Building Object-based Causal Programs for Human-like Generalization

Bonan Zhao
Department of Psychology
University of Edinburgh
b.zhao@ed.ac.uk

Christopher G. Lucas
School of Informatics
University of Edinburgh
clucas2@inf.ed.ac.uk

Neil R. Bramley
Department of Psychology
University of Edinburgh
neil.bramley@ed.ac.uk

Abstract

We present a novel task that measures how people generalize objects' causal powers based on observing a single (Experiment 1) or a few (Experiment 2) causal interactions between object pairs. We propose a computational modeling framework that can synthesize human-like generalization patterns in our task setting, and sheds light on how people may navigate the compositional space of possible causal functions and categories efficiently. Our modeling framework combines a causal function generator that makes use of agent and recipient objects' features and relations, and a Bayesian non-parametric inference process to govern the degree of similarity-based generalization. Our model has a natural "resource-rational" variant that outperforms a naïve Bayesian account in describing participants, in particular reproducing a generalization-order effect and causal asymmetry observed in our behavioral experiments. We argue that this modeling framework provides a computationally plausible mechanism for real world causal generalization.

1 Introduction

Objects appear to be a fundamental building block of our world models, appearing early in development and ubiquitously in natural language [1–3]. Observing objects interacting naturally invokes causal perceptions. For instance, in "launching" phenomena [4], when participants observe some object A moving toward a stationary object B, and if around when A touches B, A stops moving and B starts to move, participants spontaneously report that they see object A cause object B to move [see also 5–7].

While a wealth of research has been devoted to studying how children and adults acquire causal beliefs [e.g., 8–12] and generalize functional properties [e.g., 13–17], the interplay between causality, object concepts and generalization has received less attention. On the face of it, this is surprising. In reality, a key component of successful causal learning is the ability to generalize causal relations appropriately to new situations that are related but non-identical to past experiences [18–20]. Meanwhile, generalization could not be successful without tapping into what Sloman calls Nature's "invariants", the true causal laws that govern both experienced and novel situations [8]. Recent research has explored this interplay using hierarchical Bayesian models [e.g., 11, 21, 22] as a computational level account of domain knowledge [23], or formal analysis on transportability of structural causal knowledge [18–20], but these are limited in their ability to capture psychological processes due to their inherent intractability [24–26].
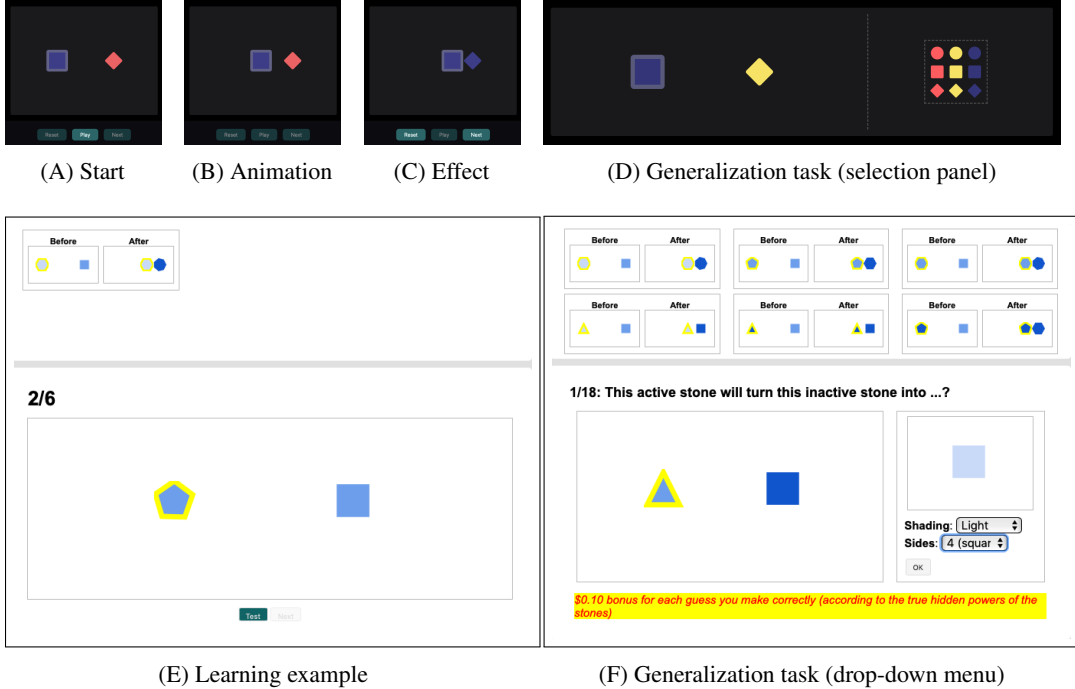
Figure 1: Task interfaces. A–D: Experiment 1. A–C step through an example learning scene animation, and D shows a generalization task consisting of novel objects (left) and a selection panel (right), in which learners select from a set of possible predictions about the appearance of the recipient after the causal interaction. E–F: Experiment 2. Summaries of previous learning examples are shown at the top of the screen. E shows one animated effect similar to A–C. In F, generalization predictions are elicited by selecting from two drop-down menus (one per feature).

In this paper, we explore how people generalize causal relations from observed interactions between pairs of simple geometric objects, and propose a computational modeling framework that has a natural "resource-rational" variant. We develop an interactive online game we call "magic stones", in which people can test how one object (the agent) acts upon another (the recipient) and brings about some change in the recipient (see Figure 1A–C), and then make predictions about new pairs of objects (Figure 1D). This game thus provides behavioral measures of how people *generalize* their causal understanding of observed objects to unseen ones.

In the following sections, we introduce our computational modeling framework for object-based causal inference with an expressive hypothesis space that captures the diverse inferences people can make [11, 27]. It draws on non-parametric approaches to category and function learning to account for similarity-based generalization predictions (Figure 2). The normative version, we call LoCaLa (Local Causal Laws), compares each generalization trial against all the learning examples in order to assign causal categories to new observations. We then describe a "resource-rational" [13, 28] variant, we call LoCaLaPro (Local Causal Laws Process), that shares causal categories among generalization trials, and only posits a new causal function and category when it cannot explain a novel observation with any existing categories.

We report on two experiments that shed light on previously unexplored inductive biases in causal learning, and so allow us to evaluate our models and the ideas that motivate them. We find that our local laws and particularly our new process model better explain our behavioral data than a purely normative account, including explaining a novel generalization-order effects observed in Experiment 1, and causal asymmetry in Experiment 2.

2   Related work

Our task generalizes the structure of standard "blicket detector" studies, in which different combinations of factors or objects are tested and an effect does or does not occur [e.g., 10, 21, 27, 29].
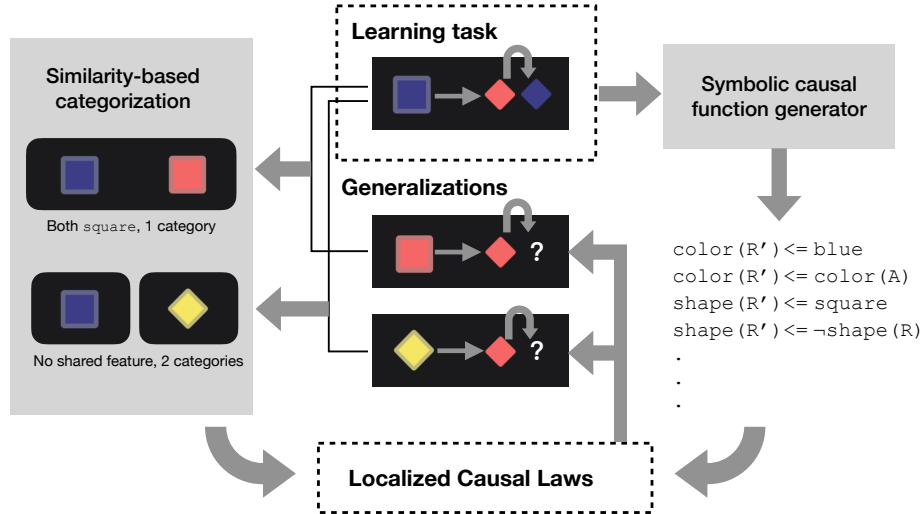
Figure 2: To model how people make object-based causal generalization predictions (middle), we combine program induction about the hidden causal laws (right) with non-parametric category inference about their domains of influence (left). Together, they form causal categories that guide generalization predictions (arrows from bottom to middle).

The collision stimuli we used in our tasks are known to evoke automatic perception of causality [4], making these an appealing way to study how people inherently reason about cause and effect. In daily life, we typically observe sequences of changes rather than independent trials [30–32], and our experiment interface can capture this pattern naturally. Unlike previous work [e.g. 21, 33], we are not constrained to binary, present/absent effects, or multiple outcomes, such as different kinds of activations [e.g. 34]. Our task can also capture higher-order causal relationships, such as rules that depend on color/shape matches between agent and recipient objects [29].

