Order effects in one-shot causal generalization

Bonan Zhao[1] (b.zhao@ed.ac.uk)**, Chris Lucas**[2] (c.lucas@ed.ac.uk)**, Neil Bramley**[1] (neil.bramley@ed.ac.uk)
[1] Department of Psychology, [2] School of Informatics, The University of Edinburgh, Scotland, EH8 9JZ UK

Abstract

We explore how people generalize from a single observation of one entity (an "agent", or cause) acting on another (a "recipient", or effect) resulting in some change to the recipient's features. In line with recent demonstrations of human capacity for few-shot concept learning, we find strong and systematic patterns of generalization that are well captured by a Bayesian inference model favoring simpler causal "rules". However, we also identify a clear order effect in which individuals' generalization patterns substantially depend on the *dissimilarity*between the observed objects and the first few new cases they need to predict. To capture this, we develop a process account in which the act of generalizing crystallizes the cognizers' causal beliefs. We demonstrate that this captures our order effect, and so outperforms the Bayesian account in match to the behavioral data while providing a computationally plausible mechanism for real world causal generalization.

Keywords: causal learning; generalization; Bayesian model; order effect;

Introduction

Reasoning about causality and generalizing from experience are two pillars of human cognition. While a wealth of research has been devoted to studying how children and adults acquire causal beliefs (e.g., Sloman, 2005; Gopnik et al., 2007; Griffiths & Tenenbaum, 2009; Kemp et al., 2012; Bramley et al., 2015) and formulate generalization rules (e.g., Shepard, 1987; Tenenbaum & Griffiths, 2001; Goodman et al., 2008; Lake et al., 2015; Wu et al., 2018), the interplay between causality and generalization—what we call "causal generalization"—has received much less attention. On the face of it, this is surprising. If causal beliefs did not frequently carry over to novel situations, they would be of limited use to us. If follows that a key aspect of successful causal learning is to generalize causal relations appropriately to new situations that are related but nonidentical to past experiences. Generalization, on the other hand, could not be successful without tapping into what Sloman calls Nature's "invariants" (2005), the true causal relations that hold in both the experienced and novel situations. While pioneering research has explored the interplay between causality and generalization using hierarchical Bayesian models (e.g., Griffiths & Tenenbaum, 2009; Goodman et al., 2011), this rational approach is limited in its ability to capture psychological processes due to its intractability.

In this paper, we investigate how people generalize causal relations from a single observed interaction between two simple objects in which one (the agent, or cause) produces some change in another (the recipient, or effect). Our experiments manipulate visual features of objects, like colors, shapes, etc. These visual features encodes both causal relations - for example *red agents can make recipients red*, and systematic generalization variations from a known case - e.g., *generalizing possible causal power of red agents to green agents*. This agent-recipient setup makes the causal directionality explicit - the animation of an agent object actively contacting the recipient object to bring the effect complies naturally with human intuition. Using abstract features to build up causal relations and generalization conditions provides rich scenarios that allow systematic and sophisticated combinatorial reasoning, and at the same time minimizes background knowledge influence.

We model our task initially using an normative Bayesian inference framework (Griffiths & Tenenbaum, 2009; Lucas & Griffiths, 2010), before exploring an algorithmic account, inspired by category learning and Bayesian approximate inference (Goodman et al., 2008; Sanborn et al., 2010) , in which generalizations crystallize the cognizer's beliefs. We show that this account better explains our behavioural data. We conclude by a discussion of limitations and future directions.

Normative Account

From a normative perspective, causal generalization involves both *learning* and *decision making*. First, reasoners must combine their prior beliefs (or inductive biases) with their observation(s) via Bayes theorem to form posterior beliefs about what forms of causal relations exist, as well as their *provenance* — i.e. what sets of entities (agents) produce particular effects and what set of entities are able to receive them (recipients).

Second, predictions about newly encountered cases — i.e. guesses about what change will occur in some new recipient when acted on by some new agent — require the reasoner to average or maximise over their posterior.

