Distinguishing Between Process Models of Causal Learning

Simon Valentin[1†] **Lucas Castillo**[2†] **Adam N. Sanborn**[2] **Chris G. Lucas**[1]
[1]University of Edinburgh, United Kingdom
[2]University of Warwick, United Kingdom
[†]The authors contribute equally to this paper.

Abstract

The mechanisms of learning stimulus-stimulus relationships are a longstanding research subject in psychology and neuroscience. Although traditional computational models provide valuable insights into learning processes, they often focus on the average behavior of a population. Individual learning trajectories, however, exhibit a diverse range of behaviors not captured by these models. In this paper, we compare sampling-based process-level models (i.e., particle filters) to representative associative and causal models (i.e., augmented Rescorla-Wagner and PowerPC) in their ability to capture individual learning behavior. We use likelihood-free inference incorporating machine-learned summary statistics for model estimation. We conduct a simulation study to demonstrate high model identifiability and test the models on an existing dataset and a newly conducted experiment which replicates and extends previous studies. We find that most participants are best explained by a particle filtering account, but more targeted experimental designs are required to estimate the best-fitting sub-type of these particle filter models.

Keywords: causal learning; Bayesian model; conditioning; process models; associative learning

Introduction

How relationships between stimuli are learned has been studied extensively since the early days of psychology and neuroscience (e.g. Pavlov, 1927). The stimulus-stimulus pairing paradigm, where two stimuli are repeatedly presented together until an association between the two is learned, has been used extensively to study learning (Shanks, 1995). The discovery of several learning phenomena has led to the development of diverse computational models that can explain them, which in turn has facilitated the discovery of new phenomena to challenge them, leading to the development of more powerful models. For example, Van Hamme and Wasserman (1994) augmented the Rescorla-Wagner model (Rescorla & Wagner, 1972) to account for backward blocking phenomena.

Another strand of research has studied the learning of these relationships from an explicitly causal perspective. While causal relationships may be learned from associative data, they rely on stronger tacit assumptions, e.g., about interventions. For example, the PowerPC model (Cheng & Novick, 1992) proposes that people learn the strength of a causal relationship as the probability of the potential cause, in the absence of all other causes, producing the effect (Danks, Griffiths, & Tenenbaum, 2002).

Finally, a recent contribution has been the emergence of sampling-based process models (Sanborn, Griffiths, & Navarro, 2010), which explain human variability as arising from stochastic approximations to optimal solutions, with departures from optimality resulting from limited cognitive resources or computation time (Abbott & Griffiths, 2011; Sanborn & Chater, 2016). Although these literatures all study how people learn associations between stimuli, they have rarely interacted (but see, e.g, Danks et al., 2002; Danks & Schwartz, 2005; Beckers, Miller, De Houwer, & Urushihara, 2006; Johnston, Hillman, & Danks, 2021, for exceptions).

One limitation of much of previous research is that, typically, models have been evaluated on the average behavior of the studied sample, which unfortunately is not always a good representation of individual learning behavior. While the average learning behavior shows smooth (and late) adaptation to newly observed data, individual learning trajectories suggest distinctly pronounced jumps in beliefs and considerable variability (Daw & Courville, 2007; Johnston et al., 2021). The issue of intra- and inter-individual variability extends beyond stimulus-stimulus learning, and is in fact ubiquitous in the study of human behavior (Kanai & Rees, 2011; Rieskamp, Busemeyer, & Mellers, 2006).

Modeling efforts are also complicated by methodological issues around model comparison. In particular, algorithmic-level accounts (Griffiths, Lieder, & Goodman, 2015; Marr, 1982) like sampling-based process models (e.g., Sanborn et al., 2010) are often impossible to evaluate using traditional likelihood-based tools, as their likelihood functions tend to be intractable.[1] In addition, we often do not know how to hand-craft summary statistics of potentially high-dimensional response data to distinguish between models or learn their parameter values (Valentin et al., 2024). Today, advancements in methodology and the availability of compute power make it possible to analyze more complex and realistic models alongside more traditional accounts under a common framework of likelihood-free inference (Lintusaari, Gutmann, Dutta, Kaski, & Corander, 2017; Cranmer, Brehmer, & Louppe, 2020).

