Title
Learning Causal Structure in Social, Statistical and Imagined Contexts

Permalink
https://escholarship.org/uc/item/8vx9r6ms

Author
Buchsbaum, Daphna

Publication Date
2013

Peer reviewed|Thesis/dissertation

Learning Causal Structure in Social, Statistical and Imagined Contexts

by

Daphna Buchsbaum

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Thomas L. Griffiths, Co-chair
Professor Alison Gopnik, Co-chair
Professor John Campbell

Fall 2013

Learning Causal Structure in Social, Statistical and Imagined Contexts

Abstract

Learning Causal Structure in Social, Statistical and Imagined Contexts

by

Daphna Buchsbaum

Doctor of Philosophy in Psychology

University of California, Berkeley

Associate Professor Thomas L. Griffiths, Co-chair

Professor Alison Gopnik, Co-chair

A major challenge children face is uncovering the causal structure of the world around them. Previous research on children's causal inference has demonstrated their ability to learn about causal relationships in the physical environment using probabilistic evidence. However, children must also learn about causal relationships in the social environment, including discovering the causes of other people's behavior, and understanding the causal relationships between others' goal-directed actions and the outcomes of those actions. In addition, many of the causal relationships children experience do not occur in the physical world at all, but instead occur in richly causal imaginary worlds.

In this dissertation, we argue that social reasoning and causal reasoning are deeply linked, both in the real world and in children's minds. Children use both types of information together and in fact reason about both physical and social causation in fundamentally similar ways. We suggest that children jointly construct and update causal theories about their social and physical environment and that this process is best captured by probabilistic models of cognition. We also argue that causal pretense may serve as a form of counterfactual causal reasoning, allowing children to explore causal "what if" scenarios in alternative imaginary worlds, and suggest a theoretical link between the development of an extended period of immaturity in human evolution and the emergence of powerful and wide-ranging causal learning mechanisms.

We investigate the complex and varied ways in which children learn causal relationships through three primary lines of research, each of which extends probabilistic models beyond reasoning about purely physical causes, while also characterizing the distinctive aspects of causal pretense and social causal reasoning. In the first set of studies, we examine how causal learning can influence the understanding and segmentation of action, and how observed statistical structure in human action can affect causal inferences. We present a Bayesian analysis of how statistical and causal cues to segmentation should optimally be combined, as well as four experiments investigating human action segmentation and causal inference. We find that both adults and our model are sensitive to statistical regularities and causal structure in continuous action, and are able to combine these sources of information in order to correctly infer both causal relationships and segmentation boundaries.

The second line of work examines how the social context influences children's causal learning, focusing particularly on children's imitation of causal actions. We define a Bayesian model that predicts children will decide whether to imitate part or all of an action sequence based on both the pattern of statistical evidence and the demonstrator's pedagogical stance. We conducted an experiment in which preschool children watched an experimenter repeatedly perform sequences of varying actions followed by an outcome. Children's imitation of sequences that produced the outcome increased, in some cases resulting in production of shorter sequences of actions that the children had never seen performed in isolation. A second experiment established that children interpret the same statistical evidence differently when it comes from a knowledgeable teacher versus a naïve demonstrator, suggesting that children attend to both statistical and pedagogical evidence in deciding which actions to imitate, rather than obligately imitating successful action sequences.

The final line of work explores the relationship between children's understanding of real-world causal structure and their pretend play. We report a study demonstrating a link between pretend play and counterfactual causal reasoning. Preschool children given new information about a causal system made very similar inferences both when they considered counterfactuals about the system and when they engaged in pretend play about it. Counterfactual cognition and causally coherent pretense were also significantly correlated even when age, general cognitive development and executive function were controlled for. These findings link a distinctive human form of childhood play and an equally distinctive human form of causal inference. We speculate that during human evolution computations that were initially reserved for particularly important ecological problems came to be used much more widely and extensively during the long period of protected immaturity.

To my grandmother Irena Hecht

Contents

List of Figures

List of Tables

Acknowledgments

"Through others we become ourselves." – Lev S. Vygotsky

When it comes to mentorship, my time at Berkeley has been an embarrassment of riches. First and foremost my advisors Alison Gopnik and Tom Griffiths have provided not only invaluable research guidance but also advice, help and moral support. I am grateful to both of them for their compassion, pragmatism and senses of humor. I would like to thank Tom for never believing that taking another math class was a bad idea, and for helping me realize that thinking about the mind computationally is more than just software engineering. Alison has nurtured my 'pink thumb' in developmental psychology, and I am especially grateful for all the informal dinner discussions and 'baby teas' at her house over the years.

