Learning Type-Based Compositional Causal Rules

Feng Cheng (fc1367@NYU.edu)
Department of Psychology, 6 Washington Pl
New York, NY 10003 USA

Bob Rehder (bob.rehder@nyu.edu)
Department of Psychology, 6 Washington Pl
New York, NY 10003 USA

Abstract

Humans possess knowledge of causal systems with deep compositional structures. For example, we know that a good soccer team needs players to fill different roles, with each role demanding a configuration of skills from the player. These causal systems operate on multiple *object types* (player roles) that are defined by features within objects (skills). This study explores how human learners perform on novel causal learning problems in which they need to infer multiple object types in a bottom-up manner, using empirical information as a cue for their existence. We model subjects' learning process with Bayesian models, drawing hypotheses from different spaces of logical expressions. We found that although subjects exhibited partial success on tasks that required learning one object type, they mostly failed at those that required learning multiple types. Our result identifies the learning of object types as a major obstacle for human acquisition of complex causal systems.

Keywords: causal cognition, causal learning, rule learning, language of thought

Introduction

Imagine you are a billionaire with a recently acquired passion for soccer. You pour your wealth into a soccer team but you know nothing about soccer. Your manager sends you profiles of potential players, each characterized by a number of skills (e.g., speed, stamina, shooting, passing, etc.). You pick players haphazardly—some are good at shooting and defending but slow, some are good at passing and dribbling but lack stamina—and jumble them into a team. Your team loses hopelessly. But this experience teaches you that rather than being independent, skills combine in useful ways to produce different *types* of players. For example, you realize that players that are fast and shoot accurately make for good *attackers*, those with stamina and that pass accurately make for good *midfielders*, and so forth. Moreover, you realize that success requires a mix of the right type of players. With this knowledge, you now create a winning soccer team.

The soccer team example a complex causal system that exhibits compositional structure at two levels. At the *object level*, objects are organized into types defined by multiple features (e.g., in soccer, any player with the right skills can be an attacker, etc.). At the *set level*, multiple object types interact to produce to a particular effect (e.g., a winning soccer team). Such systems—which we refer to as *typed compositional causal systems*, are ubiquitous in our knowledge systems. For example, the functioning of a mammal's body relies on the cooperation of several types of organs, each defined by their cell functions and organizations. The products of complex chemical processes are generally determined by the types of compounds involved, which are defined by the substructures of basic chemical elements. We argue that these systems underlie many of the complex reasoning processes that characterize human cognition.

Despite their prevalence, typed compositional causal systems have received little attention in the literature. A typical causal learning experiment investigates how subjects infer a single univariate cause from an outcome, such as whether a drug eliminates a headache or whether a block placed on a machine causes it to light up (e.g., Buehner et al. 2003; Sobel et al. 2004). Whereas such paradigms of course involve "types" (taking the drug is a type of event), here we focus on multivariate object types as the primitive component of causal rules. The influence of (and potential interactions between) multiple causes has been investigated (e.g., how plant growth is influenced by both light and fertilizer; Kemp et al. 2010, Expt 4; Lucas & Griffiths 2010; Novick & Cheng 2004; Spellman 1996; Waldmann 2007; Lucas et al. 2014), but again with univariate causes. Research has investigated causal learning that involves inducing new categories of causes, either when the new category is univariate (e.g., Lien & Cheng 2000; Marsh & Ahn 2009) or multivariate (e.g., Bramley et al. 2018; Kemp et al. 2010, Expt 3; Gopnik & Sobel 2000; Waldmann & Hagmayer 2006; Zhao et al. 2022), but subjects in these studies need only learn one new causal type. Finally, the study of Bye et al. (2023) may be most similar to ours as they considered how multiple variables could be treated as "whole causes," which in turn might be combined to yield an effect. However, in their study (a) the constituents of whole causes are not properties of objects and thus do not define object types, (b) whole causes were only assumed to combine disjunctively whereas we consider interacting object types (a soccer team needs attackers *and* midfielders), and (c) subjects only predicted how whole causes combine during a generalization phase, not learning itself. Thus, typed compositional causal systems as we've defined them are a largely unexplored learning problem.