Causal Bayes nets (CBN) [35, 36] have inspired fruitful research in causal learning and induction, among which the most relevant to causal generalization are transportability analysis [18–20] and the hierarchical Bayesian model (HBM) framework [11, 21, 27]. However, these methods have been focused on inferring statistical relationships between variables, rather than interactions between objects. Furthermore, although CBNs and HBMs can address the problem of inferring network structure together with parameter estimation, they suffer from serious scalability issues: As the number of nodes and layers increases, the number of possible causal networks increases super exponentially, making domain expectations more of a computational curse than a learning-to-learn blessing in practice [24, 25]. Therefore, more recent accounts of causal learning have treated causal inference as practically constituting a search problem in a large multi-modal theory space [11–13, 21, 22, 37], utilizing generative grammars and program induction ideas, in order to capture the representational flexibility and inductive biases of human causal hypothesis generation along with its incompleteness.

While it has been argued that we think of causal relationships as "invariant" [8], category knowledge is also integral to real world causal inference. While people refer to causal relationships when categorizing objects [38–40], they also spontaneously use featural and relational information for categorization when no causal information is available to form categories [41–43], and then make causal predictions based on these categories [21]. Since creation of such causal categories may be triggered only when required for generalization, we will present a process view such that cognitive representations are fundamentally generative, and judgments are based on samples [37, 44, 45].

3 Formulation

Causal functions We use a Probabilistic Context-Free Grammar [PCFG; 46] to define a prior over possible causal functional relationships (causal laws). Grammar $\mathcal{G}$ generates expressions that specify features of the result object. For example, if one of the features is "color", a possible causal

function could be `color(r')` $\Leftarrow$ `red` — recipient will turn red — or `color(r')` $\Leftarrow$ `color(a)` — recipient will take the agent's color, and so on. This grammar is set up to allow for arbitrarily complex expressions allowing a rule to produce conjunctions of feature changes, for example, `AND(color(r')` $\Leftarrow$ `red, shape(r')` $\Leftarrow$ `triangle)`. A causal function outputs result object(s) when particular agent and recipient objects are provided. Detailed definitions for our grammar $\mathcal{G}$ are provided in Appendix A.

Latent causal categories   While this PCFG provides an account for the set of possible causal functions that people may entertain during learning (right box in Figure 2), it cannot tell us to what extent should these causal laws apply. When making generalization predictions, if the novel objects look similar to those in the learning phase, one may consider these objects falling into one category and are thus governed by the same causal law. However, if the novel objects look very different from learning examples, these objects may belong to different categories and therefore execute different causal functions (left box in Figure 2).

We formalize the idea that the objects may fall into different causal categories with respect to featural similarities, roles in the interaction, and shared causal laws. Let $\mathbf{d}$ denote a set of observations, $\mathbf{z}$ denote a particular set of causal category memberships, and $\mathbf{w}$ some categorization parameters (weights). We use superscript (i) for the i-th observation: $d^{(i)}$ for the i-th data point, $z^{(i)}$ the causal category assigned to the i-th observation, $a^{(i)}$ the agent in the i-th data point, similarly for $r^{(i)}, r'^{(i)}$, and $\mathbf{w}_{z^{(i)}}$ for the weights associated with categorizy $z^{(i)}$; additionally, let $z^{(-i)}$ be the categorization of observations except for $d^{(i)}$, inference about the i-th observation's category is given by:

$$P(z^{(i)}|\mathbf{d}, \mathbf{w}) = P(z^{(i)}|d^{(i)}, \mathbf{w}, z^{(-i)})$$
$$\propto P(z^{(i)}|z^{(-i)})P(a^{(i)}, r^{(i)}|\mathbf{w}_{z^{(i)}})P(r'^{(i)}|a^{(i)}, r^{(i)}, \mathbf{w}_{z^{(i)}}) \tag{1}$$

Equation 1 consists of three parts: $P(z^{(i)}|z^{(-i)})$ reflects our expectations about how causal categories are distributed, $P(a^{(i)}, r^{(i)}|\mathbf{w}_{z^{(i)}})$ encodes our beliefs about object features and category membership, and $P(r'^{(i)}|a^{(i)}, r^{(i)}, \mathbf{w}_{z^{(i)}})$ marks the causal function this particular category possesses. To accommodate the fact that there could be any number of causal categories, we draw on an extended Dirichlet Process, composed by a Chinese Restaurant Process (CRP) [47], a multinomial distribution over the feature values of an unknown number of categories with a Dirichlet prior, and likelihoods defined by causal functions, in correspondence to the three parts in Equation 1. Appendix B unpacks this procedure in more details.

In total, we introduce three global parameters for our extended Dirichlet Process: a concentration parameter $\alpha > 0$ for the distribution of categories according to CRP, a Dirichlet prior $\beta \geq 0$ to control the impact of feature similarities, and a focus parameter $\gamma \in [0, 1]$ for weighting the categorization based on causal action roles (agent or recipient). In our Dirichlet Process, both the focus parameter γ and the Dirichlet prior β are embedded in a local parameter $\mu^{(z_i)}$, the mean feature vector for category $z^{(i)}$ (see Appendix B). Let $f^{(z_i)}$ be the causal function assigned to category $z^{(i)}$, we can rewrite Equation 1 as:

$$P(z^{(i)}|\mathbf{d}, \mathbf{w}) \propto P(z^{(i)}|z^{(-i)})P(a^{(i)}, r^{(i)}|\mu^{(z_i)})P(d^{(i)}|f^{(z_i)}) \tag{2}$$

Inference   The Dirichlet Process defined by Equation 2 models the acquisition of causal categories, equivalent to the learning phase of causal generalization. It is impossible to compute the posterior directly because we do not know how many categories are there in advance, but we can approximate the posterior distribution using Gibbs sampling (see also Appendix B). In the prediction stage, upon observing a partial data point $d^* = (a^*, r^*, \cdot)$, an optimal decision can be made by marginalizing over the posterior predictive distribution of each possible r'^* value:

$$P(\tilde{d}^*) \propto \int_z p(\tilde{d}^*|z)P(z|d)\mathrm{d}z$$
$$\approx \frac{1}{|\tilde{Z}|} \sum_{\tilde{z} \in \tilde{Z}} p(r'^*|a^*, r^*, f^{(\tilde{z})})P(a^*, r^*|\mu^{(z)})P(z|d) \tag{3}$$

and taking the maximum over this predictive posterior:

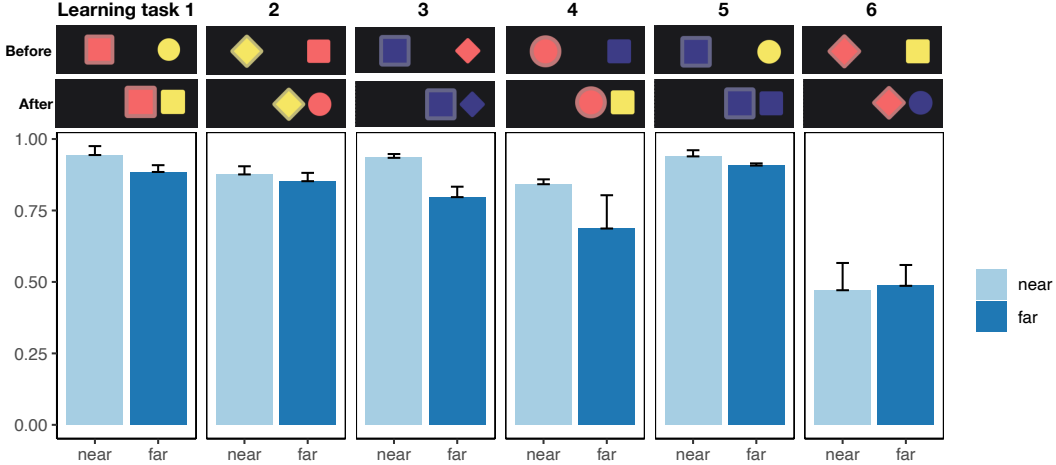$$\text{Choice} = \arg \max P(\tilde{d}^*) \tag{4}$$

Figure 3: Experiment 1 generalization consistency ρ_τ (y-axis), averaged over generalizations per each one-shot learning task (as illustrated on top of each panel) and sequence order: light blue = *near-first transfer*, dark blue = *far-first transfer*.