In non-toy contexts, this process is subject to two severe complications. (1) There is a theoretically unbounded space of possible functional forms for causal relationships — here, ways of mapping features of agents and recipients to changes in recipients' features, and (2) the world contains an unknown and also theoretically unbounded, number of potential causal categories. That is, we assume that it could be the case that several causal relationships exist in the world, applying to different kinds of objects and producing different kinds of changes.

To model the causal function space we use a Domain Specific Language (DSL, citation) capable of producing a large space of possible feature to feature change mappings and a *probabilistic context-free grammar* (PCFG) to incorporate inductive biases, in particular the principle of parsimony, favoring simpler rules over more complex ones (Bramley et al., 2018; Goodman et al., 2008). To model the causal category space, we use a Chinese restaurant process (CRP, citation) to capture the existence of between 1 and an infinite number of

Table 1: λ-abstraction Based Probabilistic Grammar.

Productions			
Start	$S \rightarrow$	$\lambda_{\phi_i} : A, \Phi$	
Bind additional	$A \rightarrow$	B	$B \wedge S$
Relation	$B \rightarrow$	$(r' = C)$	$(r' \neq C)$
Reference	$C \rightarrow$	$D1$	$D2$
Relative ref.	$D1 \rightarrow$	a_{ϕ_i}	r_{ϕ_i}
Absolute ref.	$D2 \rightarrow$	value[a]	

[a]: Sampled uniformly from the support of feature ϕ_i.

causal categories, again building in an inductive bias toward simplicity in the form of fewer, larger, categories.

Causal functions

We treat each causal relation as a function that takes an agent (A) and recipient's initial state (R) as input, and outputs the final state of the recipient (R'). Formally, $f_w(A,R) = R'$. Note that we use uppercase A, R, R' to denote "roles" in the abstract sense, and lowercase a, r, r' to refer to observations of the corresponding role.

We assume some function $f_w \in F$ describes the causal effect of A on R, where F is the (infinite) set of possible causal functions. For example, $color(R') \Leftarrow color(A)$ is a simple causal function in which the recipient inherits the agent's color property — i.e. as when paint (A) is applied to a wall (R). For simplicity, we here restrict W to those expressions that involve conjunctions of expressions that each specify a feature value of R' in terms of the observed feature values or those of A or R. For example, $(color(R') \Leftarrow color(A)) \wedge$ $(shape(R') \Leftarrow square)$. We assume that any features unspecified by w follow the principle of inertia, and remain as they were before the causal interaction, that is, $feature(R') \Leftarrow feature(R)$.

These logical expressions determine how causal functions behave. Take $w := (color(R') \Leftarrow color(A)) \wedge (shape(R') \Leftarrow square)$ for example. For an agent a = *red-circle* and a recipient r = *yellow-circle*, $f_w(a,r)$ = *red-square*. Hence, given a data-point $d = (a, r, r')$, $P(d|f_w) = 1$ if $r' = f_w(a,r)$, and $P(d|f_w) = 0$ if $r' \neq f_w(a,r)$.

Moreover, we allow f_w to be nondeterministic, as an agent might induce one of a set of possible changes in a recipient - for example changing the recipient's colour to "anything but red".[1] We can thus generalize with the above definition by applying it to some new a and r as in $f_w(a,r) = S$, where S is a set of values that R' can take from. Accordingly,

$$P(d|f_w) = \begin{cases} \frac{1}{|f_w(a,r)|} & \text{if } r' \in f_w(a,r), \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Generative process

We use a PCFG to define a space of possible functions (as well as a procedure for generating sample functions from these space. Let $G = (T, N, S, P, \Theta)$ be our PCFG, each element corresponds to a set of (T)ermination symbols, a set of (N)on-termination symbols, the (S)tart symbol, a set of (P)roduction rules, and the set of probabilities assigned to each production (Θ). Table 1 summarizes N and P of our grammar. Note that this grammar is not strictly context-free as it binds a feature to a lambda expression, and all the subsequent steps within the scope of the λ-attraction refer, again, to this feature.