Current Study

This study is a preliminary investigation of how human learners acquire typed compositional causal systems from experi-

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.*

Table 1: A list of rule families for the Bayesian model

Names	Description	Example Rule
Type-1	Rules with 1 object type that specifies 2-3 features.	$\exists o_x[Cl(o_x) = r \wedge Sh(o_x) = c \wedge Sz(o) = l]$
Type-2	Rules with 2 distinct object types that each specify 2-3 features.	$\exists o_x[Cl(o_x) = r] \wedge \exists o_y[Sh(o_y) = c \wedge Sz(o_y) = l] \wedge o_x \neq o_y$
Feature	Rules that specify exactly 1 feature.	$\exists o[Cl(o) = r]$
Conjunct*	Rules that specify a conjunction of (up to 6) features.	$\exists o[Cl(o) = r] \wedge \exists o[Sh(o) = c]$
Disjunct*	Rules that specify a disjunction of (up to 6) features.	$\exists o[Cl(o) = r] \vee \exists o[Sh(o) = c]$
Counting conjunct*	Rules that specify a conjunction of exact feature counts (up to 6).	$\exists o^{=1}[Cl(o) = r] \wedge \exists o^{=1}[Sh(o) = c]$
Prototype*	Probabilistic versions of the counting conjunct rules.	$\text{Prob}(\exists o^{=1}[Cl(o) = r])$

*Unlike Type-1 and Type-2, these rule families are insensitive to whether the features appear in the same or different objects.

ence. Our first question is whether subjects can learn a typed compositional causal rule at all. Assuming they can, our second question concerns the factors determining their acquisition difficulty. One obvious possibility is that difficulty will increase with the number of to-be-learned object types. This is so because the two-level compositional hierarchy (object types, and then how they combine) exponentiates the space of possible rules. We also ask whether learning difficulty varies with a rule's number of feature specifications, regardless of whether they are associated with the same or different objects. More feature specifications might hurt learning because they result in rules that are more complex and so less likely to be sampled (Feldman 2003).

To this end, we presented subjects with stimuli consisting of two objects, each with three binary dimensions: $Cl(\text{Color}) = \{\text{red}, \text{blue}\}$, $Sh(\text{Shape}) = \{\text{circle}, \text{square}\}$, $Sz(\text{Size}) = \{\text{large}, \text{small}\}$. Given these stimuli, we defined two families of typed causal rules, one that requires learning one object type (Type-1) and another that requires learning two (Type-2). To return to our soccer team example, Type-1 and Type-2 rules are analogous to constructing a team in a sport with one or two types of players, respectively.

For example, a rule in the Type-1 family is:

$$\texttt{ERC} \quad \equiv \quad \exists o_x[Cl(o_x) = \text{red} \wedge Sh(o_x) = \text{circle}] \tag{1}$$

Here we use "ERC" as a compact label for a rule that states that the effect occurs if within the stimulus there exists an object that is a red circle. A rule in the Type-2 family is:

$$\texttt{ER\^{}EC} \quad \equiv \quad \exists o_x[Cl(o_x) = \text{red}] \wedge \exists o_y[Sh(o_y) = \text{circle}] \wedge o_x \neq o_y \tag{2}$$

ER^EC states that the effect occurs if there exists an object that is red and there exists a different object that is a circle. Importantly, whereas ERC stipulates that the features red and circle must appear in the same object, ER^EC stipulates that they must appear in different objects.

We also tested a pair of rules ERLC and ER^ELC that, like ERC and ER^EC, differ on how the features are distributed over the objects but that specified three rather than two features:

$$\begin{aligned} \texttt{ERLC} \quad &\equiv \quad \exists o_x[C_l(o_x) = \text{red} \wedge S_h(o_x) = \text{circle} \wedge \\ &\qquad S_z(o_x) = \text{large}] \\ \texttt{ER\^{}ELC} \quad &\equiv \quad \exists o_x[C_l(o_x) = \text{red}] \wedge \\ &\qquad \exists o_y[S_h(o_y) = \text{circle} \wedge S_z(o_y) = \text{large}] \wedge \\ &\qquad o_x \neq o_y \end{aligned} \tag{3}$$

ERLC states that the effect occurs if there exists a large red circle and ER^ELC states it does so if one object is red and the other is a large circle.

