance, under certain assumptions, repeatedly asking one or a group of participants to choose which of two stimuli is more likely to be a category member can mimic the accept/reject step of Metropolis Hastings sampling. In principle, the resulting chain of trial decisions will approximate the participant’s (or group’s average) subjective function relating stimuli to probability of category membership. The current work takes a substantially different approach, proposing that people actually perform a form of MCMC in their heads when searching for new possibilities. However, the two approaches are broadly compatible: the former uses MCMC as a method for estimating the latent high level functions that people cannot report directly but which describe their behavior, while our approach uses MCMC as an explanation for why people can be described by a latent high-level function (here a distribution over possible models) yet have no reportable access to it.

An interesting alternative approach to complex model induction via local computations (Fernbach & Sloman, 2009) comes from variational Bayes (Bishop, 2006; Weierstrass, 1902). The idea behind this is that one can simplify inference by replacing an intractable distribution, here the distribution over all possible models, with a simpler one which has degrees of freedom that can be used to allow it to fit as best as possible. A common choice of simpler distribution involves factorization, with a multiplicative combination of a set of simpler distributions. Thus, for causal inference one might make a mean-field approximation (Georges, Kotliar, Krauth, & Rozenberg, 1996) and suppose the true distribution over models factorizes into independent distributions for each causal connection. Divergence between this approximation and the full model can then be minimized mathematically by updating each of the local distributions in turn (Jaakkola, 2001). This provides a different view of global inference based on local updates. Rather than a process of local search where only a single model is represented at any time, variational Bayes suggests people maintain many local distributions and try to minimize the inconsistencies between them. The biases induced by this process make the two approaches distinguishable in principle, meaning that an interesting avenue for future work may be to design experiments that distinguish between the two approaches to approximation in cognition (Sanborn, 2015).
Anchoring Neurath’s ship

The Neurath’s ship approach is related to Hogarth’s and Einhorn’s (1992) anchor-and-adjust model of sequential magnitude estimation. Hogarth and Einhorn found that when mean estimates are repeatedly elicited from participants as they see a sequence of numbers, the sequence of responses can be captured by a process whereby one stores a single value and adjusts it a portion of the way toward each new observed value. When judgments were elicited at the end of the sequence, participants behaved more like they had stored a subset of the values and averaged them at the end. In the same way, we can think of Neurath’s ship as a process in which the current model acts as an anchor, and adjustments are made toward new data as it is observed. However, the higher complexity of causal inference in comparison with mean estimation will lead to greater pressure to use a sequential strategy rather than store. Arguably, step-by-step elicitation is a closer analogue to real-world causal inference than end-of-sequence because causal beliefs are presumably in frequent use while learning instances may be spread out, with no clear start or end.

Choosing interventions aboard Neurath’s ship

The models pinned down interventions less tightly than beliefs. There are various possible reasons for this. Firstly, the models of belief change generally predicted one or few likely models, whereas there are typically many interventions of roughly equal informativeness to to an ideal learner (see Figure 3), which could be performed in many different orders. This sets the bar for predictability for interventions much lower than for the causal judgments. Secondly, to the extent that learners chose interventions based on a reduced encoding of the hypothesis space, we are also forced to average over our additional uncertainty about exactly which hypotheses or alternatives they were considering at the moment of choice (Markant & Gureckis, 2010). A third issue is that of whether and how learners represented current uncertainty, and recruited this in choosing what to focus on. In the current work we assumed that learners were somewhat able to track the current local uncertainties and use these to choose what to target next. The modeling revealed that relative to the local intervention schema, the majority of participants did tend to focus on the areas of high current uncertainty (shown by the predominantly positive $\rho$ in Figure 12c ii) but we do not yet have a model for how they did this. It is plausible that learners used a heuristic to estimate their local confidence. For example, a simple option would be to accrue confidence in an edge, (or analogously in the the descendants of a variable or in the current hypothesis) for every search step for which it is considered and
remains unchanged, reducing confidence every time it changes. In this way confidence in locales that survive more data and search become stronger, approximately mimicking reduction in local uncertainty.

We considered just three of a multitude of possible choices of local focus. However, we feel that these focuses can capture a lot of the extant proposals for human search heuristics, encapsulating modular (Markant et al., 2015) constraint seeking (Ruggeri & Lombrozo, 2014) and confirmatory (Klayman & Ha, 1989) testing, placing them within a unified schema, and showing that many learners will dynamically switch between them.

Participants’ free responses provided a complementary perspective, suggesting that even initial tests were generated as solutions to uncertainty about some specific subpart of the overall uncertainty space – often the descendants of some particular variable or the presence of some particular connection. This suggests that the most important step in an intervention selection may not be the final choice of action but the prior choice of what to focus on next. This is captured in our model, under which the values of different interventions for a chosen focus do not depend on $D^{t-1}$. This means learners need not do extensive prospective calculation on every test but can learn gradually, e.g. through experience and preplay (Pfeiffer & Foster, 2013), which interventions are likely to be informative relative to the current focus. This knowledge could then be transferred to subsequent tests, and translated to tests with different targets – e.g. if $Do[x = 1]$ is effective for identifying the effects of $x$ then $Do[y = 1]$ will be effective for identifying the effects of $y$.

It is worth noting from these data that even when participants’ interventions were relatively uninformative from the perspective of ideal or even our heuristic learners, their explanations would generally reveal that they were informative with respect to some other question or source of uncertainty. For example, participants’ tests that were uninformative with respect to identifying structure were often revealed, through our free response coding, to have been motivated by a desire to reduce uncertainty about internal $w_S$ or background $w_B$ noise. From this perspective we might think of even the completely uninformative intervention choices (e.g. fixing all the variables) as legitimate tests of illegitimate hypotheses – e.g. hypotheses that were outside of the space of possibilities we intended participants to consider – such as whether fixed variables actually always took the states they were fixed to etc. More research is needed to explicate these internal steps leading up to an active learning action, but the implication based

---

17 We might have extended the computational model of Bayesian inference to incorporate joint inference over models and parameters which would have incorporated this aspect of testing. However, this would have complicated analyses since participants were ultimately only incentivised to identify the right connections.
on the current research is that the solution will not require that the learner evaluate all possible outcomes of all possible actions under all possible models, but rather reflect a mixture of heuristics that can guide the gradual improvement of the learner’s current theory.

