



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.



THE UNIVERSITY
of EDINBURGH

A Computational Framework of Human Causal Generalization

Bonan Zhao

Doctor of Philosophy
University of Edinburgh
2023

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text. This work has not been submitted for any other degree or professional qualification.

Bonan Zhao
January 4, 2023

Lay summary

People naturally see causes and effects, and this comes in handy when we need to predict what will happen with new things. For example, you may see a match light up a candle, and when you put this glowing candle near a curtain, you may picture the candle causing a fire on the curtain. For the sake of safety, you may decide this is not a good place to put lit candles. This thesis studies this ability of generalizing cause-and-effect relationships from observed objects to new objects.

Since I am interested in the generalization of cause-and-effect relationships, I conducted experiments in which I first show people the effects of some objects interacting and then ask people to guess what happens for new interacting objects. I created several computer models to understand how people make these guesses. Overall, I found that even though these guesses are made for new objects and any answer could be correct, people tend to share strikingly similar ideas. However, interestingly, people making guesses in different orders ended up making different guesses, even though they have seen the same demonstrations (Chapter 2). This tells us that the order in which we guess things could seriously affect our future decisions. In addition, people think of cause objects and effect objects differently: One cause object can change various effect objects in the same way, but one effect object is likely to be changed in many different ways by various cause objects (Chapter 3). Next, I moved on to how people make increasingly sophisticated guesses. I offered a computer model that can grow its pool of ideas over time, making use of what it saw earlier to interpret later evidence. Inspired by this model, I designed a series of experiments and showed that people seeing the same evidence in different orders can end up with very different conclusions, and their differences are well-captured by my model (Chapter 4).

In sum, I studied how people make sense of new situations by applying what they learned in the past, especially using knowledge of cause-and-effect relationships. The computer models I developed in this thesis provide a precise account of how people make such decisions, and shed light on how to include this powerful skill in artificial intelligence agents.

Abstract

How do people decide how general a causal relationship is, in terms of the entities or situations it applies to? How can people make these difficult judgments in a fast, efficient way? To address these questions, I designed a novel online experiment interface that systematically measures how people generalize causal relationships, and developed a computational modeling framework that combines program induction (about the hidden causal laws) with non-parametric category inference (about their domains of influence) to account for unique patterns in human causal generalization. In particular, by introducing adaptor grammars to standard Bayesian-symbolic models, this framework formalizes conceptual bootstrapping as a general online inference algorithm that gives rise to compositional causal concepts.

Chapter 2 investigates one-shot causal generalization, where I find that participants' inferences are shaped by the order of the generalization questions they are asked. Chapter 3 looks into few-shot cases, and finds an asymmetry in the formation of causal categories: participants preferentially identify causal laws with features of the agent objects rather than recipients, but this asymmetry disappears when visual cues to causal agency are challenged. The proposed modeling approach can explain both the generalization-order effect and the causal asymmetry, outperforming a naïve Bayesian account while providing a computationally plausible mechanism for real-world causal generalization. Chapter 4 further extends this framework with adaptor grammars, using a dynamic conceptual repertoire that is enriched over time, allowing the model to cache and later reuse elements of earlier insights. This model predicts systematically different learned concepts when the same evidence is processed in different orders, and across four experiments people's learning outcomes indeed closely resembled this model's, differing significantly from alternative accounts.

To my parents

Acknowledgement

Most part of my PhD was during the COVID-19 pandemic. Two months into my PhD, COVID started spreading in China. After another three months, the UK went into a national lockdown that eventually lasted two years. However difficult life can be, I am fortunate to owe heartfelt gratitude to many people.

My principal advisor, Neil Bramley, for his unbeatable excellence in all aspects: as an advisor, a collaborator, a friend, and a role model to live up to. In addition to his impeccable support and guidance throughout, Neil's faith in me made me grow beyond even my own expectations. Thank you Neil.

Chris Lucas taught me how to do approximate Bayesian inference and introduced adaptor grammars to me. More importantly, Chris's attitude towards doing good science has influenced me more than just technical skills.

My lab mates, Tia Gong and Tadeq Quillien, have been my unofficial advisors ever since, teaching me how to run psychological experiments, how to analyze data, how to write, and sharing many KFC meals on the night of deadlines.

Friends in the Edinburgh CoCoSci community: Jan-Philipp Fränken, who showed me how to run MCMC and write a generative grammar in Python; Chentian Jiang and Simon Valentin, who shared many discussions about probabilistic machine learning; Steph Droop, who provided nondepletable advice on exploring Scotland; Aba Szollosi, a constant source of stimulating ideas and perspectives; and Zach Horne, who introduced to me the best craft beer, the best French bread, and the best Mexican tacos in Edinburgh.

I also owe many thanks to Frank Mollica, who always gives me the most honest advice when I'm in doubt, and Alex Dumas, who taught me that being critical and nice are not two contradictory virtues.

While I was not able to see my actual family for three years, the Chinese community in the department and my lovely flatmates made me feel at home. Thanks to Elva Peng, Zoey Zhang, Zhe Huang, Border Zhang, Zhuoni Xiao, Xinxin Zhu, Xiaomeng Zhang, Wei Xu, and so many others who cooked and shared great Chinese food with me and partied for every Chinese festival.

And the people at the ILLC: Jakub Szymanik, Sonja Smets, Johan van Benthem, Zoi Terzopoulou, Anthi Solaki, Fenrong Liu, without whom I would not have been able to do a PhD in cognitive science in the first place. My former colleagues at Screen 6, Samir Aleido and Bob van Toorn, who taught me everything about programming that gave me the confidence to work with computational models.

The CHSS/PPLS research award generously supported three years of my PhD, and the Great Britain China Educational Trust provided extra funding for my write-up period. Special thanks to Eric Schulz and the Max Planck Institute for funding my short visit to Tübingen, and the LJDM organizer, Tianshu Chen, for enabling my very first in-person talk after years of working from home.

Finally, my utmost gratitude goes to my family, for everything.

Published and submitted articles

This thesis is based on the following articles that I have worked on during my doctoral studies. Asterisks denote equal contribution.

Experiment 1 presented in Chapter 2 appears in: Zhao, B., & Bramley, N. R. (2020). Order effects in one-shot causal generalization. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 3071.

Computational models presented in Chapter 2 appear in: Zhao, B., Lucas, C. G., & Bramley, N. R. (2021). Building object-based causal programs for human-like generalization. *Causal Inference & Machine Learning Workshop, NeurIPs 2021*, 1–11.

Experiment 2 presented in Chapter 3 appears in: Zhao, B., Lucas, C. G., & Bramley, N. R. (2022). How do people generalize causal relations over objects? a non-parametric Bayesian account. *Computational Brain & Behavior*, 5, 22–44.

Experiment 3 presented in Chapter 3 appears in: Xia, Z*, Zhao, B*, Quillien, T., & Lucas, C. G. (2022). Dissecting causal asymmetries in inductive generalization. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, 592–598.

Computational models presented in Chapter 4 appear in: Zhao, B., Bramley, N. R., & Lucas, C. G. (2022). Powering up causal generalization: A model of human conceptual bootstrapping with adaptor grammars. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, 1819–1826.

Experiments 4-7 presented in Chapter 4 appear in: Zhao, B., Lucas, C. G., & Bramley, N. R. (2023). How cognition bootstraps its way to complex concepts. *Under review*.

Contents

Lay summary	1
Abstract	3
Acknowledgement	5
List of Figures	11
List of Tables	14
1 Introduction	15
1.1 Drawing boundaries for causal laws	17
1.2 Causal roles as inductive biases	19
1.3 Bootstrapping to complex causal concepts	20
2 Generalizing causal relations from single observations	23
2.1 Experiment 1: one-shot causal generalization	24
2.1.1 Method	25
2.1.2 Results	28
2.1.3 Interim discussion	30
2.2 Computational models	31
2.2.1 Universal Causal Laws (UnCaLa)	31

2.2.2	Localized Causal Laws (LoCaLa)	36
2.2.3	Local Causal Law Process model (LoCaLaPro)	42
2.2.4	Model fits	43
2.3	Discussion	49
3	Causal roles in few-shot causal generalization	52
3.1	Experiment 2: few-shot causal generalizations	52
3.1.1	Methods	53
3.1.2	Results	57
3.1.3	Model fits	62
3.1.4	Interim discussion	65
3.2	Experiment 3: Dissecting causal asymmetries	67
3.2.1	Methods	68
3.2.2	Results	70
3.2.3	Interim discussion	73
3.3	Discussion	74
4	Bootstrapping compositional causal generalization	76
4.1	Computational model	77
4.1.1	Causal programs in combinatory logic	79
4.1.2	Bootstrapping with adaptor grammars	83
4.1.3	Task and predictions	87
4.2	Experiments 4 & 5: Curriculum-order effect	88
4.2.1	Methods	89
4.2.2	Results	93
4.3	Experiments 6 & 7: Biases in compositions	97
4.3.1	Methods	97
4.3.2	Results	98

4.4	Model comparison	100
4.4.1	Alternative models	100
4.4.2	Model fits	104
4.5	Discussion	109
5	Conclusion	113
5.1	Causal representations	114
5.2	Constructive cognition	117
5.3	Final words	120
	References	122
	Appendices	137
A	Comparison with GPT-3	138
B	Open resources	143

List of Figures

2.1	The magic stone task.	24
2.2	Experiment 1 material.	26
2.3	Experiment 1 behavioral results. A. Mean ρ_τ (y-axis) for each learning scene condition and sequence order: light blue = <i>near-first transfer</i> , dark blue = <i>far-first transfer</i> . Error bars are standard errors. ρ_τ for random selection is 0. B. Density distribution of task-wise ρ_τ , per match (light blue)/new (dark blue) groups. C. Density distribution of task-wise ρ_τ , per color change (light blue)/shape change (dark blue) groups.	29
2.4	Visualization of the Localized Causal Laws (LoCaLa) model.	37
2.5	Experiment 1 results. A. Generalization patterns for all conditions visualized as proportion of participants predicting each stone type for r' (column) on each task (row). B. Fitted LoCaLaPro predictions. C-D. Example LoCaLaPro predicted proportions with small α (=0.01) and large α (=8). For both figures, $\beta = 0, \gamma = 0.5$	45
3.1	Experiment 2 interface (A) and learning material (B).	53

3.2	Experiment 2 behavioral results. All η -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 the median and the box extent is the 25 and 75 quantiles. B. Inter-person consistency per task differences. C. Inter-person consistency per role differences.	57
3.3	Experiment 2 rule guess categories.	61
3.4	Experiment 2 generalization predictions. Upper row: proportion of participants predicting each stone type for r' (column) on each task (row). Second row: Fitted LoCaLa predictions.	63
3.5	Experiment 3 cue illustrations.	69
3.6	Experiment 3 results. A. Generalization congruency per condition; y-axis is task-wise Cronbach's alpha value. B. Self-report labels with respect to which object's features were mentioned for inference.	71
4.1	Bootstrap learning. A. Forming complex concepts by extending existing concepts to account for novel observations. Without the process of bootstrapping, the world may be too complex to conceptualize. Dots represent data points. B. Visual summary of the model. G is a generative grammar and L a latent conceptual library.	78
4.2	Experiments 4-7 example task. A. Task animation. B. Visual summary of the animation.	80
4.3	Model AG components. A. Visualization for CL routers, adapted from Liang et al. (2010). B. Example program. Black arrows denote routing of the agent (magic egg), and dotted arrows for routing of the recipient (stick). C. Example frame (dotted box) and example programs. Shaded area in the bottom program reuses the program on top-right.	82

4.4	Model AG example bootstrap learning trajectories over six observations, either i-iii from simple to complex, or iv-vi the other way around.	87
4.5	Experiment 4 material and procedure. A.Generalization trials. B. Procedure in one phase. i. Test causal interactions. ii. Collect self-report. iii. Make generalization predictions by clicking on a block of segments.	90
4.6	Experiments 4-5 material and results. A. Curricula in Experiment 4 (left) and 5 (right). Texts below each phase are data-compatible causal concepts. B. Participants generalization accuracy (match to ground truth). Box plots show the first and third quantiles with lines for the medians; red dots mark the means. C. Coded self-reports.	93
4.7	Experiments 6-7 material and results. A. Curricula in Experiment 6 (left) and 7 (right). Texts below each phase are data-compatible causal concepts. B. Participants generalization accuracy (match to ground truth). Box plots show the first and third quantiles with lines for the medians; red dots mark the means. C. Coded self-reports.	98
4.8	Experiments 4-7 model fitting results. A. Total log likelihood improvement over random baseline ($y=0$), log scale. B-C. Generalization accuracy per curriculum and phase. X-axis are model predictions, y-axis people's. D. Generalization accuracy between people (black bars) and four Bayesian-symbolic models.	106
4.9	Experiments 4-7 generalization predictions by people (bars) and the best fitting AGR model (area). Rows of panels are for experimental phases, columns for conditions. In each panel, x-axis are predicted number of segments (0-16), y-axis are tasks.	109

List of Tables

2.1	Example probabilistic grammar \mathcal{G}	33
2.2	Experiment 1 model fitting results	47
3.1	Experiment 2 generalization task configurations	55
3.2	Experiment 2 self-reports coding scheme	59
3.3	Experiment 2 model fitting results	65
4.1	Model AG base terms	81
4.2	Experiments 4-7 self-reports coding scheme	95
4.3	Experiments 4-7 model fitting results	107
A.1	GPT-3 guesses about causal relationships	142

Chapter 1

Introduction

Fairy tales [...] are more than true. Not because they tell us that dragons exist, but because they tell us that dragons can be defeated.

Neil Gaiman, *Smoke and Mirrors*

A fundamental goal of cognition is to generalize from limited experience so as to behave appropriately in unpredictable future tasks and situations. We achieve this, in part, by constructing models of the environment that provide reliable predictions (Craik, 1952; Hume, 1740). While a wealth of research has been devoted to studying how children and adults acquire causal beliefs (e.g., Bramley et al., 2015; Gopnik et al., 2007; Griffiths & Tenenbaum, 2009; Kemp et al., 2012; Sloman, 2005) and generalize functional properties (e.g., Goodman et al., 2008; Lucas et al., 2015; Shepard, 1987; Tenenbaum & Griffiths, 2001; Wu et al., 2018), the interplay between causality and generalization has received less attention. On the face of it, this is surprising. If causal beliefs did not frequently extend to novel entities and situations, they would be of limited use to us. Therefore, 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" (Sloman, 2005), the true causal laws that govern both

experienced and novel situations. While research has explored the interplay between causality and generalization using hierarchical Bayesian models (e.g., Goodman et al., 2011; Griffiths & Tenenbaum, 2009; Kemp et al., 2010), this computational level approach (Marr, 1982) is limited in its ability to capture psychological processes due to its inherent intractability (Kwisthout & Van Rooij, 2020; Van Rooij, 2008). In particular, for these hierarchical approaches, the number of possible structures and parameter combinations grows super-exponentially as the number of nodes increases, even within a fixed class of structures, let alone when each node could be paired with multiple features.

This thesis explores how people generalize causal relations from observed interactions between pairs of simple geometric objects, and provides a computational modeling framework for object-based causal inference in the spirit of Griffiths and Tenenbaum (2009) and Lucas and Griffiths (2010), but with a more expressive hypothesis space that better captures the diverse inferences people can make (see Bramley, Rothe, et al., 2018). For single- and few-shot causal generalizations, it draws on non-parametric approaches to category and function learning to account for similarity-based generalization predictions. In the realm of compositional generalization, it extends current state-of-the-art Bayesian symbolic models of concept learning with adaptor grammars (Johnson et al., 2007; Liang et al., 2010), providing a formal characterization of conceptual bootstrapping that uniquely predicts a range of order-effects in causal generalization.

As you will see in the following sections and chapters, causal learning and generalization are so closely intertwined, that they constantly shape each other. I will start with a brief conceptual analysis into this topic in this chapter, and then successively look into single-shot, few-shot, and compositional causal generalization in later chapters, providing both computational models and behavioral experiments in each investigation.

1.1 Drawing boundaries for causal laws

People readily generalize from familiar causal relationships to novel ones, using the features of prospective objects as a guide. For example, if you need to pound a nail but cannot find a hammer, you might pick up a nearby brick instead, reasoning that it will “do the job”; a child who has recently discovered drawing with colored chalks on paper may then explore the extent of this new power, using them to draw on the walls, the mirror, or even their bed sheets.

However, while it has been argued that we think of causal relationships as “invariant” (Sloman, 2005), in the sense that they apply across contexts and over time, our causal beliefs are so entangled with our concepts and categories that we think of certain kinds of objects as having particular causal powers, and others as being susceptible to particular causal influences. For instance, we may well understand that a bucket of paint can cause almost any surface to take on the paint’s color, but other classes of objects, like jumpers and cables, do not make other objects take their color. Category knowledge thus seems integral to real-world causal inference. If novel encounters involve objects of familiar categories, one can generalize the causal functional relationships and predict likely effects. When objects fall into different categories, however, those causal laws that one category has are not necessarily possessed by the other category. In fact, while people refer to causal relationships when categorizing objects (Gopnik & Sobel, 2000; Rehder, 2003; Rehder & Hastie, 2001), they also spontaneously use featural and relational information for categorization when no causal information is available (Anderson, 1991; Kemp & Tenenbaum, 2008; Love et al., 2004), and then make causal predictions based on these categories (Kemp et al., 2010), suggesting the widespread assumption that features reflect hidden causal powers.

It thus makes sense to integrate theories of generalization with theories of causal learning. As Roger Shepard famously put, psychology’s first general law should be a law

of generalization (Shepard, 1987). Shepard (1987) proposed that animals generalize learned stimuli according to a geometric decay in a psychological space of similarity, such that the more similar two stimuli are, the more likely these two stimuli share the same property of interest. Tenenbaum and Griffiths (2001) formulated this problem in a Bayesian inference framework that operates over a hypothesis space of possible ways of generalizations. In fact, to infer whether a novel stimuli shares the same property with some known observations could be viewed as a problem of property induction Heit and Rubinstein (1994): Does the novel stimuli belong to an existing concept or category that carries this property? Tenenbaum et al. (2006) combined a Bayesian framework of categorization with generic feature-based priors that well-captured people’s judgment on a range of tasks. I therefore draw inspirations from non-parametric Bayesian categorization methods (Kemp et al., 2010; Sanborn et al., 2010) and model generalization through feature similarity-based categorization as a starting point. As it unfolds in later chapters, this method also speaks to a function learning view of generalization (Lucas et al., 2015; Schulz, 2017; Wu et al., 2018), where people can learn a law of generalization as a function, and then make predictions for novel inputs by applying the learned function. For object-based causal generalization, causal relations could be readily viewed as functions, while objects naturally invoke a categorization process for principled generalization.

In Chapter 2, I will provide a computational model of causal generalization that takes these concerns into consideration, and compare several variants of this model to how people generalize their causal hypotheses in novel situations. To foreshadow, I find that the process model combining causal learning with categorization best captures people’s causal generalization patterns.

1.2 Causal roles as inductive biases

Suppose you add honey to your tea and find that the tea tastes sweeter, how would you generalize this newly-found knowledge? Should you infer that putting anything in your tea will make it sweeter, or that honey makes things sweeter in general? Generalization about interactions between objects is challenging, because it also involves a credit assignment problem: You may observe interactions between objects A and B, and later encounter interactions involving A and C, or involving B and C. Which features of the first interaction do you expect to generalize? Should you attribute the effect of the first interaction to properties of object A, B, or both?

In the honey and tea case, we can turn to pre-existing causal knowledge for help. We know that being sweet is a property of food items, and that sweetness can be transmitted to other food items it adds to. Hence, we should infer that honey makes things sweeter, and not that tea (or food in general) gets sweeter when something is added to it. In general, causality is a powerful guide to inductive generalization (Gelman, 2003; Rehder & Hastie, 2001), limiting the vast space of possibilities to a handful of possible ones (Griffiths & Tenenbaum, 2009; Kemp et al., 2010; Lagnado & Sloman, 2006). People seem to naturally impose causal roles onto objects based on how they interact, construing one object as a causal “agent” and another as a passive “recipient” (Mayrhofer & Waldmann, 2015), and we seem to do this even in situations where scientific theory would not single out either of them as special. For instance, when a moving billiard ball A collides with a stationary ball B and then ball B moves while ball A becomes stationary, people tend to say that A caused B to move (Michotte, 1963), even though from the point of view of Newtonian mechanics it would be equally valid to say that ball B caused A to stop moving. White (2006) coined the term “causal asymmetry” to summarize how people tend to treat the cause object and effect object differently.

In Chapter 3, I extend the experimental design and modeling framework to few-shot cases, manipulating whether people observe the same agent interacts with various recipients, or the other way around. In one experiment, I find evidence that people anchor causal generalization predictions with respect to the agent object only, consistent with the causal asymmetry bias, and capture this effect with the same computational models proposed earlier. In a follow-up experiment, I further identify the exact cues that people use to decide causal anchors, from the perspective of how object interactions probe causal anchoring.

1.3 Bootstrapping to complex causal concepts

When multiple causes or causal relationships are at play, people seem to have this remarkable ability to *bootstrap*: we arrive at complex concepts by starting small and building upon past successes. This ability to bootstrap enables us to grow rich mental concepts that go beyond our limited cognitive resources, and is taken to be a cornerstone of cognitive development (Carey, 2004). For instance, by building from atomic concepts of small numbers one, two, three, and counting, young children seem to bootstrap to more general and abstract numerical concepts such as successor relationships and the infinite line of real numbers (Piantadosi et al., 2012). Via bootstrapping, extant hard-earned knowledge need not be re-discovered every time it is used, saving the learner time and effort in constructing new concepts that build on old concepts. By effective re-representation of existing knowledge, people can arrive at rich mental constructs incrementally (Gobet et al., 2001; Klein, 2017; Krueger & Dayan, 2009). Similar effects have also been identified in artificial learning systems, where a simple-to-complex task order can greatly improve learning performance (e.g. Bengio et al., 2009; Mao et al., 2019; Narvekar et al., 2020). Interestingly, Elman (1993) found that a recurrent neural network with increasing amount of memory (window size for text processing) per round of training

demonstrated a learning trajectory similar to human children, such that easier tasks were learned first, and then used to facilitate learning of more complex tasks. The hierarchy of memory capacity configured by Elman (1993) also induced a spontaneous ranking of task difficulty, and Elman argued that this formed some internal representations that encoded varying degrees of variances in the learning data.

While seemingly a key aspect of human and machine learning, there are relatively few behavioral studies or cognitive models that make use of this bootstrap learning to understand how people deal with compositional generalization. Piantadosi et al. (2012) first formalized bootstrapping in a Bayesian-symbolic concept learning framework, but their work has focused on the discovery of a recursive function in learning numeric concepts, and therefore leaves open the task of examining bootstrapping as a general model of online inductive inference. In fact, current state-of-the-art Bayesian-symbolic models of concept discovery is missing a mechanism for caching and later reuse or repurposing earlier discoveries, and therefore lacks this crucial ability of conceptual bootstrapping. In Chapter 4, I draw upon adaptor grammars (Johnson et al., 2007; Liang et al., 2010) and provide a formal and computational characterization of bootstrap learning complex causal concepts, using a dynamic conceptual repertoire that is enriched over time, allowing the model to cache and later reuse elements of earlier insights. This model predicts systematically different learned concepts when the same evidence is processed in different orders, without any extra assumptions about prior beliefs or background knowledge. Across four behavioral experiments, I find strong curriculum-order and conceptual garden-pathing effects, demonstrating that people’s inductive concept inferences closely resemble our model’s, and differ from those of alternative accounts. This model provides an explanation for why information selection alone is not enough to teach complex concepts, and offers a computational account of how past experiences shape future conceptual discoveries.

Chapter 5 revisits major findings and concludes with a general discussion on the modeling approaches I am taking and their implications in computational cognitive science.

Chapter 2

Generalizing causal relations from single observations

We¹ first investigate the one-shot generalization case: Given a single observation of a causal interaction, will people form consistent causal generalization predictions to new objects? Can we identify the inductive biases behind their generalization choices? Section 2.1 introduces an original causal generalization task that systematically varies the degree of similarity between learning examples and novel situations, and in Section 2.2 we present a computational modeling framework that can synthesize human-like generalization patterns, and sheds light on how people may navigate the compositional space of possible causal functions and categories efficiently. This modeling framework combines a causal function generator that makes use of object features and relations, and a Bayesian non-parametric inference process to govern the degree of similarity-based generalization. A natural “resource-rational” variant of this model outperforms a naïve Bayesian account in describing participants, in particular reproducing a novel

¹I switch to “we” in content chapters to match the tune used in the published and submitted work.

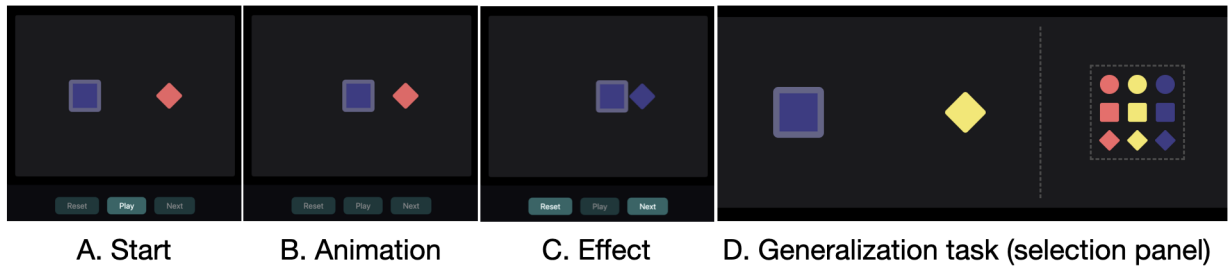


Figure 2.1: The magic stone task.

generalization-order effect in our behavioral experiments. We conclude with a discussion of our model’s scope and limitations, and highlight some potential future directions in Section 2.3.

2.1 Experiment 1: one-shot causal generalization

In order to systematically control and measure causal generalization, we developed a “magic stone” task (Figure 2.1). In it, participants test causal relationships between a magic stone (the agent) and a normal stone (the recipient) by watching the agent object moves toward the recipient object (Figure 2.1A-B), and upon touching each other the recipient object changes into a result form (Figure 2.1C). Note that we use the term *recipient* as equivalent to *patient* elsewhere in the causal learning literature, (e.g. Mayrhofer & Waldmann, 2015). Participants are asked to make predictions about new pairs of objects: “This new magic stone will turn this new normal stone into ...?” (Figure 2.1D). Observing objects interacting naturally invokes causal perceptions. For instance, Michotte (1963) reported the launching phenomenon, in which people directly perceive a causal influence connecting two objects that involved in a collision: If object A moves toward a stationary object B, and if around when A touches B, A stops moving and B starts to move, participants report that they see object A cause object B to move (see also Gordon et al., 1990; Leslie & Keeble, 1987; Scholl & Tremoulet, 2000). Similarly, the animated

agent-recipient setup in the magic stone task lays out an overtly causal framing, allowing us to probe the inductive biases and cognitive processes that are distinctive to causal reasoning. Unlike previous work in causal induction (e.g. Griffiths & Tenenbaum, 2009), this abstract setting minimizes the influence of specific domain priors and background knowledge.² Our experimental framework can be viewed as a conceptual extension to classic Blicket experiments in developmental psychology (e.g. Gopnik & Sobel, 2000; Kemp et al., 2010; Lucas & Griffiths, 2010), and we discuss this connection in detail in Section 2.3.