We here, simply assume uniform probabilities for each production rule ($\theta_i = \frac{1}{N_I}$ for each row I with options $i \in I$. The grammar inherently favors simpler sentences because the "Bind additional" rule is called with probability $\frac{1}{2^n}$. Therefore, this generative process provides prior belief about causal functions that encodes a generic inductive bias toward simplicity. Letting w_r be the frequency of appearances of production rule r in expression w, we can define the prior probability of any rule as:

$$P(f_w) = P_G(w|\Theta) = \prod_{r \in R} \theta_r^{w_r} \tag{2}$$

Inference is now fully defined. Given a data point $d = (a, r, r')$ and the space of possible causal functions F:

$$P(f_w|d) = \frac{P(d|f_w)P(f_w)}{\sum_{f'_w \in F} P(d|f'_w)P(f'_w)} \tag{3}$$

Causal Categories

While causal relations are by-definition invariant, they are not universal - not all objects have the same causal powers, and causal powers do not work on all objects. For example, in the current task environment, it could be the case that only square Agents make Recipients take their color, while non-square agents leave their recipients unchanged. It could also be the case that a causal power only works on certain recipients, for example, agents could make square recipients take their color but leave other recipients unchanged. This potentiality for limited provenance for a given causal function leads to the second complication highlighted in the introduction, such that there could be one or several causal rules that apply to specific subsets of possible As and Rs.

We formalize this with the idea that the entities fall into different causal categories, and a given causal function applies to a particular causal category, and we accommodate this possibility by drawing on another non-parametric technique, that of a *Dirichlet Process*, composed by a Chinese Restaurant Process (CRP) and a Dirichlet prior over the feature values of categories.

For a magic stone a and category Cat_i,

$$P(Cat_i|a) \propto P(a|Cat_i) \cdot P(Cat_i) \tag{4}$$

[1] But for simplicity we assume any stochastic function has uniform probability across its possible outputs.

The likelihood part expresses the intuition that stones sharing the same category have something in common, and the prior $P(Cat_i)$ encodes a tendency to favor categories with more members as they seem to be more "common".

Specifically, $P(a|Cat_i)$ is calculated based on feature values of a and the total feature values of stones already assigned to Cat_i. Let v^k_a be a binary code for feature value k of stone a, for all feature values $k_1,\ldots,k_n$, if stone a takes value k_j, $v^{k_j}_a = 1$, otherwise $v^{k_j}_a = 0$. Take a *red_square* stone for example, $v^{red}_a = 1$, $v^{circle}_a = 0$, etc. Fixing a list of possible feature values, we can use a vector $\mathbf{v_a} = \{v^{k_i}_a,\ldots,v^{k_n}_a\}$ to represent the feature values of stone a. Each category is initialized with a pseudo feature prior μ, and enriched by the stones assigned to this category: $v^k_{Cat} = \sum_{a_i\in Cat} v^k_{a_i} + \mu$. Let $\mathbf{a^k}$ be the set of feature value indexes where $v^{k_i}_a > 0$, i.e, features that stone a exhibits, likelihood of stone $a = \mathbf{v_a}$ given a category $Cat_i = \mathbf{v_{Cat_i}}$ is

$$P(a|Cat_i) = \prod_{k\in\mathbf{a^k}} \frac{v^k_{Cat_i}}{\sum_{j\in K} v^j_{Cat_i}} \qquad (5)$$

We use Chinese Restaurant Process (CRP) to assign priors. Let $|Cat_i|$ be the number of magic stones assigned to category Cat_i, and N be the number of magic stones already assigned to existing categories, $N = \sum_{i\in C} |Cat_i|$, the prior probability according to CRP is $P(Cat_i) = \frac{|Cat_i|}{N+\alpha-1}$. α is known as the concentration, or dispersion parameter - the larger α is, the more likely a new stone fells into a new category. The normalization factor marginalize on all existing categories plus a potential new category, whose prior can be calculated from CRP as $P(Cat_{new}) = \frac{\alpha}{N+\alpha-1}$.