**Scope of the theory**

We modelled causal belief change as a process of gradually updating a single representation through local, conditional edits. While we chose to focus on causal structure inference within the causal Bayes net framework here, there is no reason why this approach should be limited to this domain. By taking the *Neurath’s ship* metaphor to reveal an intuitive answer as to how people sidestep the intractability of rational theory formation (van Rooij et al., 2014), we can start to build more realistic models of how people generate the theories that they do and how and why they get stuck. We might explain the induction and adaptation of many of the rich representations utilized in cognition by analogous processes. Future work could explore the piecemeal induction of models involving multinomial, continuous (Nodelman, Shelton, & Koller, 2002; Pacer & Griffiths, 2011) or latent variables (Lucas, Holstein, & Kemp, 2014); unrestricted functional forms (Griffiths, Lucas, Williams, & Kalish, 2009); hierarchical organization (Griffiths & Tenenbaum, 2009; Williamson & Gabbay, 2005) or with temporal (Pacer & Griffiths, 2012) and spatial (Battaglia, Hamrick, & Tenenbaum, 2013; Ullman, Goodman, & Tenenbaum, 2012; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2014) semantics.

In the current modeling, we assumed that learners made updates at the level of individual directed edges. Again this is just one illustrative choice, but our model is consistent with the idea that the learners altered beliefs by making changes local to arbitrary sub-spaces of an unmanageable learning problem. We showed that so long as the learner’s updates are conditioned on the rest of their model, and are appropriately balanced, the connection to approximate Bayesian inference can be maintained through the ideas of MCMC sampling and a single particle particle filter. A sophisticated learner might be able to update several edges of their causal model at a single time, with a more complex proposal distribution. However, on a larger scale this is still likely to be a small subset of all potential relata that learner has encountered, meaning even the most sophisticated learner must lean on their broader beliefs for support. The sequential conditional re-evaluation process illustrated by our *Neurath’s ship* model shows how this radical anti-foundationalism need not be fatal for theory building in general.
Conclusions

In this paper, we proposed a new model of causal theory change, based on an old idea from philosophy of science – that learners cannot maintain a distribution over all possible beliefs, and so must rely on sequential local changes to a single representation when updating beliefs to incorporate new evidence. We showed that we can provide a good account of participants’ sequences of judgments in two experiments and argued that our model is a flexible and ubiquitous candidate for explaining how complex representations are formed in cognition. We also analyzed participants’ information-gathering behavior, finding it consistent with the thesis that learners focus on resolving manageable areas of local uncertainty rather than global uncertainty, showing cognizance of their learning limitations. Together these accounts show how people manage to construct rich, causally-structured representations through their interactions with a complex noisy world.

Acknowledgments

We would like to thank Michael Pacer, Maarten Speekenbrink, Falk Lieder and Eric Schulz for helpful comments. NB is supported by a 4-year Engineering and Physical Sciences Research Council UK stipend awarded by UCL’s Centre for Doctoral Training in Financial Computing and Analytics. DL is supported by an Economic and Social Research Council UK grant (RES 062330004). TG is supported by the Air Force Office of Scientific Research (FA9550-13-1-0170), and PD is supported by the Gatsby Charitable Foundation.
References


Lucas, C. G., Holstein, K., & Kemp, C. (2014). Discovering hidden causes using statis-
tical evidence..


A Formal specification of the models

A.1 Representation and inference

A noisy-OR parametrized causal model \( m \) over variables \( X \), with strength and background parameters \( w_S \) and \( w_B \) assigns a likelihood to each datum (a complete observation, or the outcome of an intervention) \( d \) as the product of the probability of each variable that was not intervened upon given the states of its parents

\[
P(d|m, w) = \prod_{x \in X} P(x|d_{pa(x)}, w)
\]

(20)

\[
P(x|d_{pa(x)}, w) = x + (1 - 2x)(1 - w_B)(1 - w_S)\sum_{y \in pa(x)} y
\]

(21)

where \( pa(x) \) denotes the parents of variable \( x \) in the causal model (see Figure 1 for an example).

We can thus compute the posterior probability of model \( m \in M \) over a set of models \( M \) given a prior \( P(M) \) and data \( D = \{d^i\} \) associated with interventions \( C = \{c^i\} \). We can condition on \( w_S \) and \( w_B \) if known (e.g. in Experiment 1)

\[
P(m|D, w) = \frac{P(D|m, w; C)P(m)}{\sum_{m' \in M} P(D|m', w; C)P(m')}
\]

(22)

or else marginalize over their possible values (e.g. in Experiment 2)

\[
P(m|D) = \frac{\int_w P(D|m, w; C)p(w)P(m) \, dw}{\sum_{m' \in M} \int_w P(D|m', w; C)p(w)P(m') \, dw}
\]

(23)

A.2 Intervention choice

The value of an intervention can be quantified relative to a notion of uncertainty. Here we adopt Shannon entropy (2001), for which the uncertainty in a distribution over causal models \( M \) is given by

\[
H(M) = -\sum_{m \in M} P(m) \log_2 P(m)
\]

(24)
Assuming $w$ is known, let $\Delta H(M|d, w; c)$ refer to the reduction in uncertainty going from prior $P(M)$ to posterior $P(M|d, w; c)$ after performing intervention $c$, then seeing data $d$

$$\Delta H(M|d, w; c) = \left[ - \sum_{m \in M} P(m) \log P(m) \right] - \left[ - \sum_{m \in M} P(m|d, w; c) \log P(m|d, w; c) \right]$$ \hspace{1cm} (25)