With these causal and object-based representations of the task, we open up a large space of scenarios and possibilities that demand sophisticated combinatorial reasoning, especially in terms of generalization. The relevant inference here is not about whether the agent object is the cause of the recipient object’s change or not (cf. Cheng, 1997; Jenkins & Ward, 1965; Pearl, 2000; Sloman & Lagnado, 2005; Tenenbaum & Griffiths, 2000). Instead, this is just the starting point. We are interested in a harder subsequent question: Given an observation where a particular agent object causes a particular change in a particular recipient object, how do people generalize this causal interaction to novel objects, where both agent and recipient may share more or fewer features with those in the original observation?

2.1.1 Method

Participants

One-hundred-and-twenty participants (53 female, aged 40 ± 11) were recruited from Amazon Mechanical Turk and were paid \$1.19. The task took 5.23 ± 3.17 minutes. No

²Children and adults are known to share a *shape bias* in category learning, such that they weight shape more than color/texture when generalizing category labels to novel objects (Landau et al., 1988). However, this shape bias is more sensitive to language rather than perceptual processes (Landau et al., 1992), and our experimental interface minimizes the influence of language, in line with a series of few-shot generalization tasks in cognitive psychology (e.g. Dasgupta et al., 2020; Kemp et al., 2010).

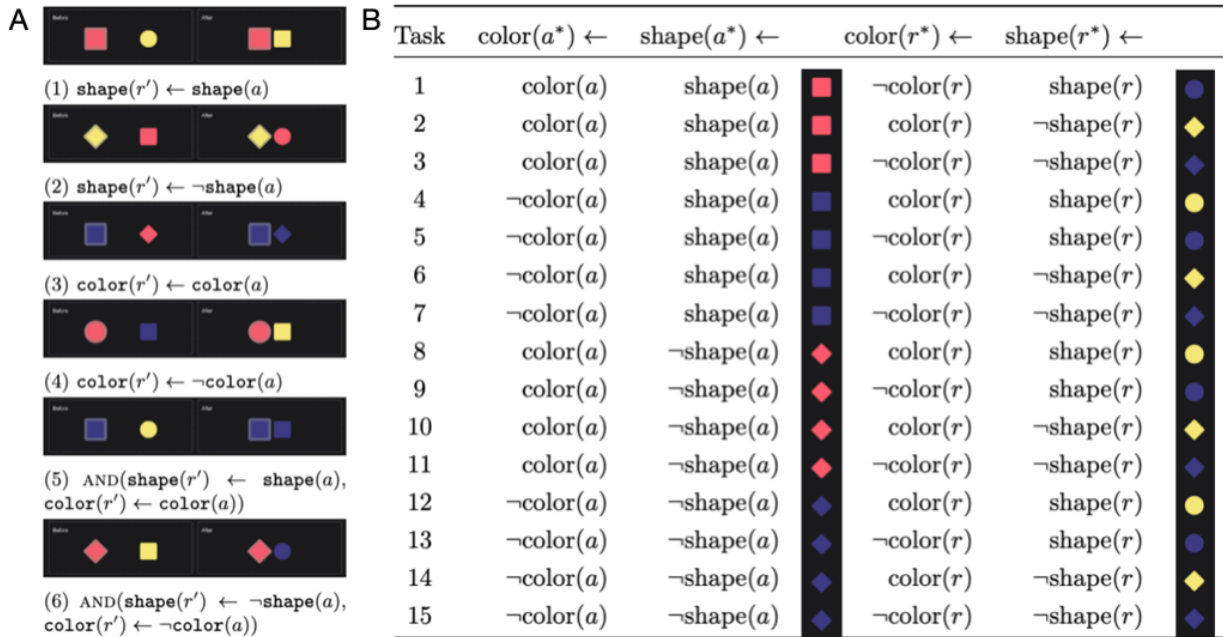


Figure 2.2: Experiment 1 material.

participant was excluded from analysis. All the experiments in this thesis were approved by the Research Ethics panel at the University of Edinburgh.

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 animations, participants observed a magic stone collide with a normal stone and appear to alter the normal stone’s color and/or shape (see Figure 2.1A-C). Magic stones had a thick border while normal stones had no border. We manipulated two object features—color $\{\text{red}, \text{yellow}, \text{blue}\}$ and shape $\{\text{circle}, \text{square}, \text{diamond}\}$, 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 2.2A)—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 2.2B). For example, A1 in Figure 2.2A depicts a *red square* agent and a *yellow circle* recipient, and according to the specifications in Figure 2.2B, generalization 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.

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 2.1A-C., Figure 2.2A), 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 2.1D). A summary of the learning example (as used in Figure 2.2A) 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.

2.1.2 Results

We were primarily interested in assessing the level of agreement between participants on each generalization task, as this gives a sense of how systematic or strong preferences for any particular patterns of generalization are.³ We used Cronbach’s alpha (Cronbach, 1943) to measure inter-person consistency. Specifically, the Kuder-Richardson Formula 21 (KR-21) (Kuder & Richardson, 1937):

$$\rho_\tau = \frac{k}{k-1} \left(1 - \frac{kp(1-p)}{\sigma_X^2} \right) \quad (2.1)$$

where k is the number of participants assigned to each condition, p is the chance probability of picking an object if responding randomly ($p = 1/9 \approx 0.11$), and X is the vector of aggregated participant selections for each option. KR-21 is a simplified version of Cronbach’s alpha known to be more conservative. The resulting consistency measure ρ_τ ranges between 0 — indicating uniform spread across all selections — and 1 — indicating perfect agreement between participants. Specifically, when $\sigma_X^2 = 0$, $\rho_\tau = 0$.

Task-wise consistency ρ_τ demonstrates that participants made systematic one-shot causal generalizations. Across 12 conditions \times 15 tasks = 180 tasks, $\rho_\tau = 0.80 \pm 0.22$. Fisher’s exact test confirmed that participants’ generalization consistency is significantly above random selections, $p < .001$. This is therefore another example of human capacity to make systematic one-shot causal generalizations (Kemp et al., 2007).

Next, we compared prediction consistency in the *near-first* and *far-first* transfer order conditions. Generalizations were more consistent overall under near-first transfer: $\rho_\tau = 0.83 \pm 0.21$, compared with far-first transfer $\rho_\tau = 0.77 \pm 0.21$, $t(89) = 3.54$, $p < .001$, 95%CI = [0.03, 0.10] (Figure 2.3A). ρ_τ was higher for near-first transfer under all learning

³Given the one-shot and causal-functional nature of the task, it is hard to measure systematicity of individual generalizations, due to the intractable space of possible causal functions people may entertain at the time. Hence we focused on group-level task-wise agreement as a starting point.

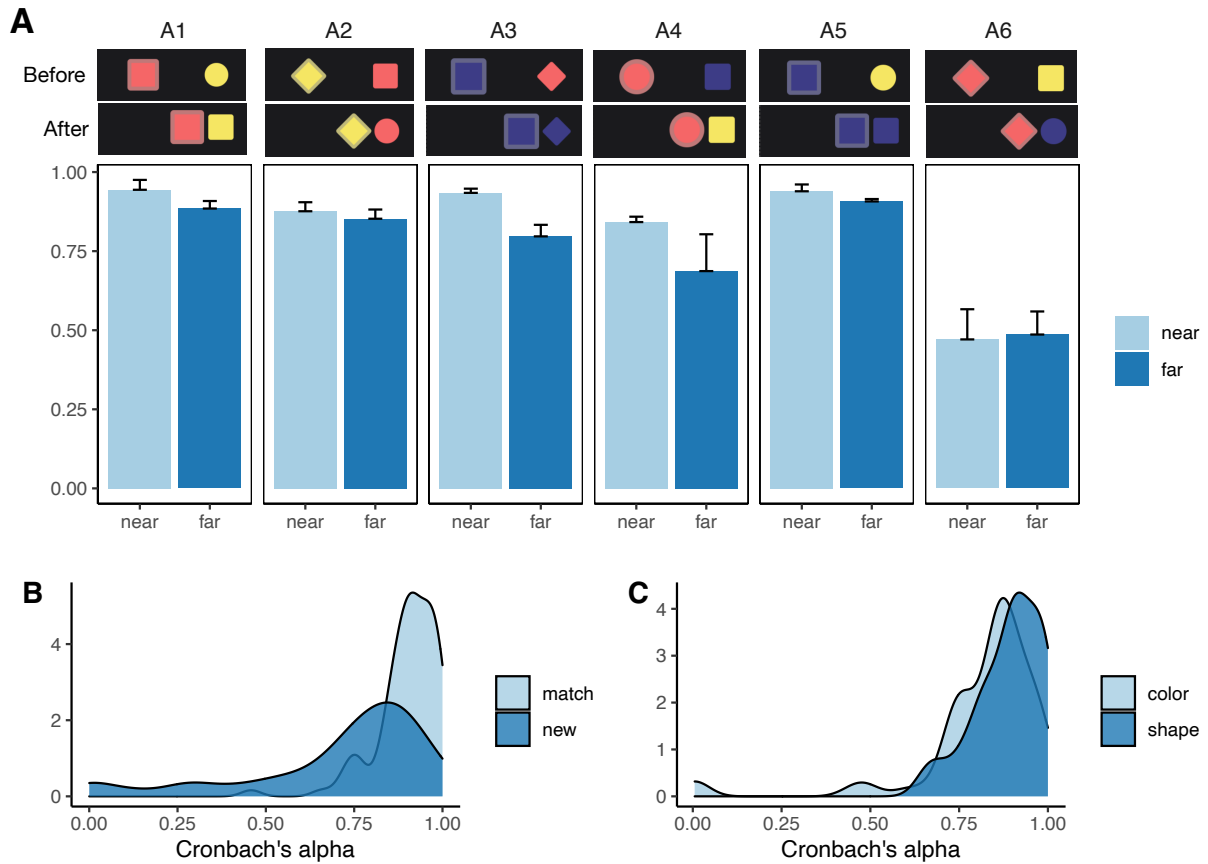


Figure 2.3: Experiment 1 behavioral results. A. Mean ρ_τ (y-axis) for each learning scene condition and sequence order: light blue = *near-first transfer*, dark blue = *far-first transfer*. Error bars are standard errors. ρ_τ for random selection is 0. B. Density distribution of task-wise ρ_τ , per match (light blue)/new (dark blue) groups. C. Density distribution of task-wise ρ_τ , per color change (light blue)/shape change (dark blue) groups.

conditions except A6 “Recipient changes to a new color and shape”, for which both transfer sequences induced low agreement.

Participants also generalized less consistently when the learning task involved new colors or new shapes (Figure 2.3B). For learning scenes A1, A3, and A5, where effect states *match* agents’ features, overall consistency was high: $\rho_\tau = 0.90 \pm 0.09$. Learning scenes A2, A4, and A6, where effects involved brand *new* values, consistency was lower: $\rho_\tau = 0.70 \pm 0.26$, differing significantly from the *match* group, $t(89) = 6.96, p < .001$, 95% CI = [0.14, 0.26]. Finally, color and shape changes were generalized to different extents despite these features appearing in symmetric and counterbalanced contexts in the task

(Figure 2.3C). Shape changes (A1, A2) induced more homogeneous predictions, $\rho_\tau = 0.89 \pm 0.09$, compared to color changes (A3, A4) $\rho_\tau = 0.81 \pm 0.19$, $t(59) = 2.88$, $p = .005$, 95%CI = [0.02, 0.13].

2.1.3 Interim discussion

On one hand, Experiment 1 demonstrates the strength and consistency of human causal priors, with participants making systematic generalizations from a single example despite these examples being compatible with a very large number of potential causal rules. On the other hand, we observed a clear departure from normativity providing a clue about cognitive processing, in the form of a generalization order effect. The *near-transfer first* conditions induced more consistent predictions (across subjects), compared to *far-first transfer*. Taking a closer look, for most conditions inter-subject consistency stayed fairly constant across all 15 generalization tasks. If there was a high level of agreement about the first generalization — as there tended to be in the *near-first transfer* conditions — participants also tended to make the same predictions as one another right through to the end, even once facing the highly dissimilar scenarios. Conversely, if initial generalizations were diverse (lower homogeneity) — as they tended to be in the *far-first transfer* condition — diversity of judgments persisted until the end, even though the objects in the final tasks were very similar to the learning example. This suggests participants might be influenced by their own generalization history in some way.

Generalizations following examples where the recipient is changed to a completely new feature value (A2, A4, A6) induced substantially more diversity in generalization predictions than those that did not (A1, A3, A5). This provides a possible explanation for the particularly low consistency measure ρ_τ in A6. Here, both of the result object's features are different from those of both the agent and the recipient. Potentially, some participants may have inferred a stochastic rule here such as that agents make recipients

take on random feature values. To the extent that participants inferred stochastic rules, we might expect varied predictions even if there is high consistency about the nature of the causal function.

2.2 Computational models

How do we evaluate, and possibly model such genuinely out-of-distribution causal generalization predictions? As discussed in Section 1.1, causal generalization involves two forms of induction: (1) Inferring what causal relationship is at work in an observed setting, known as causal learning or causal induction, and (2) Inferring the domain to which a causal law applies, closely related to categorization. In correspondence, we model the vast open space of possible causal relationships (causal laws) with a generative grammar, and account for the domain of influence for those causal relationships using a Bayesian non-parameteric categorization process. Together, they provide a principled account for causal generalization over novel interacting objects.

2.2.1 Universal Causal Laws (UnCaLa)

To a first approximation, objects are identifiable by their features and causal powers (Aristotle, 322/1998; Gopnik et al., 2004). Adults find basic features of objects, such as color, shape, and orientation to be salient cues for information selection (Treisman & Gelade, 1980; Treisman & Paterson, 1984). Therefore, we represent objects in terms of their observable features, and model interactions between objects using causal functions. For example, we can read an object’s color by `color(o)` is `red`. When an agent acts on a recipient and causes the recipient to change, we model this with a causal function f that takes the agent (a) and recipient’s initial state (r) as input, and outputs the final state of the recipient (r'), which we call the *result*. Depending on the situation, real causal interactions could result in changes to the form of the agent object as well. However, given the

examples in Figure 2.1, here we restrict our focus to $f(a, r) \Rightarrow r'$. Naturally, a causal function defines the result r' by specifying its features, potentially conditional on specific features of a and r . Take an everyday understanding of paint for an example: When applied to a wall, paint causes that wall to take the color of the paint. We can formalize this as a function $f(\text{paint}, \text{wall}) \Rightarrow \text{wall}'$, where $\text{color}(\text{wall}') \Leftarrow \text{color}(\text{paint})$. Note that arrow \Rightarrow reads as “gives” or “produces”: $f(a, r) \Rightarrow r'$ says that function $f(a, r)$ produces result object r' . Arrow \Leftarrow is an assignment operation: $\text{color}(\text{wall}') \Leftarrow \text{color}(\text{paint})$ means that color of the paint is assigned to (color of) the wall.

Griffiths and Tenenbaum (2009) proposed hierarchical Bayesian model (HBM) for causal inference where structured domain knowledge restricts the space of possible or plausible causal relationships. However, this computational level model focused on statistical relationships between variables rather than interactions between objects. Furthermore, the space of possible causal functions in natural settings is clearly intractable, posing a serious computational challenge for any bounded learner. Therefore, more recent accounts of causal learning have treated causal inference as practically constituting a search problem in a large multi-modal theory space (Bramley et al., 2017), and utilize generative grammars and program induction ideas to cover the open ended space of a learner’s possible theories and hypotheses (Bramley & Xu, 2023; Fränken et al., 2022; Goodman et al., 2008; Piantadosi et al., 2016), as well as the human preference for simpler causal explanations (cf. Feldman, 2000).

Following this approach, we use a Probabilistic Context-Free Grammar (PCFG; Ginsburg, 1966) to define a prior over possible causal relationships (causal laws, right column in Figure 2.4). A PCFG is defined by a tuple $\langle \Gamma, \Theta, T \rangle$, where Γ is a set of production rules, Θ a set of production probabilities, and T a set of transition symbols. Our example grammar \mathcal{G} (Table 2.1) has a set of transition symbols $T = \{S, A, B, C, D, E\}$, where S is the “Start” symbol by convention. Starting from symbol S , grammar \mathcal{G} follows the production rules to generate expressions, and stops when there are no transition symbols

Table 2.1: Example probabilistic grammar \mathcal{G}

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

Note: ϕ_i denotes the i -th feature in the set of all observable object features Φ . The lambda abstraction in the “bind feature” production rule samples a feature without replacement from the set of all features, and binds this feature ϕ_i to the rest of the generation: ϕ_i in D uses the same feature selected in A , and value in E is sampled uniformly from the support of feature ϕ_i . Production probabilities are omitted from the table because we assume a uniform prior.

any more in the expression. Production rules Γ define how transition symbols transform. Production probabilities Θ assign a probability distribution for each transition symbol’s possible transformations. For simplicity, we assume uniform production probabilities: let Γ_L be the set of all production rules that start with symbol $L \in T$ (i.e, production rules in the form of $L \rightarrow K$, where K can be any symbol in grammar \mathcal{G}), the transition probability for each $l \in \Gamma_L$ is simply $\frac{1}{|\Gamma_L|}$. For example, on the “Reference” row in Table 2.1, symbol C can either follow production rule $C \rightarrow D$ and produce D , or follow production rule $C \rightarrow E$ and produce E . We thus assume symbol C has 0.5 probability to become D , and 0.5 probability to become E .

Let us walk through an example for our grammar \mathcal{G} in Table 2.1. Starting from symbol S , production rule $S \rightarrow \lambda_{\phi_i} : A, \Phi$ samples a feature uniformly from the set of all observable features (in the task) and binds it to the production. Let’s assume we sampled feature `color` with probability 0.5 (out of $\Phi = \{\text{color}, \text{shape}\}$), and now the expression becomes $\lambda_{\text{color}} : A$. Symbol A leads two productions: either becomes B , or $\text{AND}(B, S)$, with uniform prior probability. Assume that with probability 0.5 we

retrieve expression B . Proceeding to row “Relation”, with probability 0.5 we could arrive at $\text{color}(r') \Leftarrow \neg C$. Then on row “Reference”, with 0.5 probability we could get $\text{color}(r') \Leftarrow \neg D$. Finally, symbol D produces either $\text{color}(a)$ or $\text{color}(r)$ equally likely, and with probability 0.5 we end up with $\text{color}(r') \Leftarrow \neg \text{color}(r)$: result object’s color is assigned to a color that’s different from the recipient’s, i.e., result object changes its color. In total, the probability of producing $\text{color}(r') \Leftarrow \neg \text{color}(r)$ is $0.5^5 = 0.03$. If at the step of “Reference” we followed production rule $C \rightarrow E$ instead, then with probability 0.33 we might sample a color `blue` (out of $\text{value}^{\text{color}} = \{\text{red}, \text{yellow}, \text{blue}\}$), and the probability of producing $\text{color}(r') \Leftarrow \neg \text{blue}$ is $0.5^4 \times 0.33 = 0.02$. By design, this grammar is inherently more likely to produce simpler expressions. This is because the “Bind additional” rule $A \rightarrow \text{AND}(B, S)$ 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.

Formally speaking, the prior for a given expression is the product of all the productions that produced it:

$$P_{\mathcal{G}}(f) = \prod_{l \in \Gamma} (\theta_l)^{c_l} \quad (2.2)$$

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

Grammar \mathcal{G} assigns a prior over a potentially infinite set of causal functions. A causal function defines the result object(s) by describing the result object’s feature values, given the particular agent and recipient object inputs. Take $\text{AND}(\text{color}(r') \Leftarrow \text{color}(a), \text{shape}(r') \Leftarrow \text{square})$ for example. For an agent a that is a `red-circle` and a recipient r that is a `blue-pentagon`, r will become r' : a `red-square`. When a causal function f involves a negation, it could have produced more than one outcome. For instance, consider a causal function $\text{shape}(r') \Leftarrow \neg \text{triangle}$, 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

$$P(d|f) = P(r'|f, a, r) = \begin{cases} \frac{1}{D(f(a, r))} & \text{if } r' \in D(f(a, r)), \\ 0 & \text{otherwise.} \end{cases} \quad (2.3)$$

Here, $D(f(a, r))$ refers to the set of all possible result objects coming out of f given agent a and recipient r (D stands for domain). We initially assume a likelihood to 0 for any observation $(a, r, r') \notin D(f(a, r))$, 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. For example, if you observe a recipient turning blue, this is more consistent with a function where the agent invariably turns the recipient blue than with one where the agent turns the recipient to either red or blue. We note that while many of these choices are somewhat arbitrary, or are made for computational convenience with respect to the current task context, for example recursing with only conjunctions leads to a rather constrained set of extensional equivalences in this given task, but the approach itself is highly general and flexible, compatible with many other more or less expressive grammars and production processes embodying stronger or weaker priors.

According to Bayes Theorem, upon seeing some learning data d , the posterior distribution over causal functions is

$$P(f|d, \mathcal{G}) \propto P(d|f)P_{\mathcal{G}}(f). \quad (2.4)$$

If causal functions apply universally to all the objects, Equation 2.4 solves the learning and generalization problems at the same time: after updating the prior of causal functions with learning data, the posterior predictive gives generalization predictions for every novel pair of objects. For instance, the animation example in Figure 2.1A-C results in a posterior over causal functions favoring $\text{color}(r') \Leftarrow \text{color}(a)$, $\text{color}(r') \Leftarrow \text{blue}$ and some other possibilities (recall the set is potentially infinite). Then, in the generalization prediction phase as in Figure 2.1D, marginalizing over that posterior leads to a prediction favoring `blue-diamond`. Formally, 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:

$$P(r'|d^*) = \sum_{f \in \mathcal{G}} P(r'|a^*, r^*, f) P(f|d, \mathcal{G}). \quad (2.5)$$

Grammar \mathcal{G} and Equations 2.2-2.5 together define our first normative model Universal Causal Laws (UnCaLa).

2.2.2 Localized Causal Laws (LoCaLa)

We formalize the idea that the pairs of objects may fall into different categories with respect to featural similarities and their roles in the interaction with a Dirichlet Process (DP). We treat one such category as a distribution over objects, and DP defines a prior over a potentially infinitely-many categories. Let \mathbf{d} denote a set of observations, \mathbf{z} denote a particular set of 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 and $z^{(-i)}$ for causal category assignment to all the other observations, $a^{(i)}$ the agent in the i -th data point, similarly for $r^{(i)}$, $r'^{(i)}$, and $\mathbf{w}_{z^{(i)}}$ for the weights associated with category $z^{(i)}$. Inference about the i -th

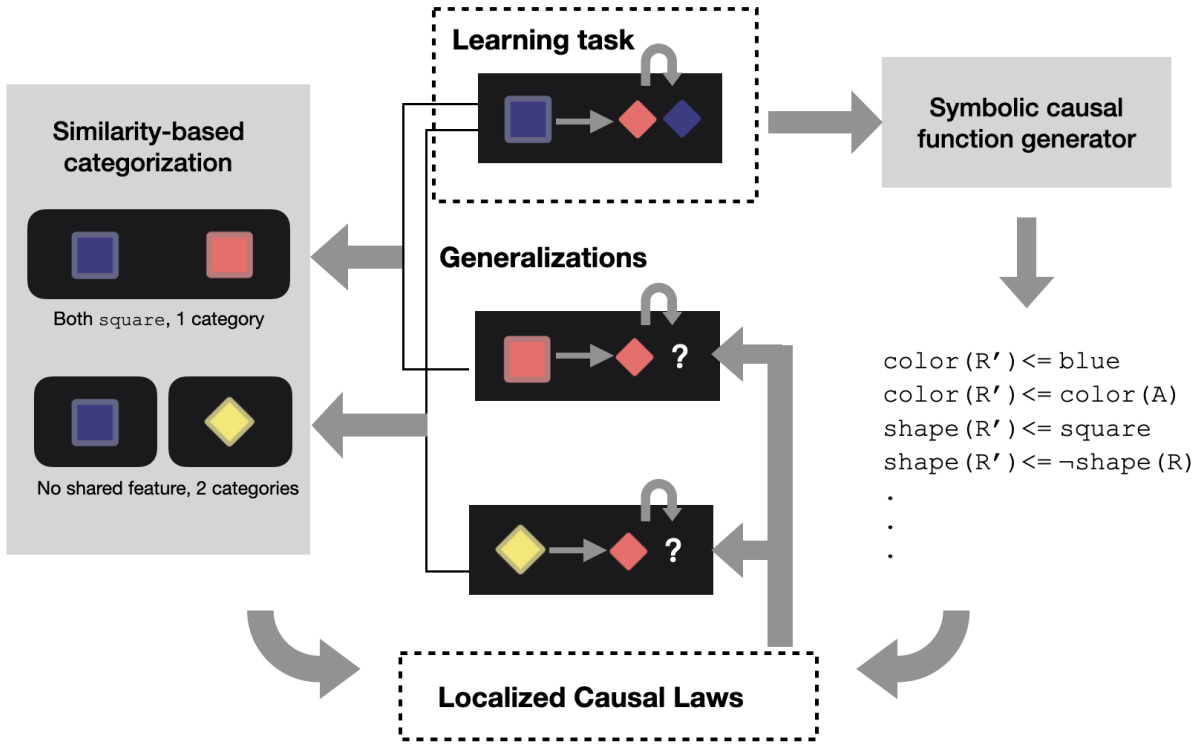


Figure 2.4: Visualization of the Localized Causal Laws (LoCaLa) model.

observation's category is given by

$$P(z^{(i)} | \mathbf{d}, \mathbf{w}) \propto P(z_i | z^{(-i)}) P(a^{(i)}, r^{(i)} | \mathbf{w}_{z^{(i)}}). \quad (2.6)$$

Equation 2.6 consists of two parts: $P(z^{(i)} | z^{(-i)})$ reflects our prior expectations about how categories are distributed, and $P(a^{(i)}, r^{(i)} | \mathbf{w}_{z^{(i)}})$ encodes our beliefs about object features and category membership.

In DP, the prior expectation of categories is given by a Chinese Restaurant Process (CRP), controlled by a concentration parameter α . A CRP is a stochastic process widely used for creating partitions among entities. 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

$$P(z^{(i)} = x | z^{(-i)}) = \begin{cases} \frac{\alpha}{i-1+\alpha} & \text{if } x \text{ is a new category} \\ \frac{|z^{(j)}|}{i-1+\alpha} & \text{if } x = z^{(j)} \end{cases} \quad (2.7)$$

where $z^{(j)}$ is an existing category, and $|z^{(j)}|$ is the number of assigned objects in category $z^{(j)}$. Parameter α 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”.

Preference for feature similarities can be modeled by a multinomial distribution over feature values. Let $[\mu_1, \dots, \mu_n]$ be the mean feature vector of a given category where each subscript k is a feature value, probability of an object assigned to a particular category according to feature similarities is given by

$$P(o^{(i)} | z^{(i)}) = \prod_{k=1}^n \text{Bernoulli}(o^{(i)}; \mu_k). \quad (2.8)$$

To compute μ_k , let $o_v = [o_{v_1}, \dots, o_{v_n}]$ 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 $z = \{o^{(i)}, \dots, o^{(m)}\}$, $z_v := \sum_{j=1}^m o_v^{(j)}$, which can be written as $z_v = [z_{v_1}, \dots, z_{v_n}]$, 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 toward any specific feature, the prior distribution over mean features is simply $\text{Dir}(\beta)$.

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.

In total, we introduce three global parameters: a concentration parameter $\alpha > 0$, a Dirichlet prior $\beta \geq 0$, and a focus parameter $\gamma \in [0, 1]$. Dirichlet prior β and focus parameter γ together decide the mean feature vector $\mu^{(z_i)}$ for category $z^{(i)}$. Equation 2.6-2.8 provide the full definition for featural similarity-based categorization (left column, Figure 2.4).

Take the generalization tasks in Figure 2.1A-C as an example again. Assuming we saw a blue-square agent causing a red-diamond to become a blue-diamond, but then we need to make predictions about a red-square agent and a yellow-diamond agent. According to the model, both square objects have a high probability of falling into the same “square agent” category and hence sharing the same causal power. However, a yellow-diamond has no shared feature with a blue-square, hence it is more likely to belong to a different category and have potentially different causal powers than a blue-square.