Together, these four experimental conditions, which were manipulated between subjects, allow us to test the effect of needing to learn one versus two object types (conditions ERC and ERLC vs. ER^EC and ER^ELC) and two versus three feature specifications (ERC and ER^EC vs. ERLC and ER^ELC). Subjects were asked to infer the ground-truth rule by predicting the effects of a series of two-object combinations as stimuli.

Although subjects' overall prediction accuracy will serve as one dependent variable, we note that high accuracy does not necessarily imply that the ground truth rule was learned. This is because learners might settle for suboptimal yet high-performing rules that approximate the ground truth. Thus, in addition to the Type-1 and Type-2 families we considered a variety of simpler rule families that do not contain object types (see Table 1). In particular, the Feature, Conjunct, and Disjunct families are rule spaces that are known to be easily accessible to human learners (Feldman 2000; Goodman et al. 2008; Haygood & Bourne Jr 1965; Piantadosi et al. 2016). The Counting Conjunct family accounts for the possibility that subjects may form a rule based on the exact number

of features (e.g., there is exactly one red object and exactly one circle; Kemp et al. 2008). Finally, the Prototype family is a probabilistic version of the Counting Conjunct family that treats a counting conjunct rule as an ideal or prototypical cause; the probability of the effect occurring is a linear function of the number of count mismatches. We will compare the performance of these rule families with that of the human subjects to help identify which rule they induced during training.

Methods

Participants

148 New York University undergraduates participated for course credit or monetary compensation.

Materials & Design

As mentioned, our stimuli are pairs of objects characterized by three binary dimensions. Because we allow object repetition within a stimulus, there are eight unique objects and 36 unique two-object combinations (ignoring order). Aside from ERC, ERLC, ERˆEC, and ERˆELC, as a control we also tested a simple single feature rule to verify that subjects understood the task and were learning during the experiment:

$$\texttt{R} \quad \equiv \quad \exists o[C_l(o) = \text{red}] \tag{4}$$

Prior research suggests that R, which state that the effect occurs if at least one of the objects is red, should be very easy to learn (Haygood & Bourne Jr 1965; Shepard et al. 1961). To ascertain that subjects' performance cannot be attributed to the salience of particular features, we also flip the feature specification of the ground-truth rules for half of the subjects (e.g., the R rule became $\exists o[C_l(o) = \text{blue}]$). (Despite this counterbalancing of features, we will continue to refer to the rules as R, ERC, etc.) This results in five between-subject conditions presented in two stimuli lists.

Procedure

Subjects were randomly assigned to one of the five rule conditions presented in one of the two stimuli lists. Each trial, presented two objects and subjects predicted whether the effect (an explosion) would occur. During training, subjects made predictions for 160 trials and received immediate feedback after each. The base rate of the effect was constant under all conditions (40%), with each unique stimulus (i.e., pair of objects) presented at least twice. In the testing phase, subjects made predictions for all 36 unique stimuli without feedback. The presentation order of trials was randomized for each subject. Subjects were also tested with 15 generalization trials in which three objects are presented as stimuli; their confidence ratings and verbal descriptions of the learned rule were also collected. Due to space limitations we omit presentation of these additional measures.

At the start of the experiment subjects were told that the effect is a deterministic function of the object pairs and that the order of the objects is irrelevant. They were also informed that the ground truth rule will remain constant. Subjects were not allowed to take notes and were instructed not to rely on memorization to solve the task.

Results

Model-Free Results

There were no significant effects of the stimuli lists so the results are collapsed over this factor. Subjects' average accuracy at predicting the outcome in the test phase is presented in Fig. 1 for each condition. Performance was above chance (.50) on all rules, ps $< 10^{-4}$. Moreover, accuracy on rule R was greater than the other four rules combined, $p < 10^{-7}$, a result that was expected given its simplicity.