In this paper, we contrast sampling-based process-level models with representative associative and causal learning models, focusing on their ability to capture individual learning behavior. We reanalyze a study asking people to judge the causal strength of a stimulus at repeated observations (Danks & Schwartz, 2005), data that we extend with our own con-

[1]That is, the likelihood function is either unavailable or very computationally expensive to calculate.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

ceptual replication. We utilize likelihood-free inference for model evaluation to allow for rigorous empirical validation, using a simulation study to evaluate model identifiability before applying the models to the two datasets. Our findings provide a more detailed view of learning behavior, suggesting that the largest proportion of participants align best with a particle filtering account. Meanwhile, more targeted empirical work is required to estimate the precise learning mechanisms as described by the particle filter model.

The Tasks

In each task, participants observed a series of cause-effect pairs and had to estimate how strongly they believed the cause generated or prevented the effect, considering all the data observed thus far in the series. The true contingencies were either non-causal (the presence of the cause does not influence the probability of the effect), generative or preventative (the presence of the cause increases or decreases the probability of the effect, respectively). Crucially, except when contingencies were non-causal, these changed at the midpoint of the series. Thus there were three types of series: some had a first-generative-then-preventative cause (Gen/Prev), or vice versa (Prev/Gen), or were non-causal throughout (NC/NC).

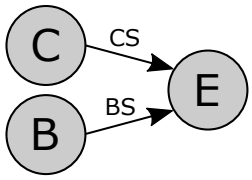
We consider this task an inference problem where participants use repeated observations of the presence or absence of the cause (C or ¬C) and the effect (E or ¬E) to infer causal strength (CS, ranging from -1 to 1; see Figure 1)[2]. Causal strength is positive if the cause is generative, negative if it is preventative, and zero if it is non-causal. The effect can also occur in absence of the cause, here formalized as the influence of background causes (B). The background strength (BS) ranges from 0 to 1, and represents the probability of the effect occurring in the absence of the cause (i.e. $BS = p(E|\neg C)$). Following Danks et al. (2002), we define the conditional probability $P(E|C)$ using a noisy-OR function when CS ≥ 0 (i.e. $p(E|C) = BS + CS - BS \cdot CS$), which encodes the assumption that both C and B have independent opportunities to produce E. For preventative causes, we use a noisy-AND-NOT function (i.e. $p(E|C) = BS - |CS| \cdot BS$), which encodes the assumption that B can produce E with probability BS, and independently C may prevent E with probability $|CS|$.

Because participants reported CS estimates in a $[-100, 100]$ range (see below), we scale simulated estimates to this range throughout.

Exp. 1: Reanalysis of Danks and Schwartz (2005)

51 participants took part in Danks and Schwartz (2005, henceforth DS05). In their experiment, participants acted as doctors researching the relationships between native plants and skin diseases found on foreign islands. In each block of trials (each block a different island), participants interviewed varying numbers of villagers (each villager a trial) who may or not have been exposed to a plant (C or ¬C) and may or not have a skin rash (E or ¬E). After each observed case, participants were asked "How much does the plant cause the rash?", and responded using a slider ranging from -100 (the plant "always prevented" the rash) to +100 (the plant "always caused" the rash), with 0 indicating no relationship between the two. The numeric value for the slider was set to 0 after each rating.

Figure 1: Directed causal graph showing the (B)ackground, (C)ause and the (E)ffect, with edges CS and BS representing the causal and background strength respectively.



Each participant completed six blocks of trials, with fixed lengths of 8, 80, 8, 32, 16, and 48 trials (presented in this order). The pairing of sequence types (Gen/Prev, Prev/Gen, or NC/NC) and lengths was random, with the constraint that participants saw every type once before any type was repeated. Contingencies were $p(E|C) = .75$ and $p(E|\neg C) = .25$ for the generative case; $p(E|C) = .25$ and $p(E|\neg C) = .75$ for the preventative case; and $p(E|C) = p(E|\neg C) = .5$ when non-causal. $p(C)$ was always .5.

Sequences participants saw were created from 6 prefabricated sequences, two each for generative, preventative and non-causal contingencies, which were then merged as appropriate (e.g. a Gen/Prev sequence of length 32 would be created by merging the first 16 items of the `gen1` sequence with the first 16 items of the `prev2` sequence).

Exp. 2: Replication and extension

Participants We recruited 186 participants from Prolific. Based on pre-registered exclusion criteria, 85 participants (45.7%) were excluded from analysis: one participant who messaged to report that they had misunderstood the instructions, one who clicked an invisible button designed to detect bots, and 83 who made 3 or more attempts in the comprehension check following instructions[3]. The sample after exclusion was N=101 (64 female, 35 male; age M = 36.4, SD = 12.2). Compensation was £5. On average, participants completed the experiment in 24.7 minutes.