Finally, we combine causal functions and object categories into causal categories. The core assumption is that objects within the same causal category share a same causal function:

$$\begin{aligned}
 P(z^{(i)}|\mathbf{d}, \mathbf{w}) &= P(z^{(i)}|d^{(i)}, \mathbf{w}, z^{(-i)}) \\
 &\propto P(z^{(i)}|z^{(-i)})P(a^{(i)}, r^{(i)}|\mathbf{w}_{z^{(i)}})P(r^{(i)}|a^{(i)}, r^{(i)}, \mathbf{w}_{z^{(i)}}) \\
 &\propto P(z^{(i)}|z^{(-i)})P(a^{(i)}, r^{(i)}|\mu^{(z_i)})P(d^{(i)}|f^{(z_i)}). \tag{2.9}
 \end{aligned}$$

Equation 2.9 adds a causal function component onto Equation 2.6. On the final line of Equation 2.9, the three products correspond to Equation 2.7, Equation 2.8, and Equation 2.3 separately. In other words, the priors for constructing causal categories are

provided by

$$\begin{aligned}
z^{(i)} | z^{(-i)} &\sim \text{CRP}(\cdot | \alpha) \\
\mu^{(i)} &\sim \text{Dir}(\cdot | \beta) \\
f^{(z_i)} &\sim \mathcal{G}(\cdot)
\end{aligned} \tag{2.10}$$

And likelihoods are given by

$$\begin{aligned}
a^{(i)}, r^{(i)} | \mu^{(z_i)} &\sim \text{Dir}(\cdot | \mu^{(z_i)}, \beta) \\
d^{(i)} | f^{(z_i)} &\sim f^{(z_i)}(a^{(i)}, r^{(i)})
\end{aligned} \tag{2.11}$$

When learning data points are abundant, it is impossible to compute the posterior directly because we do not know how many categories are there in advance. We can approximate the posterior distribution using Gibbs sampling (Geman & Geman, 1984). To achieve this, 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 2.9, and the local parameter update step follows definition of computing these parameters given objects in this category. When the number of iterations $n \rightarrow \infty$, the sampled categories \tilde{Z}_n converges to the true posterior.

With a posterior over causal categories in place, we can make normative generalization predictions to new cases. Similar to Equation 2.5, 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'^* value:

$$\begin{aligned}
 P(\tilde{d}^*) &\propto \int_z p(\tilde{d}^*|z)P(z|d)\mathbf{d}z \\
 &\approx \frac{1}{|\tilde{Z}|} \sum_{\tilde{z} \in \tilde{Z}} p(r'^*|a^*, r^*, f^{(\tilde{z})})P(a^*, r^*|\mu^{(z)})P(z|d)
 \end{aligned} \tag{2.12}$$

and taking the maximum over this predictive posterior

$$\text{Choice} = \arg \max P(\tilde{d}^*). \tag{2.13}$$

Consider the example task in Figure 2.1 again. After watching a blue-square agent turning a red-diamond object blue, the posterior distribution over causal functions provides a pool of causal functions these objects may have. For the sake of the example, assume the most salient causal function is that the blue-square object transfers its color to other objects. When making generalization predictions for a red-square object, its shared square feature with the blue-square object leads us to guess they belong to one same category, hence this red-square object might also transfer its color to other objects. When facing a yellow-diamond object, we are less certain in applying the same causal function. Thus we are more likely to draw upon the prior distribution of causal laws to account for our uncertainty. Note that since there are no further feedback on these predictions, this approach differs from semi-supervised learning (Zhu & Goldberg, 2009) where unlabelled data are used to improve learning accuracy on the labeled ones.

Algorithm 1 LoCaLaPro

```
1: Initialize an empty list of causal categories  $Z$  ▷ Initialization
2: Assign  $a^{(0)}, r^{(0)} \in d^{(0)}$  to category  $z^{(1)}$ , update  $\mu^{(1)}$  ▷ Assign to category 1
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 2.6 & 2.8
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
```

2.2.3 Local Causal Law Process model (LoCaLaPro)

To account for our conjecture that generalization predictions could be influenced by past decisions, we further develop a Local Causal Law Process model (LoCaLaPro) that commits to its own causal category allocations as it makes generalizations, instead of treating each generalization trial independent from each other as in the LoCaLa model. As a result, LoCaLaPro behaves differently when generalizations are made in a different order. To unpack, LoCaLaPro 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 $P(f|d)$. 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 2.6. If an existing causal category is selected, the model simply applies the causal function of this 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 in pseudo code.

Instead of approximating a posterior over infinitely many possible categories as the LoCaLa model, the process model LoCaLaPro maintains a small set of available categories that are created online as new generalizations are performed. Furthermore, after categorizing an observation, the LoCaLaPro model updates the list of causal categories Z with this categorization decision, reflecting a commitment to its earlier decisions. Concentration parameter α thus plays a slightly different role in the LoCaLaPro model as LoCaLa. When $\alpha \rightarrow 0$, the model becomes increasingly likely to stick with existing categories (Equation 2.7).

2.2.4 Model fits

We now further analyze people’s generalizations in Experiment 1 using our modeling framework. To do this, we fit several model variants to our choice data using maximum likelihood. We then compared them using Bayesian Information Criterion to accommodate for different numbers of parameters.

We first computed a random choice *Baseline*. This model simply predicts $P(\text{choice} = r') = \frac{1}{9}$, for the 9 candidate objects and has no parameters. We then consider three models based in the modeling framework developed above.

Universal Causal Laws (UnCaLa) model uses the causal law induction process to generate a large prior sample of possible causal functions \tilde{F} using the PCFG described earlier (Table 2.1), then filters this according to the learning example to generate a posterior of potential causal functions consistent with the training data (Equation 2.2-2.4). It then integrates over these to generate posterior predictions for each generalization task according to Equation 2.5. Essentially, UnCaLa assumes that the causal function governing the training case applies universally to all potential generalization scenarios no matter how dissimilar the objects involved may be.

Local Causal Laws (LoCaLa) model captures the idea that multiple latent relationships might be at work, and which will apply to a particular object pair depends on which causal category they fall under. Based on a sample of possible causal functions generated by the PCFG defined in Table 2.1, the LoCaLa model makes predictions about the result object in a generalization task according to Equation 2.9-2.12. Note that LoCaLa compares each generalization task with the learning example to make predictions, treating each task as an independent decision problem. Essentially, the more dissimilar a generalization scenario is to the learning example, the less likely LoCaLa thinks it is that the same causal law will apply, meaning it reverts its prediction increasingly toward the prior. How strongly it reverts depends on the concentration parameter α : with larger values producing a more drastic return to the prior (Equation 2.7). Relatedly, Dirichlet prior β captures categorization sensitivity to feature similarity (Equation 2.8) with larger values meaning less sensitivity and consequently more noise in the predicted behavioral pattern. Note that we fit α and β , but fix focus parameter $\gamma = 0.5$ in Experiment 1, because in this experiment there is no information about what causal categorization assumptions should be preferred.

Local Causal Laws Process (LoCaLaPro) further commits to previous generalization guesses in making new predictions. As a result, under the *near-first transfer* conditions, throughout the entire generalization phase this model makes predictions closely approximated by the posterior distribution after watching the learning example; in the *far-first transfer* conditions, it is likely to trigger the creation of a new category to accommodate the fact that the generalization scenario drastically differs from the learning example. Subsequent generalization predictions tend to join this newly-created category. This induces a generalization-order effect (Figure 2.5C). When α becomes very large, a new observation has a high probability of being attributed to a new category (Equation 2.7),

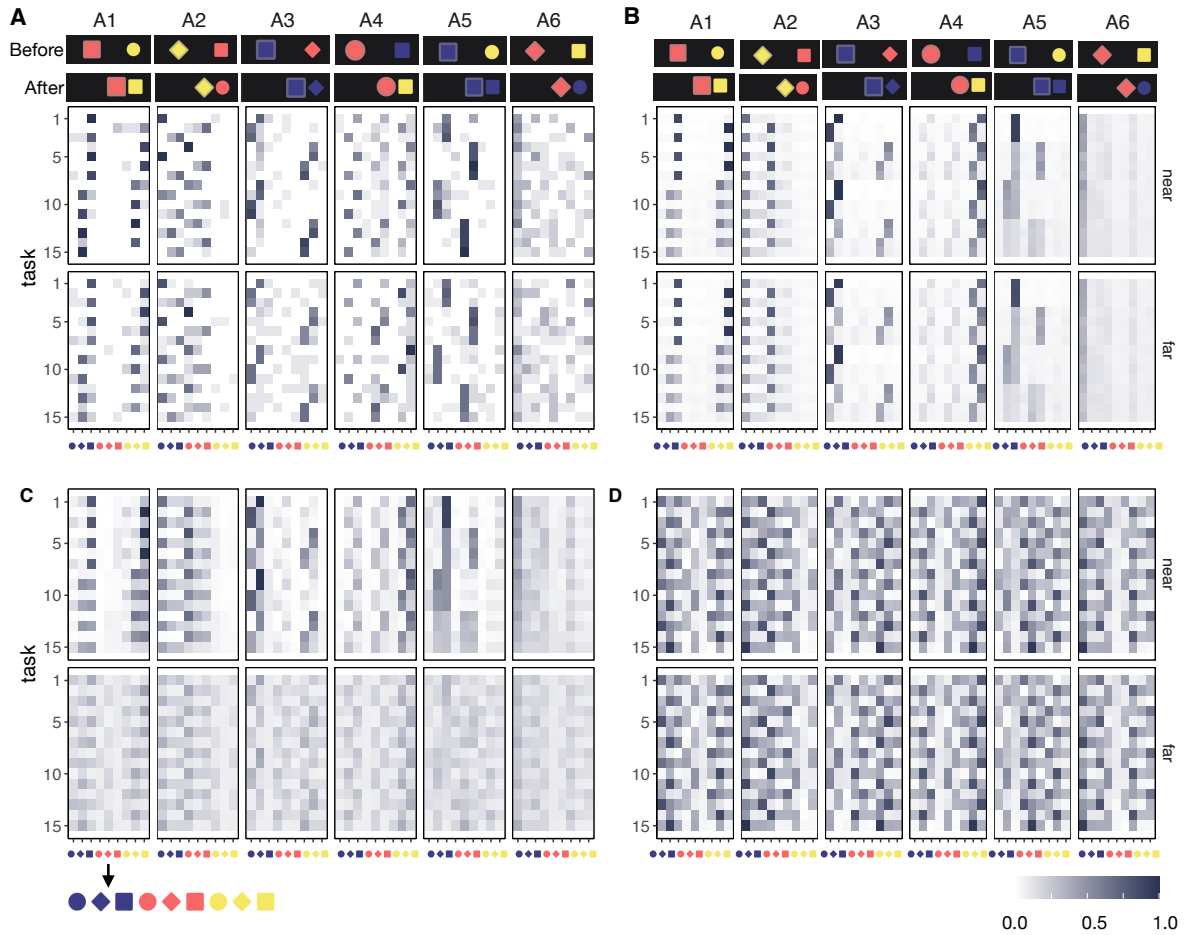


Figure 2.5: Experiment 1 results. A. Generalization patterns for all conditions visualized as proportion of participants predicting each stone type for r' (column) on each task (row). B. Fitted LoCaLaPro predictions. C-D. Example LoCaLaPro predicted proportions with small α ($=0.01$) and large α ($=8$). For both figures, $\beta = 0, \gamma = 0.5$.

and the overall generalization predictions will simply approach the prior (Figure 2.5D).

The other hyperparameters β and γ play the same role as in the LoCaLa model.

Model fitting

Each model natively provides predictive posterior probability distribution over the nine options, while participants make a single discrete prediction. Thus, for each case, we convert the model's posterior into discrete choice probabilities using a softmax function

to account for decision noise (Luce, 1959). Taking $P_m(r'|d) = \{x_{o_1}, \dots, x_{o_9}\}$ as the posterior predictive distribution over candidate objects for model m , and t as an “inverse temperature” parameter:

$$P(\text{choice}) = \frac{e^{P_m(r'|d)t}}{\sum_{x \in r'} e^{P_m(r'|d)t}}. \quad (2.14)$$

When $t \rightarrow 0$, Equation 2.14 corresponds to flattening the input distribution toward a uniform distribution while as $t \rightarrow \infty$ the input distribution is sharpened, approaching a hard maximisation over the probabilities.

We used `optim` function in R to fit the UnCaLa and LoCaLa models to behavioral data. Recall that we generate a large prior sample of possible causal functions \tilde{F} for all three models. In practice, we exploited the fact that each object is composed of two features, and therefore enumerated all the possible causal functions generated by grammar \mathcal{G} up to depth 2. Any causal function in our grammar that is syntactically more complex than those in this set is semantically-equivalent to one in this set. With a fixed set of \tilde{F} , the UnCaLa model has only one softmax parameter that can be optimized by `optim`.

LoCaLa has an analytical solution in this case because there is a single learning example, which by definition belongs to category 1. Each generalization task is then compared against the learning example independently. As a result, the chance that a generalization task belongs to category 1 can be computed straightforwardly from parameters α and β . Assuming the model applies the same α, β and softmax inverse temperature t to each generalization task, we jointly optimize all three parameters to maximise the likelihood of the data using R’s `optim` function.

For the LoCaLaPro model, since each sampling decision for one generalization task affects how future tasks will be categorized, we can only approximate its posterior distribution with simulation-based method, and optimized parameter values via grid search. Firstly, we set up a coarse grid with $\alpha = \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.5, 2, 4, 8\}$, $\beta = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 4, 8, 16, 32, 64, 128,$

Table 2.2: Experiment 1 model fitting results

	α	β	t	Log likelihood	BIC
Baseline				-3955	7910
UnCaLa			6.96	-2761	5529
LoCaLa	2.41	938.81	9.44	-2748	5518
LoCaLaPro	0.38	1	10.09	-2736	5494

256, 512, 1024}. For our single shot experiment, in the far-first condition, when $\alpha = 1$, the first generalization observation has a half-half chance to join the learning example or create its own causal category in terms of category size preference (Equation 2.7). Therefore, the presence of generalization-order effects in behavioral data indicates that α is likely to be smaller than 1. Hence we included a range of values for $0 < \alpha < 1$ with fixed-step 0.1, as well as few larger values in case of surprise. For β , we take a set of exponentially-growing large values up until 1024 in order to accommodate behavioral noise. 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.

Model fitting results

Table 2.2 summarizes the model fits. Both the Universal UnCaLa and Local LoCaLa models improve dramatically over the random Baseline, and LoCaLa outperforms UnCaLa in both likelihood and BIC. Figure 2.5B shows that these computational models indeed predict the dominant judgment patterns among participants. The process model LoCaLaPro best predicts the empirical data. Its fitted α parameter for LoCaLaPro is 0.38, confirming the presence of a dominant order effect. The fitted $\beta = 1.0$, indicating a moderate level of noise in this sample. For model LoCaLa, β parameter is fitted to a rather large value as its best attempt to account for the different kinds of categorization within all generalization tasks: Since the best fitting $\alpha = 2.41$, each new generalization tasks is likely to

fall into its own category, however we do notice the overall trend of committing to the same category for different-looking objects in the near-transfer condition, and therefore forcing β to settle on rather large values.

Overall, participants' generalization patterns were well-captured by our Bayesian inference model operating on a latent space of causal laws generated by a simple PCFG 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 (Bramley et al., 2017; Goodman et al., 2008; Kemp et al., 2010), 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. Griffiths & Tenenbaum, 2009; Lucas & Griffiths, 2010); 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.

There are also some interesting discrepancies between model predictions and people's generalizations. In Condition A2, for example, people's generalizations seem to be more sensitive to the novel objects in each task, while the model predictions are mostly focused on a rather fixed set of result objects. A similar discrepancy can be found in Condition A4 too. Conditions A2 and A4 involve changing one feature (shape in A2 and color in A4) of the recipient object to a value that is neither shared by the present agent nor recipient objects. It seems that people are more sensitive to such relative feature changes than the model, which may simply prefer a given value as a result of the rather limited space of possible feature values in this task.

2.3 Discussion

In this chapter, we demonstrated that people can make systematic causal generalizations from single observations, and introduced a Bayesian-symbolic modeling framework to capture distinct behavioral patterns we identified in the experiment: Generalization-order effects in one-shot causal generalization. While previous research has shown effects of the order in which learning examples are presented (Danks & Schwartz, 2006; Lu et al., 2016), ours is the first study to find effects of the order in which generalization predictions are made. Many order effects can be understood as a consequence of cognitive agents with limited resources updating their beliefs sequentially, for example anchoring-and-adjustment (Lieder et al., 2012), local updating (Bramley et al., 2017), or amortized computations (Dasgupta et al., 2020). These and other models predict an order effect of *evidence* — people update their beliefs sequentially as evidence arrives. However, our participants made judgments without receiving feedback. In our experiments, we did not vary the order of evidence, so there is no basis for expecting order effects under these previous models. The LoCaLaPro model assumes that people implicitly commit to their generalizations as they make them, essentially treating their earlier generalizations as evidence that must be accommodated going forward rather than uninformative guesses that may or may not line up with the ground truth. Consequently, the order in which judgments are solicited can lead to systematic changes in its inferences, even in the absence of new evidence.

Our work generalizes the structure of standard “blicket detector” studies, in which different combinations of objects are tested and an effect does or does not occur (e.g., Gopnik et al., 2007; Kemp et al., 2010; Lucas & Griffiths, 2010; Sim & Xu, 2017), making predictions about a wider family of scenarios while accommodating previous results. If we treat the recipient object’s feature change(s) as a multinomial activation outcome, this can be viewed as analogous to the blicket detector becoming active in the presence

of a blicket, and we can use our current framework, unaltered, to see how people reason about a machine’s interactions with prospective individual blickets. However, our setup puts more emphasis on causal interactions. The collision stimuli we used in our tasks are known to evoke automatic perception of causality (Michotte, 1963), making it an appealing way to study how people reason about cause and effect specifically. In contrast, many studies of causal induction involve descriptions of events that already occurred, or carefully orchestrated demonstrations where combinations of putative causes are presented simultaneously (e.g. Griffiths & Tenenbaum, 2009; Johnson & Ahn, 2015; Steyvers et al., 2003). Such approaches of simultaneously presenting causes is necessary for answering certain scientific questions, but in daily life, we typically observe sequences of changes, which tends to be more informative than an “episodic” approach (Soo & Rottman, 2018).

Our modeling approach attempts to provide a general and principled computational account for causal generalization, connecting theories of generalization and causal learning in computational terms. This formalization considers two aspects of objects at the same time—an object’s causal relationships with other interacting objects, and their feature similarities with novel observations. To unify these two aspects in one computational framework, we draw on program induction (Fränken et al., 2022; Goodman et al., 2008) to model causal functions that operate over object features, and make use of standard Bayesian non-parameteric techniques to formalize similarity-based generalization (Kemp et al., 2010; Sanborn et al., 2010). This may look like a lot of complex machinery on first glance, in contrast to the simplicity of the task, but such complexity (1) is needed to express the richness of human cognition in seemingly simple tasks—detecting dogs and cats looks like a simple task, yet it requires layers of neural networks to do barely as well as people, and (2) can be greatly reduced since our choice of methods—generative grammars and Bayesian non-parametric categorization—are both standard techniques attracting increasing attention across fields of cognitive science (Quilty-Dunn et al., 2022). In light of resource-rational approaches to cognition (Anderson, 1991;

Griffiths et al., 2015; Newell & Simon, 1972), it might look favorable to consider alternative heuristics to this computational problem such as propose-but-verify (Trueswell et al., 2013), win-stay-lose-shift sampling (Bonawitz et al., 2014; Robbins, 1952), particle filters (Gelpi et al., 2020; Thaker et al., 2017), etc. These approaches are better suited for learning tasks where participants update their posterior estimates according to feedback signals, and would require a considerable amount of assumptions and follow-up work to apply in open-ended generalization tasks. In fact, Chapter 4 looks into a particular kind of such heuristics, cache-and-reuse, in compositional causal generalization, and compares it against the above-mentioned heuristics in more detail.

Last but not least, thanks to the generality of our approach, unlike previous models (e.g. Kemp et al., 2010; Lucas et al., 2014), we are not constrained to binary, present/absent effects, or multiple outcomes, such as different kinds of activations (e.g. Schulz & Sommerville, 2006). Our model can capture higher-order causal relationships, e.g., color/shape matches between blickets and machines (Sim & Xu, 2017). The animations can be extended to investigate more subtle cases such as both agent and recipient objects change features, or agent objects change rather than the recipient object. While we have focused on, and argued for the advantages of interactions between pairs of objects, our model can also make inferences from simultaneously presented causes by marginalizing over possible orders and intermediate states. Similarly, it can be applied to non-deterministic and conjunctive causes by introducing and marginalizing over hidden features.

Chapter 3

Causal roles in few-shot causal generalization

Chapter 2 explored one-shot causal generalization, and found evidence of systematic generalization predictions between participants. In everyday life, however, we may face more than one-shot cases. We could observe several instances interact with each other, and seek to generalize causal relationships from a batch of observations to new ones. In this chapter, we extend the setup to investigate causal generalization on the basis of multiple complete observations. Experiment 2 in Section 3.1 investigates few-shot causal generalization using the same animation as in Chapter 2, while Experiment 3 in Section 3.2 looks into this particular animation’s effect on perceived causal roles.

3.1 Experiment 2: few-shot causal generalizations

Experiment 2 used the same causal generalization task as in Experiment 1 (Figure 3.1A), but provided six learning observations instead of one (Figure 3.1B). Recognizing that prediction consistency may not fully imply consistency in causal law induction, we also elicited free guesses about the nature of the causal laws.



Figure 3.1: Experiment 2 interface (A) and learning material (B).

3.1.1 Methods

Participants

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.¹ We thus analysed 102 participants (37 female, aged 35 ± 10). Each participant was paid \$0.50 plus up to \$2.30 bonus. The task took 10.4 ± 7.2 minutes.

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

¹Data for Experiment 2 was collected summer 2020 at the height of the COVID-19 pandemic. See the “Exclusion criteria” section for data quality control.

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 enlarged the space of effects and greatly enriched 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. To minimize the potential hazard of recognizing these ordinal feature values for our participants, we used two drop-down menus, one for color shades and one for the number of edges, to visualize selected feature values in real time.









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 3.1B). 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

Table 3.1: 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.

is just one of a much larger set again of possible local causal law category structures. Accordingly, there is no strict measure of being “correct” in the generalization tasks, and we take this into consideration when analyzing data.

We composed generalization tasks according to the configurations in Table 3.1. We first defined the set of constraints to describe the objects in a generalization task, such that a novel object should have different shades, and/or different number of edges. Then we picked instances that satisfy these constraints, taking all the feature values appeared during learning into account. This led to $4 \times 4 = 16$ generalization tasks for each condition. Additionally, we included two catch-trials for each condition. We randomly chose two learning examples and turned them into generalization trials by hiding the result state. This resulted in $16 + 2 = 18$ generalization tasks for each condition.

Procedure

Each participant was randomly assigned to one of the four learning conditions (Figure 3.1B). 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. In the learning phase, the six pairs of agent and recipient stones were shown

in random order, one after another. By clicking a “Test” button, participants 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 3.1A), and remained visible for the rest of the task. After the learning phase, participants were asked to write down their best guesses 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 3.1A). Participants were instructed they would receive a \$0.10 for each correct prediction. We bonused participants as described afterwards. After the main task, participants provided demographic information and feedback. A demo of the task is available at <http://bramleylab.ppls.ed.ac.uk/experiments/bnz/myst/p/welcome.html>.

Exclusion criteria

To check data quality, we screened participants’ self-reports. As with past work, we required workers with Turk approval ratings of above 90%. But in line with (Chmielewski & Kucker, 2020), we found an unusual number of suspicious responses with very fast completion rates and nonsensical text responses. We thus chose to exclude participants if they failed to provide a task relevant response on the free text guess about the rule. In addition, we checked participant accuracy on the two catch-trials, and found that while overall accuracy is 41%, far above chance (5%), the excluded batch’s accuracy is just 8%, indistinguishable from chance. The full dataset along with the analysed dataset can be found at https://github.com/bramleyccslab/causal_objects.

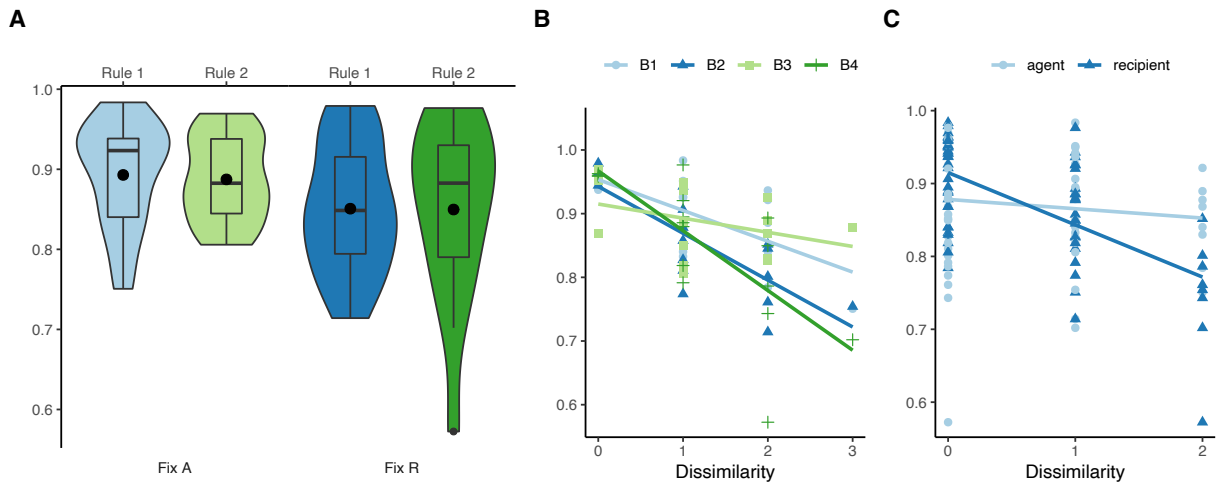


Figure 3.2: Experiment 2 behavioral results. All y -axes are Cronbach's alpha values. A. Task-wise inter-person consistency per condition. Violin plots are density. Black dots are mean Cronbach's alpha values per condition. The major bar in the box plot is the median and the box extent is the 25 and 75 quantiles. B. Inter-person consistency per task differences. C. Inter-person consistency per role differences.

3.1.2 Results

For participants' generalization predictions, we measured inter-participant consistency as in Experiment 1. To analyze free-text self-reports, we coded them into several categories and ran statistical tests on the coded labels.

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, $p < .001$. Thus, as in Experiment 1, participants produced systematic generalization patterns.

We then compared inter-person generalization consistency by condition. As illustrated in Figure 3.2A, the *fixed-agent* condition induced higher consistency ($\rho_T = 0.89 \pm 0.06$) than the *fixed-recipient* condition ($\rho_T = 0.85 \pm 0.1$), $t(31) = 2.12, p = .04, 95\%CI = [0.001, 0.08]$, while the difference in ρ_T between the ground truth condition was negligible, $t(31) = 0.22, p = \text{n.s.}$. 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 3.2B). 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 3.1). We found a significant negative relationship between task dissimilarity and generalization consistency, $\beta = -0.06, F(1, 62) = 37.48, p < .001$.