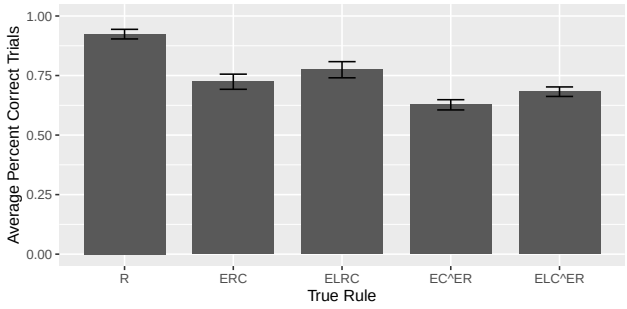

Figure 1: Mean test accuracy in the five rule conditions. Error bars are standard errors.

Focusing on the remaining four conditions, to ask whether accuracy varied with the type of rule we ran a mixed effect generalized linear model with a binomial (logit) link function and per-subject intercepts. We found effects of the number of both object types, $F(1,144) = 14.8, p < 10^{-4}$, and feature specifications, $F(1,144) = 4.23, p = .034$, and no interaction. Although we expected the finding that the rules with more object types (ERˆEC and ERˆELC) would be harder to learn, it is notable that the rules with more feature specifications (ERLC and ERˆELC) were *easier* to learn, contrary to our hypothesis.

Model-Based Results

The main goal of our model-based analysis is to determine if subjects learned the ground-truth rule and, if not, to gain insight into the kind of rules they learned instead. To this end, we fit the rule families in Table 1 to each subjects' predictions using optimal Bayesian models. Each family defines a hypothesis space of rules of the same form. As mentioned, the Type-1 family includes all rules that have exactly one object type and thus this family will fit the subjects in the ERC and ERLC conditions well if they inferred the correct rule. The same applies to the Type-2 family in the ERˆEC and ERˆELC conditions, and the Feature family in the R conditions. Note that for this reason these rule families have the ceiling accuracy of 1 in their corresponding conditions in Table 2. The remaining families are fitted to the subjects' predictions to determine if they instead learned a conjunctive, disjunctive, counting conjunctive, or prototype rule, even though doing

so would result in suboptimal performance (in Table 2 the ceiling accuracy of these models is always < 1). Although not shown in Table 1, we also fit a baseline model that simply matches the subjects' preference for one response over the other. This model will capture those subjects whose predictions were insensitive to the which objects were presented.

All rules in family $\mathcal{H}$ start with a uniform prior $P(h|\mathcal{H})$. For a stimulus on trial i, denoted s_i, its likelihood for any $h \in \mathcal{H}$ is:

$$p(s_i, O(s_i)|h) = \begin{cases} 1-m & h \text{ matches } s_i, O(s_i) \\ m & \text{otherwise} \end{cases} \quad (5)$$

where $O(s_i)$ indicates whether the effect occurs in trial i and $0 < m < 0.5$ is a free parameter that defines the penalization of mismatch for a rule. Note that, as defined above, the matching function of h in the Prototype family gives a probabilistic response instead of 0 or 1. On trial i the model predicts that the probability of an effect will be:

$$\begin{aligned} &p(O(s_i) = 1|S_{i-1}, O(S_{i-1}), \mathcal{H}) \\ &\quad = \sum_{h \in \mathcal{H}} p(s_i, O(s_i) = 1|h) p(h|S_{i-1}, O(S_{i-1})) \\ &\quad = \alpha \sum_{h \in \mathcal{H}} p(s_i, O(s_i) = 1|h) \times p(h|S_{i-1}, O(S_{i-1})) \\ &\qquad + (1-\alpha) \times \beta \end{aligned} \quad (6)$$

where $0 \leq \alpha, \beta \leq 1$ are free parameters and S_{i-1} is a vector of all past stimuli up through trial $i-1$. α determines the probability that a response is made at random and β represents the baseline preference for a positive prediction. In each trial, the model first makes a prediction about the effect, then receives the feedback and updates its posterior. Each model is ran on the same sequence of training stimuli that the subject received. In the testing phase, the posteriors are fixed and only the model's prediction are recorded. The three free parameters are fitted using a simple grid search followed by optimization that maximized the log-likelihood of the subjects' predictions given $\mathcal{H}$.