Method We made the following modifications to the DS05 design: First, each block had a fixed length of 40. Second, because we were interested in within-participant variability for the same sequence, our participants saw one block of cases for each sequence type, in a random order, and then a fourth block which was identical to the first block they had experienced. Third, we varied the cover story participants received, which was either that of a plant causing a possible rash (as in

[2] We explicitly consider the task of learning causal strength, but note that this could be combined with questions about the existence of causal relationships (Griffiths & Tenenbaum, 2005)

[3] A more lenient exclusion criterion, allowing up to 4 errors before exclusion, reduces the exclusion rate to 20.43% and does not change results: model allocation follows similar proportions to the main analysis and cover stories also do not influence judgments in the main task. We show this in the supplemental information (SI), `https://osf.io/tkjmu/`

DS05) or a chemical causing bacterial growth, medicine causing sleep, or medicine causing seizures, randomly chosen for each participant.

We also included other related tasks after these four "experiential" blocks. First, participants carried out three "descriptive" blocks, where they were given a tally of all four possible events and they had to issue a single causal judgment (we do not discuss this data here). Finally, in three "prior belief" blocks participants had to give causal strength judgments for the cover stories they had not been allocated to during the rest of the experiment, without observing any events. The experiment can be accessed at https://pmcl.netlify.app.

Computational models

Augmented Rescorla-Wagner Model

The Rescorla-Wagner model (RW; Rescorla & Wagner, 1972), one of the earliest computational models of associative learning, provides a foundational framework for our comparison. The RW model posits that learning is driven by the discrepancy between expected and actual outcomes, known as the *prediction error*. However, this model assumes a constant learning rate and does not account for certain learning effects like forward/backward blocking or overshadowing.

The Augmented Rescorla-Wagner (ARW; Van Hamme & Wasserman, 1994) model is an extension of the classic Rescorla-Wagner model. It was developed to better capture a range of empirical phenomena observed in associative learning experiments while maintaining the core principle of the original model, and thus provides a more refined representation of learning processes.

In the ARW model, each stimulus is assigned an associative strength, which represents the learned association between that stimulus and the outcome. These associative strengths are updated over time based on the prediction error. The degree to which the prediction error updates the associative strength is governed by learning rate parameters, with separate learning rates for when the cause is present or absent. Moreover, the ARW model also incorporates an associative strength for the background context ('contextual cue'), which is always present. This allows the model to account for learning about the background context separately from the specific causes. This feature can help capture phenomena such as context-specific learning, where the learned associations are tied to the specific context in which learning occurred.

The ARW model is characterized by the following parameters: α_{01} and α_{00}, which are the learning rates of the observed cause when present and absent, respectively; α_{10}, the learning rate of the always-present background cause; and β, the salience parameter. In each trial t, the causal strength estimate $\hat{CS}$ is updated as $\hat{CS}_t = \alpha_{01}\beta\varepsilon_t C_t + \alpha_{00}\beta\varepsilon_t(1-C_t)$, and the background strength estimate is updated as $\hat{BS}_t = \alpha_{10}\beta\varepsilon_t$, where ε_t is the prediction error, calculated as $\varepsilon_t = E_t - CS_{t-1}C_t + BS_{t-1}$, with E_t and C_t taking value 1 when effect and cause are present, respectively, at time t, and 0 otherwise. In simulations, we draw α_{01} and α_{10} from a uniform prior $U[0,1]$, and α_{00} from a uniform prior $U[-1,1]$. We set a constant value $\beta = .9$ as the model is unidentifiable otherwise. Initial states BS_0 and CS_0 are set to zero.

PowerPC

The PowerPC model (PPC; Cheng, 1997) builds upon the RW model by integrating the concept of "causal power". We implement the sequential version of this model proposed in (Danks et al., 2002). This model posits that learners estimate a strength parameter within a particular causal structure, altering their judgments based on the accumulation of observational data. The sequential version of the PPC model offers a dynamic perspective, tracking how judgments change over time as more data is observed. This model takes the same parameters (drawn from the same prior) as ARW above, and updates CS and BS the same way. The error ε_t is updated differently, this time as $\varepsilon_t = C(E_t - [BS_{t-1} + CS_{t-1} - BS_{t-1}CS_{t-1}]) + (1-C)(E_t - BS_{t-1})$ when $CS_{t-1} \geq 0$; $\varepsilon_t = C(E_t - [BS_{t-1} - BS_{t-1}CS_{t-1}]) + (1-C)(E_t - BS_{t-1})$ otherwise. Intuitively, this model can be understood as a version of the ARW model, but with a noisy-OR/AND-NOT functional form. As in the ARW model, initial states BS_0 and CS_0 are set to zero.