Process variant Instead of trying to approximate a global optimal distribution of latent causal categories, we further develop a process model that commits to its own causal category allocations as it makes generalizations, and assigns new observations to a new or existing category according to category sizes and objects' featural similarity:

$$P(z^{(i)}|a^{(i)}, r^{(i)}) \propto P(z_i|z^{(-i)})P(a^{(i)}, r^{(i)}|\mu^{(z_i)}). \tag{5}$$

Equation 5 is a simplified version of Equation 2, because in generalization scenarios, the resulting r' is unknown and the assignment is purely based on the basis of feature-based fit to existing categories. Instead of approximating a posterior over infinitely many possible categories, this process model maintains a small set of available categories that are created online as new generalizations are performed. Interested readers can find more implementation details in Appendix C. All code and data are openly available at https://github.com/bramleyccslab/causal_objects.

4 Evaluations

4.1 One-shot causal generalization

Data We designed a one-shot causal generalization experiment (Figure 1A–D) involving six different one-shot learning tasks (see Appendix D). We collected data from one-hundred-and-twenty participants (53 female, aged 40 ± 11) from Amazon Mechanical Turk. Each participant faced a single learning task and made 15 generalization predictions, leading to 1800 generalizations in total.

In addition, we distinguished two generalization sequences for each one-shot learning task: a *near-first transfer* and a *far-first transfer*. In the near-first transfer condition, generalizations start with cases that differ by only one feature from the learning example and progress to cases in which all of the features differ. In the far-first transfer condition, generalizations are first made about sets of objects that are completely different from those in the learning examples and progress back to the more similar cases.

Behavioral results Since there is no ground truth in this task to measure accuracy against, we use Cronbach's alpha [48] to measure inter-participant generalization consistency ρ_τ for each generalization under each learning task and order condition. Since our design is completely between-subject, this tests how peaked the distribution of responses was on average across the sample of participants in that condition and task (e.g. each corresponding to the rows in Figure 3A). Fisher's exact test confirms that participants' generalization consistency ($\rho_\tau = .80 \pm .22$) is significantly above chance, $p < 0.001$, demonstrating a robust human capacity to make systematic one-shot causal generalizations [49].
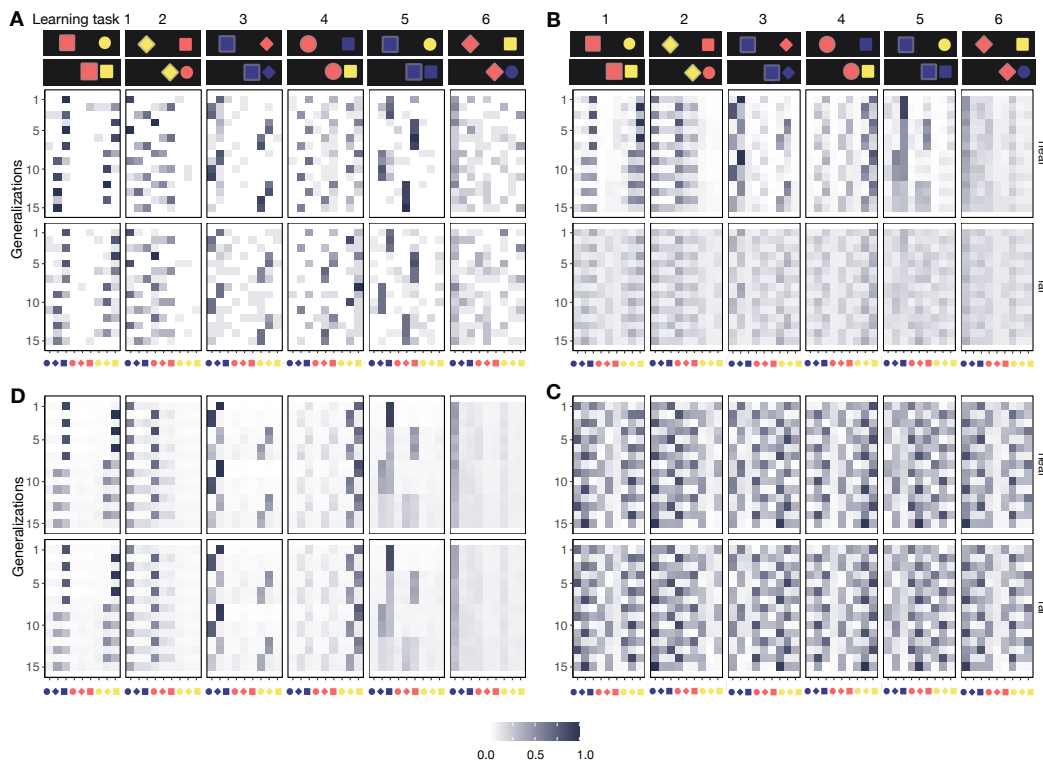
5

Figure 4: Experiment 1. A. Generalization consistency patterns for all conditions visualized as proportion of participants predicting each stone type for r' (column) on each task (row). B-C. Example LoCaLaPro predicted proportions with small α (=0.01) and large α (=8). For both figures, $\beta = 0, \gamma = 0.5$. D. Fitted LoCaLaPro predictions.

Near-first transfers induced more consistent predictions across subjects ($\rho_\tau = .83 \pm .21$), compared with far-first transfers ($\rho_\tau = .77 \pm .21$), $t(89) = 3.54, p < .001$, 95% CI = [0.03, 0.10]. Inter-person generalization consistency ρ_τ was higher for near-first transfer under all learning conditions except A6 *"Recipient changes to a new color and shape"*, for which both transfer sequences induced low agreement (Figure 3). This generalization-order effect suggests that participants may be influenced by their own generalization history in some way.

Models We fit several model variants to our choice data (Figure 4A) using maximum likelihood, and then compared them using Bayesian Information Criterion to accommodate for different numbers of parameters. The random choice *Baseline* model simply predicts $P(\text{choice} = r') = 1/9$, for the 9 candidate objects and has no parameters. The *Universal Causal Laws (UnCaLa)* model uses the PCFG-generated causal functions, assuming that the causal function governing the training case applies universally to all potential generalization scenarios, no matter how dissimilar the objects involved may be. The *Local Causal Laws (LoCaLa)* model implements the joint inference of latent causal categories, combining both symbolic causal laws and similarity-based categorizations. Finally, the *Local Causal Laws Process (LoCaLaPro)* model is the averaged predictions of our process version of the LoCaLa model.

For the two causal category models LoCaLa and LoCaLaPro, we fit hyper parameters α and β, but fixed the focus parameter $\gamma = 0.5$ because there is no information about what causal categorization assumptions should be preferred in this experiment. For all of the non-random models, we applied a softmax with a "inverse temperature" parameter t on the posterior predictives to account for response noise [50].

Model fits Table 1 summarizes model fits. Both the UnCaLa and LoCaLa models improve dramatically over the random Baseline, and the LoCaLa outperforms the UnCaLa in both log likelihood and

Table 1: Model fitting results

	α	β	γ	t	Log likelihood	BIC
			Experiment 1			
Baseline					-3955	7910
UnCaLa				6.96	-2761	5529
LoCaLa	2.41	938.81	(0.5)	9.44	-2748	5518
LoCaLaPro	0.38	1	(0.5)	10.09	**-2736**	**5494**
			Experiment 2			
Baseline					-4889	9778
UnCala			(0.5)	3.19	-3706	7417
LoCaLa	9	256	1	9.5	**-3462**	**6942**

BIC. The process model LoCaLaPro best predicts the empirical data, and as shown in Figure 4D, it indeed predicts the dominant judgment patterns among participants.

Figure 4B-C demonstrate that when concentration parameter α is small, it can reproduce a strong generalization-order effects, because each new observation will stickily join previously assigned categories (Figure 4B). When α becomes very large, however, a new observation has a high probability of being attributed to a new category (Equation 8), and the overall generalization predictions will simply approach the prior (Figure 4C). The fitted α parameter for LoCaLaPro is 0.38, confirming the presence of a dominant order effect.

4.2 Multi-shot causal generalization

Data We extended the setup in Experiment 1 to investigate inference from multiple complete observations, where each participant tested six pairs of objects before making generalizations. We controlled whether participants observe the same agent object paired with various recipient objects (fixed-agent conditions), or *vice versa* (fixed-recipient conditions). We employed two ground truth rules between-subject to counterbalance between shape and shading features (see Appendix E). One-hundred-and-sixty-three participants were recruited from Amazon Mechanical Turk. Sixty-one participants were excluded before analysis for failure to provide task-relevant responses, resulting in 102 participants (37 female, aged 35 ± 10) $\times$ 16 generalization predictions = 1632 generalizations.

Behavioral results As with Experiment 1, we measured inter-person generalization consistency ρ_T, and Fisher's exact test confirmed that across all experimental conditions, participants produced systematic generalization patterns against random guesses, $p < 0.001$. In particular, the *fixed-agent* condition induced higher consistency ($\rho_T = 0.89 \pm 0.06$) than the *fixed-recipient* condition ($\rho_T = 0.85 \pm 0.1$), $t(31) = 2.12, p = 0.04, 95\%\text{CI} = [0.001, 0.08]$, while the difference in ρ_T between the ground truth condition was negligible, $t(31) = 0.22, p = $ n.s. (see also Appendix E.3-E.4). This suggests that participants made more homogeneous predictions after observing the same agent acting on a range of recipients, and diverged more having observed different agents interacting on the same recipient, echoing a well-known inductive bias—causal asymmetry—in physical causation [51].