Given this objective, we can define the value of an intervention as the expected reduction in uncertainty after seeing its outcome. To get the expectancy, we must average, prospectively, over the different possible outcomes $d \in D_c$ (where $D_c$ is the space of possible outcomes of intervention $c$) weighted by their marginal likelihoods under the prior, giving

$$\mathbb{E}_{d \in D_c} \left[ \Delta H(M|d, w; c) \right] = \sum_{d \in D_c} \left( \Delta H(M|d, w; c) \sum_{m \in M} P(d|m, w; c)P(m) \right)$$ \hspace{1cm} (26)

For a greedily optimal sequence of interventions $c^1, \ldots, c^t$, we take $P(M|D^{t-1}, w; C^{t-1})$ as $P(M)$ and $P(M|D^t, w; C^{t-1}, c^t)$ as $P(M|d, w; c)$ in Equation 25. The most valuable intervention at a given time point is then

$$c^t = \arg \max_{c \in C} \mathbb{E}_{d \in D_c} \left[ \Delta H(M|d, D^{t-1}, w; C^{t-1}, c) \right]$$ \hspace{1cm} (27)

If $w$ is unknown, we must use the marginal distribution, replacing Equation 25 with

$$\Delta H(M|d; c) = \left[ - \sum_{m \in M} P(m) \log P(m) \right] - \left[ - \sum_{m \in M} \int_w P(m|d, w; c)p(w) \, dw \log \int_w P(m|d, w; c)p(w) \, dw \right]$$ \hspace{1cm} (28)

### A.3 An algorithmic-level model of sequential belief change

Let $E$ be an adjacency matrix such that the upper triangle entries where $E_{ij}$ (if $i < j \leq N$) denotes the state of edge $i-j$ in a causal model $m$. Any model $m \in M$ corresponds to a setting for all $E_{ij}$ where $i < j \leq N$, to one of three edge states $e \in \{1: i \rightarrow j, 0: i \leftrightarrow j, -1: i \leftarrow j\}$.

By starting with any hypothesis and iteratively sampling from the conditional distributions on edge states $P(E_{ij}|E_{\setminus ij}, d^t, w; c^t)$ (Goudie & Mukherjee, 2011) using the following equation:

$$P(E_{ij} = e|E_{\setminus ij}, d^t, w; c^t) = \frac{P(E_{ij} = e|E_{\setminus ij}, d^t, w; c^t)}{\sum_{e' \in E_{ij}} P(E_{ij} = e'|E_{\setminus ij}, d^t, w; c^t)}$$ \hspace{1cm} (29)

we can cheaply generate chains of dependent samples from $P(M|d^t, w; c^t)$. This can be done systematically (cycling through all edges $\in i < j \leq N$), or randomly selecting the next edge sample with $P(\frac{1}{|E_{\setminus ij}|})$ where $|i, j|$ is the number of edges in the graph. Here we assume random
sampling for simplicity. Thus, on each step, the selected $E_{ij}$ is updated using the newest values of $E_{ij}$. Specifically, we assume that after each new piece of evidence arrives:

1. The learner begins sampling with edges $E_{ij}^{(0)}$ for all $i$ and $j$ set as they were in their previous judgement $b^{t-1}$.
2. They then randomly select an edge $E_{ij}$ in $i < j \leq N$ to update.
3. They resample $E_{ij}^{(1)}$ using Equation 29.
4. The learner repeats this $k$ times, with their final edge choices $E^{(k)}$ constituting their new belief $b^t$.

We assume for simplicity that $b^0$, before any data has been seen is an unconnected graph, but test this assumption in Appendix B.2.

**Resampling, hill climbing or random change**  We also consider generalizations of Equation 29 allowing transitions to be governed by higher powers of $P(E_{ij} = e|E_{\setminus ij}, d^t, w; c^t)$

$$P^\omega(E_{ij} = e|E_{\setminus ij}, d^t, w; c^t) = \frac{P^\omega(E_{ij} = e|E_{\setminus ij}, d^t, w; c^t)}{\sum_{e' \in E_{ij}} P^\omega(E_{ij} = e'|E_{\setminus ij}, d^t, w; c^t)}$$

yielding stronger preference for the most likely state of $e_{ij}$ if $\omega > 1$ and more random sampling if $\omega < 1$.

**A distribution over search lengths**  We assume that for each update, the learner’s length of search $k$ is drawn from a Poisson distribution with average $\lambda \in [0, \infty]$

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

**Putting these together**  To calculate the probability distribution of new belief $b^t$ given $d^t$, $b^{t-1}$ search behavior $\omega$ and a chain of length $k$, we first construct the transition matrix $R^\omega_t$ for the Markov search chain by averaging over the conditional distributions associated with the choice of each edge, weighted by the probability of selecting that edge

$$R^\omega_t = \sum_{i<j \leq N} P^\omega(E_{ij} = e|E_{\setminus ij}, d^t, w; c^t) \times \frac{1}{|i,j|}$$

for each possible belief $b$.

\[\text{Edge changes that would create a cyclic graph always have a probability of zero.}\]
By raising this transition matrix to the power $k$ (i.e. some search length) and selecting the row corresponding to starting belief $[(R^w_t)^k]_{b^t-1}$, we get the probability of adopting each $m \in M$ as new belief $b^t$ (see Figure 2 for a visualization) at the end of the $k$ length search

$$P(B^t|d^t, b^{t-1}, \omega, k; c^t) = [(R^w_t)^k]_{b^t-1,m} \quad (33)$$

Finally, by averaging over different possible chain lengths $k$, weighted by their probability Poisson($\lambda$) we get the marginal probability that a learner will move to each possible new belief in $B$ at $t$

$$P(B^t|d^t, b^{t-1}, \omega, \lambda; c^t) \approx \sum_0^\infty \frac{\lambda^k e^{-\lambda}}{k!} [(R^w_t)^k]_{b^t-1,m} \quad (34)$$