John Campbell has consistently provided an incisive and interesting outside perspective on my work, and I would like to thank him for being a committee member for my qualifying examination, as well as this dissertation. Fei Xu and Trevor Darrell have provided me with a variety of thougful commentary and advice on my research, as well as being committee members for my qualifying examination and informal advisors on my dissertation proposal. Over the years, Silvia Bunge has gone out of her way to provide me with both academic and non-academic advice, and I want to thank her for her compassion and support, and her wonderful work as graduate advisor. Dare Baldwin's support and interest in my work since I first started my PhD has also meant a lot to me.

The research in this dissertation could never have been accomplished without the help of my collaborators, the thoughtful feedback and guidance of my labmates and colleagues, and a small army of dedicated undergraduate research assistants. This work was developed along with my co-authors Alison Gopnik (chapters 1, 2, 3, and 4), Tom Griffiths (chapters 2 and 3), Elizabeth Seiver (chapter 1), Dare Baldwin (chapter 2), Pat Shafo (chapter 3), Sophie Bridgers (chapters 1 and 4) and Deena Skolnick Weisberg (chapter 4), and I would like to thank them for their contributions to this research.

I would like to thank Meredith Meyer for discussions and feedback on the studies presented in chapter 2, including help with experimental design and providing experimental stimuli, as well as Jennifer Ng, Kimmy Yung, James Bach, Mia Krstic, Jonathan Lesser and Sophie Bridgers for their help with data collection. I would also like to thank Cari Kaufman for fruitful discussions on the model design for the work presented in chapter 3, and Kimmy Yung, Mia Krstic, Rachel Kim, Sophie Bridgers and Elon Ullman for help with data collection and coding. In addition, Jennifer Ng, Adrienne Wente, Erin Klein, Matthew Yanus, Laura Hazlett, Francesca Ucciferri, Augustine Lombera, Ryan Woo, Swe Tun and Gabriela Libin all assisted with data collection and coding for the studies presented in chapter 4. Finally, Kimmy Yung, Jennifer Ng, Natasa Vrachimis, Emily Margolin, Aly DiRocco, Sophie Bridgers, Andrew Whalen and Adrienne Wente deserve extra gratitude for their excellent contributions to many projects.

I would also like to thank the members of the Computational Cognitive Science and Gopnik labs, especially Chris Lucas, Elizabeth Seiver and Anna Waismeyer who formed my welcoming committee at UC Berkeley, and Joe Austerweil who has been a friend and compatriot throughout this process. Also Joseph Williams, for asking impossible questions. Sophie Bridgers and Andrew Whalen have set the bar for my future students, I hope they've learned half as much from me as I've learned from them. Jane Hu, Caren Walker and Kevin Canini have all been not only great labmates but also great collaborators. Chris Holdgraf is a man. Stephanie Denison has never failed to give me good advice. Anna Rafferty, Jing Xu, Kevin Uttich, Nick Gwynne, Mike Pacer, Josh Abbott, Naomi Feldman and Liz Bonawitz have provided friendship, moral support and inspiration.

I am also grateful to the funding sources that made this work possible – the NSF Graduate Research Fellowship, the McDonnell Foundation Causal Learning Initiative, Grant FA9550-07-1-0351 from the Air Force Office of Scientific Research, grant FA-9550-10-1-0232 from the Air Force Office of Scientific Research, grant IIS-0845410 from the National Science Foundation and Grant BCS-1023875 from the National Science Foundation.

I have never been known as organized, and I would like to thank the administrative staff of the UC Berkeley Psychology department, especially Michael Ortt and John Schindell for helping me with all my last minute paperwork over the years.

Going back in time, if there is any single person I can blame for starting me down this path then it is Jon Waage and his squirrel-chasing course BIO 45 Animal Behavior. Doug Morse is a true field ecologist who continues to shape my thinking about evolution, and inspire an obsessive need to know the name of every insect and arachnid. John "Spike" Hughes introduced me to computer science, and with it a new way of thinking. Bruce Blumberg and the other members of the late and great Synthetic Characters group not only helped me cultivate interests from robotics to cognitive ethology, they changed me from a beginning programmer into a software engineer. Eric Bonabeau, Paolo Gaudiano, Pablo Funes, Julien Budynek and my other Icosystem friends and coworkers will be delighted to know that there is a bit of the Chinese Postman Problem in this dissertation.

The California Rescue Dog Association has been an incredibly meaningful part of my time in Berkeley. I would especially like to thank the members of the NEBay and EBay training groups for their support, advice and friendship. Kathryn Stewart deserves special acknowledgment for, among other things, her insistence that I graduate and get a job. I always wanted a dog, and my search partner Pumpkin is exactly the dog I always wanted.