Particle filter

Particle Filters (PF; see Doucet, Freitas, & Gordon, 2001 for a technical introduction) are Monte Carlo sampling methods, which are commonly used, e.g., to perform Bayesian inference on filtering problems, where the objective is to infer latent states from a series of observations. They have also been suggested as a process model for how people learn stimulus-stimulus relations (e.g. Abbott & Griffiths, 2011) as well as perform other tasks such as categorization (Sanborn et al., 2010).

The PF model maintains a set of particles, each representing a possible state of the world; that is, a potential set of causal and background strengths in our context. At each time step, the particles are updated based on the observed data. In addition, if some particles hold a relatively unlikely estimate, they may be discarded and replaced through a resampling process. This model operates in a sequential and Bayesian manner, capable of handling non-linear and non-Gaussian functional forms, making it a powerful tool for capturing individual learning behavior (we provide pseudo-code for the PF Algorithm in the SI linked in Footnote 3).

Initial beliefs Unlike in ARW and PPC, strength estimates in the PF are distributed among N particles, with a given estimate at time t being computed as a weighted average of the point estimates of each particle. Initially, each particle has weight $1/N$. In our implementation, each particle's initial estimate for CS and BS is drawn from a "sparse and strong" prior (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008), which encodes the following assumptions: that if the cause is generative, then either CS or BS is high, but rarely both; and that if the cause is preventative, then BS will proba-

bly be high and CS will either have small influence or strongly prevent the effect (see Lu et al., 2008, for details).

Dynamic model After each data point, a new set of particles is proposed following a random walk. In our case, following Abbott and Griffiths (2011), both CS_t and BS_t are drawn from a beta distribution with a free parameter λ, with larger values of λ leading to greater similarity between the particles at time $t-1$ and time t:

$$s_t = sgn(s_{t-1})Beta(\lambda|s_{t-1}|+1,\ \lambda-\lambda|s_{t-1}|) \qquad (1)$$

where $sgn(\cdot)$ is the sign function.

Resampling If all initial particles were maintained, only very few would be good explanations of the data after several iterations (with the others having negligible weights). To avoid this, particles are resampled if the effective sample size falls under a given threshold θ. When this happens a new set of particles is obtained by sampling from the current set with replacement, with each particle having probability equal to its weight. Then, weights are reset.

MCMC-Rejuvenation To restore diversity into the set of particles, an additional rejuvenation step is carried out when particles are resampled. The CS and BS estimates are updated by using Markov Chain Monte Carlo (MCMC) with the joint likelihood of all the data observed thus far as the posterior (Chopin, 2002; Abbott & Griffiths, 2011). In our implementation, we chose Metropolis-Hastings as our MCMC algorithm, but others have been used to model human performance elsewhere (Castillo, León-Villagrá, Chater, & Sanborn, 2024; Zhu, León-Villagrá, Chater, & Sanborn, 2022).

Random Responses

We augment ARW, PPC and PF with a stochastic component: random responses. For all models, we include an additional parameter $\varepsilon \sim Beta(0.1,1)$, which encodes the probability of a response at time t being drawn from a uniform distribution $U[-1,1]$, rather than $\hat{CS}_t$ being reported. Crucially, both $\hat{BS}_t$ and $\hat{CS}_t$ are still tracked even if the response is corrupted. We primarily included this additional step because we found that ARW and PPC having $\hat{BS}_0 = \hat{CS}_0 = 0$ meant that it was difficult for these models to explain potential big departures from 0 at the first CS judgment.

Model evaluation

Previous accounts of causal learning or associative learning have focused on tractable models with known likelihoods, or employed more qualitative ways of model comparison (like visualization techniques) for more complex models. A quantitative comparison involving models with intractable likelihoods, such as PF, requires that we use likelihood-free inference.