Generalization decisions

With a posterior over causal functions and their respective categories in place, a normative agent can make generalizations to new cases. In the decision stage, upon observing a partial data point $d^* = (a^*, r^*, \cdot)$, an optimal decision can be made by aggregating the posterior predictive distribution of each possible $r^{*\prime}$ value:

$$P(\tilde{d}^*) \propto \int_{f_w} p(\tilde{d}^*|f_w)P(f_w|d)\mathrm{d}f_w \qquad (6)$$

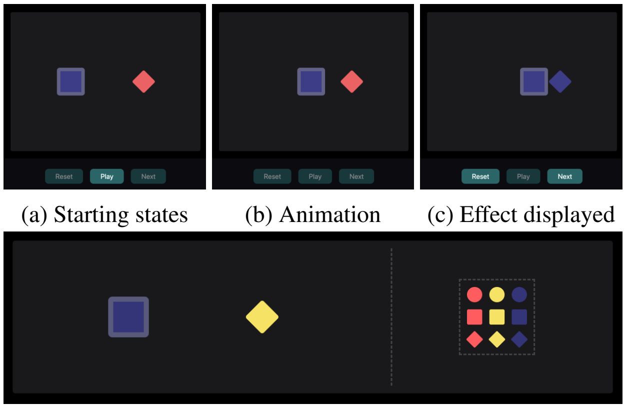
and taking the maximum over this predictive posterior

$$\text{Choice} = \arg\max P(\tilde{d}^*) \qquad (7)$$

We will analyse participants judgments in our experiment using this normative model as a benchmark before considering a more computationally tractable alternative.

Experiments

We developed an inference causal game called "Magic stones". Participants learn that in the Magic stones world, there are two types of stones: *magic stones* that can affect other stones (Figure 1, indicated by a thick border) and *normal stones* (no border) that cannot. Participants watch an animation in which the magic stone moves and touches the normal stone (Figure 1b) whereupon the normal stone's color and or shape may change (Figure 1c). Then they complete a generalization (test) phase (Figure 1d), where they make a sequence of predictions about how new magic stones will affect new normal stones, selecting in each case from the possible final appearances of the normal stone. A demo of the task is available at `http://bramleylab.ppls.ed.ac.uk/experiments/bnz/magic_stones/index.html`.



(a) Starting states (b) Animation (c) Effect displayed

(d) Generalization task

Figure 1: Magic stones task interface. (a–c) show an example learning scene animation. (d) shows a generalization task.

Methods

Participants One-hundred-and-twenty participants (53 female, aged 40±11) were recruited from Amazon Mechanical Turk and were paid \$1.19 for their time. The task took 5.23±3.17 minutes.

Materials and procedure We manipulated two object features - color and shape, and three values for each feature in this experiment. Colors were sampled from $\{$*red, yellow, blue*$\}$, and shapes from $\{$*circle, square, diamond*$\}$. Participants first completed an instruction phase, and had to successfully complete a comprehension quiz before starting the main task. In the main task, each participant experienced a single learning scene followed by fifteen generalization tasks. Table 2 shows the six types of learning scene we varied between subjects. These learning scenes investigate a range of example causal effects differing in whether one or both features of the normal stone, and whether the new values are a match to the magic stone's features. The initial learning scene was displayed at all times and the animation was replayed once between each task to ensure it was not forgotten.

Generalization task sequence Based on the the stones used for each learning scene, we constructed 15 generalization tasks per learning scene (Table 3). For each task, the new magic (normal) stone takes value with respect to the magic (normal) stone in the learning scene. For example, L1 (see Table 2) depicts a *red square* magic stone and a *yellow cir-*

Table 2: Learning Conditions

Cond.	Magic stone	Normal stone	After-state of normal stone	Possible summary
L1	red-square	yellow-circle	yellow-square	Normal stone takes magic stone's shape
L2	yellow-diamond	red-square	red-circle	Normal stone changes to a new shape
L3	blue-square	red-diamond	blue-diamond	Normal stone takes magic stone's color
L4	red-circle	blue-square	yellow-square	Normal stone changes to a new color
L5	blue-square	yellow-circle	blue-square	Normal stone becomes same as the magic stone
L6	red-diamond	yellow-square	blue-circle	Normal stone changes to new colors and shapes

Table 3: Generalization Tasks

Conditions	N_C, N_S	$\neg N_C, N_S$	$N_C, \neg N_S$	$\neg N_C, \neg N_S$
$M_C,\ M_S$	Task→	1	2	3
$\neg M_C,\ M_S$	4	5	6	7
$M_C, \neg M_S$	8	9	10	11
$\neg M_C, \neg M_S$	12	13	14	15

M: magic stone, N: normal stone, C: color, S: shape, ¬: not, relative to the learning scenes. Numbers represent task ID.

cle normal stone. According to the specifications in Table 3, task 1 for L1 has a *red square* magic stone, and a *blue circle* normal stone.