Finally, we fit a linear regression model predicting ρ_T with task dissimilarity, evidence balance, and ground truth, $F(3, 60) = 15.63, p < .001$. This revealed main effects of dissimilarity ($\beta = -0.06, p < .001$) and evidence balance (*fixed-recipient*, $\beta = -0.04, p = .01$), but not ground truth (*rule 2*, $\beta = -0.003, p = \text{n.s.}$). As depicted in Figure 3.2B, 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 3.2C). Holding recipient dissimilarity constant, increasing agent dissimilarity does not predict prediction

Table 3.2: Experiment 2 self-reports coding scheme

Scheme	Code	Definition
rule_type	specific	Predict an exact result state.
	fuzzy	Predict more than one possible result states.
	tacit	Leave one feature mentioned (tacit overwrites fuzzy).
	universal	Did not categorize causal relationships.
categorization	A	Group observations according to the mysterious stone.
	R	Group observations according to the normal stone.

consistency significantly, $F(1, 62) = 0.77, p = \text{n.s.}$; however, recipient dissimilarity does, $F(1, 62) = 38.8, p < .001$.

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 detailed coding scheme can be found in Table 3.2. The first coder categorized all free responses, and 15% of the categorized responses were then compared against the second coder’s. Agreement level was 92%. The full set of free responses 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”). Our ground truth rules all belong to the *specific* class of response. 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”). Note that we did not instruct participants on the specificity of self-reports, because we were primarily interested in the intuitions people had regarding the learning stimuli. Since participants could report from a rather under-specified range of productions, this makes the self-report data more naturalistic and the computational problem more challenging and exciting.

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*. 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 3.3 illustrates the coding results by learning condition. Guess specificity is summarized in Figure 3.3A. 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 ($\beta = 0.09, p = .008$), while evidence balance is not. Neither of these two factors are significant for the *fuzzy* type. Figure 3.3B summarizes participants’ guesses in terms of localization. No participant localized their rule in terms of both Agent and

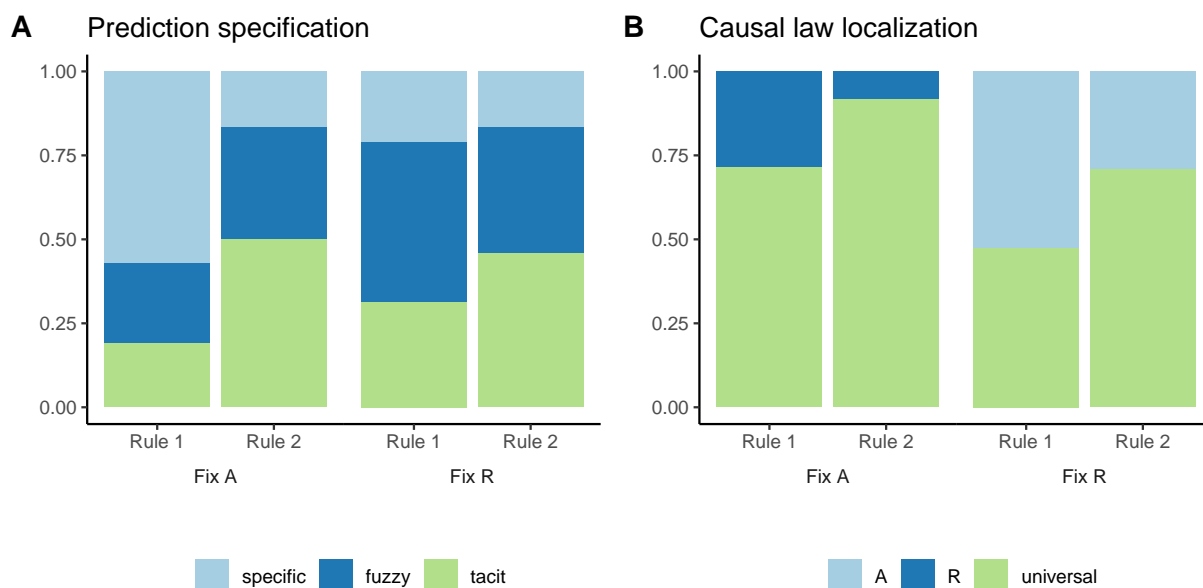


Figure 3.3: Experiment 2 rule guess categories.

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*, $\beta = -1.21, z = -2.3, p = .02$) and ground truth (*rule 2*, $\beta = 1.17, z = 2.3, p = .02$) were associated with more universal rules. There was no evidence for an interaction, $z = -0.5, p = \text{n.s.}$

These self-reports provided additional information to the generalization predictions participants made. We further coded each self-report into a corresponding causal function, and compared how consistent people’s generalizations were with their self-reports. Out of 102 participants, 27 were consistent throughout all the generalization tasks, and 9 participants never made a prediction that matched their self-reported causal relations. Most people are in-between. Since we elicited these self-reports before participants made any generalization predictions, these statistics are in line with our discovery in the previous chapter: Guesses may fall under the influence of other guesses. Therefore, we

should take both the generalization predictions and these self-reports into account, and view them as complementary to each other.

Causal asymmetry

Taking both generalization predictions and self-reports into account, these results reveal an asymmetry in causal generalization: The two underlying causal relationships induced feature changes in fact depend critically on both the agent and the recipient, but participants responses suggested they more readily identified the causal effect with the agent object. Consistency was higher for the fixed-agent condition where learners saw the same agent acting on various recipients (B1, B3) than conditions where agent was varied and the recipient was constant (B2, B4). Generalization consistency decayed more slowly when agents became more dissimilar to the training cases than for the matched degree of dissimilarity in terms of the recipient. Self-reported causal laws showed a much higher share of universal causal laws in the conditions where the agent was fixed, and more localization of causal laws were posited when the agents were varied. In fact, causal asymmetry is a well-known inductive bias in physical causation. White (2006) argues that people tend to judge the “cause” object to be more responsible for bringing the effect even when both objects play equally critical roles. For example, we more naturally think of a moving billiard ball as causing a previously static one to move rather than the static ball causing the moving one to slow down or stop even though the interaction is mathematically symmetric and jointly determined. This experiment thus supports the idea that there is a fundamental causal asymmetry to our causal generalizations.

3.1.3 Model fits

We extended the grammar introduced in Section 2.2 to cover a larger space of ordered feature relationships. Concretely, we introduced $+1$, -1 , $>$, $<$ at the “bind relation” step

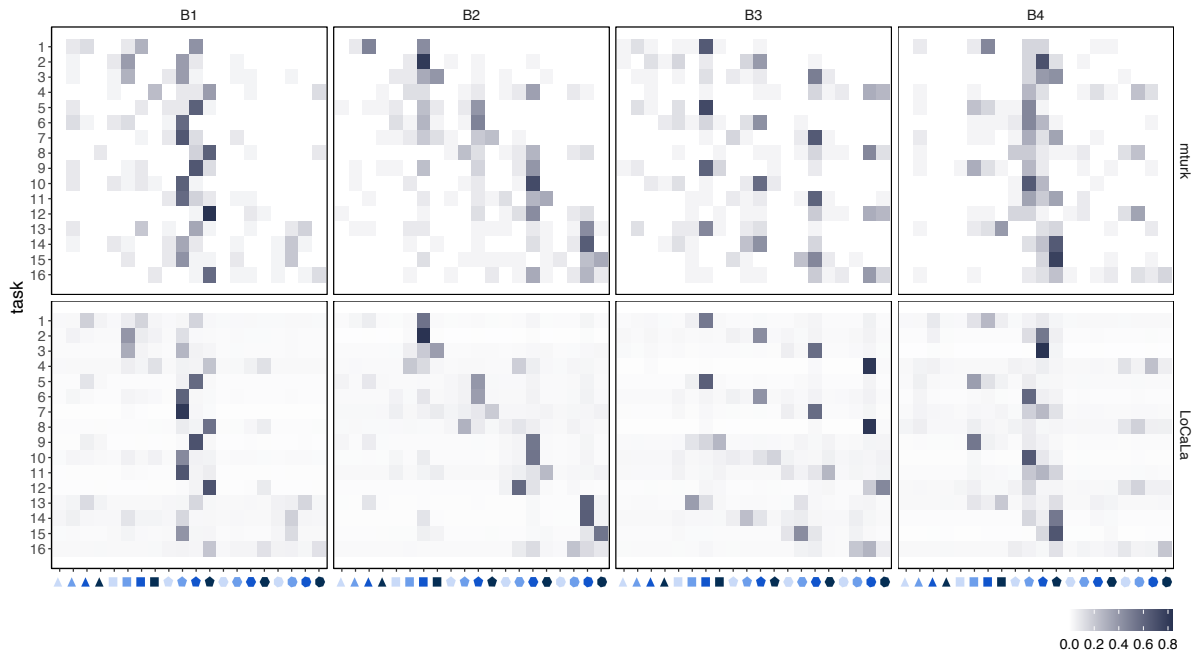


Figure 3.4: Experiment 2 generalization predictions. Upper row: proportion of participants predicting each stone type for r' (column) on each task (row). Second row: Fitted LoCaLa predictions.

to accommodate potential assertions about the ordering of feature values used in this experiment. 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 LoCaLaPro, so we focus on comparison between UnCaLa and LoCaLa.

Similarly as in Experiment 1, the UnCaLa model is fitted using the `optim` function in R with one softmax inverse temperature parameter t . However, different from the single-shot setup in Experiment 1, in Experiment 2 our LoCaLa model runs over six learning examples with potentially infinite categorizations. Therefore, we used Gibbs sampling to estimate the predictions under each parameterization, and optimised the parameters with a coarse grid search. On each iteration of the Gibbs sampler, one observation is

sampled and compared against the other five observations. According to Equation 2.7, when $\alpha = 5$ this observation has a 0.5 chance to create its own category in terms of size preference. This probability grows as α increases. Therefore, we centered the support values for α around 5, with an exponential increase for larger values, resulting in consideration of $\alpha \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 16, 32, 64, 128, 256\}$. 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.

We fit UnCaLa and LoCaLa to all $102 \times 16 = 1632$ data points taken together. Results are summarized in Table 3.3. 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. Recall that for the conditions where agent was varied, almost half of the participants reported non-universal causal rules, and when agent was fixed, very few participants responses suggested categorization. Here, $\gamma = 1$ together with $\alpha = 9$ captures this pattern: 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, this indicates a substantial heterogeneity across participant data taken together. As Figure 3.4 shows, our best fitting model indeed visually reproduces participants' generalization patterns.

Table 3.3: Experiment 2 model fitting results

	α	β	γ	t	Log likelihood	BIC
Baseline					-4889	9778
UnCala				3.19	-3706	7417
LoCaLa	9	256	1	9.5	-3462	6942

3.1.4 Interim discussion

Experiment 2 found strong causal asymmetry in generalizing few-shot causal observations, supporting a bias toward anchoring on the agent object in participants’ generalizations. Our localized causal laws model (LoCaLa) truthfully reflects this bias through its fitted focus parameter $\gamma = 1$. Here is how. When $\gamma = 1$, the model anchors its categorization process with the agent object, i.e., only considers the agent object’s features when deciding how general a causal law should be. When observing six interactions that always involve the same agent object along with a range of different recipient objects, the model will be biased to assume that all these interactions are ruled by the same causal law, and hence use all of the six observations to infer what this causal law consists of. Contrast to this case, when observing six interactions where the agent object is different every time and the recipient object is the same, then the model will (by assumption) infer that each interaction is determined by its own causal law, and only have one trial’s worth of evidence to infer the content of each causal law. The prediction, then, is that in the first condition (where the agent object is always the same) people should agree with each other more when they make predictions about what should happen in novel interactions, because they have abundant evidence to infer a single causal law, while in the second condition (where the agent object is varied) people are more uncertain and show less agreement with each other during generalization. If people are not biased toward anchoring on the agent object, however, then we should not observe a difference in inter-participant agreement between the two conditions.

Furthermore, causal asymmetry also presents implicitly in the experiment stimuli: Across our experiments, the agent object stayed the same and only the recipient object went through feature changes. As a result, one may feel that the agent object is more “powerful”, and recipients are “weaker” and susceptible to changes. Participants may have strong prior beliefs that an object being “active” suggests properties of agents are more likely to be communicated. In fact, the animation in Experiment 2 (Figure 3.1) differed from the recipient objects along three dimensions: (1) the agent object was marked by a glowing yellow border; (2) it moved toward the recipient object, which had no border; (3) when the agent object touched the recipient object, the recipient object would change into the result form, while the agent object remained unchanged. Each of these factors has theoretical reasons to induce the observed asymmetry in generalizations:

Movement As demonstrated by Michotte (1963), people watching simple physical interactions between two objects report that the moving object *causes* the state-change of the other object. In Figure 3.1, the fact that the object on the left moved might have led participants to consider that the object on the left was causally responsible for what happened to the object on the right.

Stability Change of state is another possible marker for introducing a causal asymmetry. Soo and Rottman (2018) discovered that in time series data, people are more likely to think that the object that remains stable is the cause and the object that changes is the effect. Again, in Figure 3.1, the agent object does not change during the interaction, while the recipient object changes after contact with the agent object. This asymmetry in state change (stability) may be another reason for biasing the focus parameter toward the agent.

Indicator In Figure 3.1, the agent object was marked by a glowing yellow border, and participants were instructed that a glowing yellow border means the object is “active”,

and that active stones can change the other inactive stones. If participants assume that such instructions are relevant (Grice, 1975; Sperber & Wilson, 1986), they might have constructed causal laws that are anchored in the objects labeled as being active. In addition, the glowing yellow border might also have led participants to pay more visual attention to the agent objects.

3.2 Experiment 3: Dissecting causal asymmetries

In this experiment, we attempt to disentangle different kinds of asymmetries coming from physical movement, state change, or nominal indication. We test one of the three cues (movement, change of state/stability, and visual-nominal indicator) separately, and measure their effect on the level of inter-person agreement in causal generalizations using a similar “keeping one object constant” design. As explained above, if the agent object (the object on the left in the original design) embodies cues that people use to anchor categorization, then we should observe higher inter-participant agreement in the condition where the agent object remains the same across interactions than in the condition where it varies. While we could in principle also look at the proportion of participants’ correct responses (i.e. responses that match the ground truth used to generate the training examples), this information is less helpful because there are many possible hypotheses, in addition to the ground truth rule, that are consistent with the data.

3.2.1 Methods

Participants

Two-hundred-and-two participants were recruited from Amazon Mechanical Turk (82 females, $M_{\text{age}} = 37.6 \pm 10.1$). Twenty-eight participants were excluded from analysis because they failed to provide task-relevant responses in free-text inputs, leading to one-hundred-and-seventy-four participants in total. Participants were paid both for their time and a performance-based bonus. The task took 12.5 ± 10.1 minutes.

Materials and design

Objects in this experiment are composed of a shading feature, ranging in {light, medium, dark, very dark} shades of blue, and number of edges, ranging from three (triangle) to seven (heptagon). The ground truth causal relationship we used to generate the training examples is the same across four experiments: the recipient object becomes one shade darker than itself and gains one more edge than the agent object (Figure 3.5). Note that the final state of the recipient object is a function of both its own features and those of the agent object. Therefore, the ground truth used in generating training examples does not pre-suppose asymmetries. During the learning phase, participants could observe causal interactions between six pairs of objects.

Cue 1, original is a replication of Experiment 2, in which we used the same animation (Figure 3.5A) as in Figure 3.1. Here, the object on the Left is intuitively seen as a causal agent, and the object on the Right is intuitively seen as the causal recipient.

Cue 2, movement aims to dissect the movement factor from the original animation. We designed an animation as in Figure 3.5B, where the Left object remains static while the Right object moves. When the Right object touches the Left object, the moving

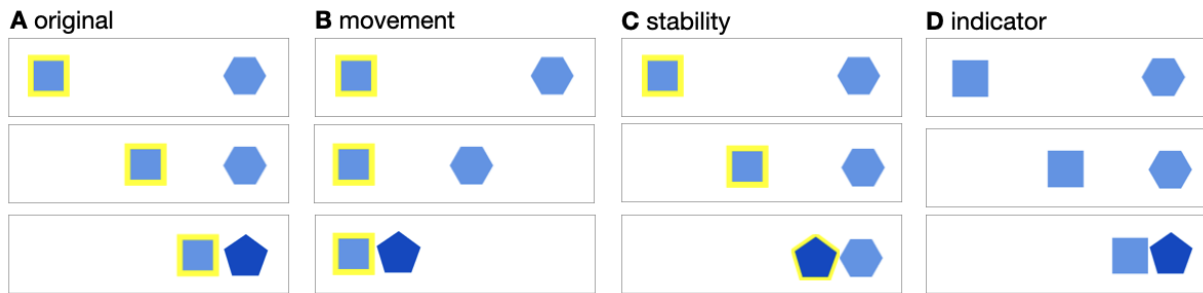


Figure 3.5: Experiment 3 cue illustrations.

object Right changes according to the ground truth causal relationship, while the Right object stays unchanged as in the original animation. By this animation, we removed the movement cue from the Left object.

Cue 3, stability uses an animation that removes the stability cue from the Left object (Figure 3.5C). While keeping the indicator and movement factors identical to the original animation, it is now the Left object, rather than the original Right object, that changes into the result form after the interaction.

Cue 4, indicator removes the glowing yellow border from the Left object (Figure 3.5D), while keeping the movement and stability factors as identical to the original animation.

For all four cues, we manipulated whether the Left object stayed the same across the six interactions while the Right object varied (fixed-L condition), or whether the Left object varied across interactions while the Right object stayed the same (fixed-R condition). In total, this leads to $4 \times 2 = 8$ between-subject conditions.

Procedure

Each participant is randomly assigned to one of the eight conditions described above. After reading instructions and passing a comprehension quiz, participants proceeded to a learning phase, where they were invited to test six pairs of objects' causal interactions

by clicking a “test” button and watched the animated outcomes. A visual summary of each tested pair was shown after the test on top of the screen, and remained visible until the end of the experiment. Next, participants were asked to write down their best guesses about the causal relationship between those objects. After that, participants went into the inductive generalization phase, where they made sixteen generalization predictions about novel pairs of objects. Each generalization task was presented sequentially and in random order. Participants composed their predictions by selecting from two drop-down menus, one for the shading feature and another for shape.

3.2.2 Results

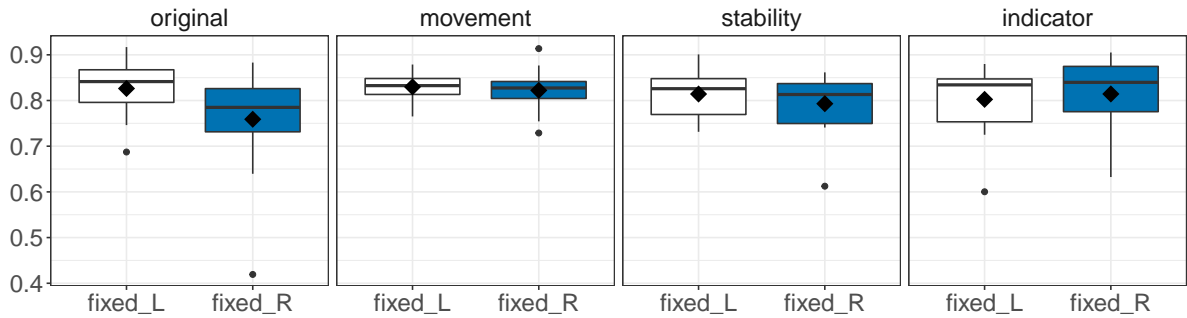
Systematic generalization

As analyzed earlier, our key dependent measure is the inter-participant agreement in generalization, which we measure using Cronbach’s alpha over how many participants in a given condition agree with each other in their predictions. For a total $8 \times 16 = 128$ generalization tasks, the mean consistency $\rho_\tau = 0.80 \pm 0.074$ with $\max = 0.91$ and $\min = 0.39$, demonstrating a high level of agreement between participants. Fisher’s exact test confirmed that for all eight between-subject conditions, participants’s generalizations are not random, $p < .001$. Therefore, we conclude that participants made systematic generalization predictions in all eight conditions, even though there were just six learning data points, no strict ground truth, and potentially misleading animation cue types.

Causal asymmetry in generalizations

Figure 3.6A summarizes task-wise consistency measures aggregated per condition. Cue 1 (original) replicates the causal asymmetry as in Experiment 2: participants in the fixed-L condition (original fixed-agent) made more homogeneous predictions across 16

A Cronbach's alphas



B Self-report labels

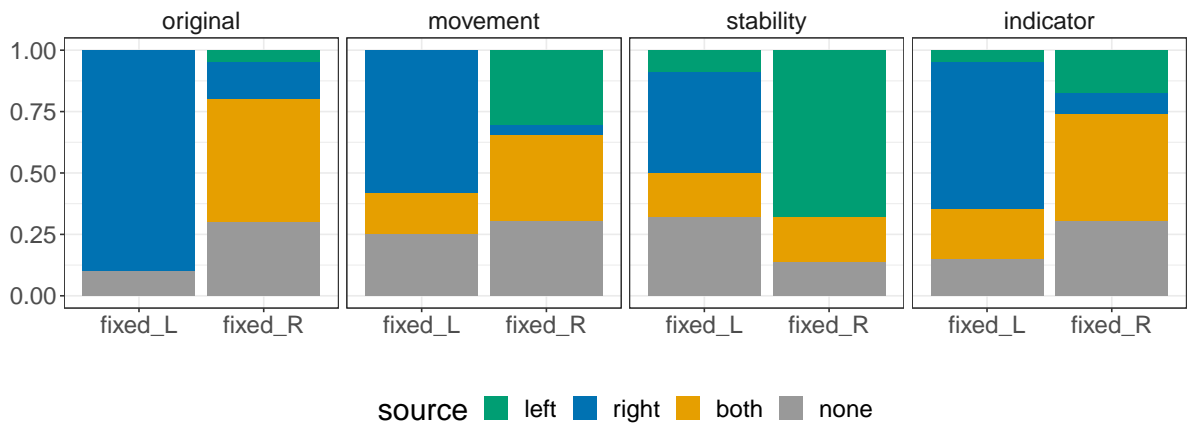


Figure 3.6: Experiment 3 results. A. Generalization congruency per condition; y-axis is task-wise Cronbach's alpha value. B. Self-report labels with respect to which object's features were mentioned for inference.

generalization tasks ($M_{\rho_r} = 0.83 \pm 0.06$), and those in the fixed-R condition (original fixed-recipient) made more diverse predictions ($M_{\rho_r} = 0.76 \pm 0.11$), $t(15) = 1.92, p = .04$.

However, none of the other three cues exhibits any causal asymmetry (Figure 3.6A), cue 2 movement $p = .29$, cue 3 stability $p = .64$, and cue 4 indicator $p = .18$. For these three cues, mean consistency measures are at similar levels between fixed-L and fixed-R conditions, and no significant difference was detected. This indicates that all three cues contribute together to the original causal asymmetry effect, and removing any one of them from the Left object leads people to treating both the agent and recipient equally in generalizations.

Focuses in categorization

To understand how people focus their categorization processes under different interaction cues, we analyzed participants' free text self-reports collected at the end of the learning phase. We coded these self-reports using *left*, *right*, *both* and *none* to represent which object people referred to when describing a causal relationship. For example, “become darker than itself” is classified as *right* for cues 1, 2 and 4, but as *left* for cue 3 (see Figure 3.5); “become darker than the moving stone” would be classified as *right* for cues 1, 3 and 4 (and as *left* for cue 2). Self-reports that took both objects into account are classified as *both*, such as “becomes one shade darker and converts into a shape with one more side than the active stone”. Those that do not refer to objects, not consistent with data, or makes no sense are classified as *none*.

Figure 3.6B visualizes percentages of coded self-reports for all four cues per fixed-L/R. For cue 1 original, 90% of participants in the fixed-L/fixed-agent condition reported causal relationships referring to the recipient object's feature only, while those in the fixed-R/fixed-recipient conditions showed a more diverse pattern: 50% mentioned both objects, 15% recipient-only, and 5% referring to just the agent objects' properties. A linear model predicting label *both* using the fixed condition as predictor confirms its significance, $\beta_{\text{fixed-R}} = 0.5, p < .001$.

Strikingly, only participants observing cue 1 original showed such difference between fixed-R and fixed-L conditions. For all the other three cues, participants showed no significant difference for label *both* in the two fixing conditions (cue 2: $\beta_{\text{fixed-R}} = 0.18, p = .16$; cue 3: $\beta_{\text{fixed-R}} = 0, p = 1$; cue 4: $\beta_{\text{fixed-R}} = 0.23, p = .11$). We fitted a multinomial regression model predicting self-report labels with this 4 cues \times 2 fixed-L/R mixed design, and took label *right* and the *original* cue as baselines. The fitted model revealed that interaction cue is indeed a significant predictor: between original and indicator cues, label *left* differs significantly, $\beta = 2.16, p = .03$; between original and movement cues, label *left*

($\beta = 2.66, p = .008$) and label *none* ($\beta = 2.19, p = .03$) both differ significantly, and between original and stability cues, all the other three labels *left* ($\beta = 4.33, p < .001$), label *both* ($\beta = 2.51, p < .001$), and label *none* ($\beta = 2.83, p < .001$) differ significantly. Fixed-L/R also appears to be a significant predictor for all three labels *left* ($\beta = -5.16, p < .001$), *both* ($\beta = -4.29, p < .001$), and *none* ($\beta = -3.93, p < .001$), but this is due to the difference between either the left or right object changes in the animations. In sum, these coded self-reports revealed that removing either factor from the original animation shifts participants' focus to both objects in the causal interaction, and as a result exhibits symmetry in generalizations as in the experiment design.

3.2.3 Interim discussion

In this experiment, we systematically examined what cues in causal interactions shape people's anchor of categorization in generalization. While successfully replicating the causal asymmetry in Experiment 2, we found that this asymmetry is sensitive to a mix of factors: object movement, stability in state changes, and visual and nominal causal role indicators. The original causal asymmetry depends on all three factors working together, and removing either one of them will shift the focus of categorization, leading people to assume that the causal law that determines what happens in that interaction is a joint function of both objects.

People's tendency to parse interactions in terms of a causal "agent" and "recipient" is often derided as an irrational bias. For instance, researchers scold lay people for saying that, in a physical collision, it is the moving ball that exerted a force on the static ball, when Newtonian mechanics tell us all forces in the scene are symmetric (White, 2006). We suggest that attributing causal agency to certain objects may serve functional roles, for example efficient generalizations, and people do take into account multiple factors when making that attribution decision. The demolition of causal asymmetry for cues

2-4 demonstrates that people can be fully aware of the symmetric ground truth causal relationship when they put equal focus toward both objects in the causal interaction. The fact that cue 1 original replicates causal asymmetry reinforces that an overly strong causal framing may effectively structure the kind of causal laws that people tend to conclude (Gopnik et al., 2004; Griffiths & Tenenbaum, 2009; Lucas & Griffiths, 2010; Mayrhofer & Waldmann, 2015), reflected both in self-report data and inter-participant generalization agreement levels.

3.3 Discussion

Via two experiments, we went deep into a causal asymmetry bias (White, 2006) in few-shot causal generalization. Our results support this bias: Experiment 2 had pairs of conditions (B1/B2, B3/B4) that shared the same underlying causal relationships, but swapped the dominant presence of the agents and the recipients. If agents and recipients were treated equivalently, this swapping would have had led to symmetric patterns of generalization. In contrast, we observed that participants in conditions B1 and B3 had significantly higher inter-person generalization consistency, and reported inferring fewer, more widely applicable rules.