A.4 A local uncertainty schema

A.4.1 Edge focus

Relative to a focus on an edge $E_{xy}$, intervention values were calculated using expected information as in Appendix A.2, but assuming prior entropy as that of a uniform distribution over the three possible edge states

$$H(E_{xy}|E_{\backslash xy}) = -3 \left( \frac{1}{3} \log_2 \frac{1}{3} \right) \quad (35)$$

and calculating posterior entropies for the possible outcomes $d \in D$ using

$$H(E_{xy}|E_{\backslash xy}, d, w; c) = - \sum_{z \in \{-1,0,1\}} P(E_{xy} = z|E_{\backslash xy}, d, w; c) \log_2 P(E_{xy} = z|E_{\backslash xy}, d, w; c) \quad (36)$$

A.4.2 Effect focus entropy

Relative to a focus on the effects of variable $x$, intervention values were calculated using expected information as in Appendix A.2 but using prior entropy, calculated by partitioning a uniform prior over models $M$ into sets of models $Mo(z)$ corresponding to each descendant set $z \subseteq De(x)$

$$H(De(x)) = - \sum_{z \subseteq De(x)} \left( \sum_{m \in Mo(z)} \frac{1}{|M|} \right) \log_2 \left( \sum_{m \in Mo(z)} \frac{1}{|M|} \right) \quad (37)$$
Posterior entropies were then calculated by summing over probabilities of the elements in each Mo(z) for each z ⊆ De(x)

\[ H(De(x)|d, w; c) = - \sum_{z \subseteq De(x)} \left( \sum_{m \in Mo(z)} P(m|d, w; c) \right) \log_2 \left( \sum_{m \in Mo(z)} P(m|d, w; c) \right) \]  

(38)

A.4.3 Confirmation focus entropy

Relative to a focus on distinguishing current hypothesis \( b^t \) from null hypothesis \( b^0 \), intervention values were calculated using expected information as in Appendix A.2 but prior entropy was always based on a uniform prior over the two hypotheses

\[ H(\{b^t, b^0\}) = -2 \left( \frac{1}{2} \log_2 \frac{1}{2} \right) \]  

(39)

and posterior entropies were calculated using

\[ H(\{b^t, b^0\}|d, w; c) = - \sum_{z \in \{0, t\}} \frac{P(b^z|d, w; c)}{\sum_{z' \in \{0, t\}} P(b^{z'}|d, w; c)} \log_2 \frac{P(b^z|d, w; c)}{\sum_{z' \in \{0, t\}} P(b^{z'}|d, w; c)} \]  

(40)

B Additional modeling details

B.1 Details of the model fitting

All models were fit using maximum likelihood. Maximum likelihood estimates were found using Brent (for one parameter) or Nelder-Mead (for several parameters) optimization, as implemented by R’s \texttt{optim} function. Convergence to global optima was checked by repeating all optimizations with a range of randomly selected starting parameters.

\( k \) For averaging across different values of \( k \) in the belief models, we capped \( k \) at 50 and renormalized the distribution such that \( P(k \geq 0 \land k \leq 50) = 1 \). This made negligible difference to the fits since the probabilities of \( P(B^k|d^t, b^{t-1}, \omega, k; c') \) for values of \( k \gg N \) (where \( N \) is the number of variables) were very similar.

\( \epsilon \) In Experiment 1, subjects occasionally made a model judgment that had been perfectly ruled out in the deterministic and semi-deterministic conditions (1 - 6). To avoid assigning these participants zero likelihoods we included a lapse parameter \( \epsilon \) – i.e., a parametric amount of decision noise \( \epsilon \in [0, 1] \) – so that the probability of a belief would be a mixture of that predicted by the model and uniform noise.
We assume for simplicity that people’s starting belief, $b_0$, before any data has been seen, is an unconnected graph. But, to test this assumption we also fit all models to data on trials $2 : T$ on each problem (Tables 6 and 7, below).

**Marginalization** For all modeling in Experiment 2, we had to average over the unknown noise $w$. To do this, we drew 1000 paired uniformly distributed $w_S$ and $w_B$ samples and averaged over these when computing marginal likelihoods and posteriors. These marginal priors and posteriors were used for computing expected information gain values.

**Evaluating fits** *Baseline* acts as the null model for computing BIC’s (Schwarz, 1978) and pseudo-$R^2$’s (Dobson, 2010) for all other models.

**Relaxing assumptions** The assumption that participants’ starting hypothesis on each problem $b_0$ should be an unconnected graph was quite arbitrary. Therefore, to test whether this privileged the NS model in the results, we also fit the models to the data from $t = 2 : T$ on each problem, such that even on the first fitted trial there would already be a previous judgment $b_1$. This actually strengthened the fit for NS. Results are available in Appendix B.2, Tables 6 and 7.

**Code** Finally, code used for analyzing the data and fitting the models is available upon request.

### B.2 Additional fitted models

Here we present additional tables giving the fits of variants of the NS model presented in the main text.

Table 6: Experiment 1 fitted belief models. As in Table 3 but fit to data from trials 2:T only.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda_{\text{med}}$</th>
<th>$\lambda_{\text{SD}}$</th>
<th>$\omega_{\text{med}}$</th>
<th>$\omega_{\text{SD}}$</th>
<th>$\theta_{\text{med}}$</th>
<th>$\theta_{\text{SD}}$</th>
<th>$\epsilon_{\text{med}}$</th>
<th>$\epsilon_{\text{SD}}$</th>
<th>N</th>
<th>M score</th>
<th>logL</th>
<th>$R^2$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Baseline</td>
<td>0.7</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 NS, $\omega=0$</td>
<td>0.07</td>
<td>0.17</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 NS</td>
<td>0.9</td>
<td>60</td>
<td>136</td>
<td>298</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Rational</td>
<td>13</td>
<td>723</td>
<td>0.16</td>
<td>0.38</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