We call the sequence of tasks from 1 to 15 "near-first transfer" because they have the same magic stone as in the learning scenes, and from task 15 to 1 the "far-first transfer" sequence, because their sets of stones are completely different from those in the learning scenes. Within each sequence, color-different tasks and shape-different tasks (task 1 & 2, 5 & 6, 9 & 10, 13 & 14, 4—7 & 8—11) are shuffled to counterbalance feature order. Participants were randomly assigned to one of 6 learning scenes × 2 task sequences = 12 conditions, leading to 10 ± 1.6 participants per condition.

Results

We are primarily interested in the level of agreement among participants on each generalization test, as this gives a sense of how systematic or strong preferences for any particular patterns of generalization are. To measure this, we define a homogeneity measure η for task t:

$$\eta_t := \frac{\sigma^2(\mathbf{d}_t)}{m} \tag{8}$$

where $\mathbf{d}_t = \{fN_{r'_1}, \ldots, fN_{r'_n}\}$, the set of relative frequency of selecting r' object for task t, aggregated on all participants. $\sigma^2(\mathbf{d}_t)$ measures how various participant selections are, regardless of sample size. $m = \sigma^2(\{1, \underbrace{0, ..., 0}_{n-1}\})$ is the maximal level of agreement, scaling η to a range of [1,0], where 1 represents maximal agreement. η approaches 0 when everyone selects randomly.

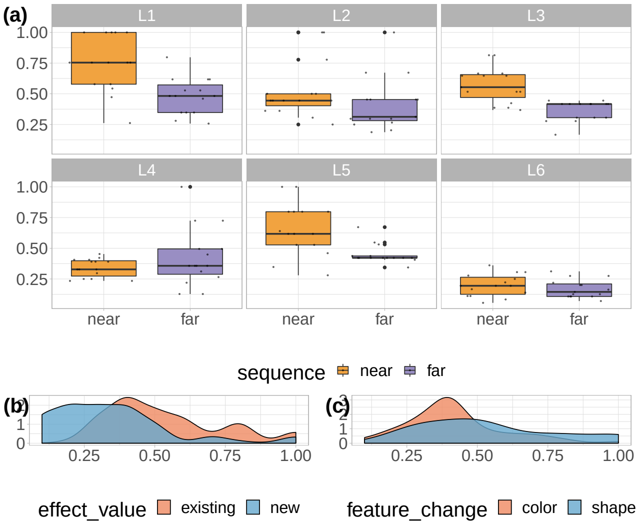


Figure 2: Behavioral results. Figure (a): η per task sequence type for each learning scenes, near-first transfer in yellow, far-first in purple. Figure (b), (c): density distribution of η.

We found that people make systematic causal generalization predictions under single observations. Across 6 learning scenes × 15 trials = 90 tasks, the overall homogeneity measure $\eta = .41 \pm .19$, with min .03 (L6, task 6), and max 1.00 (L2, task 4). A simulated random baseline induces $\eta_{\text{random}} = 0.05 \pm 0.02$, differs significantly, $t(89) = 17.87, p < .001$, 95%CI = [0.32, 0.40].

Next, we compared the homogeneity of generalizations depending on the order they were probed (i.e. in the *near-first* or the *far-first* transfer order. Generalizations were more homogeneous for $\eta_{\text{near-first}} = .51 \pm .25$, compared with $\eta_{\text{far-first}} = .38 \pm .18$, $t(89) = 5.12, p < .001$, 95%CI = [0.07, 0.17].