Models As with Experiment 1, we compared participants generalizations to a random *Baseline* model, a *Universal Causal Laws (UnCaLa)* and a *Local Causal Laws (LoCaLa)* model, again using maximum likelihood and BIC to account for different numbers of parameters. Since we randomized the presentation of both evidence and generalization trials between subjects, we do not expect systematic effects of the sort accommodated by our process model LoCaLaPro, so focus on comparison between UnCaLa and LoCaLa. Different from Experiment 1, values of $\gamma = 1, 0.5$ and 0 are of particular theoretical interest here, representing categorization based on just the agent, agent and recipient equally, or just the recipient. We also included $\gamma = 0.25$ and $\gamma = 0.75$ consistent with a mixed focus biased toward either agent or recipient.

Model fits As summarized in Table 1, both models improve substantially over the random Baseline, with LoCaLa fitting better than UnCaLa as in Experiment 1. Within LoCaLa, the best fitting γ value was 1, indicating that causal categorization was dominated by features of the agents, in line with the

asymmetric causal attribution bias suggested by our regression analyses. The fitted α for LoCaLa is 9, above chance-level probability of assigning a new causal law to each new observation, confirming the behavioral tendency to create multiple causal categories to account for the evidence. Here, $\gamma = 1$ together with $\alpha = 9$ captures the causal asymmetry observed in behavioral data: When observing multiple different agents, participants imputed many local causal laws. When seeing a single agent interact with multiple recipients, they tended to impute a single causal law. The fitted β parameter was quite large, as in Experiment 1, indicating a substantial heterogeneity across participant data taken together.

5   Discussion

In this paper, we investigated causal generalization based on observations of interactions between objects. Our two experiments demonstrated that people make systematic causal generalizations from one or a few observations and revealed some of the inductive biases that drive these. Participants' generalization patterns were well-captured by our Bayesian inference model operating on a latent space of causal laws generated by a simple Probabilistic Context Free Grammar prior favoring parsimony, and an extended Dirichlet Process that localized causal laws according to the interacting objects' features as well as their causal behaviors. Separately, these ideas extend previous work in causal inference and categorization [13, 21, 37], and in combination they give the first precise formal account of how people (1) partition the world according to causal behavior without relying on innate knowledge – an essential feature of any general model of causal learning [e.g. 11, 27]; and (2) do so in a way that is resource-efficient, requiring modest attention and memory, and supporting snap judgments, albeit at the expense of inducing order effects.

Our framework integrates a symbolic approach to represent causal law generation, with non-parametric Bayesian categorization to model latent categories, emphasizing the constructive nature of causal belief formation, in which both the content and extension of our causal concepts are generated rather than pre-specified. The constructive nature of the PCFG calls upon a potentially infinite set of possible causal functions, yet is governed by the preference for parsimony, and encourages systematic composition [see also 13, 52]. The extended Dirichlet Process for category construction goes beyond a hierarchical Baysian modeling approach where categories are pre-defined as inductive biases [e.g. 11, 22], and thus better captures the flexibility of human generalization behaviors [see also 21]. This method draws a close link with probabilistic program induction models [e.g. 14, 52–54], where causal beliefs and concepts can be viewed as programs, and accurate generalizations can be viewed as evidence for successful program synthesis whereby these programs increasingly reflect the true causal laws of nature. Moreover, our constructive computational modeling framework balances between learning a single causal law versus making generalization predictions based on multiple causal categories, and with the "creating new categories only on demand" assumption for a process account, our model successfully reproduces the generalization-order effects in behavioral data.

However, our simple task and specific modeling choices just mark the starting point for exploring causal generalization. Our modeling framework is compatible with many other options. For example, one may to extend the symbolic approach to cover the categorization process as well, or use causal Bayes net as an alternative representation for causal functions. Our task setup can be adapted to investigate causal interactions between multiple objects, reversed agent-object roles, or active learning and planning tasks. In our current work-in-progress, we are extending the existing framework to use adaptor grammar [55–57] to model organic compositional causal generalization under boostrapping conditions. We hope this line of work opens up more insights into how causality permeates the cognitive representations we use to predict, explain, and act in the world.

References

[1] Renee Baillargeon. Physical reasoning in infancy. *The Cognitive Neurosciences*, pages 181–204, 1995.

[2] Elizabeth S Spelke. Principles of object perception. *Cognitive Science*, 14(1):29–56, 1990.

[3] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental Science*, 10(1):89–96, 2007.

[4] Albert Michotte. *The perception of causality*. Basic Books, Oxford, England, 1963.

[5] Ian E Gordon, Ross H Day, and Erica J Stecher. Perceived causality occurs with stroboscopic movement of one or both stimulus elements. *Perception*, 19(1):17–20, 1990.

[6] Alan M Leslie and Stephanie Keeble. Do six-month-old infants perceive causality? *Cognition*, 25(3):265–288, 1987.

[7] Brian J Scholl and Patrice D Tremoulet. Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8):299–309, 2000.

[8] Steven A Sloman. *Causal models: How people think about the world and its alternatives*. Oxford University Press, 2005.

[9] Neil R Bramley, David A Lagnado, and Maarten Speekenbrink. Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(3):708, 2015.

[10] Alison Gopnik, Laura Schulz, and Laura Elizabeth Schulz. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press, 2007.

[11] Thomas L Griffiths and Joshua B Tenenbaum. Theory-based causal induction. *Psychological Review*, 116(4):661, 2009.

[12] Charles Kemp, Patrick Shafto, and Joshua B Tenenbaum. An integrated account of generalization across objects and features. *Cognitive Psychology*, 64(1-2):35–73, 2012.

[13] Noah D Goodman, Joshua B Tenenbaum, Jacob Feldman, and Thomas L Griffiths. A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154, 2008.

[14] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[15] Roger N Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

[16] Joshua B Tenenbaum and Thomas L Griffiths. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640, 2001.

[17] Charley M Wu, Eric Schulz, Maarten Speekenbrink, Jonathan D Nelson, and Björn Meder. Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, 2(12):915–924, 2018.

[18] Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *Twenty-fifth AAAI conference on artificial intelligence*, 2011.

[19] Elias Bareinboim and Judea Pearl. A general algorithm for deciding transportability of experimental results. *Journal of causal Inference*, 1(1):107–134, 2013.

[20] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[21] Charles Kemp, Noah D Goodman, and Joshua B Tenenbaum. Learning to learn causal models. *Cognitive Science*, 34(7):1185–1243, 2010.

[22] Noah D Goodman, Tomer D Ullman, and Joshua B Tenenbaum. Learning a theory of causality. *Psychological Review*, 118(1):110, 2011.

[23] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 1982.

[24] Johan Kwisthout and Iris Van Rooij. Computational resource demands of a predictive Bayesian brain. *Computational Brain & Behavior*, 3(2):174–188, 2020.

[25] Iris Van Rooij. The tractable cognition thesis. *Cognitive Science*, 32(6):939–984, 2008.

[26] Simon Valentin, Bonan Zhao, Chentian Jiang, Neil R Bramley, and Chris Lucas. Symbolic and sub-symbolic systems in people and machines. In *Proceedings of the 43th Annual Meeting of the Cognitive Science Society*, 2021.

[27] Christopher G Lucas and Thomas L Griffiths. Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, 34(1):113–147, 2010.

[28] Adam N Sanborn, Thomas L Griffiths, and Daniel J Navarro. Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144–1167, 2010.

[29] Zi L Sim and Fei Xu. Learning higher-order generalizations through free play: Evidence from 2-and 3-year-old children. *Developmental Psychology*, 53(4):642, 2017.

[30] Samuel GB Johnson and Woo-kyoung Ahn. Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*, 39(7):1468–1503, 2015.

[31] Kevin W Soo and Benjamin M Rottman. Causal strength induction from time series data. *Journal of Experimental Psychology: General*, 147(4):485, 2018.

[32] Mark Steyvers, Joshua B Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum. Inferring causal networks from observations and interventions. *Cognitive Science*, 27(3):453–489, 2003.

[33] Christopher G Lucas, Sophie Bridgers, Thomas L Griffiths, and Alison Gopnik. 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, 2014.

[34] Laura E Schulz and Jessica Sommerville. God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development*, 77(2):427–442, 2006.

[35] Judea Pearl. *Causality: Model, Reasoning, and Inference*. Cambridge University Press, 2000.

[36] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[37] Neil R Bramley, Peter Dayan, Thomas L Griffiths, and David A Lagnado. Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, 124(3):301, 2017.

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

[39] Bob Rehder and Reid Hastie. Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, 130(3):323, 2001.