77
Table 7: Experiment 2 fitted belief models. As in Table 3 but fit to data from trials 2:T only.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda_\text{med}$</th>
<th>$\lambda_\text{SD}$</th>
<th>$\omega_\text{med}$</th>
<th>$\omega_\text{SD}$</th>
<th>$\theta_\text{med}$</th>
<th>$\theta_\text{SD}$</th>
<th>$\epsilon_\text{med}$</th>
<th>$\epsilon_\text{SD}$</th>
<th>N rem</th>
<th>N diss</th>
<th>M score</th>
<th>logL</th>
<th>$R^2$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 baseline</td>
<td>2</td>
<td>5.0</td>
<td>-12505</td>
<td>0</td>
<td>25011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 $\text{NS}_{\omega=0}$</td>
<td>0.8</td>
<td>2</td>
<td>0.09</td>
<td>0.27</td>
<td>9</td>
<td>5</td>
<td>8.5</td>
<td>-8450</td>
<td>0.32</td>
<td>17688</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 $\text{NS}$</td>
<td>1.4</td>
<td>99</td>
<td>18.1</td>
<td>264</td>
<td>0.04</td>
<td>0.20</td>
<td>28</td>
<td>25</td>
<td>12.6</td>
<td>-7343</td>
<td>0.41</td>
<td>15870</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Rational</td>
<td>14.2</td>
<td>229</td>
<td>0.78</td>
<td>0.24</td>
<td>16</td>
<td>26</td>
<td>15.5</td>
<td>-8131</td>
<td>0.35</td>
<td>17052</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
List of Tables

1 The true underlying structure for each problem and the three most frequent final judgments. In all cases but the 3-variable chain the most frequent final judgment was the correct model. ∅ indicates a final judgment that there were no connections. ................................................................. 28

2 Experiment 1 belief models. Columns: med = median estimated parameter (across participants), SD = standard deviation of parameter estimate across participants, N fit = number of participants best fit by each model (/150), M score = average score of those participants (/45), LogL = total log likelihood of model over all data, $R^2 = \text{McFadden’s pseudo-}R^2$, BIC = Bayesian information criterion. Best fitting model denoted with boldface. ........................................... 51

3 Experiment 2 belief models. As in Table 2. ...................................... 51

4 Experiment 1 intervention models. Columns: as in belief model Tables 2 and 3 . 54

5 Experiment 2 intervention models. ................................................... 54

6 Experiment 1 fitted belief models. As in Table 3 but fit to data from trials 2:T only. 77

7 Experiment 2 fitted belief models. As in Table 3 but fit to data from trials 2:T only. 78
Table 1: The true underlying structure for each problem and the three most frequent final judgments. In all cases but the 3-variable chain the most frequent final judgment was the correct model. ∅ indicates a final judgment that there were no connections.

<table>
<thead>
<tr>
<th>N variables</th>
<th>True model</th>
<th>Most frequent</th>
<th>N</th>
<th>2nd</th>
<th>N</th>
<th>3rd</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Single $x \rightarrow y$</td>
<td>$x \rightarrow y$</td>
<td>40</td>
<td>$x \rightarrow y, z \rightarrow y$</td>
<td>15</td>
<td>$z \rightarrow x, z \rightarrow y$</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>Common cause $x \rightarrow y, x \rightarrow z$</td>
<td>$x \rightarrow y, x \rightarrow z$</td>
<td>43</td>
<td>$x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>14</td>
<td>$x \rightarrow y, z \rightarrow y, x \rightarrow z$</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>Chain $x \rightarrow y, y \rightarrow z$</td>
<td>$x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>30</td>
<td>$x \rightarrow y, y \rightarrow z$</td>
<td>28</td>
<td>$x \rightarrow y, x \rightarrow z$</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>Common effect $x \rightarrow z, y \rightarrow z$</td>
<td>$x \rightarrow z, y \rightarrow z$</td>
<td>41</td>
<td>$x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>15</td>
<td>$y \rightarrow x, z \rightarrow x, y \rightarrow z$</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>Fully connected $x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>$x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>39</td>
<td>$x \rightarrow y, x \rightarrow z$</td>
<td>20</td>
<td>$x \rightarrow z, y \rightarrow z$</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>Single $w \rightarrow x$</td>
<td>$w \rightarrow x$</td>
<td>26</td>
<td>∅</td>
<td>9</td>
<td>$w \rightarrow x, w \rightarrow z$</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>Common cause $w \rightarrow x, w \rightarrow y, w \rightarrow z$</td>
<td>$w \rightarrow x, w \rightarrow y, w \rightarrow z$</td>
<td>26</td>
<td>$y \rightarrow w, y \rightarrow x, y \rightarrow z$</td>
<td>3</td>
<td>$w \rightarrow x, y \rightarrow x, z \rightarrow x, w \rightarrow y, z \rightarrow y, w \rightarrow z$</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Chain $w \rightarrow x, x \rightarrow y, y \rightarrow z$</td>
<td>$w \rightarrow x, x \rightarrow y, y \rightarrow z$</td>
<td>7</td>
<td>$w \rightarrow x, x \rightarrow y, x \rightarrow z, y \rightarrow z$</td>
<td>6</td>
<td>$w \rightarrow x, w \rightarrow y, x \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Common effect $w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>$w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>25</td>
<td>∅</td>
<td>3</td>
<td>$y \rightarrow z$</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Fully connected $w \rightarrow x, w \rightarrow y, x \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>$w \rightarrow x, w \rightarrow y, x \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>18</td>
<td>$w \rightarrow x, w \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>7</td>
<td>$w \rightarrow x, w \rightarrow y, x \rightarrow y, w \rightarrow z, x \rightarrow z, y \rightarrow z$</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 2: Experiment 1 belief models. Columns: med = median estimated parameter (across participants), SD = standard deviation of parameter estimate across participants, N fit = number of participants best fit by each model (/150), M score = average score of those participants (/45), LogL = total log likelihood of model over all data, $R^2 =$ McFadden’s pseudo-$R^2$, BIC = Bayesian information criterion. Best fitting model denoted with boldface.