Humans excel at generalizing from sparse data, in part because they use assumptions about causality as inductive biases to guide generalizations (Gelman, 2003; Rehder & Hastie, 2001). Experiment 3 found that, in intuitive perceptual causality settings, people rely on interaction cues such as whether an object is moving, or remains stable throughout the interaction, to decide whether the object has causal agency, and anchor their future generalization based on this. Different from verbal stimuli where the cause and effect can be communicated directly, perceptual causal stimuli needs such intuitive probing of causal relationships (Bramley, Gerstenberg, et al., 2018; Gopnik & Sobel, 2000; Ullman et al., 2017). These results suggest that any study that aims to measure causal

reasoning involving animated feature changes need to take these interaction cues seriously. However, since our goal in Experiment 3 was trying to dissect each cue from the original design of Experiment 2, we recognize that the new animations we designed here still mix two factors on one object at a time. Future research could expand on these results by employing a fully factorial design, manipulating the presence or absence of each cue independently. Other kinds of experimental techniques, such as iterated learning (Griffiths et al., 2008; Kirby, 2001; Yeung & Griffiths, 2015) could provide further evidence for the roles that potential cues of agency may play in object-based categorization.

In addition, we noted that participants were more likely to describe shape-related changes and leave color changes unmentioned, and this tendency can prevail the agent/recipient evidence balance control. In object cognition, it is well established that shapes and colors are perceived differently (Landau et al., 1988; Treisman & Gelade, 1980), and shape is thought to be taken to be a more fundamental feature than color (Wilcox, 1999). Our behavioral data demonstrate this pattern in a causal setting: Experiment 1 in Chapter 2 revealed that in one-shot causal generalization, participants made more systematic generalizations given shape-related effects than color-related effects, indicating that causal laws that are thought to induce more fundamental changes were generalized more consistently; here in Experiment 2, *tacit* rule guesses were more common for ground truth rule 2, and 91 percent of these *tacit* responses (29/32) described only the edge property — the shape feature, and left color changes unmentioned. These findings echo those in the developmental literature suggesting shape is perceived as a more fundamental or “essential” feature (Landau et al., 1988), and therefore more likely to be critical for an object’s causal powers. However, since such shape bias is sensitive to language (Landau et al., 1992) and can vary across cultures (Li et al., 2009), this finding may be restricted to this particular sample of English-speaking participants. Further studies are required to investigate the scope of this effect on the wider population.

Chapter 4

Bootstrapping compositional causal generalization

After looking at single-shot and few-shot causal generalization, we move on to more complex scenarios in this chapter. Following the same experimental setup, we extend the space of causal concepts to include compositions of multiple causes. Here, we take a constructive and compositional view, and suggest that people acquire rich mental representations by building on existing knowledge structures and enriching them with insights from new observations, which eventually leads to the generation of new concepts (Figure 4.1A). We first present a computational model of causal conceptual bootstrapping in Section 4.1, and then test this model's predictions in four pre-registered online experiments (Sections 4.2 & 4.3). We compare our bootstrap learning model with several alternative accounts in Section 4.4, and conclude with a discussion on the contribution and constraints of our approach in Section 4.5.

4.1 Computational model

So far, we have treated concepts as reusable, modular, and functional programs, and use a probabilistic grammar \mathcal{G} to generate such mental programs. Given observational data $D = \langle X, Y \rangle$ with inputs X and outputs Y , we can evaluate a program z using the inputs, checking whether $z(X)$ matches the actual observed output Y . If so, the likelihood of program z producing data D , $P(D|z) = 1$, and otherwise $P(D|z) = 0$. Using Bayes rule, one can infer a posterior distribution over programs Z given data D :

$$P(Z = z|D) \propto P_{\mathcal{G}}(z) \cdot P(D|z). \quad (4.1)$$

This Bayesian-symbolic concept learning framework has successfully characterized human judgments in a range of feature-based concept and category learning tasks (e.g. Bramley, Rothe, et al., 2018; Fränken et al., 2022; Goodman et al., 2008; Zhao et al., 2021, 2022). However, some core assumptions of this modeling framework limit its ability to bootstrap. Specially, the choice of generative grammar \mathcal{G} in these models is usually context-free. That is, when generating programs, grammar \mathcal{G} samples conceptual primitives without considering their contexts, or the other programs they combine with. As a result, when a target concept involves genuinely complex combination of conceptual primitives, to generate this target concept demands either an intractable enumeration with the grammar, or an infeasibly time-consuming sampling process (Van Rooij, 2008). Having previously discovered a sub-part of a complex target concept therefore does not help this kind of model, because it must reinvent the sub-concept on each use.

Even if such issues of tractability could be sidestepped with approximation methods, the context-free assumption innately contradicts a key inductive bias in human learning: When some concepts go together frequently, it makes sense to expect that the entire

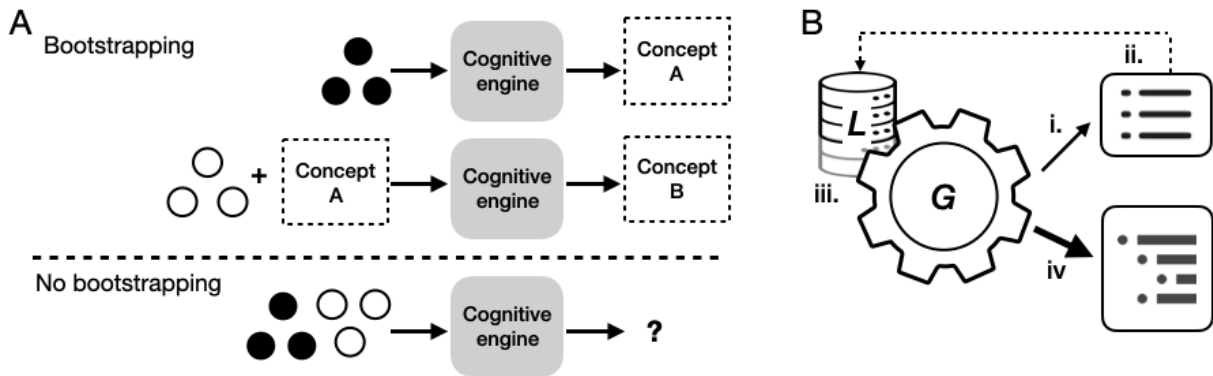


Figure 4.1: Bootstrap learning. A. Forming complex concepts by extending existing concepts to account for novel observations. Without the process of bootstrapping, the world may be too complex to conceptualize. Dots represent data points. B. Visual summary of the model. G is a generative grammar and L a latent conceptual library.

ensemble will be common in the future. This is a key idea that backends the learning-to-learn paradigm in computational cognitive science (e.g. Dasgupta et al., 2020; Griffiths & Tenenbaum, 2009; Kemp et al., 2010; Lake et al., 2015), as it provides an explanation to how people generalize and adapt to new situations so efficiently — by bringing in domain-specific expectations rather than starting from scratch. However, methods like hierarchical Bayesian models used in Griffiths and Tenenbaum (2009) usually make use of preset domain knowledge, and hence lack the space for such knowledge to grow. Approximation methods like particle filters (Gelpi et al., 2020; Thaker et al., 2017) mostly aim to approximate a target posterior, and therefore cannot provide the machinery for constructing adaptive priors.

Here, we loosen this context-free constraint in generative grammars, replacing a fixed set of conceptual primitives with a latent concept library L that can adapt to learning experiences over time (Fig 4.1B). Similar to other Bayesian-symbolic models, this bootstrap learning model also makes use of a generative grammar G to compose concepts (Fig 4.1B.i). Unlike existing methods, if some generated concepts are useful in explaining part of the learning data, the model will cache these concepts into its library

L (Fig 4.1B.ii-iii), allowing them to be drawn as primitives in explaining other observations. Hence, grammar \mathcal{G} can effectively compose more sophisticated concepts by sampling from this enriched library (Fig 4.1B.iv). Going from a fixed set of conceptual primitives to an ever-evolving concept library, this cache-and-reuse mechanism enables a cognitively-bounded learner to arrive representations at search depths far beyond their search capacity.

Such pervasive reuse posits a challenge to PCFGs. For PCFGs, while it is possible to learn an informed distribution over primitives and production rules, there is no clear way how to include generated fragments. Fortunately, we can draw upon techniques from adaptor grammars (AG; Johnson et al., 2007) to solve this problem. As a Bayesian-symbolic model, our formalization shares all of the virtues of the PCFG framework, but crucially, supports abstraction and reuse in ways that a PCFG framework does not. While adaptor grammar is not the only option that offers a solution—methods like fragment grammars (O’Donnell et al., 2009) or an arbitrary term rewriting system (Bezem et al., 2003) may also satisfy such requirement, existing research like Liang et al. (2010) provides a clear guidance on implementation details that greatly facilitates our formalization.

The task interface is similar as in Chapters 2 & 3: In each learning trial, participants see an agent object collide with a stationary recipient object, which consequently transforms in some way (Figure 4.2). Participants are instructed to reason about the causal relationship between features of the agent and recipient objects, and the resulting changes in the recipient. We refer to this task when introducing our formalization.

4.1.1 Causal programs in combinatory logic

Since we expect modular reuse of program fragment, we formalize programs in combinatory logic (CL; Schönfinkel, 1924) to bypass variable binding problems in generating

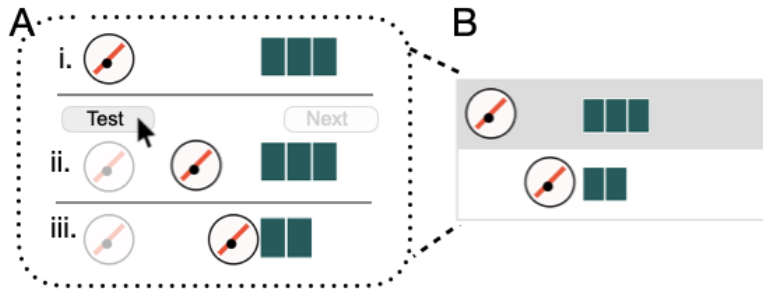


Figure 4.2: Experiments 4-7 example task. A. Task animation. B. Visual summary of the animation.

functional programs (Crank & Felleisen, 1991), as well as echoing recent work by Piantadosi (2021) arguing that CL provides a unified low-level coding system for human mental representations.

Functional terms

CL programs are composed of *terms* and input *variables*. Terms are interpreted as functions by definition, and can be composed iteratively to generate new terms. Starting with our assumption that relevant features are salient to the learner, we let function $getFeature(o) = v$ take an object o as input and return its feature value v ; function $setFeature(o, v) = o'$ sets object o 's feature value to v , returning an updated object o' . For our task, as illustrated in Figure 4.2, we consider a minimal set of base terms: $getSpot()$, $getStripe()$, $getSegment()$, and $setSegment()$. Since numbers of spots, stripes or segments are all numerical, we include some operations over these feature values as additional base terms (or chunks that are salient from past experience): addition $add(v, u) = v + u$, subtraction $sub(v, u) = v - u$, and multiplication $mult(v, u) = v \times u$. We additionally consider four primitive integers 0, 1, 2 and 3 because these are the quantities involved in the learning examples.

Types

Since terms are functions, they are naturally constrained by their input domains and output co-domains, known as being “typed”. Taking object (obj) and integer (int) as

Table 4.1: Model AG base terms

Terms	Type signature
<i>getSpot, getStripe, getSegment</i>	$\text{obj} \rightarrow \text{int}$
<i>add, sub, mult</i>	$\text{int} \rightarrow \text{int} \rightarrow \text{int}$
<i>setSegment</i>	$\text{obj} \rightarrow \text{int} \rightarrow \text{obj}$

base types, a type t for a term is written as $t_I \rightarrow t_O$, where t_I and t_O are types for the input and output respectively. Table 4.1 lists the type signatures for the primitive terms we introduced earlier. Conventionally, type signatures are written as subscripts, like $\text{getSpot}_{\text{obj} \rightarrow \text{int}}$. Type signatures are critical for ensuring valid compositions. For instance, we can plug in any subprogram that returns a number as one argument for *add*. However, we cannot use *setSegment* as an argument for *add*, because *setSegment* returns an object while *add* requires numbers as inputs. Conveniently, we use $t(z)$ to read off the type of term z , e.g., $t(\text{getSpot})$ is $\text{obj} \rightarrow \text{int}$.

Routing variables

When evaluating nested functions, it is essential to make sure input variables are sent to the right place. In lambda calculus, for example, this is done by ensuring the unique and uniform use of symbols representing the same variable throughout the nested layers. As a result, when composing subprograms, additional machinery is necessary to determine what symbols are re-used and where. To solve this variable binding problem, CL introduces some terms that serve as “routers” (Figure 4.3A): For a tree-like structure $[\text{router}, x, y]$, router **B** routes an incoming variable z to the right of the tree— z is first fed to the right-hand y , and the result of this is then sent to x . Similarly, router **C** routes z to the left, router **S** sends z to both sides, and router **I** is an identity function that returns an input as it is. For N input variables, we can concatenate N routers in corresponding order.

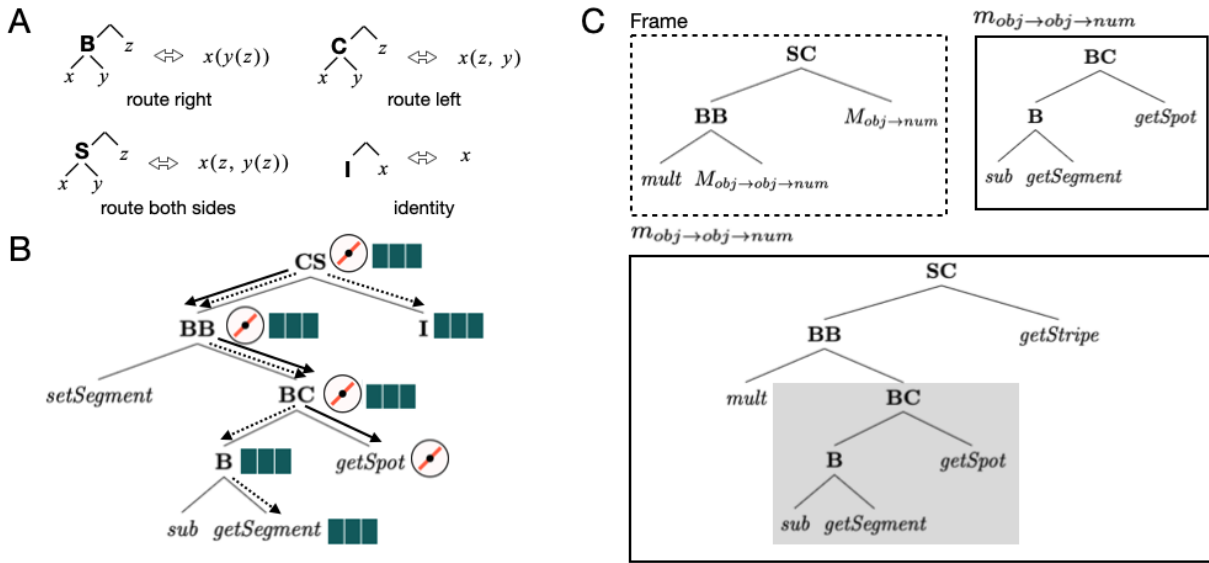


Figure 4.3: Model AG components. A. Visualization for CL routers, adapted from Liang et al. (2010). B. Example program. Black arrows denote routing of the agent (magic egg), and dotted arrows for routing of the recipient (stick). C. Example frame (dotted box) and example programs. Shaded area in the bottom program reuses the program on top-right.

Causal programs

With variables, terms, types and routers all at hand, we are now all set to consider an example program as unpacked in Figure 4.3B:

$$[\text{CS } [\text{BB } \text{setSegment } [\text{BC } [\text{B } \text{sub } \text{getSegment}] \text{getSpot}]] \text{I}]$$

Evaluating this example program with the agent and recipient objects in Figure 4.2 as input, the first router **CS** routes the agent object to the left (solid arrows), and routes recipient object to both sides (dotted arrows), so on and so forth. After instantiating all the variables, this program reads as: “take the recipient and make its number of segments to be its original number of segments minus agent’s number of spots”, and outputs a result object with two segments.

Program evaluation

Let observational data $D = \langle X, Y \rangle$ where X are input data and Y the output. The likelihood function of program m_t producing observational data D is given by:

$$P(m_t|D) = \begin{cases} 1 & \text{if } m_t(X) = Y \\ 0 & \text{otherwise.} \end{cases} \quad (4.2)$$

We here use a simple deterministic likelihood function as a starting point, and later introduce a soft version that takes noise/uncertainty into account. Taking the agent and recipient objects in Figure 4.2 as input, the example program in Figure 4.3C returns the result object that is a stick of two segment, hence its likelihood for producing the example task is 1.

4.1.2 Bootstrapping with adaptor grammars

The core difference between AGs and PCFGs is that AGs allow caching: a generated program can be added to the library of “primitives” for later reuse; program generation can result from either composing a new program, or sampling directly from the cache. We may think about these cached programs as “concepts”: They possess some internal complexity, serve certain functional aims, and more importantly, can be reused directly without having to be “rediscovered” by regenerating all the internal parts again. The caching mechanism of AGs thus facilitates bootstrapping via chunking useful subprograms and reusing them as building blocks anywhere that their type constraints allow (Liang et al., 2010).

Algorithm 2 Adaptor Grammar $AG(\tau, X)$

Require: Type $\tau = t_0 \rightarrow \dots \rightarrow t_k$ **Require:** Variables $X = \{x_0, \dots, x_n\}$ Sample $\lambda \sim U(0, 1)$ **if** $\lambda \leq \lambda_1$ **then** $z_L \sim \{z \mid t(z)_{\text{output}} = t_k\}$ $\mathbf{r} \sim \mathbf{R}^{|X|}$ $i \leftarrow |t(z_L)|$ **while** $i > 0$ **do** $X' = r(X)$ $\tau' = t(X') \rightarrow t(z_L)_{i-1}$ $AG(\tau', X')$ $i \leftarrow i - 1$ **end while****else**Return* $z \in C_\tau$ with probability λ_2 **end if**

▷ Construct new hypothesis

▷ Sample a term, e.g., `mult`▷ Sample a router, e.g., `SC`

▷ Grow RHS branches

▷ Get routed variables

▷ Get type constraints

▷ Compose recursively

▷ Fetch existing hypothesis

Generative process

As PCFGs, AGs implicitly define a distribution over programs via a generative process. Let \mathcal{L} be a program library consisting of base terms and/or some programs, with probability λ_1 grammar \mathcal{G} constructs new programs of type t , and otherwise it returns a cached program of type t with probability λ_2 .

We employ a tail recursion for the construction step as in (Dechter et al., 2013) in order to efficiently satisfy type constraints in Table 4.1. As demonstrated in Algo. 2, for a given target type $\tau = t_o \rightarrow \dots t_k$, and a set of input variables $X = \{x_0, \dots, x_n\}$, we start by sampling a left-hand side term LHS whose output type is the same as the output type of t . Based on how many variables are fed to this stage, grammar \mathcal{G} then samples a router \mathbf{r} of corresponding length that sends these variables to either/both branches. Since both LHS and router \mathbf{r} are given, now the type signature for the right-hand side of the tree is fully specified, because it has all the input types (routed by \mathbf{r}) and a required output type (to feed into LHS). Therefore, we apply the same procedure iteratively to

get this right-hand side subprogram RHS, returning the final program [r LHS RHS]. The constructed program [r LHS RHS] is then added to the program library \mathcal{L} (caching).

Each step in this generative process comes with a probability distribution. For the starting program library \mathcal{L} , we assume a uniform distribution over terms that share the same type signature. We also assume a uniform distribution over routers sharing the same number of variables to route. Following the notation in (Liang et al., 2010), for a collection of terms C_t of type signature t , let N_t be the number of distinct elements in C_t , and M_z the number of times z occurs in C_{z_t} :

$$\lambda_1 = \frac{\alpha_0 + N_t d}{\alpha_0 + |C_t|}, \quad \lambda_2 = \frac{M_z - d}{|C_t| - N_t d}. \quad (4.3)$$

Hyper-parameters $\alpha_0 > 0$ and $0 < d < 1$ control the amount of sharing and reuse. Since λ_1 is proportional to $\alpha_0 + N_t d$, the smaller α_0 and d are, the less construction and more sharing we have. Similarly, λ_2 is proportional to M_z , hence the more frequently a program is cached, the higher weight it gets, regardless of its internal complexity.

Approximate Bayesian inference

Given this probabilistic model, we are faced with the challenge of efficiently approximating a posterior distribution over latent programs given learning data, according to the prior distribution (Equations 4.3) and likelihood function (Equation 4.2). Following previous work suggesting that human learners make inferences by sampling from an approximate posterior instead of tracking the entire posterior space of possibilities (Bramley et al., 2017), we use known methods for sampling from Pitman-Yor processes (Pitman & Yor, 1997), such that conditional on a program library at any given moment, learners can make appropriate inferences about the probabilities of different explanations for new or salient events.

Concretely, we use a Gibbs sampler for program library L : for the i -th iteration, conditional on the library from previous iteration L_{i-1} , sample an updated library L_i and add it to the collection of samples. For the sampling step, let library L_{i-1} generate programs with probabilities defined above and calculate their likelihoods with respect to learning data D . The caching mechanism of AG will add consistent programs into library L_{i-1} , or increase the counter for those already present in L_{i-1} , resulting in an updated library L_i . In practice, our learning data is very sparse, hence we adopt both breadth-first search (Knuth, 1973) and beam search (Hayes-Roth et al., 1977) to facilitate search for programs that can produce learning data. For the outer loop, we use “frames” for intermediate programs built with typed placeholders (Figure 4.3C). Fixing a generation depth, we first enumerate a set of frames \mathcal{F} . Next, sample a frame from \mathcal{F} according to generation probabilities. The sampled frame can then be unfolded, replacing its placeholders with programs of required types, yielding a set of fully-articulated programs M (Figure 4.3C). If some programs $M^* \subseteq M$ produce learning data with likelihood 1, we stop the search; otherwise, we sample another frame from \mathcal{F} and repeat. If no programs are consistent with data after depleting frame set \mathcal{F} , we increase depth by 1 and repeat until a maximal cap is met. Because of this comprehensive search-check-sample procedure, we expect our Gibbs sampler to approximate the true posterior quickly and without the need for extensive burn-in.

Generalization predictions

We can run the generative procedure of grammar \mathcal{G} using the sampled libraries 10,000 times to approximate a distribution $Dist_M$ over latent causal programs, and make generalization predictions about new partially observed data $D^* = \langle X^*, ? \rangle$, producing a predicted distribution $Dist_P$ over generalizations.

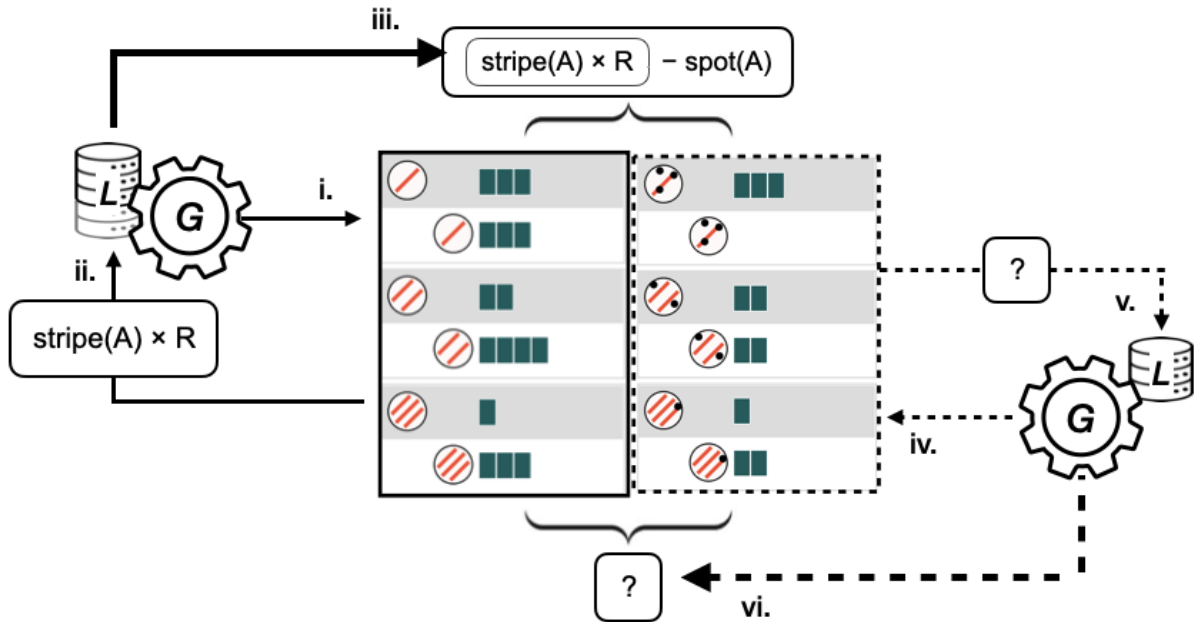


Figure 4.4: Model AG example bootstrap learning trajectories over six observations, either i-iii from simple to complex, or iv-vi the other way around.

4.1.3 Task and predictions

Model AG predicts that successful search for a complex target concept is heavily reliant of having good, previously-learned abstractions. To test these predictions, we designed a series of few-shot causal learning and generalization tasks inspired by Zhao, Lucas, et al. (2022). In these tasks, participants see several examples in which a causal agent A interacts with and changes the features of a recipient object R , into a result form R' (Figure 4.2). Since the recipient and result objects in these tasks only involve segment number changes, for simplicity we use R as a shorthand for $getSegment(Recipient)$, and R' for $getSegment(Result)$. Similarly, $stripe(A)$ is short for $getStripe(Agent)$, and $spot(A)$ for $getSpot(Agent)$. To further improve readability, we will write functions $mult()$, $add()$, and $sub()$ as \times , $+$, $-$, respectively in later texts.

Let us consider a cognitively-bounded learner that can only search for one simple concept at a time. Given the six learning observations in Figure 4.4, this learner could first reason about the three observations on the left (solid border, Figure 4.4). This might lead

them to conjecture that “stripes on the causal agent can multiply the number of segments on the recipient object”, such that the result number of segments $R' \leftarrow \text{stripe}(A) \times R$ ¹ (Figure 4.4.i). With this concept cached and ready for reuse (Figure 4.4.ii), when facing all six observations, this learner would be able to compose a complex concept that “spots on the causal agent subtract segments on the recipient object on top of what stripes can do”, such that the result number of segments $R' \leftarrow \text{stripe}(A) \times R - \text{spot}(A)$ (Figure 4.4.iii). This compositional concept involves two different features of the causal agent, each of which plays a different causal function, exceeding the learner’s search depth capacity.

This effect of bootstrap learning is further demonstrated if we swap the processing order of the same six learning observations in Figure 4.4. Even though the three observations on the right (dashed border, Figure 4.4) favor concept $R' \leftarrow \text{stripe}(A) \times R - \text{spot}(A)$, this is beyond what our bounded learner is able to discover (Figure 4.4.iv), and fail to learn reusable bits (Figure 4.4.v). When further challenged generating a concept that can explain all six observations, this learner will continue to struggle (Figure 4.4.vi).

4.2 Experiments 4 & 5: Curriculum-order effect

Experiment 4 ($N = 165$) used the target concept analyzed in the “Task and prediction” section as ground truth, i.e., $R' \leftarrow \text{stripe}(A) \times R - \text{spot}(A)$. In a follow-up Experiment 5 ($N = 165$), we flipped the roles of the stripes and spots of the agent object.

¹We omit the `setSegment()` wrapper here too for readability.

4.2.1 Methods

Participants

For Experiment 4, 165 participants ($M_{\text{age}} = 31.8 \pm 9.9$) were recruited from Prolific Academic, according to a power analysis for three between-subject conditions with fixed effects of 0.31 effect size. Participants received a base payment of £1.25 and performance-based bonuses (highest paid £1.93). The task took 9.69 ± 4.47 minutes. No participant was excluded from analysis.