The model fitting results are presented in Table 2, in which each cell presents the ceiling accuracy, average BIC, test phase model loss (the absolute difference between the predictions of the subjects and the model), and the number of best-fitted subjects. Starting with the R rule, we found, unsurprisingly, that the Feature family was the best-fitting family for almost all subjects. The small test phase model loss reflects that by the test phase these subjects' judgments were almost entirely consistent with the ground-truth R rule.

In the ERC and ERLC conditions, we also found that the ground truth rule family (in this case, Type-1) achieved the lowest loss and best fitted the majority of subjects, indicating that subjects had considerable success learning these rules. Moreover, subjects perform slightly better ERLC than in ERC, consistent with their difference in accuracy (Fig. 1). Nevertheless, the substantial average loss of the ground-truth model, plus the fact that 14/30 of the ERC subjects and 7/30 ERLC subjects were best fit by an alternative model, indicates that the Type-1 family was not a good description of the judgments of many subjects. Those subjects either learned little or nothing (5/30 subjects were best fit by the baseline model in both conditions) or inferred a suboptimal rule (e.g., three ERC subjects were best fit by a Conjunctive rule and two ERLC subjects were best fit by a Feature rule).

Finally, in the ER^EC and ER^ELC conditions we found that the ground-truth rule family (Type-2) yielded the best fit for only a small portion of the subjects (4/29 and 5/30, respectively). More subjects were better fit by a Conjunctive rule (8/29 and 9/30); specifically, the judgments of the 8 ER^EC subjects were best characterized by the rule R^C (red and circle) and those of the 9 ER^ELC subjects were best characterized by L^R^C (large and red and circle). That is, instead of learning two object types, these subjects induced a presumably more familiar conjunctive rule at the cost of accuracy. Note that the Type-2 family in the ER^ELC condition yielded a slightly lower model loss than the ER^EC condition, and a slightly better BIC than the conjunctive family, results consistent with our model-free analysis (Fig. 1) indicating better performance on ER^ELC versus ER^EC.

These analyses confirmed our first hypothesis that learning two object types is harder than one, as indicated by the fact subjects induced a non-optimal (usually conjunctive) rule in the former conditions. And, similar to our model-free analysis, we found that subjects performed better when the ground-truth has more feature specifications (ERLC and ER^ELC better than ERC and ER^EC). We will return to these results in our discussion.

Stochasticity in Rule Learning

Although the Bayesian model fits presented above provide useful information regarding the rules induced by subjects, we believe they make unrealistic assumptions regarding the learning process. Such models assume that learners maintain a posterior probability distribution over all hypothesized rules, which gets optimally updated after every learning trial. These assumptions imply that the posterior will shift smoothly over time toward the hypothesis that is the best account of the observed data. But as noted by many (e.g., Bonawitz et al. 2014), maintaining a large number of hypotheses is psychologically implausible, lending credence to alternative models that assume that learners track a small number of rules (often one) and stochastically switch to a new rule when required by new evidence. Indeed, our subjects' verbal reports indicated qualitative shifts in their favored hypothesis during the course of learning.

As a preliminary investigation of the dynamics of learning, we carried out a sliding-time window analysis that asked which rule best described a subject's judgment during that window. We took all 196 training and test trials and segmented them into 12 overlapping time windows each consisting of 80 trials where each subsequent window was advanced by 10 trials, resulting in windows 1:80, 11:90, 21:100, ..., 111:196. (The last window consisted of 86 trials to cover

Table 2: Fits of the Bayesian model to subject's training and test data.