Likelihood-free inference

We adopted a simulation-based perspective called Approximate Bayesian Computation (ABC; Palestro, Sederberg, Osth, Van Zandt, & Turner, 2018). In ABC, generative models $\mathbf{m}$ are used to sample synthetic data $\mathbf{y}$, given values of their model parameters $\boldsymbol{\theta}$ and the experimental designs $\mathbf{d}$ (i.e. the sequences participants observed). For each sequence of judgments we obtained summary statistics $S(\cdot)$ (see below for which summary statistics were used). For each participant and sequence, we compared the summary statistics of the observed data ($S(\mathbf{x})$) to all of the simulated data ($S(\mathbf{y})$) from the same sequence, using the Euclidean distance ρ (after normalizing the summary statistics). Model simulations whose summary statistics had a distance to the observed data higher than a pre-defined tolerance τ were discarded. The posterior probability of a model $\mathbf{m}$ is the proportion of simulated data points generated by that model in the region defined by the tolerance, i.e. the relative frequency of data generated by $\mathbf{m}$ out of all data for which $\rho \leq \tau$. For convenience, here we defined pseudo-tolerances τ^* which establish the proportion of nearest samples kept after discarding for each participant/sequence pair (Biau, Cérou, & Guyader, 2015).

Summary statistics

A key issue in performing likelihood-free inference is the choice of summary statistics. These are functions of the data that capture relevant information about the models $\mathbf{m}$ and parameters $\boldsymbol{\theta}$. While typically these summary statistics are selected by hand, we aimed to maximize the information about $\mathbf{m}$ and $\boldsymbol{\theta}$ by adding learned summary statistics as well.

Hand-crafted For a given sequence of causes, effects and judgments a participant or model produced, we calculated: the first judgment, the judgment before the midpoint of the sequence, the judgment after the midpoint, the final judgment, the minimum, maximum, mean, variance, and autocorrelation at lags one to five. This resulted in a rich set of summary statistics that capture different aspects of the data. The choice of these was guided by their potential relevance to $\mathbf{m}$ and $\boldsymbol{\theta}$, but it is ultimately an empirical question which summary statistics are most useful for inference.

Learned ABC is typically carried out with only hand-crafted summary statistics. However, as we could not guarantee that these were sufficient[4], we also obtained learned summary statistics by using feed-forward neural networks (Chen, Zhang, Gutmann, Courville, & Zhu, 2021). These networks were trained on simulated data y sampled from the prior (over models and their parameters) to predict which generating model $\mathbf{m}$ produced the data. We obtained 6 learned features this way, which were the values of the last layer of the neural network prior to the softmax layer. These learned statistics are complex, being based on the entire sequence of responses, and so we do not attempt to show them here.

[4]Summary statistics are considered *sufficient* for a set of parameters $\boldsymbol{\theta}$ if they capture all the information in the data that is relevant to the estimation of $\boldsymbol{\theta}$. Formally, a statistic $T(\mathbf{y})$ is sufficient for $\boldsymbol{\theta}$ if the conditional probability distribution $P(\mathbf{y}|T(\mathbf{y}),\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$. In other words, once the summary statistics are known, knowing the full data does not provide any additional information about $\boldsymbol{\theta}$.

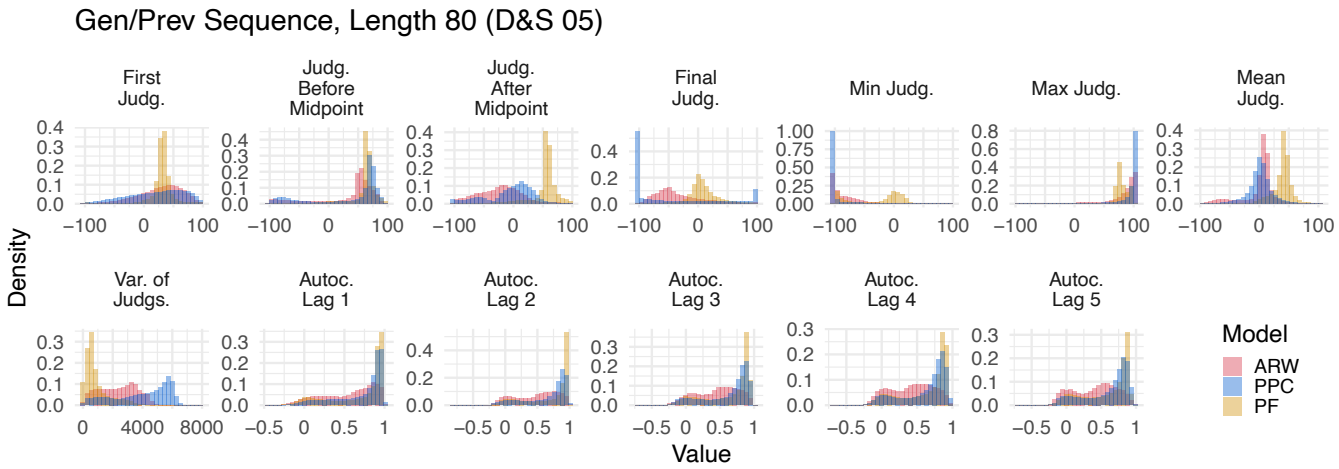


Figure 2: Prior predictive histograms for hand-crafted summary statistics in a generative-preventative sequence of length 80. Throughout, the three models here proposed exhibit different behaviours for these measures. Plots for other sequences are available in the SI.