Participants generalized less homogeneously when the learning task involved new colors or new shapes (Figure 2b). For leaning scenes L1, L3, and L5, where effect states match feature values from the magic stones, overall homogeneity measure $\eta_{\text{existing-value}} = .51 \pm .12$. Learning scenes L2, L4, and L6, where effects contain a brand new value, overall homogeneity measure $\eta_{\text{new-value}} = .30 \pm .20$, differs significantly from the *existing-value* group, $t(88) = 6.31, p < .001$, 95% CI = [0.15, 0.28].

```
Data: $D_L = \{d_0\}, D_G = \{d_1, \ldots, d_n\}$
Result: $N_G = \{n_1, \ldots, n_n\}$
Assign $d_0$ to category $Cat_1$;
Sample $f_{Cat_1}$ for $Cat_1$ from the posterior;
for $d_i \in D_G$ do
    if $d_i$ belongs to an existing category $Cat_j$ then
        $n_i = f_{Cat_j}(d_i)$;
    else
        Create $Cat_{new}$;
        Sample $f_{Cat_{new}}$ from the prior;
        $n_i = f_{Cat_{new}}(d_i)$;
    end
end
```

Algorithm 1: Process model steps

Color and shape changes were generalized to different extents despite these features appearing in symmetric and counterbalanced contexts in the task (Figure 2b). In general, shape changes induced more homogeneous predictions, $\eta_{\text{shape-changes}} = .50 \pm .19$, compared to color changes $\eta_{\text{color-changes}} = .39 \pm .13$, $t(58) = 2.56, p = 0.013$, 95%CI = [0.02, 0.19], in line with developmental literature finding that shapes are perceived as more "fundamental" features (Landau et al., 1988), and therefore more likely to be critical for causal effects (see discussion).

Modeling results

We model the experiments using the normative account introduced in the previous section. The normative model predicts the same behavioral pattern exhibited by the experiment (see Figure 3): systematic generalization agreement, stronger agreement for conditions involving matching existing feature values and higher diversity for conditions with brand new feature values. The most-favored predictions for each task according to the normative model agrees with participants' in 73.3% of the cases.

The normative model predicts that generalization agreement is stronger when new stones share more feature values with the learning scene, and weaker when generalizations are very different from the learning scene. However we do not observe this in participants' selections. In addition, normative model fails to predict order effects.

Process Model

The normative model can capture the predominant choice of participants pretty well, but it fails to predict the order effects demonstrated by experimental data. There are several reasons why the near-transfer sequence induces more homogeneous predictions while the far-transfer sequence induces more diverse ones, and here we explore one of them: when stones are presented in near-transfer sequence, participants are likely to stick with whatever causal relation they figured out during learning; in the far-transfer sequence, however, when participants are first faced with very different stones, they may assign novel causal relations to these novel observations, as their link to the learning data point is relatively weak.

Table 4: Model comparison

Model	Likelihood	BIC	R^2
Random baseline	-3955	7910	
Normative model*	-2687	5389	.63
Process model*	-2642	5299	.65

*With initial parameters $\mu = 0.1, \alpha = 0.1$.

To have good control of mixing causal relations with types of objects, we take insights from work on categorization (Sanborn et al., 2010). We will assign causal functions to causal categories, and assume that different causal categories have different causal functions. The model progresses as in Algorithm 1: first, we assign the magic stone in the learning data point to an initial category Cat_1, and sample a causal function for it from the posterior distribution of all possible causal functions. For each generalization trial, we first compare whether the magic stone in this trial belongs to any existing category, if so, we will apply the causal function tied to this category and make a prediction; and if not, we create a new category, and sample a causal function for this category from the prior distribution of all causal functions.

Results

The process model both predicts major participant selections, and reproduces order effects (see Figure 3). results. Table 4 summarizes how well both the normative model and process model explain participant data - both models perform a lot better than random baseline, while the two models have competing performance in terms of likelihood. As modeling fitting is still in-progress, here we compare the model performances with a set of initial parameters: $\mu = 0.1, \alpha = 0.1$. These parameters produce a relative medium level of order effects.

Discussion

Using the "magic stones" interface, we introduced a one-shot causal generalization task. This task demonstrated that people were able to make systematic generalizations for complex causal relations under limited learning data. We also identified a strong order effect in people's generalization patterns—tasks given in a *near-first transfer* sequence induced higher agreement among participants, while people tended to entertain more diverse options when the tasks are given in the *far-first transfer* sequence. Participants' generalization patterns are well-captured by Bayesian inference models operating on a pool of causal functions generated by a simple PCFG that favors simpler "rules". Order effects is captured by a simulation model that assigns causal functions according to how similar a novel object is to the known ones.