Rule Condition	Measurement	Rule Family: Type-1	Type-2	Feature	Conjunction	Disjunction	Counting Conjunction	Prototype
ERC	Ceiling Accuracy	**1**	0.806	0.694	0.889	0.500	0.694	0.579
	Average BIC	**209.1**	252.5	261.1	243.6	290.9	283.6	288.9
	Test Phase Model Loss	**0.316**	0.395	0.419	0.367	0.470	0.451	0.457
	No. of Subjects	**16/30**	0/30	3/30	3/30	1/30	0/30	2/30
ERLC	Ceiling Accuracy	**1**	0.882	0.500	0.833	0.306	0.756	0.556
	Average BIC	**172.5**	236.9	267.0	233.0	290.7	283.6	293.6
	Test Phase Model Loss	**0.255**	0.366	0.431	0.352	0.468	0.451	0.466
	No. of Subjects	**21/30**	1/30	2/30	0/30	0/30	0/30	1/30
ERˆEC	Ceiling Accuracy	0.778	**1**	0.694	0.889	0.500	0.694	0.580
	Average BIC	275.4	274.3	272.3	**270.2**	294.1	291.8	292.8
	Test Phase Model Loss	0.433	0.452	0.433	**0.416**	0.474	0.465	0.464
	No. of Subjects	2/29	4/29	5/29	**8/29**	0/29	0/29	2/29
ERˆELC	Ceiling Accuracy	.806	**1**	0.500	0.833	0.306	0.756	0.556
	Average BIC	254.9	**257.8**	269.2	257.9	291.3	286.8	291.4
	Test Phase Model Loss	0.392	**0.391**	0.431	0.392	0.469	0.456	0.462
	No. of Subjects	7/30	5/30	3/30	**9/30**	0/30	0/30	1/30
R	Ceiling Accuracy	0.694	0.711	**1**	0.806	0.806	0.722	0.611
	Average BIC	182.4	189.8	**89.0**	147.0	267.4	220.2	259.3
	Test Phase Model Loss	0.274	0.286	**0.102**	0.203	0.440	0.332	0.396
	No. of Subjects	0/29	0/29	**28/29**	0/29	0/29	0/29	0/29

Note. The rows are experimental conditions and the columns are Bayesian models representing different rule families. The best fitting model is depicted with bold text. Text is blue if the best fit is achieved by the family with the ground truth rule and red otherwise. Subject counts do not sum to the total because some are best fit by the Baseline model (not shown in the table).

all trials.) For every subject, window, and rule defined by the families in Table 1, we computed a BIC score reflecting how well that rule matched the subject's judgments. We then chose the best rule in each family and computed the posterior model weights for those rules. Fig. 2 presents the results from six example subjects, three each from the ERC and ERˆEC conditions. Each condition includes a *learner*, who achieved perfect test accuracy, a *non-optimal learner*, whose test accuracy was < 1 but $> .80$, and a *non-learner* whose test accuracy was near chance. In each panel the lines reflect the posterior model weights for the best rule in each family; the best rule itself appears as floating text at the top of the panel.

All six subjects exhibit distinct periods in which their judgments were characterized by one rule followed by a transition to another. For example, starting in window 21:100 the ERC learner's judgments were first best characterized by a conjunctive rule (LˆRˆC: large and red and circle), which was the followed by the correct ERC rule. And, the ERC non learner's judgments were first dominated by a disjunctive rule (L|B: large or blue), then a conjunct (LˆR: large and red), then a single feature (L: large), and finally ended with another conjunct (LˆBˆC: large and blue and circle). In summary, the majority of subjects exhibited evidence of testing specific rules and switching to alternative hypotheses as needed. We return to this point below.

Discussion

In this study, we asked subjects to learn various forms of typed compositional causal systems to investigate how human learners acquire complex causal knowledge. First, we found that whereas such systems are learnable, they are hard, as revealed by the fact that both the model-free and the model-based results indicated they are more difficult than a rule based on a single feature (the R condition). We also found that the number of object types in the system contributes to its acquisition difficulty. While a majority of subjects could learn rules with one object type, only a small minority could learn rules with two. These rules were instead approximated with simpler rules, such as conjunctions.