Simulation study

Prior predictive simulations

We generated 10^4 prior simulations per model per sequence and evaluated the summary statistics for these, in order to show the different qualitative patterns the models display. For succinctness, we plot the results of the hand-crafted summary statistics of one sequence here (Figure 2), but equivalent plots for other sequences can be found in the SI (see Footnote 3).

Model identifiability

To ensure that our model evaluation procedure was accurate and unbiased, we performed 10^3 cross validation (cv) steps per model per sequence. In each cv step, one simulated data point $\mathbf{y}$ is removed from the set of simulated data, and treated as a pseudo-observed data point $\mathbf{x}^*$. The inference procedure is carried out as normal, which allows us to evaluate its accuracy. We tested this for a range of pseudo-tolerances $\tau^* \in \{.05, .1, .15, .2, .25\}$.

We found that, for the tolerance that performed best (the smallest), accuracy scores were generally very good, with no model/sequence combination falling below 70.3% accuracy, and with average accuracy 87.3%. Using mixed-effects logistic regression, we found that that accuracy increased with sequence length, and was lower when the true model was PPC (Figure 3; see SI for model results). When model allocation erred, ARW was the wrongly selected model 70% of the time (when the true model was ARW, PPC and PF were selected 55% and 45% of the time respectively). Finally, we tested whether adding learned summary statistics had improved performance, by comparing these to an additional set of cv results obtained without the learned summary statistics. The average accuracy with only hand-crafted summary statistics was overwhelmingly lower (73.6%, $t(87) = 10.96$, $p < .001$, $BF_{10} > 10^8$). From these results, we decided to carry out inference with $\tau^* = .05$ and using hand-crafted and learned summary statistics. We note that results might be less reliable for short sequences.

Results

Experiment 1

We found that 70.6% ($n = 36$) participants were best-explained by the PF model, 17.6% ($n = 9$) by the ARW model and 11.8% ($n = 6$) by the PPC model (see Figure 4 for allocation of individual sequences). As expected by our model identifiability results, the certainty of the prediction increased with sequence length: the shorter the sequence, the greater the standard deviation of posterior probabilities given to the three models ($t(251) = 8.03$; $p < .001$; $BF_{10} > 10^6$). For the three longest sequences, 32 (62.7%) participants were allocated the same model in all occasions, well above the value expected by chance (11.1%; Binomial Test $p < .001$, $BF_{10} > 10^{14}$.

Experiment 2

Effects of Cover Story In the "prior belief" block, different cover stories had different expected causal strengths: In this block the average judgments were 5.1, 30.4, -12.6, and 15.2 for the bacteria, plant rash, seizure, and sleep contexts respectively, which were credibly different ($F(3, 244.13) = 26.84$, $p < .001$, $BF_{10} > 10^{13}$). In spite of this, we found no evidence for an effect of cover story on judgments in the main task (all $ps > .05$, all $BF_{10} < 1/3$). Cover story also did not influence the hand-crafted summary statistics of a sequence (all $ps \geq .10$, all $BF_{10} < 1/1.5$).

Model discrimination We found that 70.3% ($n = 71$) participants were best-explained by the PF model, 21.8% ($n = 22$) by the ARW model and 7.9% ($n = 8$) by the PPC model. These proportions did not depend on cover story ($\chi^2(3) = 0.99$, $p = .80$). 35 (34.65%) participants were allocated the same model in all occasions, well above the value expected by chance (3.7%; Binomial Test $p < .001$, $BF_{10} > 10^{20}$). Analyzing only the first and last block, where participants

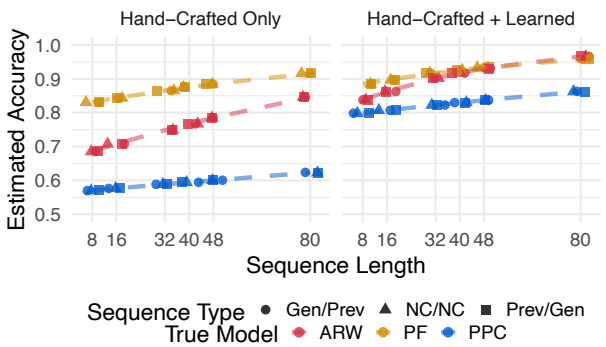


Figure 3: Estimated Model recovery accuracy for $\tau^* = .05$ (dashed lines) as a function of sequence length and true model, for sequences used here and in DS05. In the left panel ABC is performed with hand-crafted summary statistics only, while in the right panel the whole set is used. Each dot is a sequence, with shape indicating sequence type. Accuracy increases with sequence length, and PPC accuracy becomes worse more quickly when the sequence is short. Note that the y axis starts at .5, and not 0.