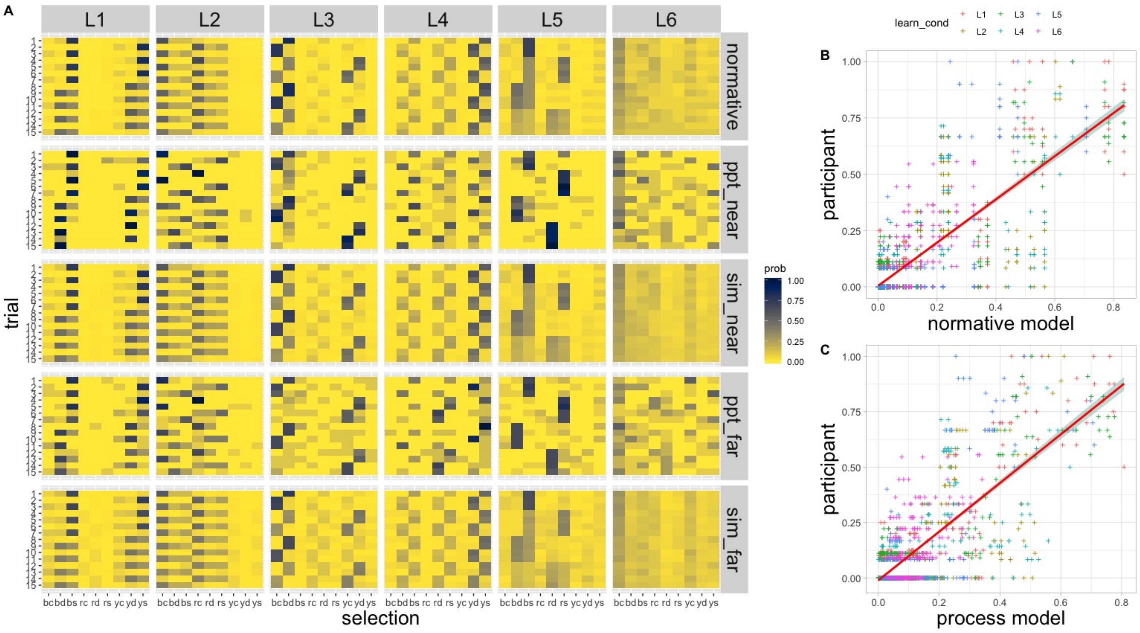


Figure 3: Model predictions in comparison to participant selection. Figure **A**: Model predicted probabilities for each R' object and participant selections for each R' in terms of relative frequency, grouped by learning scenes and models/conditions. Darker cell is higher probability/frequency. Figure **B,C**: linear regression plots of model predictions against participant selections, each point is a R' object in a generalization trial, x-axis is model predicted probability, y-axis is relative frequency by participants.

Behavioral data also suggested that shape-related effects induced more systematic generalization patterns, compared with color-related effects. People perceive shapes and colors differently (Treisman & Gelade, 1980; Landau et al., 1988), and unsurprisingly generalize causal rules with regard to them differently. A better model of causal generalization needs to take care of such "dimensional" differences, in support of the claim that prior knowledge shapes causal generalization judgments (Griffiths & Tenenbaum, 2009).

The "Magic stones" interface is very rich and flexible. Content-wise, expanding the existing fixed-value two-feature setup to allow richer feature dynamics will help us have a better understanding of how people achieve causal generalization in this bustling noisy world. Task-wise, besides the one-shot causal generalization tasks used by this experiment, we plan to use it for exploring self-directed exploration, discovery, iterated learning, etc. Planning, in particular, meaning goal-directly causal generalization, warrants further investigation.

References

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Bramley, N. R., 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.*

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.

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

Gopnik, A., Schulz, L., & Schulz, L. E. (2007). *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological review*, *116*(4), 661.

Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, *64*(1-2), 35–73.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321.

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

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.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(4), 629–640.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924.