Stimuli

The agent object was visualized as a circle that moved in from the left of screen and collided with the recipient (Figure 4.2). The agent object varied in its number of stripes and randomly positioned spots. The recipient object took the form of a stick made up of a number of cube-shaped segments. During learning, all feature values were between 0 and 3. The rule we used to determine the recipient’s final number of segments was $R' \leftarrow \text{stripe}(A) \times R - \text{spot}(A)$. Learning materials were shown as in Figure 4.6A, divided into two phases. We used the example concept analyzed earlier in “Task and predictions” as ground truth, i.e., $R' \leftarrow \text{stripe}(A) \times R - \text{spot}(A)$. We examined three curricula between-subjects: *construct*, *de-construct*, and *combine*, as shown in Fig. 4.6A’s left column. The *construct* curriculum demonstrates the stripe feature’s multiplicative power in Phase I, and then introduces the spot feature’s subtracting power in Phase II. The *de-construct* curriculum takes the same learning examples as in *construct*, but swaps Phase I and Phase II. The *combine* curriculum shares the same Phase I as in *construct*, but when introducing the spot feature’s subtracting power in Phase II, it keeps $\text{stripe}(A) = 1$ throughout, making it ambiguous how $\text{stripe}(A) \times R$ and $R - \text{spot}(A)$ should be combined. Overall, at the end of Phase II, both the *construct* and *de-construct*

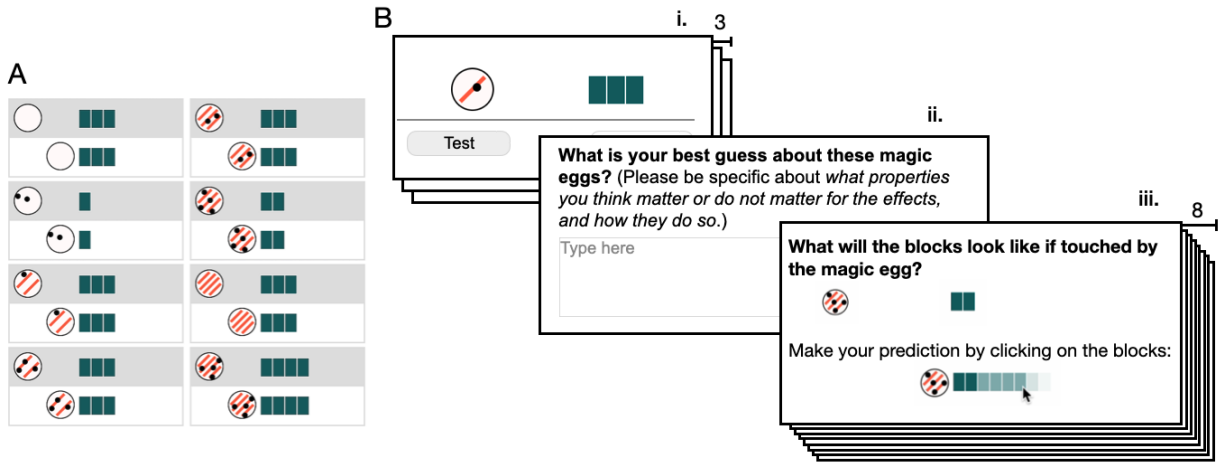


Figure 4.5: Experiment 4 material and procedure. A. Generalization trials. B. Procedure in one phase. i. Test causal interactions. ii. Collect self-report. iii. Make generalization predictions by clicking on a block of segments.

curricula provide enough evidence to favor the ground truth, and the *combine* curriculum is indifferent between the ground truth and a commensurately complex and equally evidence-compatible alternative $R' \leftarrow \text{stripe}(A) \times (R - \text{spot}(A))$.

Generalization trials were selected via a greedy entropy — minimizing search in order to select a set that distinguished well between a set of hypotheses favored by model AG. Precisely, for 5 possible stripe values (0-4), 5 possible spot values (0-5) and 4 possible recipient segment values (1-4), there are $5 \times 5 \times 4 = 100$ possible Agent (a) – Recipient (r) pairs. As a starting point, we ran a version of model AG that allows up to two exceptions in Phase I and four in Phase II, resulting in a large group of candidate programs M . We then grouped this very large M into 216 equivalence classes. That is, for two programs $m_1, m_2 \in M$, if $m_1(a, r) = m_2(a, r)$ for all the 100 possible pairs, then m_1 and m_2 belong to the same equivalence class. We kept the shortest program m_s in each equivalence class to be the class label, and recorded the size of each equivalence class to be their weight. Next, after excluding the learning pairs, we ran a greedy maximization of expected information gain for the rest of the pairs. Precisely, we started with selecting the Agent–Recipient pair that best distinguishes all these 216 programs,

and then greedily selected the next best, taking previously-chosen pairs into consideration. To measure how well a pair distinguishes between the programs, we computed the expected information gain (EIG) for this pair over all possible programs, taking the normalized program weights from their corresponding equivalence classes as the prior:

$$\text{EIG}(m, d) = H(m) - H(m|d), \quad (4.4)$$

where $H(\cdot)$ is the Shannon entropy:

$$H(X) = - \sum_{x \in X} p(x) \log p(x). \quad (4.5)$$

After running greedy maximisation over EIG, we settled on a list of ordered Agent–Recipient pairs. We then picked the top eight of them, and replaced a four-stripe, zero-dot agent with the zero-stripe, zero-dot agent because we were curious about how people would react to this limit case. This led to the eight generalization trials used in this experiment, shown in Figure 4.5A. For these generalization tasks, an arbitrary segment number (0 to 16) could be selected putting a nominal eyes-closed floor level of performance at $1/17 = 5.88\%$.

Live demos are available at <https://bramleylab.ppls.ed.ac.uk/experiments/bootstrapping/p/welcome.html>, and pre-registration at <https://osf.io/ud7jc>.

Procedure

Each participant was randomly assigned to one of the three learning conditions, *construct*, *de-construct*, and *combine*. After reading instructions and passing a comprehension quiz, participants went through experiment Phase I and then Phase II (Figure 4.5B). In each phase, a participant tested three learning examples as shown in Figure 4.6A,

each appearing sequentially and as ordered in Figure 4.6A. Participants watched the animated causal interactions by clicking a “Test” button (Figure 4.5B.i). Once tested, a visual summary of the learning example including the initial and final state of the recipient was added to the screen and remained visible until the end of the experiment. After the learning stage, participants were asked to write down their guesses about the underlying causal relationships (Figure 4.5B.ii), and make generalization predictions for eight pairs of novel objects (Figure 4.5B.iii). Generalization trials (Figure 4.5A) appeared sequentially. Once a prediction was made, the trial was replaced by the next one. The generalization object pairs in both Phase I and Phase II were the same, but their presentation order was randomized for each participant and in each phase.

Experiment 5

Experiment 5 is a feature counterbalanced replication of Experiment 4, using true rule $R' \leftarrow \text{spot}(A) \times R - \text{stripe}(A)$. Another 165 participants ($M_{\text{age}} = 33.8 \pm 10.1$) who did not participate in Experiment 4 were recruited from Prolific Academic. The task took 9.8 ± 5.2 minutes. No participant was excluded from analysis. Payment scale (highest paid £1.95) and procedure are identical to Experiment 4. Stimuli and pre-registration are available at <https://osf.io/k5dc3> and in SI. We conducted a two-way ANOVA to analyze the effect of feature counterbalancing and curriculum design on Phase II generalization accuracy. While both factors had significant main effects (curriculum design: $F(2) = 35.75, p < .001$, feature counterbalancing: $F(1) = 36.86, p < .001$), there was no significant interaction, $F(2) = 2.769, p > .5$. This indicates that people may be treating stripe and spot features differently, but this difference does not dramatically interfere with our main results.

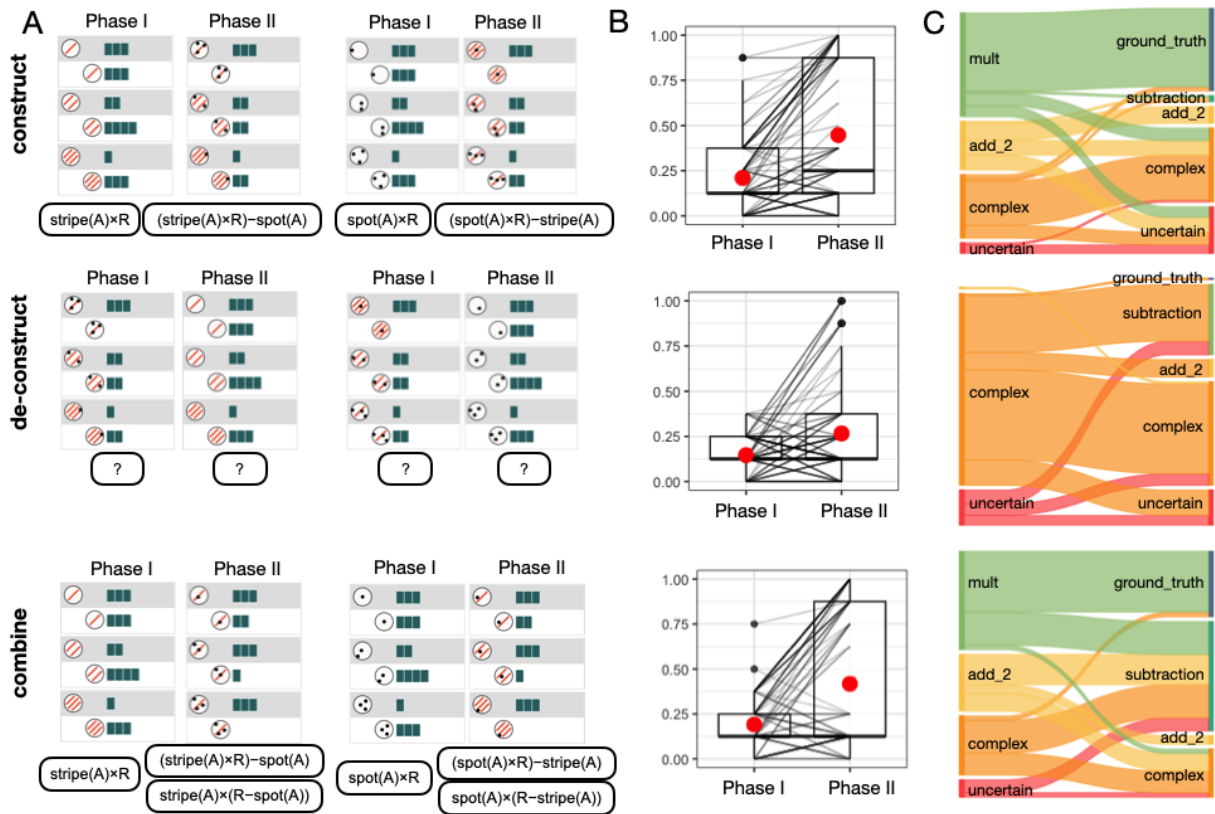


Figure 4.6: Experiments 4-5 material and results. A. Curricula in Experiment 4 (left) and 5 (right). Texts below each phase are data-compatible causal concepts. B. Participants generalization accuracy (match to ground truth). Box plots show the first and third quantiles with lines for the medians; red dots mark the means. C. Coded self-reports.

4.2.2 Results

First, we observed a significant difference in Phase II generalization accuracy (i.e. match to ground truth) between the *construct* and *de-construct* curricula. As illustrated in Figure 4.6B, participants under the *construct* curriculum achieved an accuracy of $44.7\% \pm 38.3\%$, significantly higher than those with the *de-construct* curriculum of only $22.6\% \pm 27.5\%$, $t(1717) = 8.13, p < .001, 95\%CI = [.14, .24], d = 0.38$ (chance accuracy: $1/17$

= 5.88%). The large standard deviations here imply a wide-spread individual difference in causal generalizations, which crystallizes when looking at participants' self-reports (Figure 4.6C). We coded participants' self-reports according to whether the content matches the ground truth, describes an operation such as multiplication, subtraction, or addition, is uncertain, or involves complex reasoning patterns drawing upon conditionals, positions of spots or relative quantities (Table 4.2).² Two coders categorized participants self-reports independently. The first coder categorized all free responses, and 15% of the categorized self-reports were then compared against the second coder's. Agreement level was 97.6%. For Phase II self-reported guesses, 36.4% of participants in the *construct* curriculum were classified as describing the ground truth (Figure 4.6C), and only one participant (0.85%) in the *de-construct* condition did so, $t(113.09) = 7.49, p < .001, 95\%CI = [.26, .45]$. A deeper dive into those self-reports revealed that, for those who induced that one feature multiplies in Phase I, 75.5% subsequently landed on ground truth in Phase II, showing a clear bootstrap learning trajectory. Recall that at the end of Phase II in both *construct* and *de-construct* curricula, participants have seen identical learning information (Figure 4.6A), hence this substantial difference in final learning performance coheres with our main claim that people reuse sub-concepts to compose more complex ones. Merely observing evidence that favors a target concept is not sufficient to induce this concept.

The absence of any matches with the ground truth in self-reports in the *de-construct* curriculum also reflects a strong garden-path effect (Bever, 1970). Notably, more than eighty percent (83.8%) of participants in the *de-construct* condition came up with guesses classified as “complex” in Phase I. For example, one participant wrote: “*If there are more stripes than dots the stick is reduced in length. If there are equal stripes and dots*”

²The distinction between Ground truth, Alternative, and Comp lies at whether participant explicitly reported the order of operations using words as “then”, “after that”, etc. Note that Participant 461 is classified as reporting the Alternative rule because the second half of their report implies multiplication over the subtracted number of segments.

Table 4.2: Experiments 4-7 self-reports coding scheme

Code	Definition	Example
Ground truth	Equivalent to the ground truth concept.	<i>the length is multiplied by the number of lines and then the number of dots is subtracted</i> (Participant 43)
Alternative	Equivalent to the alternative causal relation in each experiment.	<i>the dots subtract from the segments by their number, and the number of lines is multiplied by the amount of segments</i> (Participant 461)
Comp	Unclear how two sub causal concepts should be combined.	<i>the lines multiply the segments and the dots subtract the segments</i> (Participant 451)
Add 2	Add two segments to the recipient object (nothing happens if the agent object's feature value is 1).	<i>adds 2 segments to the stick only if there are 2 or more stripes on the egg</i> (Participant 35)
Mult	One feature of the agent object multiplies the recipient object.	<i>the number of stripes multiplies the number of segments</i> (Participant 59)
Subtraction	One feature of the agent object is a subtractor to the recipient object.	<i>each spot on the egg takes away one stick</i> (Participant 100)
Complex	Describe the stimuli without generalizing a rule, or report a different rule for each observation.	<i>3 dots means the sticks disappear, 2 dots means 2 sticks, 1 dot means add another stick</i> (Participant 161)
Uncertain	Not knowing, unsure, or confused about the learning stimuli.	<i>i don't have a clue!</i> (Participant 57)

then the stick stays the same. If there are more dots than stripes the stick increases in length.” The average length of Phase I guesses for the *de-construct* curriculum was 168 ± 145 characters, significantly longer than answers in the *construct* curriculum's 112 ± 68.1 characters, $t(168.09) = -3.76, p < .001, 95\%CI = [-85.65, -26.72]$. These longer and more complex initial guesses appeared to influence the second phase of the experiment. In *de-construct* Phase II, after seeing the simpler examples, 46.9% of

the complex-concept reporters either stuck with their initial complex guesses or embellished them even more, resulting in 45.3% complicated self-reported causal concepts in Phase II. This proportion was significantly higher than both those given the *construct* (32.7%, $p = .05$) and *combine* (20.8%, $p < .001$) curricula. Furthermore, only 30.8% of participants in Phase II of the *de-construct* curriculum described that one feature multiplies, again significantly lower than the 45.8% of *construct* curriculum participants after Phase I ($t(216.03) = -2.32, p = .02, 95\%CI = [-0.28, -0.02]$). These results show that people frequently fall prey to learning traps in which initial complex examples prohibit them from arriving at the ground truth (Gelpi et al., 2020; Rich & Gureckis, 2018). As we will see, this pattern is consistent with the hypothesis that participants reuse their own phase I ideas in order to bootstrap their learning in phase II.

Finally, participants in the *combine* condition overwhelmingly favored ground truth over the alternative, despite them being equally complex and compatible with the data. In Phase II self-reports, 28.3% of participants in the *combine* condition reported the ground truth, and zero reported the alternative (Figure 4.6C). The Phase II generalization accuracy of the *combine* curriculum ($41.8\% \pm 38.5\%$) did not differ significantly from that in the *construct* curriculum ($44.7\% \pm 38.3\%$) $t(1702) = 1.25, p = .2$. On the one hand, it seems that many people directly reused the multiplication sub-concept in Phase I as a modular concept in composing their final guesses in Phase II, hinting at a sequential bootstrap learning pattern. On the other hand, it could be that people just “glued” the two sub-concepts together additively, since $(\text{stripe}(A) \times R) + (-\text{spot}(A))$ is logically equivalent to ground truth. To disentangle these concerns, we further conducted two experiments.

4.3 Experiments 6 & 7: Biases in compositions

To investigate whether participants really reused their already-learned sub-concept as a conceptual primitive in Phase II, or simply glued two atomic concepts together, we designed a new curriculum, *flip*, which swaps Phase I and Phase II of *combine*. In this *flip* curriculum, if people reuse the concept they inferred in Phase I as a conceptual primitive in Phase II, they should conclude $R' \leftarrow \text{stripe}(A) \times (R - \text{spot}(A))$, the data-consistent alternative not favored by the previous *combine* condition. If people rather use “add” as their default or dominant compositional mode, then in *flip* Phase II we would expect that they will still favor the original ground truth. As in Experiment 4 and 5, in Experiment 6 stripes were multipliers and spots were subtractors, and in Experiment 7 we reversed the causal powers between the stripe and spot features and otherwise replicated Experiment 6 (Figure 4.7A).

4.3.1 Methods

Experiment 6 recruited 120 participants ($M_{\text{age}} = 35.4 \pm 10.9$) to test the *combine* and *flip* curricula (Figure 4.7A, left). We initially recruited $165 \div 3 \times 2 = 110$ participants to match group sizes in Experiments 4 and 5, but were faced with an imbalance between the two curricula (*combine*: 47, *flip*: 63) due to the random number generator the experiment used to assign participants. To even out the samples, we recruited another 10 participants on Prolific on the same day, all to the *combine* curriculum, and ensured that these extra batch did not contain participants from Experiments 4, 5 and current Experiment 6. All 120 participants were paid at the same scale as in Experiments 5 and 6 (highest paid £1.85). The task took 10.7 ± 4.5 minutes. The procedure was otherwise identical to Experiments 4 and 5. No participant was excluded from analysis. Pre-registration for this experiment is available at <https://osf.io/mfxa6>.

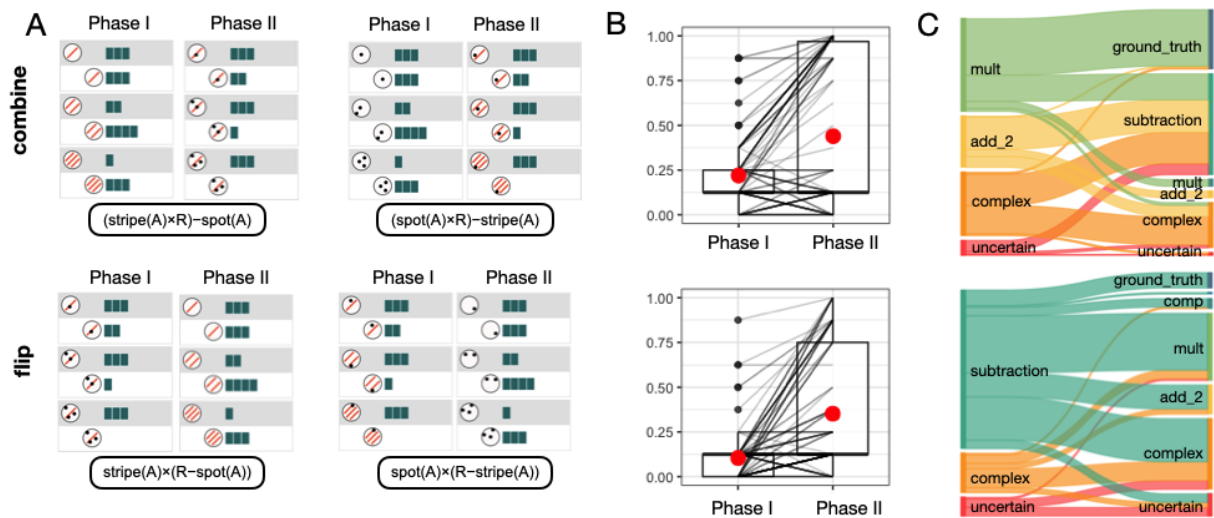


Figure 4.7: Experiments 6-7 material and results. A. Curricula in Experiment 6 (left) and 7 (right). Texts below each phase are data-compatible causal concepts. B. Participants generalization accuracy (match to ground truth). Box plots show the first and third quantiles with lines for the medians; red dots mark the means. C. Coded self-reports.

Experiment 7 was a feature counterbalanced replication of Experiment 6. We recruited another 120 participants ($M_{\text{age}} = 34.0 \pm 12.6$) on Prolific, who did not participate in Experiments 4-6. Here the roles of the stripe and spot features was reversed (Figure 4.7A, right). Participants were paid at the same scale as in Experiments 4-6 (highest paid £1.83). The task took 9.2 ± 4.4 minutes. The procedure was identical to Experiments 4-6. No participant was excluded from analysis. Pre-registration is available at <https://osf.io/swde5>. As above, a two-way ANOVA on feature-counterbalancing and curriculum-design predicting Phase II generalization accuracy revealed main effects on both factors (feature-counterbalancing: $F(1) = 53.543, p < .001$; curriculum-design: $F(1) = 15.894, p < .001$), but no interaction, $F(1) = 1.271, p > .05$. While people indeed treat stripe and spot features differently, our main results hold for both experiments.

4.3.2 Results

We found that people indeed favored the ground truth less often in the *flip* curriculum (Figure 4.7B). For generalization accuracy, here defined as match to the original ground

truth, participants in *flip* Phase II was at $35.2\% \pm 34.3\%$, while participants in *combine* achieved $44\% \pm 41.8\%$, $t(1881.9) = 3.93, p < .001, 95\%CI = [0.04, 0.13]$. In addition, only 7.14% of participants in the *flip* curriculum reported ground truth in Phase II, compared to 25.4% in the *combine* condition, $t(179.52) = 3.89, p < .001, 95\%CI = [0.09, 0.28]$. These results are in line with our previous finding that constructing, caching and later reusing the key sub-concept is crucial for acquiring the complex target concept.

However, a further examination suggests that the drop in producing ground truth in *flip* was not primarily driven by turning to the alternative. Participants' generalization accuracy in terms of matching the alternative concept was $28.8\% \pm 17.3\%$, lower than the level of agreement with the predictions of the original ground truth. In fact, as illustrated in Figure 4.7C, only one participant in *flip* Phase II reported the alternative concept (1.2%), in comparison with nine guessing the ground truth (14.3%), $\chi^2(2) = 41.1, p < .001$. This suggests that additive compositional form is still quite a prevalent inductive bias, and it interacts with sequential bootstrap learning in phased reasoning tasks. Putting it another way, people may be choosing which phase to chunk according to their inductive bias on compositional form, and this might override the order that evidence was actually presented in the experiments.

In our experimental interface, at the end of Phase II, all six pairs of learning examples were available on the screen, and participants could freely scroll up and down to revisit any earlier pairs. Such revisiting could induce orders of cache-and-reuse that are different from the ones designed by us experimenter. In fact, since we encouraged participants to synthesize causal relationships that can explain all six pairs, this may consequently encourage deliberate revisits. By revisiting evidence, in the *flip* curriculum, a strong inductive bias on additive compositional form could lead to preferring ground-truth over the alternative. In the *de-construct* curricula in Experiments 4 and 5, some participants may have revisited Phase I after observing Phase II, and therefore discovered the ground

truth accordingly, reflected by the little bump in Phase II generalization accuracy than Phase I in *de-construct* (Figure 4.7B).

4.4 Model comparison

We now examine predictions and simulations from a range of computational models comparing their ability to reproduce participants' generalization patterns. First, we considered a bootstrap learning model based on adaptor grammars AG as described above. Model AG first processes Phase I learning examples, acquiring an updated library, and then processes Phase I and II altogether with the updated library. Next, to account for the fact that participants were able to scroll up and down and re-access Phase I after reasoning about Phase II, we considered a variant of AG, Adaptor Grammar with Re-processing (AGR). This model mixes predictions \hat{y}_{\rightarrow} from Phase I to II, and predictions \hat{y}_{\leftarrow} from Phase II to I, with a weight parameter $\lambda \in [0, 1]$, getting a mixed prediction $\hat{y}_r \propto \lambda \cdot \hat{y}_{\rightarrow} + (1 - \lambda) \cdot \hat{y}_{\leftarrow}$. For hyper-parameters in models AG and AGR, we set $\alpha_0 = 1$ and $d = 0.2$, the same values used in Liang et al. (2010).

4.4.1 Alternative models

For comparison, we also examined a “rational rules” model (RR) based on Goodman et al. (2008). This assumed the same conceptual primitives as the adaptor grammar models, and used a deep generation depth cap to approximate exhaustive search of the prior. For model RR with the same search depth constraints as models AG and AGR, it will never land on the ground truth or alternative concepts (see Zhao, Bramley, & Lucas, 2022). Since we evaluate models using generalizations, we also implemented several sub-symbolic models capable of generalization but not explicit rule guesses. Here we included a similarity-based categorization model Tversky (1977), a linear regression

model (LinReg), a multinomial regression model (Multinom) and a Gaussian process regression (GpReg) model with radial basis function kernels (one per feature).

Rational rules model

Following Bramley, Rothe, et al. (2018), Goodman et al. (2008), and Zhao, Lucas, et al. (2022), we implemented a Probabilistic Context-Free Grammar $\mathcal{G}_r = \{S, T, N, \Theta\}$, where S is the starting symbol, T a set of production rules, N the set of terminal nodes, and Θ the production probabilities. In order to retain a close match with the adaptor grammar’s initial concept library, we considered production rules as follows:

$$S \rightarrow \text{add}(A, A) \mid \text{sub}(A, A) \mid \text{mult}(A, A)$$

$$A \rightarrow S \mid B$$

$$B \rightarrow C \mid D$$

$$C \rightarrow \text{stripe} \mid \text{spot} \mid \text{segment}$$

$$D \rightarrow 0 \mid 1 \mid 2 \mid 3$$

The pipe symbol $|$ represents “or”, meaning that the symbol on the left-hand side of the arrow symbol \rightarrow can transform to either of the symbols on the right-hand side of \rightarrow . As with the adaptor grammar models, we assigned uniform prior production probabilities: let Γ_L be the set of production rules all starting with L , i.e. any production rule $\gamma \in \Gamma_L$ is of the form $L \rightarrow K$, where K can be any symbol in grammar \mathcal{G}_r , the production probability for each $\gamma \in \Gamma_L$ is $\frac{1}{|\Gamma_L|}$. Since grammar \mathcal{G}_r can produce infinitely complex causal concepts, we fixed a generation depth $d = 40$ in our implementation to cover the ground truth concepts. If d is set too small, like the same constraint we set to the adaptor grammar models, \mathcal{G}_r cannot land on the ground truth by design and therefore not so useful in model comparison (see Zhao, Bramley, & Lucas, 2022). As in the adaptor grammar

models, we used a deterministic likelihood function to evaluate each concept generated by grammar \mathcal{G}_r , essentially discarding all generated concepts that fail to explain all the evidence. We set $n = 100,000$ to have good coverage of rules up to and beyond the degree of complexity seen in human responses. Generalization predictions were made following the same procedure as the adaptor grammar models: Apply the approximated posterior rules with the partially observed data $D^* = \langle A^*, R^*, ? \rangle$ in generalization tasks, and marginalize over the predicted R^* as an approximated posterior predictive.