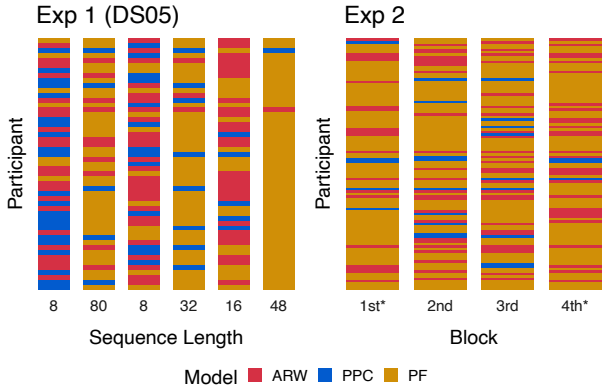


Figure 4: Model allocation results. Each row represents a different participant and each column a different sequence (in order of presentation). Note that the 4th block of Exp. 2 was the same sequence as the 1st block. Note also that results from shorter sequences are less reliable (Figure 3).

had the same sequence, showed that 75 (74.26%) participants were allocated the same model, again well above the value expected by chance (33.33%; Binomial Test $p < .001$, $BF_{10} > 10^{13}$).

Discussion

In the present study, we investigated how people learn the causal relationships between stimuli, comparing models from different research traditions that are rarely tested head-to-head. To do so, we employed a powerful method to evaluate models that lack tractable likelihood functions: we complemented Approximate Bayesian Computation with summary statistics learned by a neural network. In addition to providing tractable approximate likelihoods, this approach greatly increased predictive accuracy, as shown by our model evaluation results. This approach has the additional advantage of combining summary statistics that can highly discriminate between models (learned) with more interpretable summary statistics (hand-crafted). Crucially, our analyses were based on individual learning trajectories, allowing for insights into inter- and intra-individual differences. Across two datasets, we found a particle filtering account offered a better fit to the data than other models. Responding to a challenge identified in prior work (Johnston et al., 2021), our model comparisons inherently penalize model complexity through Bayesian Ockham's razor.

Our results showcased that individual participants were remarkably stable in terms of the model that best explained their judgments across trials, with a considerable proportion of participants being best fit by the same model every block. This stability highlights that our likelihood-free inference methods, combined with ample data from individual participants, can distinguish systematic variability from noise. Future work, covering different tasks and theories, may benefit from such an approach to more systematically compare models and investigate individual differences.

Our study focused only on discriminating which model best explains a participant's data. There are psychologically interpretable parameters in our models and further work is needed to identify the best-fitting parameters for every person, and the stability of those within a participant. Different parameter values can reveal qualitatively different behaviors: for example, different learning rates in ARW and PPC can make these models showcase a primacy or recency bias, while for the PF changes in the drift rate parameter might capture greater or lesser "forgetfulness" as well as robustness to environmental change, or the MCMC rejuvenation could consider the likelihood of fewer data points, consistent with more limited memory capacity. Future work should examine these subfamilies of models and whether participants are self-consistent when fit into these subtypes. Future work may also add optimal models to the model comparison: e.g., a model using the maximum *a posteriori* probability estimate using Lu et al.'s (2008) strong-and-sparse priors; may add similar existing datasets to the analysis (e.g. Danks & Schwartz, 2006). Future research may also generalize these process models to how participants learn the structure of causal graphs (Griffiths & Tenenbaum, 2005; Bramley, Dayan, Griffiths, & Lagnado, 2017) and the functional form (Lucas & Griffiths, 2010) that connects causal variables.

In summary, we have used a powerful likelihood-free inference approach to compare models in a robust and systematic way that would be intractable using traditional models. This evaluation reveals that particle filter models account well for individual-level patterns of belief-updating, and individual participants are stable in their behavior as captured by our models. Our framework shows promise for distinguishing signature patterns in people's behavior from noise in model assignment across individuals.