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Parameter</th>
<th>$\lambda$ med</th>
<th>$\lambda$ SD</th>
<th>$\omega$ med</th>
<th>$\omega$ SD</th>
<th>$\theta$ med</th>
<th>$\theta$ SD</th>
<th>$\epsilon$ med</th>
<th>$\epsilon$ SD</th>
<th>N fit</th>
<th>M score</th>
<th>logL</th>
<th>$R^2$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Baseline</td>
<td></td>
<td>0</td>
<td>-52268</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>104535</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 NS, $\omega=0$</td>
<td></td>
<td>0.7</td>
<td>49</td>
<td></td>
<td>0.15</td>
<td>0.18</td>
<td>41</td>
<td></td>
<td>15.0</td>
<td>-26674</td>
<td>0.49</td>
<td>54622</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 NS</td>
<td></td>
<td>1.0</td>
<td>61</td>
<td>136</td>
<td>299</td>
<td></td>
<td></td>
<td>0.13</td>
<td>0.18</td>
<td>107</td>
<td>27.0</td>
<td>-24502</td>
<td>0.53</td>
<td>50278</td>
</tr>
<tr>
<td>4 Rational</td>
<td></td>
<td>17</td>
<td>728</td>
<td>0.16</td>
<td>0.38</td>
<td>2</td>
<td></td>
<td></td>
<td>42.0</td>
<td>-43702</td>
<td>0.16</td>
<td>88678</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Experiment 2 belief models. As in Table 2.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\lambda_{med}$</th>
<th>$\lambda_{SD}$</th>
<th>$\omega_{med}$</th>
<th>$\omega_{SD}$</th>
<th>$\theta_{med}$</th>
<th>$\theta_{SD}$</th>
<th>$\epsilon_{med}$</th>
<th>$\epsilon_{SD}$</th>
<th>$N_{rem}$</th>
<th>$N_{diss}$</th>
<th>$M_{score}$</th>
<th>logL</th>
<th>$R^2$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>2</td>
<td>5.0</td>
<td>-15006</td>
<td>0</td>
<td>30013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS, $\omega=0$</td>
<td>0.9</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.10</td>
<td>0.29</td>
<td>9</td>
<td>5</td>
<td>8.4</td>
<td>-10877</td>
<td>0.28</td>
<td>22585</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NS</td>
<td>1.6</td>
<td>82</td>
<td>20.2</td>
<td>277</td>
<td>0.04</td>
<td>0.20</td>
<td>28</td>
<td>29</td>
<td>13.0</td>
<td>-9360</td>
<td>0.38</td>
<td>19963</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rational</td>
<td>16.0</td>
<td>175</td>
<td>1.00</td>
<td>0.32</td>
<td>16</td>
<td>22</td>
<td>15.3</td>
<td>-10308</td>
<td>0.31</td>
<td>21447</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

82
Table 4: Experiment 1 intervention models. Columns: as in belief model Tables 2 and 3

<table>
<thead>
<tr>
<th>Model</th>
<th>$\eta_{med}$</th>
<th>$\eta_{sd}$</th>
<th>$\rho_{med}$</th>
<th>$\rho_{sd}$</th>
<th>$\theta_{med}$</th>
<th>$\theta_{sd}$</th>
<th>N fit</th>
<th>M score</th>
<th>logL</th>
<th>$R^2$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>25</td>
<td>13.5</td>
<td>-41198</td>
<td>82396</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edges</td>
<td>6.7</td>
<td>13</td>
<td>11.4</td>
<td>680</td>
<td>5</td>
<td>23.8</td>
<td>0.09</td>
<td>75859</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effects</td>
<td>6.3</td>
<td>20</td>
<td>8.9</td>
<td>721</td>
<td>38</td>
<td>26.0</td>
<td>0.19</td>
<td>67779</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confirmatory</td>
<td>4.5</td>
<td>27</td>
<td></td>
<td></td>
<td>21</td>
<td>17.4</td>
<td>0.05</td>
<td>78575</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>12.0</td>
<td>109</td>
<td>11.6</td>
<td>232</td>
<td>28</td>
<td>32.4</td>
<td>0.20</td>
<td>67577</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rational</td>
<td>3.2</td>
<td>6</td>
<td>33</td>
<td></td>
<td>33</td>
<td>26.2</td>
<td>0.17</td>
<td>68894</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5: Experiment 2 intervention models.

<table>
<thead>
<tr>
<th>Model</th>
<th>η med</th>
<th>η sd</th>
<th>ρ med</th>
<th>ρ sd</th>
<th>θ med</th>
<th>θ sd</th>
<th>N fit</th>
<th>M score</th>
<th>logL</th>
<th>R²</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
<td>-15365</td>
<td>30730</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edges</td>
<td>4.0</td>
<td>9</td>
<td>6.8</td>
<td>98</td>
<td></td>
<td></td>
<td>25</td>
<td>-12944</td>
<td>0.18</td>
<td>26718</td>
<td></td>
</tr>
<tr>
<td>Effects</td>
<td>3.2</td>
<td>9</td>
<td>15.4</td>
<td>410</td>
<td></td>
<td></td>
<td>8</td>
<td>-13000</td>
<td>0.13</td>
<td>26829</td>
<td></td>
</tr>
<tr>
<td>Confirmatory</td>
<td>2.5</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>-14180</td>
<td>0.04</td>
<td>28776</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>3.8</td>
<td>9</td>
<td>6.4</td>
<td>705</td>
<td></td>
<td></td>
<td>18</td>
<td>-12418</td>
<td>0.19</td>
<td>25665</td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>3.0</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>36</td>
<td>-12850</td>
<td>0.14</td>
<td>26114</td>
<td></td>
</tr>
</tbody>
</table>
List of Figures