We also asked how learnability is impacted by the number of feature specifications. Contrary to our expectation, the results suggest that the rules become *easier* as more features

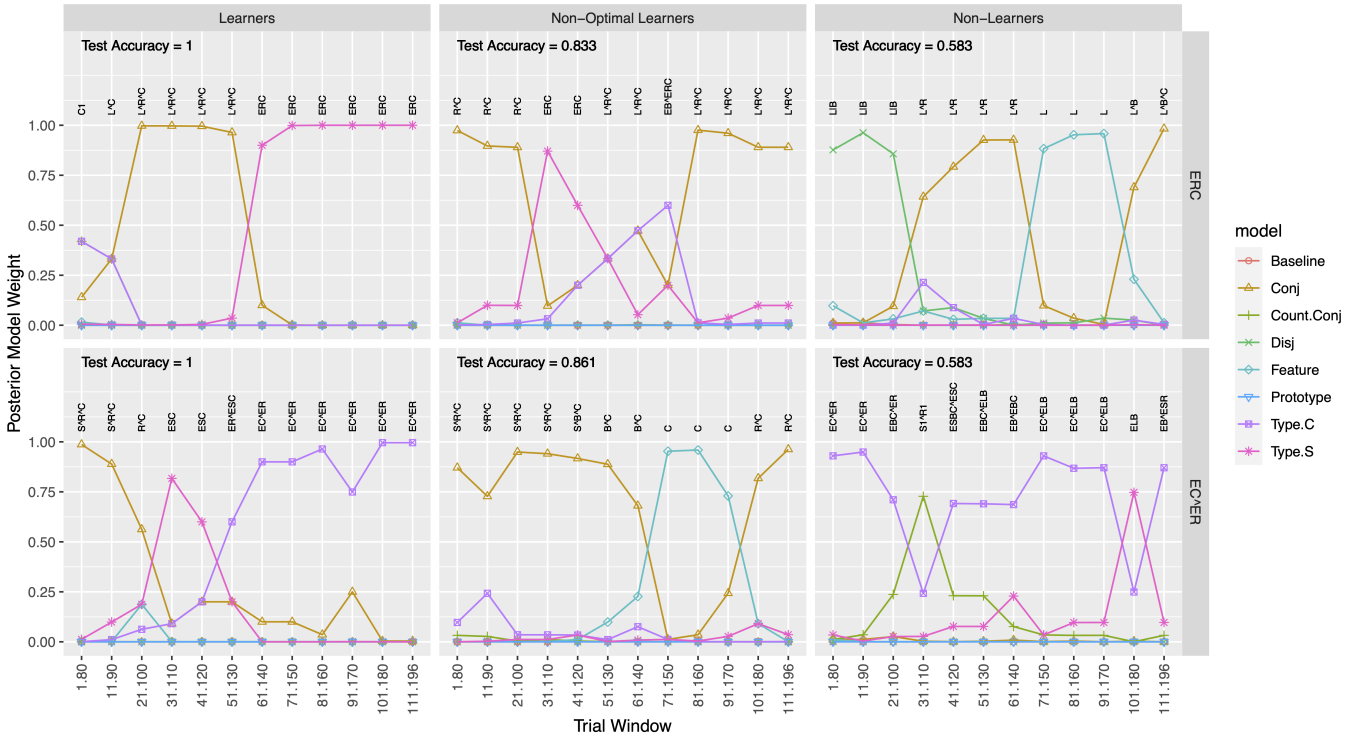

Figure 2: Example subjects in the sliding time window analysis, organized into two conditions (ERC and ER∧EC) and three learner groups. The x-axis is the start:end of the time window (in terms of trial number) and the y-axis encodes the posterior model weight of the best rule in each family. The floating texts denotes the best-fitting rule in each time window.

are specified. One possible explanation is that the sampling probability of a given rule is sensitive to the starting location (Bramley et al. 2018). That is, if subjects start from a rule representing a positive instance and continuously modify parts of the rule until ground-truth is reached, then rules with more feature specifications are more likely to be sampled. However, more testing is needed to warrant this explanation, which we intend to pursue in the future.