[40] Bob Rehder. Categorization as causal reasoning. *Cognitive Science*, 27(5):709–748, 2003.

[41] John R Anderson. The adaptive nature of human categorization. *Psychological Review*, 98(3):409, 1991.

[42] Bradley C Love, Douglas L Medin, and Todd M Gureckis. Sustain: a network model of category learning. *Psychological Review*, 111(2):309, 2004.

[43] Charles Kemp and Joshua B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008.

[44] Nick Chater. *The mind is flat: the illusion of mental depth and the improvised mind*. Penguin UK, 2018.

[45] Neil Stewart, Nick Chater, and Gordon DA Brown. Decision by sampling. *Cognitive Psychology*, 53(1):1–26, 2006.

[46] S Ginsburg. *The mathematical theory of context free languages*. McGraw-Hill Book Company, 1966.

[47] David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

[48] Lee J Cronbach. On estimates of test reliability. *Journal of Educational Psychology*, 34(8):485, 1943.

[49] Charles Kemp, Noah D Goodman, and Joshua B Tenenbaum. Learning causal schemata. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, 2007.

[50] R Duncan Luce. *Individual choice behavior*. Wiley, 1959.

[51] Peter A. White. The causal asymmetry. *Psychological Review*, 113(1):132–147, 2006.

[52] Neil R Bramley, Anselm Rothe, Josh Tenenbaum, Fei Xu, and T Gureckis. Grounding compositional hypothesis generation in specific instances. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, 2018.

[53] Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sablé-Meyer, Lucas Morales, Luke Hewitt, Luc Cary, Armando Solar-Lezama, and Joshua B Tenenbaum. Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, pages 835–850, 2021.

[54] Brenden M Lake and Steven T Piantadosi. People infer recursive visual concepts from just a few examples. *Computational Brain & Behavior*, 3(1):54–65, 2020.

[55] Percy Liang, Michael I Jordan, and Dan Klein. Learning programs: A hierarchical Bayesian approach. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 639–646, 2010.

[56] Timothy J O'Donnell, Joshua B Tenenbaum, and Noah D Goodman. Fragment grammars: Exploring computation and reuse in language. 2009.

[57] Forrest Briggs and Melissa O'neill. Functional genetic programming with combinators. In *Proceedings of the Third Asian-Pacific workshop on Genetic Programming, ASPGP*, pages 110–127, 2006.

Supplementary material for paper "Building Object-based Causal Programs for Human-like Generalization"

A PCFG for symbolic causal functions

Let f be a causal function that takes the agent object a and the recipient object r as input, and outputs the result object r'. We assume a prior over possible causal functions in the form of a PCFG $\mathcal{G} = (\Gamma, \Theta)$, where Γ is a set of production rules and Θ is a set of production probabilities (see Table 2). Let ϕ_i denote the i-th feature in the set of all observable object features Φ. Grammar $\mathcal{G}$ thus generates expressions that specify features of the result object. For example, if one of the features is "color", a possible causal function could be $\texttt{color}(r') \Leftarrow \texttt{red}$ — recipient will turn red — or $\texttt{color}(r') \Leftarrow \texttt{color}(a)$ — recipient will take the agent's color, and so on. The grammar is set up to allow for arbitrarily complex expressions through the the "bind additional" production rule (Table 2, row 2), allowing a rule to produce conjunctions of feature changes, for example, $\texttt{AND}(\texttt{color}(r') \Leftarrow \texttt{red}, \texttt{shape}(r') \Leftarrow \texttt{triangle})$.

We assume that any features unspecified by a causal function follow the principle of inertia, and remain as they were before the causal interaction.

Note that although grammar $\mathcal{G}$ is very similar to a PCFG, it is not context-free strictly speaking: the "bind feature" production rule (Table 2, row 1) binds a feature to a lambda expression, and the subsequent steps within the scope of the λ-abstraction all refer to this feature.

For simplicity, we assume uniform transition probabilities for each production rule. i.e., $\theta_l = \frac{1}{|I|}$ for each row I with production rules $l \in I$. By design, this grammar is inherently more likely to produce simpler expressions. The "bind additional" rule is called with probability 0.5, and thus the number of conjunctions in the final expression follows a geometric decay with only 50% combining more than one assertion, 25% containing more than two, and so on. The prior for a given expression is thus simply the product of all the productions that produced it:

$$P_{\mathcal{G}}(f) = \prod_{l \in \Gamma} (\theta_l)^{c_l} \tag{6}$$

where $l \in I$ is the transition probability for production rule $l \in \Gamma$, and c_l is how many times rule l was used for generating this causal function.

A causal function outputs result object(s) when particular agent and recipient objects are provided. Take $\texttt{AND}(\texttt{color}(r') \Leftarrow \texttt{color}(a), \texttt{shape}(r') \Leftarrow \texttt{square})$ for example. For an agent a that is a $\texttt{red-circle}$ and a recipient r that is a $\texttt{blue-pentagon}$, r will become r': a $\texttt{red-square}$. When a causal function f involves a negation, it could have produced more than one outcome. For example let w be $\texttt{shape}(r') \Leftarrow \neg\texttt{triangle}$, any object that is not triangular (and share the same color as r) is a possible option for being r'. We further assume for simplicity that the different potential outcomes are equally probable, and thus likelihood of a data point $d = (a, r, r')$ generated by a causal function f is given by

$$P(d|f) = P(r'|f, a, r) = \begin{cases} \frac{1}{D(f(a,r))} & \text{if } r' \in D(f(a,r)), \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where D stands for *domain* and $D(f(a,r))$ refers to the set of all possible result objects coming out of f given agent a and recipient r. We initially assume a likelihood to 0 for any observation a, r, r' incompatible of $f(a,r)$, but later consider "soft" variants in which functional relationships are somewhat fallible.

This framework naturally favors deterministic causal functions that are consistent with the evidence: if a causal function predicts a specific result, when that outcome is indeed observed, likelihood will be 1. In contrast, a causal function that predicts a range of outcomes will inevitably assign a lower likelihood to any one of these.

Table 2: Example probabilistic grammar $\mathcal{G}$

Production rules			
Bind feature	$S \rightarrow$	$\lambda_{\phi_i} : A, \Phi$	
Bind additional	$A \rightarrow$	B	$\text{AND}(B, S)$
Relation	$B \rightarrow$	$\phi_i(r') \Leftarrow C$	$\phi_i(r') \Leftarrow \neg C$
Reference	$C \rightarrow$	D	E
Relative reference	$D \rightarrow$	$\phi_i(a)$	$\phi_i(r)$
Absolute reference	$E \rightarrow$	$\texttt{value}^{\phi_i}$	

Note: "Bind feature" samples a feature without replacement from the set of all features. ϕ_i in D uses the feature selected in A, and value in E is sampled uniformly from the support of the feature selected in A.

B   Unpacking latent causal categories

A CRP is a stochastic process widely used for creating partitions among entities [47]. It draws on an analogy of sequentially seating infinite incoming customers to infinitely many tables in a Chinese restaurant, where each table is also of infinite capacity. The first observation $d^{(1)}$ is always assigned the first category $z^{(1)}$; when $i > 1$, the probability for assigning category $z^{(i)}$ is given by

$$P(z^{(i)} = x|z^{(-i)}) = \begin{cases} \frac{\alpha}{i-1+\alpha} & \text{if } x \text{ is a new category} \\ \frac{|z^{(j)}|}{i-1+\alpha} & \text{if } x = z^{(j)} \end{cases} \tag{8}$$

where $z^{(j)}$ is an existing category, and $|z^{(j)}|$ is the number of assigned objects in category $z^{(j)}$. Parameter α ($\alpha > 0$) is known as the concentration, or dispersion parameter—the larger α is, the more likely a new object falls into a new category. Holding the same α, categories with more members are preferred as they seem to be more "common".

Objects in a category characterize shared feature similarities, modeled by a multinomial distribution over a finite number of feature values. Let $\mu^{(z_i)} = [\mu_1, \ldots, \mu_n]$ be the mean feature vector of a given category z_i, where each subscript k in μ_k refers to a feature, μ_k is the mean value of feature k for all the objects in category z_i, the probability that an object is assigned to a particular category according to feature similarities is given by

$$P(o^{(i)}|\mu^{(z_i)}) = \prod_{k=1}^{n} \text{Bernoulli}(o^{(i)}; \mu_k) \tag{9}$$

To compute μ_k, let $o_v = [o_{v_1}, \ldots, o_{v_n}]$ be the feature values of an object o, where each v represents a feature value, $o_{v_i} = 1$ if object o has this feature value and $o_{v_i} = 0$ otherwise. For a category $z = \{o^{(i)}, \ldots, o^{(m)}\}$, $z_v := \sum_{j=1}^{m} o_v^{(j)}$, which can be written as $z_v = [z_{v_1}, \ldots, z_{v_n}]$, where $z_{v_i} = \sum_{j=1}^{m} o_{v_i}^{(j)}$. Mean feature $\mu_k := \frac{z_{v_k}}{\sum_{l=1}^{n}(z_{v_l})}$. We assign a Dirichlet prior to this multinomial distribution in order to capture how important feature similarity is in forming categories. Without leaning towards any specific feature, the prior distribution over mean features is simply $\text{Dir}(\beta)$, $\beta \geq 0$.