1 Causal model representation. a) An example causal Bayesian network, parametrized with strength $w_S$ and base rate $w_B$. The tables give the probability of each variable taking the value 1 conditional on its parents in the model and the omnipresent background noise rate $w_B$. b) Visualisation of intervention $Do[y = 1]$. Setting $y$ to 1 renders it independent of its normal causes as indicated by the scissors symbols. ................................................................. 7

2 An illustration of NS model of causal belief updating. a) An example search path: The learner starts out with a singly connected model at the top ($x \rightarrow y$ connection only). They update their beliefs by resampling one edge at a time $e \in \{\rightarrow, \leftrightarrow, \leftarrow\}$. Each entry $i, j$ in the matrices gives the probability of moving from model in the row $i$ to the model in the column $j$ when resampling the edge marked with the colored question mark. Lighter shades of the requisite color indicate low transition probability, darker shades indicate greater transition probability; yellow is used to indicate zero probabilities. Here the learner stops after resampling each edge once, moving from $b^{-1}$ of $[x \rightarrow y]$ to $b^1$ of $[x \rightarrow y, x \rightarrow z, y \rightarrow z]$. b) Assuming the edge to resample is chosen at random, we can average over the different possible edge choices to derive a 1-step Markov chain transition matrix $R^{\omega}$ encompassing all the possibilities. By raising this matrix to higher powers we get the probability of different end points for searches of that length. If the chain is short (small $k$) the final state depends heavily on the starting state (top) but for longer chains (large $k$), the starting state becomes less important, getting increasingly close to independent sampling from the desired distribution (bottom). .................................................. 14
An illustrative example of local focused uncertainty minimization a) Three possible “local” focuses b) Expected value of 19 different interventions at the start of learning (i.) and after several tests have been performed (ii.) assuming: global expected information gain from the true prior (green squares, and shaded), effects of $z$ focus (red circles), the relationship between $x$ and $y$ (blue triangles) and confirming $b^{t-1}$ (yellow diamonds), assuming a uniform prior over the requisite possibilities and a known $w_S$ and $w_B$ of .85 and .15. c) The value of these choices of focus according to their current uncertainty Equation 6. Note that confirmation is undefined at the start of learning where both current and null hypothesis are that there are no connections in the model. 

The true models and aggregated results. a) The problems faced by participants. b) Weighted average final judgments by participants. Darker arrows indicate that a larger proportion of participants marked this edge in their final model. c) Bayes-optimal final marginal probability of each edge in $P(M|D^T, E^T w)$, averaged over participants’ data. 

Experiment 1 procedure. a) Selecting a test b) Observing the outcome c) Updating beliefs d) Getting feedback. 

Experiment 1 results. 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 with mean search length of 1. Red squares: random guessing. b) Bars show average number of edits (additions, subtractions or reversals of connections) between all $t-1$ and $t$ judgments, as compared to Bayesian, Neurath’s ship and random choice simulations. c) Probability of data $d^t$ given final model judgement $b^T$ on three and four variable problems, ± standard errors. Increase in likelihood for more recent data points (e.g. those closer to $T$), is indicative of recency. d) Boxplot of best fitting $w^*_S$ and $w^*_B$ parameters assuming learners soft-maximised over $P(M|D^t, w^*; C^t)$. 

---

86
Running head: FORMALIZING NEURATH’S SHIP

7 Intervention choices on three variable problems in Experiment 1 compared with matched active learning simulations. a) Across the three strength ($w_S$) and three background noise ($w_B$) conditions, b) Problems by trial. c) Comparing the overall distributions and global information of participants selections to simulations using the three proposed local focuses. ........................................ 34

8 Experiment 2 additional measures - a) Outcome expectation sliders b) Outcome informativeness elicitation c) Edge confidence sliders. .......................... 37

9 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 $\int_w P(M|D^T;C^T)p(w)\, dw$ averaged over participants’ data assuming a uniform independent prior over $w \in [0,1]^2$. ........................................ 38

10 Experiment 2 results. a) Performance by condition in Experiment 2. White circles, crosses, plusses: benchmark (greedy expected information gain maximizing) Bayesian learner assuming a uniform and uniform (UU), strong and uniform (SU) and sparse and strong (SS) prior. Blue diamonds, crosses, plusses: Bayesian learner that maximizes over posterior after seeing participants’ interventions, assuming UU, SU and SS priors. Crosses: assuming a “strong” prior on $w_S$ but a uniform prior on $w_B$. Green triangles Neurath’s ship simulation with mean search length $\lambda$ of 2 assuming UU prior (see Section 3). Red squares: random guessing. b) Participants final judgments of amount of background noise ($w_B$) and strength ($w_S$), rescaled from 100 point scale to 0-1, and best-fitting $w_S^*$ and $w_B^*$ estimates assuming ideal Bayesian updating, error bars $\pm$ SD. c) Visualizations of Uniform-Uniform, Strong-Uniform and Strong-Sparse priors on $w_S$ and $w_B$. d) Plot of the likelihood of $d^1 \ldots d^T$ under $b^T \pm$ standard errors. ........ 41

11 Free explanations for interventions agreed codes over the six tests in Experiment 2, problem 7’. ................................................................. 46