More generally, our intent is to pursue the "Language of Thought" (LOT) question as regards causal learning. That is, we want to identify the primitive hypotheses that learners bring to bear in a causal learning situation (Bramley et al. 2018, Zhao et al. 2022). In our current modeling, we simply predefined a set of plausible rule families to identify which of the tested causal rules were learnable and which were approximated by simpler rules. In contrast, state-of-the-art models (e.g., Goodman et al. 2008; Piantadosi et al. 2016) typically uses probabilistic context-free grammars (pCFG) to define large hypothesis spaces. Since the probability of each production rule can be manipulated independently, these models have precise control over the prior probability of individual rules in the hypothesis space. So, a pCFG that readily generates rules with one object type but not two or more may prove to be one straightforward account of our empirical results. Of course, a complete grammar will need to generate a far wider variety of rules than we've considered here. For example, in addition to conjunctions and disjunctions, people certainly can learn relational rules such as "the objects are the same color," "one object is larger than the other", and so forth.

In addition to the primitives that learners have available, another question concerns the psychological validity of our model. As we've explained, our ideal Bayesian model cannot explain the stochasticity and sequential bias subjects demonstrated in our study. In the future, we plan to develop a rational process model of causal rule learning with approximation methods such as a particle filters (Bonawitz et al. 2014; Speekenbrink 2016) and instance-driven generators (Bramley et al. 2018) to account for subjects' learning patterns.

The major conclusion of this study is that subjects struggle to bootstrap multiple object types at the same time, at least in a single experimental setting. Perhaps this should not be surprising: After all, there is a reason soccer coaches are paid handsomely for their know-how of team building. Complex compositional causal systems, despite their prevalence, are not the low-hanging fruit of human learning experiences. Our design might not be able to offer subjects the necessary depth and width of experiences to fathom the ground-truth, and we plan to modify our designs to facilitate better learning. We believe that an understanding of the learning process behind these systems is necessary for a more comprehensive theory of human causal knowledge, and we anticipate more research in this area in the future.

References

Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive psychology*, *74*, 35–65.

Bramley, N., Rothe, A., Tenenbaum, J., Xu, F., & Gureckis, T. (2018). Grounding compositional hypothesis generation in specific instances. In *Proceedings of the 40th annual conference of the cognitive science society.*

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: a test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1119.

Bye, J. K., Chuang, P.-J., & Cheng, P. W. (2023, January). How do humans want causes to combine their effects? the role of analytically-defined causal invariance for generalizable causal knowledge. *Cognition*, *230*, 105303.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Feldman, J. (2003). The simplicity principle in human concept learning. *Curr. Dir. Psychol. Sci.*.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*(5), 1205–1222.

Haygood, R. C., & Bourne Jr, L. E. (1965). Attribute-and rule-learning aspects of conceptual behavior. *Psychological Review*, *72*(3), 175.

Kemp, C., Goodman, N. D., & B Tenenbaum, J. (2008). Theory acquisition and the language of thought. *In Proceedings of 30th Annual Meeting of the Cognitive Science Society*, Austin, TX: Cognitive Science Society.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185–1243.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*(2), 87–137.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014, May). When children are better (or at least more open-minded) learners than adults: developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, *34*(1), 113–147.

Marsh, J. K., & Ahn, W.-k. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 334.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review*, *111*(2), 455.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive Science*, *28*(3), 303–333.

Speekenbrink, M. (2016). A tutorial on particle filters. *Journal of Mathematical Psychology*, *100*(73), 140–152.

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science*, *7*(6), 337–342.

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science*, *31*(2), 233–256.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology*, *53*(1), 27–58.

Zhao, B., Lucas, C. G., & Bramley, N. R. (2022). How do people generalize causal relations over objects? a nonparametric bayesian account. *Computational Brain & Behavior*, *5*(1), 22–44.