It is not obvious whether mean features should be drawn from the agent object, recipient object, or both, therefore we introduce one more hyper parameter γ, referring to the probability that mean feature is purely based on the agent: when $\gamma = 1$, categorization is only grounded on the agent objects, when $\gamma = 0$, only recipient's features are considered for categorization, and when $\gamma = 0.5$, both agent and recipient are considered equally. We thus consider $0 \leq \gamma \leq 1$.

In sum, a Dirichlet Process that creates a distribution over causal category distributions according to Equation 2 has the priors:

$$\begin{aligned} z^{(i)}|z^{(-i)} &\sim \text{CRP}(\cdot|\alpha) \\ \mu^{(i)} &\sim \text{Dir}(\cdot|\beta) \\ f^{(z_i)} &\sim \mathcal{G}(\cdot) \end{aligned} \tag{10}$$

Algorithm 1 Process model

1: Initialize an empty list of causal categories Z ▷ Initialization
2: Assign $a^{(0)}, r^{(0)} \in d^{(0)}$ to category $z^{(1)}$, update $\mu^{(1)}$ ▷ Learning example goes to the first
     category
3: Sample $f^{(1)}$ from the learning posterior
4: Record $z^{(1)}$ in list of causal categories Z
5: **for each** $d^{(i)} \in D_G$ **do**
6:      sample $z^{(i)} \propto P(z^{(i)}|z^{(-i)})P(a^{(i)}, r^{(i)}|\mu^{(z_i)})$ ▷ Equation 5
7:      **if** $z^{(i)} \in Z$ **then**   ▷ If current object belongs to an existing category
8:          $r'^{(i)} \sim f^{(i)}(a^{(i)}, r^{(i)})$ ▷ Make prediction
9:          Add $a^{(i)}, r^{(i)}$ to $z^{(i)}$: update $\mu^{(i)}$ ▷ Update Z
10:      **else**
11:          Assign $a^{(i)}, r^{(i)}$ to a new category $z^{(k)}$: update $\mu^{(k)}$ ▷ Create a new category
12:          Sample $f^{(k)}$ from the prior
13:          $r'^{(i)} \sim f^{(k)}(a^{(i)}, r^{(i)})$ ▷ Make prediction
14:          Add $z^{(k)}$ to Z ▷ Update Z
15:      **end if**
16: **end for each**

And likelihoods are given by

$$
\begin{aligned}
a^{(i)}, r^{(i)}|\mu^{(z_i)} &\sim \text{Dir}(\,\cdot\,|\mu^{(z_i)}, \beta) \\
d^{(i)}|f^{(z_i)} &\sim f^{(z_i)}(a^{(i)}, r^{(i)})
\end{aligned}
\tag{11}
$$

where $\mu^{(z)}$ is the mean feature vector, and $f^{(z)}$ the assigned causal function.

To approximate the posterior with Gibbs sampling, we construct a chain of samples where for each iteration, we sample a causal category for a random observation $d^{(i)}$ while fixing the category assignment to the other observations, and a sampled causal category $z^{(i)}$ will then update the category parameters $\mu^{(z_i)}$ and $f^{(z_i)}$. The category sampling step of this Gibbs sampler follows Equation 2, and the local parameter update step follows definition of computing these parameters given objects in this category. When the number of iterations $n \to \infty$, the sampled categories $\tilde{Z}_n$ coverges to the true posterior.

C   Process variant

The process model first assigns the object-pair in the learning example to an initial causal category $z^{(1)}$ governed by a causal law sampled from the posterior distribution $P(f|d)$. Crucially, for each generalization task, it then assigns the encountered object pair scenario to either an existing causal category or a new category according to Equation 5. If an existing causal category is selected, the model simply applies the requisite causal law category to make its prediction. If a new category is sampled, however, a new causal law will be assigned to this category. Since there is no evidence about what causal law may apply to this new category, this new causal law is sampled from the prior. Algorithm 1 shows this process.

Instead of approximating a posterior over infinitely many possible categories as the normative model, the process model maintains a small set of available categories that are created online as new generalizations are performed. Furthermore, after categorizing an observation, the process model updates the list of causal categories Z with this categorization decision, reflecting a commitment to its earlier decisions. Hyperparameter α thus plays a slightly different role in the process model. When $\alpha \to 0$, the process model becomes increasingly likely to stick with existing categories (Equation 8).
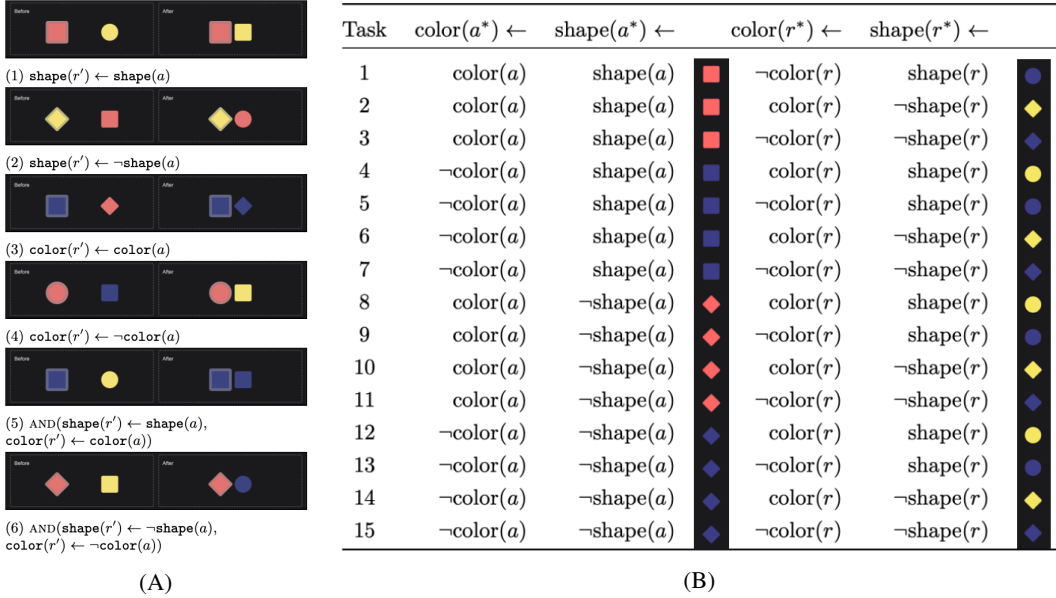
(A)

(1) shape(r') ← shape(a)

(2) shape(r') ← ¬shape(a)

(3) color(r') ← color(a)

(4) color(r') ← ¬color(a)

(5) AND(shape(r') ← shape(a), color(r') ← color(a))

(6) AND(shape(r') ← ¬shape(a), color(r') ← ¬color(a))

(B)

Task	color(a^*) ←	shape(a^*) ←	color(r^*) ←	shape(r^*) ←
1	color(a)	shape(a)	¬color(r)	shape(r)
2	color(a)	shape(a)	color(r)	¬shape(r)
3	color(a)	shape(a)	¬color(r)	¬shape(r)
4	¬color(a)	shape(a)	color(r)	shape(r)
5	¬color(a)	shape(a)	¬color(r)	shape(r)
6	¬color(a)	shape(a)	color(r)	¬shape(r)
7	¬color(a)	shape(a)	¬color(r)	¬shape(r)
8	color(a)	¬shape(a)	color(r)	shape(r)
9	color(a)	¬shape(a)	¬color(r)	shape(r)
10	color(a)	¬shape(a)	color(r)	¬shape(r)
11	color(a)	¬shape(a)	¬color(r)	¬shape(r)
12	¬color(a)	¬shape(a)	color(r)	shape(r)
13	¬color(a)	¬shape(a)	¬color(r)	shape(r)
14	¬color(a)	¬shape(a)	color(r)	¬shape(r)
15	¬color(a)	¬shape(a)	¬color(r)	¬shape(r)

Figure 5: Experiment 1 stimuli. (A): Learning conditions, showing objects before and after a causal interaction. (B): Generalization task configurations a^*, r^* are the agent and recipient in each generalization task; a and r are the agents and recipients in the learning example. Example stones are for learning condition A1.