12 Model fitting results. a) Experiment 1, i. Participants’ BIC values for the fitted belief change models (lower is better), best overall model indicated by rectangle, participant’s task performance coded by point colour (high = red, low = blue) b) Same plots for Experiment 2. c) Density estimates for mean chain lengths ($\lambda$) for belief update models and local focus rationality parameter ($\rho$) for intervention models. ........................................ 55
Model and free response correspondence. Each plot is for trials assigned a particular free response code, each bar is for the number of trials for which that local focus was most likely given the intervention choice. Effect and edge coded queries were also diagnosed as such by the model fitting while confirmatory coded queries were most likely to be diagnosed as querying the effects of the root node(s) in the true model which always was (or included) $x$. . . . . . . . . . 56
Figure 1: Causal model representation. a) An example causal Bayesian network, parametrized with strength $w_S$ and base rate $w_B$. The tables give the probability of each variable taking the value 1 conditional on its parents in the model and the omnipresent background noise rate $w_B$. b) Visualisation of intervention Do[$y = 1$]. Setting $y$ to 1 renders it independent of its normal causes as indicated by the scissors symbols.
Figure 2: An illustration of NS model of causal belief updating. a) An example search path: The learner starts out with a singly connected model at the top \((x \rightarrow y)\) connection only. They update their beliefs by resampling one edge at a time \(e \in \{\rightarrow, \leftrightarrow, -\}\). Each entry \(i, j\) in the matrices gives the probability of moving from model in the row \(i\) to the model in the column \(j\) when resampling the edge marked with the colored question mark. Lighter shades of the requisite color indicate low transition probability, darker shades indicate greater transition probability; yellow is used to indicate zero probabilities. Here the learner stops after resampling each edge once, moving from \(b_x^{y-1}\) of \([x \rightarrow y]\) to \(b'\) of \([x \rightarrow y, x \rightarrow z, y \rightarrow z]\). b) Assuming the edge to resample is chosen at random, we can average over the different possible edge choices to derive a 1-step Markov chain transition matrix \(P_{t}^{c}\) encompassing all the possibilities. By raising this matrix to higher powers we get the probability of different endpoints for searches of that length. If the chain is short (small \(k\)) the final state depends heavily on the starting state (top) but for longer chains (large \(k\)), the starting state becomes less important, getting increasingly close to independent sampling from the desired distribution (bottom).
Figure 3: An illustrative example of local focused uncertainty minimization a) Three possible “local” focuses b) Expected value of 19 different interventions at the start of learning (i.) and after several tests have been performed (ii.) assuming: global expected information gain from the true prior (green squares, and shaded), effects of $z$ focus (red circles), the relationship between $x$ and $y$ (blue triangles) and confirming $b^{1-1}$ (yellow diamonds), assuming a uniform prior over the requisite possibilities and a known $w_S$ and $w_B$ of .85 and .15. c) The value of these choices of focus according to their current uncertainty Equation 6. Note that confirmation is undefined at the start of learning where both current and null hypothesis are that there are no connections in the model.
b) Participants’ averaged final judgments Experiment 1

Figure 4: The true models and aggregated results. a) The problems faced by participants. b) Weighted average final judgments by participants. Darker arrows indicate that a larger proportion of participants marked this edge in their final model. c) Bayes-optimal final marginal probability of each edge in $P(M|D^T, E^T W)$, averaged over participants’ data.
Figure 5: Experiment 1 procedure. a) Selecting a test b) Observing the outcome c) Updating beliefs d) Getting feedback.
Figure 6: Experiment 1 results. 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 with mean search length of 1. Red squares: random guessing. b) Bars show average number of edits (additions, subtractions or reversals of connections) between all \( t - 1 \) and \( t \) judgments, as compared to Bayesian, Neurath’s ship and random choice simulations. c) Probability of data \( d^t \) given final model judgement \( b^T \) on three and four variable problems, ± standard errors. Increase in likelihood for more recent data points (e.g. those closer to \( T \)), is indicative of recency. d) Boxplot of best fitting \( w_S \) and \( w_B \) parameters assuming learners soft-maximised over \( P(M|D^t; w^*; C^t) \).
a) By condition

b) By trial

Figure 7: Intervention choices on three variable problems in Experiment 1 compared with matched active learning simulations. a) Across the three strength ($w_S$) and three background noise ($w_B$) conditions, b) Problems by trial. c) Comparing the overall distributions and global information of participants selections to simulations using the three proposed local focuses.
Figure 8: Experiment 2 additional measures - a) Outcome expectation sliders b) Outcome informativeness elicitation c) Edge confidence sliders.
Figure 9: 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 \( \int_w P(M|D^T, C^T)p(w) \, dw \) averaged over participants’ data assuming a uniform independent prior over \( w \in [0, 1]^2 \).
Figure 10: Experiment 2 results. a) Performance by condition in Experiment 2. White circles, crosses, plusses: benchmark (greedy expected information gain maximizing) Bayesian learner assuming a uniform and uniform (UU), strong and uniform (SU) and sparse and strong (SS) prior. Blue diamonds, crosses, plusses: Bayesian learner that maximizes over posterior after seeing participants’ interventions, assuming UU, SU and SS priors. Crosses: assuming a “strong” prior on $w_S$ but a uniform prior on $w_B$. Green triangles Neurath’s ship simulation with mean search length $\lambda$ of 2 assuming UU prior (see Section 3). Red squares: random guessing. b) Participants final judgments of amount of background noise ($w_B$) and strength ($w_S$), rescaled from 100 point scale to 0-1, and best-fitting $w_S^*$ and $w_B^*$ estimates assuming ideal Bayesian updating, error bars ± SD. c) Visualizations of Uniform-Uniform, Strong-Uniform and Strong-Sparse priors on $w_S$ and $w_B$ d) Plot of the likelihood of $d^1 \ldots d^T$ under $b^T$ ± standard errors.
Figure 11: Free explanations for interventions agreed codes over the six tests in Experiment 2, problem 7'.
Figure 12: Model fitting results. a) Experiment 1, i. Participants' BIC values for the fitted belief change models (lower is better), best overall model indicated by rectangle, participant's task performance coded by point colour (high = red, low = blue) b) Same plots for Experiment 2. c) Density estimates for mean chain lengths ($\lambda$) for belief update models and local focus rationality parameter ($\rho$) for intervention models.
Figure 13: Model and free response correspondence. Each plot is for trials assigned a particular free response code, each bar is for the number of trials for which that local focus was most likely given the intervention choice. *Effect* and *edge* coded queries were also diagnosed as such by the model fitting while *confirmatory* coded queries were most likely to be diagnosed as querying the effects of the root node(s) in the true model which always was (or included) $x$. 