Similarity-based model

Let d_l be a learning example data point, consisting of an agent, a recipient object, and a result object; d_g a generalization task data point, consisting of only an agent and a recipient objects. Let $\text{stripe}(x)$ be the number of stripes of object x , we can measure the similarity between learning example d_l and generalization task d_g in terms of stripes by taking the absolute difference $|\text{stripes}(A)_{d_l} - \text{stripes}(A)_{d_g}|$, denoted by $\delta_{\text{stripes}}(d_l, d_g)$. Taking all three features stripes, spots and segments into account, the feature difference Δ between learning example d_l and generalization task d_g can be measured by $\Delta(d_l, d_g) = a \cdot \delta_{\text{stripe}}(d_l, d_g) + b \cdot \delta_{\text{spot}}(d_l, d_g) + c \cdot \delta_{\text{segment}}(d_l, d_g)$. With these measures, we defined a similarity score

$$\sigma_{\text{sim}}(d_l, d_g) = e^{-\Delta(d_l, d_g)}$$

such that the more similar d_l and d_g are (smaller distance Δ), the higher the similarity σ_{sim} . When the two data points share the same agent and recipient objects, similarity score σ_{sim} reaches its max of $= 1$. When making generalization predictions, this model first computes similarity score σ_{sim} between the current generalization task g_i with all the available learning examples $\{l_1, \dots, l_k\}$, resulting in $S = \{\sigma_{\text{sim}}(d_{l_1}, d_{g_i}), \dots, \sigma_{\text{sim}}(d_{l_k}, d_{g_i})\}$. Now for this generalization task g_i , it mimics $\text{result}(d_{l_k})$ with confidence $\sigma_{\text{sim}}(d_{l_k}, d_{g_i})$.

Let $n = \text{result}(d_{l_k})$, task g_i predicts $p(n) = \text{result}(d_{l_k}) \cdot \sigma_{\text{sim}}(d_{l_k}, d_{g_i})$. Marginalizing over all possible result segment values n gives the distribution over task g_i 's predicted result segment values.

Linear regression model

Let the number of stripes, spots and segments in each learning example be the independent variables, and the resulting stick length R' be the dependent variable. We fit a linear regression model after each phase of the experiment with formula

$$R' \sim a \cdot \text{stripe}(A) + b \cdot \text{spot}(A) + c \cdot R + \epsilon.$$

We made generalization predictions using fitted parameters and the requisite generalization task's feature values. We rounded the predicted result segment number to the two nearest integers in order to match the required prediction output.

Multinomial logistic regression model

We treated each possible result segment value as categorical value (instead of continuous as in the linear regression case), and fit a multinomial logistic regression model to predict the probability of each result segment value using a formula same as the one used in the linear regression model, with the `nnet` package in R. After fitting the model, we used the `pred` function to gather probabilistic predictions about the possible result segment values for each trial. We then normalized this probabilistic prediction to ensure this is a probabilistic distribution.

Gaussian process model

Treating each learning example as three-dimensional input (stripes, spots, segments) with a one-dimensional output (result segments), we fit a Gaussian Process (GP) regression model with radial basis function kernels, each per feature x_f :

$$K(x_f, x'_f) = \exp\left(-\frac{\|x_f - x'_f\|}{2\sigma^2}\right)$$

We used the `GPY` package in Python to fit the model. Conditioning on the three dimensional input for each generalization task, the fitted GP regression model outputs a Gaussian distribution over possible segment lengths $\mathcal{N}(\mu, \sigma^2)$. We then binned this distribution over the possible discrete segment values for comparison with empirical data.

4.4.2 Model fits

Cross validation

We used cross validation to evaluate models against behavioral data in generalization tasks on log likelihood fits. To do this, we collapsed data from all four experiments by curriculum c , keeping how many people n chose which segment number $y \in [0, 16]$ in each task i , resulting in data $\mathcal{D} = \{n_{ciy}\}$. We then let each computational model generate a distribution P_{ci} over all possible segment numbers $Y = \{0, 1, \dots, 16\}$ for task i in curriculum c . Since many model predictions are point estimates, or centered on only a few segment numbers, we considered a trembling hand noise parameter $h \in (0, \frac{1}{|Y|})$ such that for a probability distribution $P(Y)$:

$$P^h(Y = y) = \frac{P(Y = y) + h}{1 + h|Y|}. \quad (4.6)$$

Essentially, we add noise h to each random variable in set Y to avoid 0 likelihoods. The denominator ensures $P^h(Y)$ is still a probability. Different from softmax functions, $P^h(Y)$ stays close to the shape of $P(Y)$ when h is small, and therefore best maintains each model’s raw degree of confidence on those 1 or 2 predictions. Log likelihood of a model producing data \mathcal{D} is thus given by:

$$LL = \sum_{c=c_1}^{c_k} \sum_{i=t_1}^{t_j} \sum_{y=y_1}^{y_m} \ln(P_{ci}^h(Y = y)) \cdot n_{ciy}. \quad (4.7)$$

For each run of the cross validation, we hold out one curriculum c_{test} , and fit the noise parameter h on the other three curricula using maximum likelihood estimation (MLE) with the `optim` function in R. Note that for model AGR, an additional weight parameter λ is jointly fitted. Then we compute LL_{test} on curriculum c_{test} with the fitted parameters. Summing over LL_{test} for all four curricula serves as the total log likelihood fit LL for the model. As a baseline, choosing randomly yields $LL_{\text{rand}} = 570 \times 16 \times \ln(\frac{1}{17}) = -25838.91$, for there were 570 participants, each completing $8 \times 2 = 16$ tasks, where in each task there were 17 possible responses (final stick lengths, including 0) to choose from. Any value smaller than LL_{rand} is improvement over an eyes-closed baseline.

Results

Figure 4.8A shows each model’s improvement over baseline, $\Delta_{\text{model}} = LL_{\text{model}} - LL_{\text{rand}}$. Model AGR achieves the greatest improvement, with the three Bayesian-symbolic models (AGR, AG, RR) easily outperforming similarity-based or regression models. With fitted model parameters, Figure 4.8B-C plots generalization accuracy in each phase for each curriculum between model and people. In line with overall model fits, AGR best predicts people’s performance across all cases, and the non-symbolic models fail to match people’s predictions.

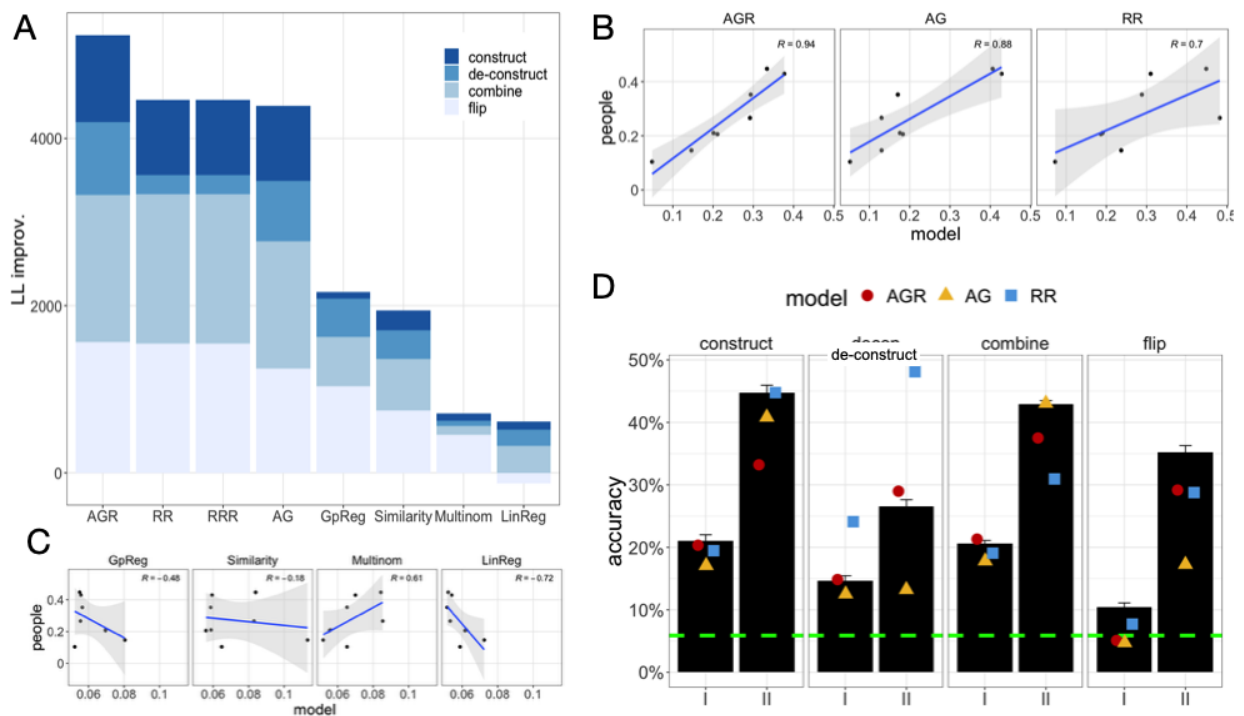


Figure 4.8: Experiments 4-7 model fitting results. A. Total log likelihood improvement over random baseline ($y=0$), log scale. B-C. Generalization accuracy per curriculum and phase. X-axis are model predictions, y-axis people's. D. Generalization accuracy between people (black bars) and four Bayesian-symbolic models.

Note that in Figure 4.8A, even though there is a close tie between model AG and RR, only AG is able to improve significantly in the *de-construct* curriculum. We further plot generalization accuracies for models AGR, AG, and RR against behavioral data in Figure 4.8D, showing that RR fails to reproduce the curriculum-order effects between the *construct* and *de-construct* curricula. This is because model RR is likely to have figured ground truth after seeing all the data, even for the *de-construct* curriculum, and thus deviating from how people process phases of information. Model AG, on the other hand, is defeated by the learning trap as many people were, exhibiting no accuracy improvement in Phase II relative to Phase I. Model AGR mixes model AG with some re-processing, and is therefore able to capture participants' modest improvement in *de-construct* Phase II generalizations. Furthermore, RR achieves lower accuracy than people in the *combine*

Table 4.3: Experiments 4-7 model fitting results

Model	Const.	De-con.	Comb.	Flip	Total	Improv.	N best fit	N match
AGR	-3807	-4436	-8210	-4149	-20602	5237	150	45/64
RR	-3955	-5076	-8183	-4166	-21380	4459	93	37/64
AG	-3955	-4579	-8455	-4462	-21451	4388	141	45/64
GpReg	-4770	-4840	-9386	-4678	-23674	2165	69	11/64
Similarity	-4616	-4959	-9364	-4962	-23901	1938	20	13/64
Multinom	-4761	-5244	-9864	-5255	-25124	715	39	17/64
LinReg	-4758	-5106	-9652	-5834	-25350	489	32	12/64
Random	-4850	-5304	-9973	-5712	-25839	0	26	9/64

Phase II, because it assigns as much posterior probability to the intended ground truth as to the equivalent-consistent alternatives.

Table 4.3 lists all model comparison results in detail. Following the specifications above, columns Construct, De-construct, Combine and Flip contain cross validation results on each corresponding held-out curricula. The Total NLL (NLL stands for negative log likelihood) column sums over the four curricula. The Improvement column takes each model’s total NLL, and subtracts the Random baseline model’s total NLL. All numbers in the above mentioned columns are log likelihoods, and a change in unit 1 reflects exponential scale of difference.

Column “N best fit” in Table 4.3 is the number of participants best fitted by the corresponding model. To evaluate this, for each participant, we computed the Bayesian information criterion (BIC) for all the models, and selected the model with the lowest BIC to be the model that best fits this participant. We then computed how many participants each model best fits, serving as the “N best fit” measure. Model AGR best fits the most number of participants ($N = 150$), with model AG on a close match ($N = 141$), followed by the rational rules model ($N = 93$) and Gaussian Process regression ($N = 69$).

Since we forced a single-prediction per generalization task in the experiments, we compared how often a model’s forced single-prediction matches people’s most selected single-prediction in each task. To do so, for each model’s distribution over predicted

number of segments, we took the one with highest probability to be its single choice. There were no ties for all tasks and all the models we considered. Next, in the aggregated selection in each task from people, we took the most selected one. There were no ties either. We then computed how many of each model’s single choice match with people’s most favored option, being the “N match” measure. In total there are $8 \text{ task} \times 2 \text{ phase} \times 4 \text{ curriculum} = 64$ unique tasks. As Table 4.3 shows, model AGR and model AG match the most number of these forced single-predictions, N match = 45, constituting 70% of all tasks.

Overall, the adaptor grammar models AG and AGR provided a much better account of people’s behavioral patterns in the experiments than the other models we considered. More generally, this means that curriculum-order effects and garden-pathing effects exhibited by people, can be explained as consequences of a cache-and-reuse mechanism expanding the reach of a bounded learning system. Critically, these phenomena cannot be explained by a standard Bayesian-symbolic model out of the box, or by familiar sub-symbolic categorization models, showcasing that a cache-and-reuse mechanism is central to human-like inductive inference to compositional concepts.

Figure 4.9 shows the best fitting AGR model’s predictions in each generalization task with participant data showing a close match. We note one interesting discrepancy in generalization task 1, which asked about an agent with no spots or stripes: While many participants predicted the disappearance of segments, since $R' \leftarrow \text{stripe}(A) \times R$ and $0 \times 3 = 0$, many participants also predicted that the result number of segments would stay the same. This could be due to participants concluding that absent features meant that nothing would happen. Future work could investigate how people reason about these kinds of edge cases.

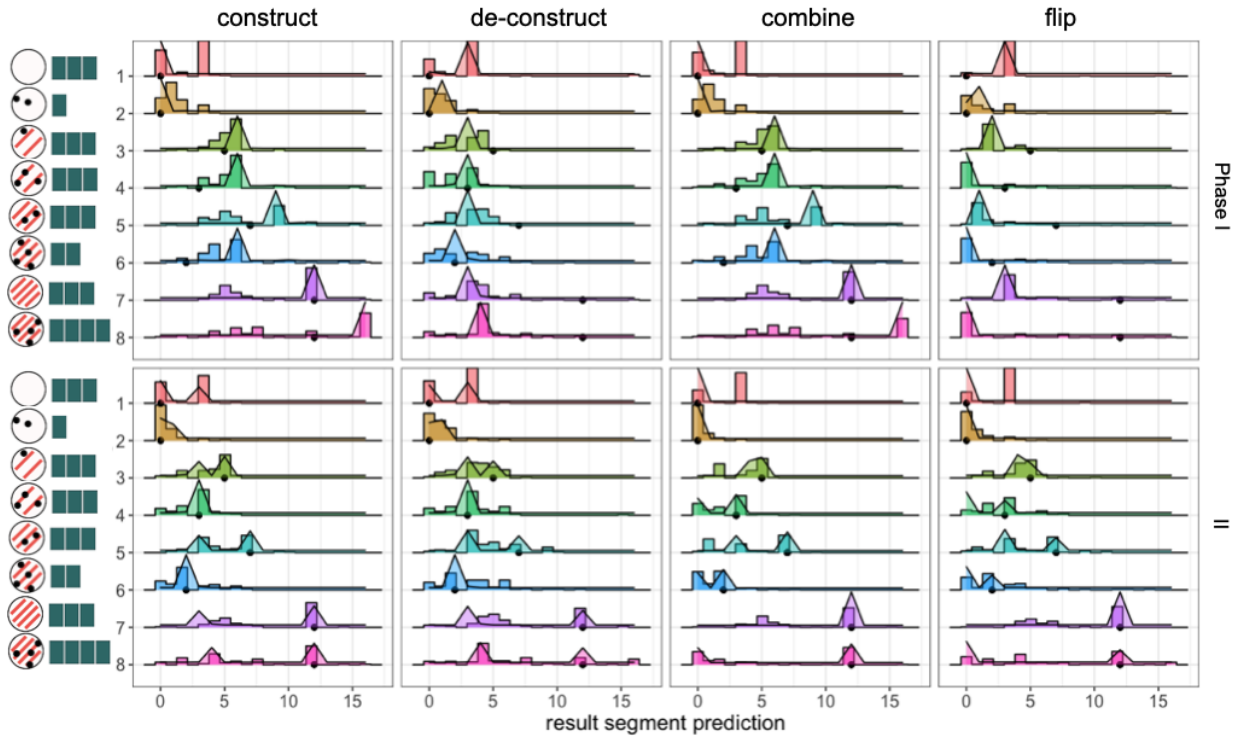


Figure 4.9: Experiments 4-7 generalization predictions by people (bars) and the best fitting AGR model (area). Rows of panels are for experimental phases, columns for conditions. In each panel, x-axis are predicted number of segments (0-16), y-axis are tasks.

4.5 Discussion

We proposed a formalization of bootstrap learning that supercharges Bayesian-symbolic concept learning frameworks with an effective cache-and-reuse mechanism. This model replaces a fixed set of conceptual primitives with a dynamic concept library enabled by adaptor grammars, facilitating incremental discovery of complex concepts under helpful curricula in spite of finite computational resources. We showed how compositional concepts evolve as cognitively-bounded learners bootstrap over batches of data, and how this process gives rise to systematically different interpretations of the same evidence.

Our method differs from previous sampling-based approaches that our model can bootstrap and develop compositional concepts by caching and reusing learned conceptual chunks, and hence uniquely captures this essential aspect of how people reason

about the world (Carey, 2004). Previous attempts to synthesize human-like order effects in concept learning tended to use constraints on the number of samples the learner could draw to capture how learning trajectories differ from each other (Bonawitz et al., 2014; Gelpi et al., 2020; Lieder et al., 2018; Thaker et al., 2017). While these approaches indeed reproduce key order effects like the anchoring-effect (Lieder et al., 2018) and conceptual garden-pathing (Gelpi et al., 2020; Thaker et al., 2017), they aim still asymptotically to approximate a fixed posterior distribution. Human learners, however, can constantly refresh their conceptual repertoire and redefine what a posterior looks like. Recent work in program induction points out that learning experiences can shape a learner’s conceptual libraries (Tian et al., 2020; Wong et al., 2022), and therefore unlock different interpretations of evidence. Our formalization addresses a less-explored yet critical part of this computational enterprise—reuse and re-presentation (Cheyette & Piantadosi, 2017). Effective reuse balances between inference and memory (Gershman & Goodman, 2014), since caching useful chunks can save computational cost. Developed to solve exactly this problem of finding optimal caches, adaptor grammar and other fragment grammar methods provide a rigorous, elegant, and powerful toolset in tasks such as word-segmentation (Johnson et al., 2007), phrasal parsing (O’Donnell et al., 2009), and text-editing (Liang et al., 2010). Our work pushes across discipline boundaries and unites adaptor grammars with theories of human concept learning. The resulting model is able to synthesize human-like bootstrap learning and compositional generalization attributing these abilities to an evolving concept library that goes beyond parsing structural regularities to effectively constructing richer world models.

This interaction between our evolving mental concepts and the environment they seek to reflect outlines several interesting future directions. For example, Experiments 6 & 7 suggest that, instead of being passive information receivers, people may actively choose which subset of a complex information flow they find the easiest to process first, and then build gradually to make sense of the whole picture. Future work may extend our

framework to active learning scenarios to study such information-seeking behaviors and self curriculum-design patterns (e.g. Bramley & Xu, 2023). Moreover, cache-and-reuse is a useful way to refactor representations. Liang et al. (2010) introduced a sub-tree refactoring method for parsing shared sub-programs, providing natural future extensions on studying refactoring as a cognitive inference algorithm (Rule et al., 2020).

Our current work also has several limitations that future work could address. For instance, we assumed a deterministic likelihood function, but this does not handle vague concepts like the stick decreases or increases very well. A grammar and likelihood able to capture softer constraints could capture a larger range of people’s guesses and predictions. Since, for simplicity, we did not include conceptual primitives for conditionals, our model could not express all self-reports people made under overwhelmingly complex information. In particular, there were a number of ‘divide and conquer’ type responses (e.g. “If there are more dots than sticks it removes segments. If there are more sticks than dots it adds segments. If there are the same number of sticks and dots the number of segments stays the same.” by participant 49). We could include these by assuming an `ifElse` base concept with corresponding generation rules. Piantadosi (2021) argued that base primitives in combinatorial logic suffice to provide a fundamental formalization of our mental representations and computations. In our case, using natural-language-like base terms is rather for their computational and expressive convenience. All of the base primitives and learned concepts could be decomposed into using solely combinatorial logic bases. One other limitation of our current model is that it does not handle forgetting by default. Forgetting is known to be a critical feature of human memory and learning (Della Sala, 2010; Gravitz, 2019; Nørby, 2015). To extend our formalization to model life-long learning, it would be important to incorporate a mechanism through which obsolete concepts get cleared or recycled. Even though our formalization of cache-and-reuse benefits from the rigor and elegance of adaptor grammars, future work may also explore using simplified versions of reuse with fewer representational constraints,

for example an ad-hoc cache-and-normalize algorithm proposed by Cheyette and Piantadosi (2017).

One interesting behavioral pattern that deserve future attention is the substantial individual variances in the self-reports. Like visualized in Figure 4.6C and Figure 4.7C, many participants concluded an *add 2* causal concept instead of multiplication. The *add 2* concept states that if the agent has more than one stripe (in Experiments 1 and 3, or spot in Experiments 2 and 4), then add two segments to the recipient, under the assumption that nothing happens if the agent object has exactly one stripe (or spot). Participants who reported this *add 2* concept demonstrated weaker cache-and-reuse effect than those reporting a multiplicative concept: Most of the *add 2* reporters ended up with *complex* concepts in Phase II, while a small portion reported the *subtraction* concepts or locally extended their *add 2* concept with a *subtraction* concept. While our participants were recruited from a pool of adult native English speakers on Prolific, further measurement on math literacy or meta-cognitive abilities would be of interest to investigate the systematicity of such variance.

Chapter 5

Conclusion

This thesis studied how people generalize causal laws from observations of interactions between objects. Overall, participants made systematic causal generalizations after one (Experiment 1, Chapter 2), several (Experiments 2-3, Chapter 3), and batches of (Experiments 4-7, Chapter 4) observations despite there being a large number of potentially compatible explanations. In addition, these experiments identified several intriguing behavioral patterns. Experiment 1 found that the order in which people generalize to new scenarios affects the final set of concepts they form. Experiment 2 demonstrated that people treat properties of agents and recipients asymmetrically in constructing causal hypotheses. These causal inductive biases are highly sensitive to interaction cues, as revealed by Experiment 3. Experiments 4 and 5 further showcased robust curriculum-order and conceptual garden-pathing effects: While people can successfully acquire a complex causal concept when they have an opportunity to cache a key sub-concept, simply reversing the presentation order of the same learning examples induces dramatic failures, and leads people to complex and ad hoc concepts. Two subsequent Experiments 6 and 7 revealed an interaction between inductive biases and curriculum orders.

Throughout this thesis, I used probabilistic generative grammars (Johnson, 1998) as a framework to capture the richness of human causal explanations, treating causal hypotheses as compositional mental functions or programs (Chater & Oaksford, 2013). I integrated this generation mechanism with nonparametric categorization (Navarro et al., 2006) to model human-like one- and few-shot causal generalization. A process variant of this model, embodying a form of bounded-rationality, accounts for the generalization-order effect by treating one’s own earlier judgments as evidence when making new generalizations. This model also captures the causal asymmetry bias, by preferentially localizing causal laws based on the properties of the agent objects. With the help of adaptor grammars (Johnson et al., 2007), I was able to extend this modeling framework to provide a neat mathematical characterization for cache-and-reuse, predicting when people will form different interpretations of the same evidence, and explaining how human cognition reaches so far beyond its grasp.

5.1 Causal representations

In my formalization, causal representations are no longer networks of statistical association (cf. Pearl, 2009), but fundamental cognitive models (Chater & Oaksford, 2013) for predicting, explaining and controlling the world (Gopnik et al., 2007; Griffiths et al., 2010; Sloman, 2005). Generative grammars produce causal functions that explicitly describe the consequence of causal interaction on the recipient object’s features, allowing these causal functions to take absolute feature values like $\text{color}(r') \Leftarrow \text{blue}$, as well as values relative to the agent or recipient’s pre-interaction features such as $\text{color}(r') \Leftarrow \text{color}(a)$ or $\text{edge}(r') \Leftarrow \text{edge}(r) + 1$. These causal functions natively capture many kinds of causal theories people may entertain, as confirmed by their self-reports and our model fits (see also Bramley, Rothe, et al., 2018; Goodman et al.,

2008; Lake & Piantadosi, 2020). Moreover, by grounding causal functions in such object-based representations, these causal functions naturally generalize to novel objects.

Besides being inherently causal and efficiently generalizable, these causal functions are also compatible with the flexibility of human causal reasoning. As Mayrhofer and Waldmann (2015) pointed out, agent-recipient roles and cause-effect roles are separate concepts. Even though agents are usually taken to be the cause, in some cases the static, passive recipients (patients) are actually seen as more causal of an outcome, for instance a red traffic light being the cause for an active pedestrian to stop walking. Symbolic grammars used in this thesis makes no assumption about whether agent or recipient object features determine the result, rather, they treat agent and recipient objects equally in its grammar generation process because of uniform priors (e.g. Table 2.1, row “Relative reference”). In the categorization process, I introduced the focus parameter γ that interpolates between considering only the agent, or only the recipient as relevant for what causal function applies. γ is later fit to empirical data and yielded an best-fitting value of 1, corresponding to categorization by features of the agent only, confirming the hypothesized causal asymmetry. As a result, this framework is applicable for further investigation into the intricate relationship between agent–recipient concepts and cause–effect roles. For instance one might estimate an inductive bias controlling the balance of agent and recipient roles in the grammar, or modeling γ conditional on learning data.

The applicability of the symbolic grammar generator approach goes beyond these particular causal functions applications. Though generative grammars have been most-traditionally used for modeling language processing (Johnson, 1998), they can be created for many tasks involving symbolic representations (e.g. Fränken et al., 2022; Goodman et al., 2008; Mollica & Piantadosi, 2021; Rule et al., 2020). For the grammars used in this paper, I included a minimal set of primitives that simply cover the features participants were told about in the instructions. However, recent work has also explored question of whether there is an optimal set of primitive domain-specific-languages (Ellis

et al., 2021; Piantadosi et al., 2016). In fact, the adaptor grammar model presented in Chapter 4 can be adapted to address the problem of deciding primitives for a symbolic model. A model with caching mechanisms like ours could learn a set of conceptual fragments most useful for composing new concepts in adaption to the learning data, serving as a principled way to create domain specific languages.

One important future extension to this formalization is to incorporate active learning and interventions, the key characteristics of causal learning (Bramley et al., 2015; Sloman, 2005; Steyvers et al., 2003). While Pearl (2000) developed Do-calculus to model interventions in causal Bayes nets, active learning in an open-domain symbolic conceptual framework remains rather unexplored. One might apply measures like expected information gain to model sequential hypothesis testing (e.g. Bramley & Xu, 2023), or test existing heuristics such as positive testing (Klayman & Ha, 1987) or win-stay-loose-shift (Robbins, 1952) in predicting interventions. Emerging research in Bayesian optimal experimental design (Lindley, 1956; Ryan et al., 2016) also provides promising future directions, in the same spirit of people as naive scientists as in the developmental literature (Gopnik, 1996), where one could compare whether lay people’s choice of causal interventions align or deviate from the optimal experimental design.

Nevertheless, these modeling choices are not the only way to represent human causal cognition. This modeling framework is open to, and compatible with, many other options. For example, one may choose to extend the symbolic approach to cover the categorization process as well, or incorporate causal Bayes nets as a representation for causal functions among multiple relata (Griffiths & Tenenbaum, 2009; Kemp et al., 2010; Lucas & Griffiths, 2010; Pearl, 2000, 2009).