D   Experiment 1 setup

D.1   Stimuli and design

Participants were told that they were making predictions about the behavior of a magic world containing magic stones (agents) and normal stones (recipients). In short videos, participants observed a magic stone collide with a normal stone and appear to alter the normal stone's color and/or shape (see Figure 1). Magic stones had a thick border while normal stones had no border. We manipulated two object features—color {*red, yellow, blue*} and shape {*circle, square, diamond*}, leading to $3 \times 3 = 9$ possible configurations for each object and a nominal $9 \times 9 \times 9 = 729$ configurations of agent and of recipient both pre- and post- the causal interaction.

We used a 6×2 between-subject design. There were six learning examples varied between subjects (Figure 5A)—each participant saw one. Each learning example demonstrates a causal effect differing in whether it results in a change to one or both features of the recipient object, and whether either or both of these new values match the agent object's features. Note that the function descriptions were not shown to participants and are by no means the only possible way to characterise the causal relationship being displayed.

For each learning example, we constructed 15 generalization tasks by varying object features systematically from the learning example (Figure 5B). For example, A1 (see Figure 5A) depicts a *red square* agent and a *yellow circle* recipient, and according to the specifications in Figure 5A, task 1 for A1 has a *red square* agent, and a *blue circle* recipient. We call the sequence of tasks from 1 to 15 "near-first transfer" because this sequence of tasks starts with those that differ by only one feature from the learning example and progress to scenes in which all of the features differ. Conversely, we call the sequence of tasks 15 to 1 the "far-first transfer" sequence, because it starts with sets of stones that are completely different from those in the learning examples and progresses back to the more similar cases. Within each sequence, whether the set of different-color tasks or the set of different-shape tasks appeared first (task 1 & 2, 5 & 6, 9 & 10, 13 & 14, 4—7 & 8—11) was shuffled to counterbalance feature order.

(1) Rule 1, fixed agent				(2) Rule 1, fixed recipient		
(3) Rule 2, fixed agent				(4) Rule 2, fixed recipient		

Figure 6: Experiment 2 learning conditions.

D.2 Procedure

After instructions, participants had to pass a comprehension quiz to start the main task. The main task contained a learning phase and a generalization phase. During learning, participants watched one specific magic stone's effect on a normal stone (Figure 1A–C), and they could replay the effect as many times as they wanted. After that, participants were asked to make predictions for 15 new pairs of magic stones and normal stones sequentially, by selecting from a panel of 9 possible stones (Figure 1D). A summary of the learning example (as used in Figure 5A) was displayed at all times and the animation was replayed once between each generalization task to ensure it was not forgotten. A demo of the task is available at `http://bramleylab.ppls.ed.ac.uk/experiments/bnz/magic_stones/index.html`.

D.3 Model fits

Both the UnCala and LoCaLa models were fit to the behavioral data using the `optim` function in R. As for the LoCaLaPro model, since it approximates posterior distribution with simulation-based method, we optimized parameter values via grid search. Firstly, we set up a coarse grid with α = {0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.5, 2, 4, 8}, β ={0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024}. After running this coarse grid and locating an optimal area, we ran another search over a finer grid for $\alpha = \{0.28, 0.30, 0.32, 0.34, 0.36, 0.38, 0.4, 0.42, 0.44, 0.46, 0.48, 0.5, 0.52\}$ (β is the same as previously) to improve precision.

Table 3: Experiment 2 generalization task configurations

	For the fixed object	Instance	For the varied object	Instance
$o^* =$	$\text{shade}(o), \text{edge}(o)$		$\text{shade}(o), \neg\text{edge}(o)$	
	$\neg\text{shade}(o), \text{edge}(o)$			
	$\text{shade}(o), \neg\text{edge}(o)$		$\neg\text{shade}(o), \text{edge}(o)$	
	$\neg\text{shade}(o), \neg\text{edge}(o)$			

o^* is the object in generalization tasks, o is the object shown during learning. For the varied object, $\neg\text{shade}(o)$ means picking a shade that has not appeared during the learning phase, and we chose two instances for it.

E Experiment 2 setup

E.1 Stimuli and design

Similar to Experiment 1, we varied the shape and color properties of the objects. However, instead of using categorical values, we introduced intuitively ordinal feature values. Shapes were all equilateral and differed in terms of their number of sides: 3 (triangle), 4 (square), 5 (pentagon), 6 (hexagon), and 7 (heptagon); colors were of identical hue and saturation (blue) but differed in lightness varying between: 1 (light blue `#c9daf8`), 2 (medium blue `#6d9eeb`), 3 (dark blue `#1155cc`), and 4 (very dark blue `#052e54`). Staying within the features' observed values this leads to $4 \times 4 = 16$ possible configurations for each object, and a nominal $16^3 = 4096$ possible configurations for objects both pre- and post- the causal interaction. These ordinal features enlarge the space of effects and greatly enriches the space of plausible rules, for example allowing causal laws in which a recipient stone becomes *darker* or *lighter* when acted upon, gaining or losing sides, as well as those involving copying or taking specific or random values.

During learning, each participant observed six causal interactions between different pairs of agent and recipient before making generalizations. We included 2 (evidence-balance) $\times$ 2 (ground truth) between-subject factors (see Figure 6). with respect to evidence-balance, for *fixed-agent* conditions B1 and B3, an identical agent was shown in all learning examples, while the recipients it acted on were varied systematically; in the *fixed-recipient* conditions B2 and B4, the recipient object was always identical but was acted on by six different agents. We designed the evidence to be consistent with two "ground truth" rules that counterbalance between the roles of the shape and the color features:

Rule 1 (B1/B2) The recipient gets one increment darker and takes the agent's shape plus one edge
$\text{AND}(\text{edge}(r') \leftarrow \text{edge}(a) + 1, \text{shade}(r') \leftarrow \text{shade}(r) + 1)$

Rule 2 (B3/B4) The recipient gains an edge and takes the agent's shade plus one shade increment
$\text{AND}(\text{shade}(r') \leftarrow \text{shade}(a) + 1, \text{edge}(r') \leftarrow \text{edge}(r) + 1)$

Note that these "ground truth" rules are just one of an unbounded set of possible universal causal relations consistent with the six learning trials, and a single universal category is just one of a much larger set again of possible local causal law category structures.

We composed generalization tasks according to the configurations in Table 3. In total there were $4 \times 4 = 16$ generalization tasks for each condition.

E.2 Procedure

After completing instructions, participants had to pass a comprehension quiz to proceed to the main task, consisting of a learning phase, self-report, and a generalization phase. After the main task, participants provided demographic informatio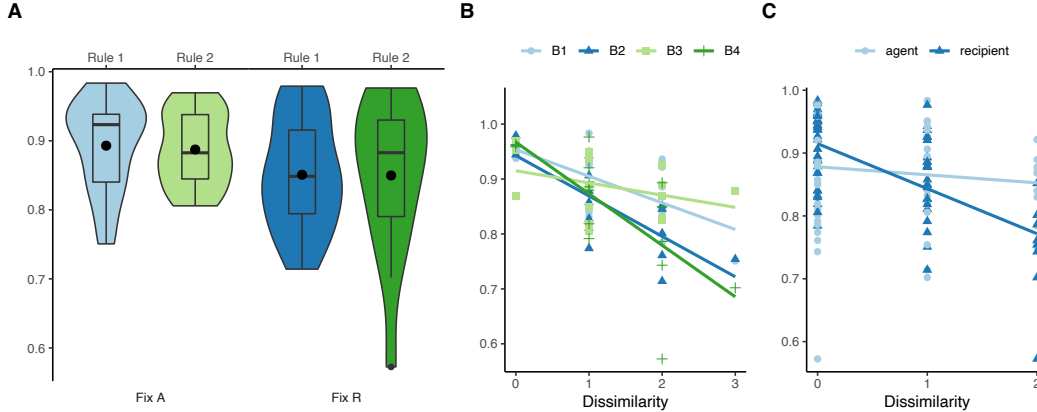n and feedback. A demo of the task is available at `http://bramleylab.ppls.ed.ac.uk/experiments/bnz/myst/p/welcome.html`.

Figure 7: Behavioral results of Experiment 2. All y-axes are Cronbach's alpha values. A. Task-wise inter-person consistency per condition. Violin plots are density. Black dots are mean Cronbach's alpha values per condition. The major bar in the box plot is median and box extent is the 25 and 75 quantiles. B. Inter-person consistency per task differences. C. Inter-person consistency per role differences.

Each participant was randomly assigned to one of the four learning conditions (Figure 6). The six pairs of agent and recipient stones were shown in random order, one after another. By clicking a "Test" button, they could watch the causal interaction as many times as they wanted. After each object pair was tested, a summary visualization of the agent, recipient and the result was added to the top of the page (see Figure 1E–F), and remained visible for the rest of the task. After the learning phase, participants were asked to write down their best guess about how the mysterious stones worked, and told they would receive a $0.50 bonus if they described the true underlying causal law. In the generalization phase, participants faced the 18 generalization trials sequentially in random order. For each, participants predicted the result recipient by selecting a number of edges and the shade of blue from two drop-down menus (see Figure 1F). Participants were instructed they would receive a $0.10 for each correct prediction. We bonused participants as described afterwards.