Acknowledgments

We thank David Danks and Laila Johnston for providing data, David Danks for valuable feedback on our early experimental designs, and Peter Dayan and David Shanks for valuable comments on plausible models of associative belief updating. Any errors are our own. This work was supported by a European Research Council grant (817492-SAMPLING).

References

Abbott, J. T., & Griffiths, T. L. (2011). Exploring the Influence of Particle Filter Parameters on Order Effects in Causal Learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, pp. 2950–2955).

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: Forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology: General*, *135*(1), 92. doi: 10.1037/0096-3445.135.1.92

Biau, G., Cérou, F., & Guyader, A. (2015). New insights into Approximate Bayesian Computation. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, *51*(1), 376–403. doi: 10.1214/13-AIHP590

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338. doi: 10.1037/rev0000061

Castillo, L., León-Villagrá, P., Chater, N., & Sanborn, A. (2024). Explaining the flaws in human random generation as local sampling with momentum. *PLOS Computational Biology*, *20*(1), e1011739. doi: 10.1371/journal.pcbi.1011739

Chen, Y., Zhang, D., Gutmann, M. U., Courville, A., & Zhu, Z. (2021). Neural approximate sufficient statistics for implicit models. In *International conference on learning representations (ICLR)*.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405. doi: 10.1037/0033-295X.104.2.367

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*(2), 365–382. doi: 10.1037/0033-295X.99.2.365

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, *89*(3), 539–551. doi: 10.1093/biomet/89.3.539

Cranmer, K., Brehmer, J., & Louppe, G. (2020). The frontier of simulation-based inference. In *Proceedings of the National Academy of Sciences* (Vol. 117, pp. 30055–30062). doi: 10.1073/pnas.1912789117

Danks, D., Griffiths, T., & Tenenbaum, J. (2002). Dynamical causal learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems* (Vol. 15). MIT Press.

Danks, D., & Schwartz, S. (2005). Causal Learning from Biased Sequences. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 27).

Danks, D., & Schwartz, S. (2006). Effects of Causal Strength on Learning from Biased Sequences. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *28*, 1180–1185.

Daw, N., & Courville, A. (2007). The pigeon as particle filter. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems* (Vol. 20, pp. 369–376). MIT Press.

Doucet, A., Freitas, N., & Gordon, N. (Eds.). (2001). *Sequential Monte Carlo Methods in Practice*. New York, NY: Springer New York. doi: 10.1007/978-1-4757-3437-9

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational Use of Cognitive Resources: Levels of Analysis Between the Computational and the Algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229. doi: 10.1111/tops.12142

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384. doi: 10.1016/j.cogpsych.2005.05.004

Johnston, L., Hillman, N., & Danks, D. (2021). Individual Differences in Causal Learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, pp. 1935–1941).

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience*, *12*(4), 231–242. doi: 10.1038/nrn3000

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. (2017). Fundamentals and recent developments in approximate Bayesian computation. *Systematic biology*, *66*(1), e66–e82. doi: 10.1093/sysbio/syw077

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955–984. doi: 10.1037/a0013256

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–147. doi: 10.1111/j.1551-6709.2009.01058.x

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman.

Palestro, J. J., Sederberg, P. B., Osth, A. F., Van Zandt, T., & Turner, B. M. (2018). *Likelihood-Free Methods for Cognitive Science*. Cham: Springer International Publishing. doi: 10.1007/978-3-319-72425-6

Pavlov, I. P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. Oxford University Press.

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Theory and Research* (pp. 64–99). New York: Appelton-Century-Crofts.

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the Bounds of Rationality: Evidence and Theories of Preferential Choice. *Journal of Economic Literature*, *44*(3), 631–661. doi: 10.1257/jel.44.3.631

Sanborn, A. N., & Chater, N. (2016). Bayesian Brains without Probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. doi: 10.1016/j.tics.2016.10.003

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. doi: 10.1037/a0020511

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511623288

Valentin, S., Kleinegesse, S., Bramley, N. R., Seriès, P., Gutmann, M. U., & Lucas, C. G. (2024). Designing optimal behavioral experiments using machine learning. *eLife*, *13*, e86224. doi: 10.7554/eLife.86224

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and motivation*, *25*(2), 127–151. doi: 10.1006/lmot.1994.1008

Zhu, J.-Q., León-Villagrá, P., Chater, N., & Sanborn, A. N. (2022). Understanding the structure of cognitive noise. *PLoS Computational Biology*, *18*(8), e1010312. doi: 10.1371/journal.pcbi.1010312