5.2 Constructive cognition

This modeling framework lines up with a range of recent symbolic accounts of inductive and creative reasoning (e.g. Fränken et al., 2022; Goodman et al., 2008; Griffiths & Tenenbaum, 2009; Kemp et al., 2010). This framework emphasizes 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 these Bayesian-symbolic methods calls upon a potentially infinite set of possible causal functions, yet is governed by the preference for parsimony, and encourages systematic composition (see also Bramley, Rothe, et al., 2018). The extended Dirichlet Process for category construction goes beyond a hierarchical Bayesian modeling approach where categories are pre-defined as inductive biases (e.g. Goodman et al., 2011; Griffiths & Tenenbaum, 2009), and thus better captures the flexibility of human generalization behaviors (see also Kemp et al., 2010). This 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 when on demand” assumption for a process account, our model successfully reproduces the generalization-order effects in behavioral data. This constructive view of cognition is not unique to causal cognition. Generative grammars have been proven useful in many other fields such as concept learning and category induction (Goodman et al., 2008; Lake et al., 2015; Piantadosi et al., 2016). Symbolic approaches enable compositionality and systematicity, while the sub-symbolic techniques, especially the fast, incremental approximations, make this more scalable to real-world data (Bramley et al., 2017).

People have a remarkable ability to develop rich and complex concepts despite limited cognitive capacities. On the one hand, there is abundant evidence that people are bounded reasoners (Griffiths et al., 2015; Kahneman et al., 1982; Newell & Simon, 1972; Van Rooij, 2008), entertain a rather small set of mental options at a time (Bonawitz et

al., 2014; Cowan, 2001; Sanborn & Chater, 2016; Sanborn et al., 2010; Vul et al., 2014), and generally deviate from exhaustive search over large hypothesis spaces (Acerbi et al., 2014; Bramley et al., 2017; Chater, 2018; Fränken et al., 2022; Gelpi et al., 2020). On the other hand, these bounded reasoners can develop richly structured conceptual systems (Gopnik & Meltzoff, 1997; Kemp & Tenenbaum, 2008; Quine & Ullian, 1978), produce sophisticated explanations (Craik, 1952; Keil, 2006; Lombrozo, 2012), and push forward complex scientific theories (Kuhn, 1970). How are people able to create and grasp such complex concepts that seem so far beyond their reach?

Newton gave a famous answer to this question: “If I have seen further, it is by standing on the shoulders of giants.” (Newton, 1675). This reflects the intuition that people are bounded yet blessed with a capacity to not just learn from others, but to extend and re-purpose existing knowledge to create new and more powerful ideas. Our formalization for conceptual bootstrapping gives a cognitive basis to a wide range of order effects. Besides the behavioral patterns we identified in our experiments—a strong curriculum-order effect in swapping two phases of learning the same material—it is well-established that people fall prey to confirmation bias (Edwards & Smith, 1996), motivated reasoning (Redlawsk, 2002), and exhibit path-dependence (Mahoney & Schensul, 2006). While previous research suggested that different prior beliefs can lead to different evaluations of the same piece of information (e.g. Jern et al., 2014), the formalization presented here demonstrates that drastically different conceptualization of the same data may stem from the same priors and the same cognitive process. In particular, all the experiments in this thesis tested causal learning and generalization in abstract settings, rather than subjective opinions such as political attitude, and therefore serves a friendly reminder that an objective rule is not guaranteed to prevail, even among capable cognizers scrutinizing the same data. Being limited and adaptive information processors, people may develop biased interpretation of features (Searcy & Shafto, 2016), and fall for various kinds of

learning traps in category-based generalization, related to assumptions about stochasticity, similarity, or selective attention (Rich & Gureckis, 2018). This work provides a new set of evidence and a computational account about learning traps in complex causal reasoning tasks (see also Gelpi et al., 2020), rationalizing learning traps as consequences of optimal adaptation to learning complexities.

While here the cache-and-reuse process applies to objects drawn from the same category, it is alluring to envision how this process interplays with other forms of causal generalization, such as the categorization process discussed in Chapters 2 and 3. Cache-and-reuse may apply to causal laws attached to each category, where different categories update their causal properties independently from each other; or, cache-and-reuse may be applied globally, and thus interferes with the categorization process itself, leading to learning traps in categories learning (Rich & Gureckis, 2018). Future work could look into these different predictions in either behavioral experiments or computational equivalents, and this will paint us a more precise landscape of how people conceptualize the world in the flow of information.

One may attempt to interpret the process of making generalization predictions as some versions of self-supervised learning (Mikolov et al., 2013), in the sense that people construct learning targets themselves to better understand the underlying causal dynamics of the world. While this intriguing analogy warrants careful future investigation, especially in the realm of human concept learning where observations are extremely sparse and representations are usually symbolic and discrete, the curriculum-order effects we reported in compositional causal generalization point out a clear discrepancy: People reply on what they understood earlier to synthesize later information, and it would be fascinating to explore to what extent such conceptual garden-pathing could be overridden by the underlying environmental dynamics.

Overall, our modeling framework draws a close link with probabilistic program induction models (e.g. Bramley, Rothe, et al., 2018; Ellis et al., 2021; Lake & Piantadosi,

2020; Lake et al., 2015), where causal beliefs and concepts can be viewed as programs, and accurate generalizations can be viewed as a evidence for successful program synthesis whereby these programs increasingly reflect the true causal laws of nature. Our modeling framework is open to broader generalization cases beyond causal cognition, and contributes to the collective effort for a hybrid approach in understanding human cognition (Lake et al., 2017; Oaksford, Chater, et al., 2007; Valentin et al., 2021).

5.3 Final words

Usually, successful learning refers to predicting the correct outcomes given some inputs. Recovering or approximating an underlying true relationship can help in making such predictions, and the closer this recovered relationship is to the ground truth, the better predictions it can make in scenarios that differ significantly from training. This view has been tremendously helpful for thinking about problems like over- and underfitting, when there are clear-cut right or wrong answers in both the learning and generalization tasks. Unfortunately, to define a problem and criteria of being correct can be genuine challenges, especially in everyday causal reasoning (Hayes, 1981). First of all, how do we decide what matters? Among all the wibbly-wobbly timey-wimey stuff, all the progression of events, changes and non-changes, salient or hidden features, how do we pick out that handful of meaningful variables to ponder at a given time? Second, given that we are faced with a selection of factors and some ideas about their relationships, how do we evaluate the success of these ideas? Sometimes we are happy with a good-enough approximation, and feel safe allowing for some “exceptions”. Sometimes we can provide a precise characterization, and are confident to hold that it was the data-collection rather than the theory that went wrong (Feynman, 1992). Sometimes, we rely so much on existing theories to seek out evidence, that we end up in local optima and are never able to come out (Redlawsk, 2002). Nevertheless, there is perhaps no way to

figure out whether we are currently in one just now (Quine & Ullian, 1978). All of these scenarios seem to suggest a dynamic picture of how cognition constructs causal models of the world, and the criteria for their success may change over time and evolve over generations.

I hope my investigations into causal generalization make at least one point clear: Life is a series of barely controlled experiments, and cognition is set up to deal with this situation. In a way, what Bayesian models take as their prior, or probabilistic models call uncertainties, are honest reflections of the vast possibilities of probabilities, and we treat these possibilities seriously when making decisions (Lewis, 1986). I opened this thesis with Neil Gaimen's remark about dragons in fairy tales, as this sentence beautifully reflects how our minds commit to causal generalization: Dragons are not something we have met, but we readily situate dragons in a rich web of causal relations, finding no trouble picturing their destroying power, and at the same time coming up with ways to fight, even tame, these imaginary creatures. What's more, the hope and courage readers could harvest from these fairy tales, might then cast significant influence on how these readers face challenges and obstacles in their future life. As we have seen in this thesis, predictions generated by our existing generalizable causal models could guide us what to see, how to engage with the environment, and may lead to iterations of new causal models building on existing ones. Therefore, while causality seems to indicate invariances across generalizations, it warrants careful thoughts of how possible generalizations shape the way we think about causal relations, let alone how causal theories evolve over time and on top of each other.

Nothing summarizes this better than Google's dedication to the year in search 2022:

To everyone who sees not what the world is, but what it can be.

References

- Acerbi, L., Vijayakumar, S., & Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS Computational Biology*, *10*(6), e1003661.
- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.
- Aristotle, A. (322/1998). *The metaphysics*. Penguin Classics.
- Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning*, 41–48.
- Bever, T. G. (1970). The cognitive basis for linguistic structures. In R. Hayes (Ed.), *Cognition and language development* (pp. 279–362). Wiley & Sons, Inc.
- Bezem, M., Klop, J. W., & de Vrijer, R. (2003). *Term rewriting systems*. Cambridge University Press.
- Bonawitz, E., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating Bayesian inference. *Cognitive Psychology*, *74*, 35–65.
- Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338.
- Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9–38.

- 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), 708–731.
- Bramley, N. R., Rothe, A., Tenenbaum, J., Xu, F., & Gureckis, T. (2018). Grounding compositional hypothesis generation in specific instances. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, 1390–1395.
- Bramley, N. R., & Xu, F. (2023). Active inductive inference in children and adults: A constructivist perspective. *In revision*.
- Carey, S. (2004). Bootstrapping & the origin of concepts. *Daedalus*, 133(1), 59–68.
- Chater, N. (2018). *The mind is flat: The illusion of mental depth and the improvised mind*. Penguin UK.
- Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*, 37(6), 1171–1191.
- Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104(2), 367–405.
- Cheyette, S., & Piantadosi, S. (2017). Knowledge transfer in a probabilistic language of thought. *Proceedings of the 39th Annual Meeting Cognitive Science Society*.
- Chmielewski, M., & Kucker, S. C. (2020). An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4), 464–473.
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1), 87–114.
- Craik, K. J. W. (1952). *The nature of explanation* (Vol. 445). CUP Archive.
- Crank, E., & Felleisen, M. (1991). Parameter-passing and the lambda calculus. *Proceedings of the 18th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, 233–244.

- Cronbach, L. J. (1943). On estimates of test reliability. *Journal of Educational Psychology*, 34(8), 485–494.
- Danks, D., & Schwartz, S. (2006). Effects of causal strength on learning from biased sequences. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, 28, 1180–1185.
- Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. (2020). A theory of learning to infer. *Psychological Review*, 127(3), 412–441.
- Dechter, E., Malmaud, J., Adams, R. P., & Tenenbaum, J. B. (2013). Bootstrap learning via modular concept discovery. *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 1302–1309.
- Della Sala, S. (2010). *Forgetting*. Psychology Press.
- Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of personality and social psychology*, 71(1), 5–24.
- Ellis, K., Wong, C., Nye, M., Sablé-Meyer, M., Morales, L., Hewitt, L., Cary, L., Solar-Lezama, A., & Tenenbaum, J. B. (2021). Dreamcoder: Bootstrapping inductive program synthesis with wake-sleep library learning. *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, 835–850.
- Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1), 71–99.
- Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, 407(6804), 630–633.
- Feynman, R. P. (1992). The 7 percent solution. In E. Hutchings (Ed.), *“Surely you’re joking, Mr Feynman!” Adventures of a curious character as told to Ralph Leighton* (pp. 247–255). Vintage Books, London.
- Fränken, J.-P., Theodoropoulos, N. C., & Bramley, N. R. (2022). Algorithms of adaptation in inductive inference. *Cognitive Psychology*, 137, 101506.

- Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought*. Oxford University Press, USA.
- Gelpi, R., Prystawski, B., Lucas, C. G., & Buchsbaum, D. (2020). Incremental hypothesis revision in causal reasoning across development. *Proceedings of the 42th Annual Conference of the Cognitive Science Society*.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6), 721–741.
- Gershman, S., & Goodman, N. (2014). Amortized inference in probabilistic reasoning. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 517–522.
- Ginsburg, S. (1966). *The mathematical theory of context free languages*. McGraw-Hill Book Company.
- Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6), 236–243.
- 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), 110–119.
- Gopnik, A. (1996). The scientist as child. *Philosophy of Science*, 63(4), 485–514.
- Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1), 3–32.
- Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. MIT Press.
- Gopnik, A., Schulz, L., & Schulz, L. E. (2007). *Causal learning: Psychology, philosophy, and computation*. Oxford University Press.

- Gopnik, A., & Sobel, D. M. (2000). Detectingblickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71*(5), 1205–1222.
- Gordon, I. E., Day, R. H., & Stecher, E. J. (1990). Perceived causality occurs with stroboscopic movement of one or both stimulus elements. *Perception, 19*(1), 17–20.
- Gravitz, L. (2019). The forgotten part of memory. *Nature, 571*(7766), S12–S12.
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, volume 3: Speech acts* (pp. 41–58). New York: Academic Press.
- Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357–364.
- Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science, 32*(1), 68–107.
- Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science, 7*(2), 217–229.
- Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116*(4), 661–716.
- Hayes, P. J. (1981). The frame problem and related problems in artificial intelligence. In B. L. Webber & N. J. Nilsson (Eds.), *Readings in artificial intelligence* (pp. 223–230). Morgan Kaufmann.
- Hayes-Roth, P., Fox, M., Gill, G., DJ, M., & Reddy, R. (1977). Speech understanding systems: Summary of results of the five-year research effort. *Department of Computer Science*.
- Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(2), 411.

- Hume, D. (1740). *A treatise of human nature*. Courier Corporation.
- Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and applied*, 79(1), 1–17.
- Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, 121(2), 206.
- Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics*, 24(4), 613–632.
- Johnson, M., Griffiths, T. L., Goldwater, S., et al. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in neural information processing systems*, 19, 641–648.
- Johnson, S. G., & Ahn, W.-k. (2015). Causal networks or causal islands? the representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*, 39(7), 1468–1503.
- Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.
- Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57, 227–254.
- Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*.
- Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, 34(7), 1185–1243.
- Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, 64(1-2), 35–73.
- Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31), 10687–10692.

- Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2), 102–110.
- Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2), 211.
- Klein, G. A. (2017). *Sources of power: How people make decisions*. MIT press.
- Knuth, D. E. (1973). Sorting and searching. CUMINCAD.
- Krueger, K. A., & Dayan, P. (2009). Flexible shaping: How learning in small steps helps. *Cognition*, 110(3), 380–394.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151–160.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (Vol. 111). Chicago University of Chicago Press.
- Kwisthout, J., & Van Rooij, I. (2020). Computational resource demands of a predictive Bayesian brain. *Computational Brain & Behavior*, 3(2), 174–188.
- Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3), 451–460.
- Lake, B. M., & Piantadosi, S. T. (2020). People infer recursive visual concepts from just a few examples. *Computational Brain & Behavior*, 3(1), 54–65.
- Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266), 1332–1338.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.
- Landau, B., Smith, L. B., & Jones, S. (1992). Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language*, 31(6), 807–825.

- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3(3), 299–321.
- Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3), 265–288.
- Lewis, D. (1986). *On the plurality of worlds*. Blackwell.
- Li, P., Dunham, Y., & Carey, S. (2009). Of substance: The nature of language effects on entity construal. *Cognitive Psychology*, 58(4), 487–524.
- Liang, P., Jordan, M. I., & Klein, D. (2010). Learning programs: A hierarchical Bayesian approach. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 639–646.
- Lieder, F., Griffiths, T. L., & Goodman, N. D. (2012). Burn-in, bias, and the rationality of anchoring. *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2699–2707.
- Lieder, F., Griffiths, T. L., Huys, Q. J., & Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic bulletin & review*, 25(1), 322–349.
- Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4), 986–1005.
- Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *The oxford handbook of thinking and reasoning*. Oxford University Press.
- Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. *Psychological Review*, 111(2), 309–332.
- Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2016). A Bayesian theory of sequential causal learning and abstract transfer. *Cognitive Science*, 40(2), 404–439.

- Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.
- Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, *34*(1), 113–147.
- Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.
- Luce, R. D. (1959). *Individual choice behavior*. Wiley.
- Mahoney, J., & Schensul, D. (2006). Historical context and path dependence. In R. Goodin & C. Tilly (Eds.), *The oxford handbook of contextual political analysis* (pp. 454–471). Oxford University Press.
- Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *International Conference on Learning Representations*.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.
- Mayrhofer, R., & Waldmann, M. R. (2015). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science*, *39*(1), 65–95.
- Michotte, A. (1963). *The perception of causality*. Basic Books.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.
- Mollica, F., & Piantadosi, S. T. (2021). Logical word learning: The case of kinship. *Psychonomic Bulletin & Review*, *29*, 766–799.
- Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E., & Stone, P. (2020). Curriculum learning for reinforcement learning domains: A framework and survey. *arXiv preprint arXiv:2003.04960*.

- Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology*, *50*(2), 101–122.
- Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hall.
- Newton, I. (1675). Letter to Robert Hooke.
- Nørby, S. (2015). Why forget? on the adaptive value of memory loss. *Perspectives on Psychological Science*, *10*(5), 551–578.
- Oaksford, M., Chater, N. et al. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.
- O'Donnell, T. J., Tenenbaum, J. B., & Goodman, N. D. (2009). Fragment grammars: Exploring computation and reuse in language.
- Pearl, J. (2000). *Causality: Model, reasoning, and inference*. Cambridge University Press.
- Pearl, J. (2009). *Causality*. Cambridge University Press.
- Piantadosi, S. T. (2021). The computational origin of representation. *Minds and Machines*, *31*(1), 1–58.
- Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199–217.
- Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392–424.
- Pitman, J., & Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Annals of Probability*, *25*, 855–900.
- Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2022). The best game in town: The re-emergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 1–55.

- Quine, W. V. O., & Ullian, J. S. (1978). *The web of belief* (Vol. 2). Random house New York.
- Redlawsk, D. P. (2002). Hot cognition or cool consideration? testing the effects of motivated reasoning on political decision making. *Journal of Politics*, *64*(4), 1021–1044.
- Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, *27*(5), 709–748.
- Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, *130*(3), 323–360.
- Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, *147*(11), 1553–1570.
- Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*, 527–535.
- Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in Cognitive Sciences*, *24*(11), 900–915.
- Ryan, E. G., Drovandi, C. C., McGree, J. M., & Pettitt, A. N. (2016). A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, *84*(1), 128–154.
- Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.
- 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.
- Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309.

- Schönfinkel, M. (1924). Über die bausteine der mathematischen logik. *Mathematische Annalen*, (92), 305–316.
- Schulz, E. (2017). *Towards a unifying theory of generalization* (Doctoral dissertation). UCL (University College London).
- Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development*, 77(2), 427–442.
- Searcy, S. R., & Shafto, P. (2016). Cooperative inference: Features, objects, and collections. *Psychological Review*, 123(5), 510–533.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820), 1317–1323.
- Sim, Z. L., & Xu, F. (2017). Learning higher-order generalizations through free play: Evidence from 2- and 3-year-old children. *Developmental Psychology*, 53(4), 642–651.
- Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives*. Oxford University Press.
- Sloman, S. A., & Lagnado, D. (2005). The problem of induction. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 95–116).
- Soo, K. W., & Rottman, B. M. (2018). Causal strength induction from time series data. *Journal of Experimental Psychology: General*, 147(4), 485–513.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Citeseer.
- Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27(3), 453–489.
- Tenenbaum, J. B., & Griffiths, T. L. (2000). Structure learning in human causal induction. *Advances in Neural Information Processing Systems 13 (NIPS 2000)*, 59–65.
- Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4), 629–640.

- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.
- Thaker, P., Tenenbaum, J. B., & Gershman, S. J. (2017). Online learning of symbolic concepts. *Journal of Mathematical Psychology*, *77*, 10–20.
- Tian, L., Ellis, K., Kryven, M., & Tenenbaum, J. (2020). Learning abstract structure for drawing by efficient motor program induction. *Advances in Neural Information Processing Systems*, *33*, 2686–2697.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.
- Treisman, A. M., & Paterson, R. (1984). Emergent features, attention, and object perception. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(1), 12–31.
- Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.
- Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.
- Valentin, S., Zhao, B., Jiang, C., Bramley, N. R., & Lucas, C. (2021). Symbolic and sub-symbolic systems in people and machines. *Proceedings of the 43th Annual Meeting of the Cognitive Science Society*, 1–2.
- Van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, *32*(6), 939–984.
- Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.
- White, P. A. (2006). The causal asymmetry. *Psychological Review*, *113*(1), 132–147.

- Wilcox, T. (1999). Object individuation: Infants' use of shape, size, pattern, and color. *Cognition*, 72(2), 125–166.
- Wong, C., McCarthy, W. P., Grand, G., Friedman, Y., Tenenbaum, J. B., Andreas, J., Hawkins, R. D., & Fan, J. E. (2022). Identifying concept libraries from language about object structure. *arXiv preprint arXiv:2205.05666*.
- 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.
- Xia, Z*, Zhao, B*, Quillien, T., & Lucas, C. G. (2022). Dissecting causal asymmetries in inductive generalization. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, 592–598.
- Yeung, S., & Griffiths, T. L. (2015). Identifying expectations about the strength of causal relationships. *Cognitive Psychology*, 76, 1–29.
- Zhao, B., & Bramley, N. R. (2020). Order effects in one-shot causal generalization. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 3071.
- Zhao, B., Bramley, N. R., & Lucas, C. G. (2022). Powering up causal generalization: A model of human conceptual bootstrapping with adaptor grammars. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*, 1819–1826.
- Zhao, B., Lucas, C. G., & Bramley, N. R. (2021). Building object-based causal programs for human-like generalization. *Causal Inference & Machine Learning Workshop, NeurIPS 2021*, 1–11.
- Zhao, B., Lucas, C. G., & Bramley, N. R. (2022). How do people generalize causal relations over objects? a non-parametric Bayesian account. *Computational Brain & Behavior*, 5, 22–44.
- Zhao, B., Lucas, C. G., & Bramley, N. R. (2023). How cognition bootstraps its way to complex concepts. *Under review*.

Zhu, X., & Goldberg, A. B. (2009). *Introduction to semi-supervised learning*. Morgan & Claypool Publishers.

Appendices

Appendix A

Comparison with GPT-3

In response to a growing interest in large language models being able to solve reasoning tasks like people, I transcribed stimuli used in Experiments 4-7 in natural language, and probed GPT-3 to provide its best guesses about the causal relationship between observed causal agent and recipient objects. Overall, while GPT-3 is able to produce responses that flows naturally, it lacks some crucial inductive biases such as the multiplicative operation, and hence cannot discover the ground truth rule as most people do. Furthermore, there is no evidence for bootstrap learning in GPT-3's responses to either of the four experiments.

A list of GPT-3's responses is presented in Table A.1. For each entry in Table A.1, I provided a verbal description of the observed agent's effect on the recipient (see below), and asked the GPT-3 prompt to complete a guess about the underlying causal relationship. The first guess I got from each test was provided in Table A.1. The verbal description for each causal interaction follows a template

Observation: a magic egg with one stripe , touching three segments

Resulting segments: three

where the corresponding number of stripes, spots, segments, and resulting segments faithfully reflect those used in the experiments. After inputting three observations in Phase I and six in Phase II, I provided a question

How are stripes and spots changing the segments?

The text-completion responses from GPT-3 after this input is then collected and presented in Table A.1.

I further exploited this work flow to collect generalization predictions. To do so, after providing verbal description of observations, instead of asking the above-mentioned question, I presented a generalization task like so:

Observation: a magic egg with no stripe and no spot, touching three segments

Resulting segments:

where the number of stripes, spots and segments are taken according to each generalization task used in the experiments. By doing so, GPT-3 auto-completes a number, serving as its generalization prediction in this generalization task. Using the free OpenAI API, I was able to collect not only a single prediction, but also the top four predictions with their corresponding log-probabilities. The sum of these top four predictions' probabilities P usually constitute above 98% of total probabilities. I then equally divided $1 - P$ among the other possible segment numbers, leading to a posterior prediction over the entire selection space. The log likelihood fits for this posterior on participant data is $NLL_{GPT-3} = -26642.5$, surprisingly worse than the random baseline -25838.9 .

Exp.	Cur.	Ph.	GPT-3 completion
1	construct	1	Stripes are adding one segment for each stripe, while spots are subtracting one segment for each spot.
1	construct	2	The stripes and spots are not changing the segments.
1	decon	1	The stripes and spots on the egg are not affecting the segments.
1	decon	2	The stripes and spots on the magic egg are changing the resulting segments by adding or subtracting from the number of segments that are touching the egg.
1	combine	1	The number of stripes on the egg determines how many segments the egg will touch. The number of spots on the egg does not appear to have any effect.
1	combine	2	The stripes and spots on the magic egg are changing the segments because they are touching different segments. The egg with one stripe and one spot is touching two segments, while the egg with one stripe and three spots is touching three segments.
2	construct	1	The stripes and spots on the magic egg are changing the number of segments that the egg is touching. The more stripes and spots there are, the fewer segments the egg is touching.
2	construct	2	The stripes and spots are changing the segments by making the egg more magical. The more stripes and spots there are, the more magical the egg becomes, and the more segments it can touch.
2	decon	1	The stripes and spots on the magic egg seem to be affecting the segments in a way that is not yet known.

Exp.	Cur.	Ph.	GPT-3 completion
2	decon	2	The stripes and spots on the magic egg are changing the segments by adding or subtracting from the total number of segments. If there is one stripe and one spot, it will add two segments. If there are two stripes and two spots, it will subtract four segments.
2	combine	1	The stripes and spots on the egg seem to be dividing the segments into smaller parts. The more stripes or spots there are, the smaller the segments become.
2	combine	2	The stripes and spots on the egg seem to be dividing the segments into halves.
3	combine	1	The stripes and spots on the magic egg are changing the segments by adding or subtracting from the total number of segments.
3	combine	2	The stripes are dividing the segments into thirds, while the spots are halving the segments.
3	flip	1	The stripes and spots are not changing the segments.
3	flip	2	The stripes and spots on the egg seem to be affecting the number of segments the egg is touching. If the egg has one stripe and one spot, it is touching two segments. If the egg has one stripe and two spots, it is touching one segment. If the egg has one stripe and three spots, it is not touching any segments.
4	combine	1	The stripes and spots on the magic egg are changing the segments by breaking them up into smaller pieces. The egg is essentially dividing the segments into smaller sections, which makes it easier for the egg to touch more segments at once.

Exp.	Cur.	Ph.	GPT-3 completion
4	combine	2	It appears that stripes and spots are both reducing the number of resulting segments.
4	flip	1	The stripes and spots are not changing the segments.
4	flip	2	The stripes and spots on the magic egg are changing the segments by adding or subtracting from the total number of segments. One stripe subtracts one segment, while one spot adds one segment.

Table A.1: GPT-3 guesses about causal relationships

Appendix B

Open resources

Experiment 1

- Demo: http://bramleylab.ppls.ed.ac.uk/experiments/bnz/magic_stones/index.html
- Data: https://github.com/bramleyccslab/causal_objects
- Models: https://github.com/bramleyccslab/causal_objects

Experiment 2

- Demo: <http://bramleylab.ppls.ed.ac.uk/experiments/bnz/myst/p/welcome.html>
- Data: https://github.com/bramleyccslab/causal_objects
- Models: https://github.com/bramleyccslab/causal_objects

Experiment 3

- Data: <https://osf.io/en9uy/>

Experiments 4-7

- Pre-registrations:
 - Experiment 4: <https://osf.io/ud7jc/>
 - Experiment 5: <https://osf.io/k5dc3/>
 - Experiment 6: <https://osf.io/mfxa6/>
 - Experiment 7: <https://osf.io/swde5/>
- Demo: <https://bramleylab.ppls.ed.ac.uk/experiments/bootstrapping/p/welcome.html>
- Data: <https://osf.io/9awhj/>
- Models: https://github.com/bramleyccslab/causal_bootstrapping
- Analysis: <https://bramleylab.ppls.ed.ac.uk/experiments/bootstrapping/analysis.html>