E.3 Generalization consistency

As with Experiment 1, we measured inter-person consistency in generalization predictions computing ρ_T for the sixteen generalization tasks per condition (excluding the two catch-trials), totalling $4 \times 16 = 64$ values. Mean consistency was $\rho_T = 0.87 \pm 0.08$, with min $\rho_T = 0.57$, max $\rho_T = 0.98$. To compare generalization consistency against random selections, for each condition we conducted Fisher's exact test on the contingency table of selecting each possible result per trial. For all four conditions, $p < 0.001$. Thus, as in Experiment 1, participants produced systematic generalization patterns.

We then compared inter-person generalization consistency by condition. As illustrated in Figure 7A, the *fixed-agent* condition induced higher consistency ($\rho_T = 0.89 \pm 0.06$) than the *fixed-recipient* condition ($\rho_T = 0.85 \pm 0.1$), $t(31) = 2.12, p = 0.04, 95\%\text{CI} = [0.001, 0.08]$, while the difference in ρ_T between the ground truth condition was negligible, $t(31) = 0.22, p = $ n.s.. No interaction was detected. In short, participants made more homogeneous predictions after observing the same agent acting on a range of recipients, and diverged more having observed different agents interacting on the same recipient.

Generalization consistency decreased as objects in the generalization tasks become more distinct from those in the learning examples (Figure 7B). To show this, we constructed a rough measure of *dissimilarity*, by counting the features of generalization trials that took novel values never observed in the learning phase. Formally, let F_L be the set of unique feature values of all the objects appeared during learning, and F_i be the set of unique feature values of objects in a generalization trial i, dissimilarity score $DS = |F_i \setminus F_L|$. By design, dissimilarity scores $DS \in \{0, 1, 2, 3\}$ (Table 3). We found a significant negative relationship between task dissimilarity and generalization consistency, $\beta = -0.06, F(1, 62) = 37.48, p < 0.001$.
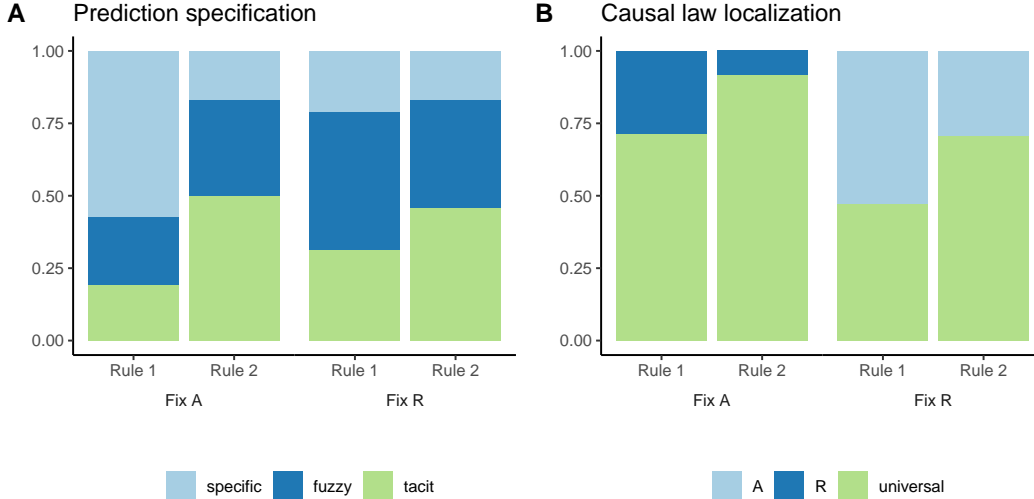
A Prediction specification

B Causal law localization

Figure 8: Rule guess categories.

Finally, we fit a linear regression model predicting ρ_T with task dissimilarity, evidence-balance and ground truth, $F(3, 60) = 15.63, p < 0.001$. This revealed main effects of dissimilarity ($\beta = -0.06, p < 0.001$) and evidence-balance (*fixed-recipient*, $\beta = -0.04, p = 0.01$), but not ground truth (*rule 2*, $\beta = -0.003, p =$ n.s.). As depicted in Figure 7B, consistency of judgments in the fixed-agent conditions (B1 & B3, lighter lines) decreased slower than the fixed-recipient conditions as dissimilarity increased (B2 & B4, darker lines).

Not only did the evidence-balance condition have a significant effect on generalization consistency, dissimilarity of the agent or recipient objects in the generalization tasks was also associated with lower consistency (Figure 7C). Holding recipient dissimilarity constant, increasing agent dissimilarity does not predict prediction consistency significantly, $F(1, 62) = 0.77, p =$ n.s.; however, recipient dissimilarity does, $F(1, 62) = 38.8, p < 0.001$.

E.4 Self-reports

In Experiment 2, we asked participants to provide an explicit free-text guess about the nature of the causal relationship(s) being tested after they completed the learning phase. Eighty-six percent of these total responses (88/102) were compatible with the relevant learning observations, and here we only analyze these. Two independent coders categorized participants guesses according to their specificity and implicit localization of causal powers. The first coder categorized all free responses, and the categorized 15% were then compared against the first coder's. Agreement level was 92%. The full set of free responses and the detailed coding scheme are available at https://github.com/bramleyccslab/causal_objects.

Since our ground truths are not the only rules consistent with the learning data, we analyzed participant self-reports not according to whether they got the ground truths right, but whether their own rules were consistent with the learning data, as well as the level of generality in the reports. Hence, we first defined three exclusive and exhaustive response specificity categories: *specific*, *fuzzy*, and *tacit*. A *specific* self-report would predict a unique result object for any potential combination of agent and recipient (for example "The inactive shape is always changed to a pentagon & its shade is changed to one step darker than the active stone"). A *fuzzy* rule was one that left open for more than one possible result objects (for example "It will be different colors and shapes"). We distinguished a second form of under-specified self-report, *tacit*, if it left a feature unmentioned, which depending on background assumptions might be taken to imply that feature remained unchanged but could also be compatible with it taking some new or random value (for example "The active stone adds a side to the inactive stone").

We also had the coders categorize responses according to whether and how a self-report localized the domain of the causal law asserted. Concretely, we included four labels *A, R, AR*, and *universal*.

A8

If a response mentioned a specific context of influence, typically using an *if...* clause, we labelled this according to whether the context mentioned the Agent (e.g. "If the active stone is darker than the inactive stone, it turns the inactive stone darker"), Recipient (e.g. "The active stone causes the other stones to change into a pentagon shape, unless it is already a pentagon shape, in which case it makes it darker"), or both. If a response made no localization or context (e.g. "The active stone cause inactive stones to five sided stone") then it was labeled as *universal*.

Figure 8 illustrates the coding results by learning condition. Guess specificity is summarized in Figure 8A. We fit a multinomial logistic regression model predicting specificity by evidence-balance and ground truth factors, and found that when taking the *specific* self-report type as baseline, the ground truth factor is a significant predictor for the *tacit* type ($\beta = 1.54, p = 0.008$), while evidence-balance is not ($\beta = 0.78, p = $ n.s.). Neither of these two factors are significant for the *fuzzy* type. Figure 8B summarizes participants' guesses in terms of localization. No participant localized their rule in terms of both Agent and Recipient. Unsurprisingly, whenever localization occurred, it was applied with respect to the object that varied during the learning phase. A logistic regression predicting universal rule probability by condition showed that both evidence-balance (*fixed-recipient*, $\beta = -1.21, z = -2.3, p = 0.02$) and ground truth (*rule 2*, $\beta = 1.17, z = 2.3, p = 0.02$) were associated with more universal rules. There was no evidence for an interaction, $z = -0.5, p = $ n.s..

E.5   Model fits

We extended the grammar used in Experiment 1 to cover a larger space of ordered feature relationships. Concretely, we introduced +1, -1, >, < at the "bind relation" step to accommodate potential assertions about the ordering of feature values used in this experiment. As in Experiment 1, LoCaLa was expensive to evaluate so we optimised its parameters using a coarse grid search. Since there were six data points, during each iteration of the Gibss sampler, when $\alpha = 5$ this observation has a half-half chance to create a new causal category or join the rest, in terms of category size preference, and this chance grows as α increases (Equation 8). Therefore, we centered the support values for α round 5, with an exponential increase for larger values, resulting in $\alpha \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 16, 32, 64, 128, 256\}$. β takes the same range of values as in fitting the models in Experiment 1. For γ, values of $\gamma = 1, 0.5$ and 0 are of particular theoretical interest, representing localization based on just the agent, agent and recipient equally, or just the recipient. We also included $\gamma = 0.25$ and $\gamma = 0.75$ consistent with a mixed focus biased toward either agent or recipient.