Over the years I have been incredibly lucky to not have to leave home to find my friends. My parents Maya and Gershon Buchsbaum have always supported my academic aspirations. My sisters Nilly and Talia are amazing, and know what it means to "pull a Buchsbaum". The residents (and honorary residents) of 46/48 East George, 41 Burnside and 845 The Alameda are some of my favorite people.

Last time I wrote some acknowledgments, Dan Goldwater was merely a monkey. These days, he is Monkeylectric. Thank you for lighting up my life (see what I did there?). Finally, my son Yoshi reminds me every day that PhD or not I know nothing about cognitive development.

Chapter 1

Introduction

1.1 Overview

In the past 10 years, the probabilistic modeling approach to cognitive development, also known as rational constructivism (e.g., Xu & Kushnir, 2012), has begun to be applied to many aspects of children's development, particularly their causal inference and learning. In the first wave of this research, however, the focus was squarely on real-world physical knowledge, such as the relation between blickets and blicket detectors (or the workings of other physical machines). In these types of studies, for example, an experimenter may place a series of blocks on top of a machine. Some blocks are "blickets" and make the machine produce an effect (e.g., lighting up and playing music), while other blocks do not. Children are then asked to make causal inferences from the evidence they see, such as which block was a blicket or which new block should make the machine go. This research program has demonstrated that children possess sophisticated causal reasoning abilities, including making rational inferences from probabilistic input (e.g., Gopnik et al., 2004; Kushnir & Gopnik, 2005, 2007; Schulz, Bonawitz, & Griffiths, 2007; Schulz, Gopnik, & Glymour, 2007; Sobel & Kirkham, 2006; Sobel, Tenenbaum, & Gopnik, 2004).

These initial studies were generally limited to investigating how children learn by observing causal relationships in their physical environment and did not take the child's social environment into account. From an early age, children are exquisitely sensitive social beings and their causal learning takes place in a rich social context. A natural question is therefore how social interaction informs and influences children's causal learning and how causal reasoning influences children's social inferences.

In a similar vein, while previous work examined how children represent and act on causal relationships they observe in the real world, it has had relatively little to say about how children use causal relationships in pretend scenarios. Pretend play is a characteristic and pervasive behavior of human children (see for instance Fein, 1981; Harris, Kavanaugh, Well-

[1] Portions of this chapter were adapted from the co-authored work Buchsbaum, Seiver, Bridgers, and Gopnik (2012).

man, & Hickling, 1993; Leslie, 1987; Singer & Singer, 1990; Gosso, Otta, Morais, Ribeiro, & Bussab, 2005, but also Lillard et al., 2013 for a more critical view), in which children frequently act out imagined causal secenarios and also observe and respond to the causal consequences of others' pretend actions (Harris et al., 1993). How do children navigate and learn from the causal consequences of actions taken in imaginary worlds?

In this dissertation, we present three lines of research that apply the ideas of probabilistic modeling to causal cognition and explore the complex and interdependent relationship between social and causal learning. We examine how the social context, in the form of both demonstrations and testimony, influences children's causal learning. We also examine how causal learning can influence the understanding and segmentation of action and how observed statistical structure in human action can affect causal inferences. In the final set of studies, we examine how children might use causal pretend play as a mechanism for exploring causal counterfactuals, in the same way that they use real-world exploratory play to infer physical causal structure (Schulz & Bonawitz, 2007; Cook, Goodman, & Schulz, 2011).

1.2 The Social Context of Causal Reasoning

Data about "purely physical" causes does not exist in a vacuum – blickets are not putting themselves on the machine, after all. There is a social and psychological component to the causal learning that results from our interactions with other people. Even in the relatively simple context of a blicket detector experiment, the child not only must consider the physical evidence of the machine's activation but also must make inferences about the experimenter's actions and mental states. Did she put the blicket on the machine in the right way? She says she knows what makes the machine go, but does she? Is she just trying to make the machine go or does she also want to teach me how it works? Children can use the physical blicket evidence to make social inferences (the block did not work, so she must not know what she is doing) or use the experimenter's testimony and actions to make inferences about the blickets (since she says she knows what she is doing, she must be teaching me about which blickets I should use, so I will pick the same one).

In general, social and physical causation will be inextricably linked in most real-life causal learning, especially since the goal-directed actions of others lead to many of the causal outcomes children observe. In fact, even infants and toddlers seem to expect that the causally relevant events they observe in the world will have been produced by the actions of social agents (e.g., Bonawitz et al., 2010; Meltzoff, Waismeyer, & Gopnik, 2012; Saxe, Tenenbaum, & Carey, 2005; Saxe, Tzelnic, & Carey, 2007).

We argue that children jointly construct theories about both the physical and the social world, which in turn generate higher-order theories that shape children's interpretation of future events. This natural learning process parallels the scientific method, and thus, we can characterize children's learning with the metaphor of children as intuitive scientists. This metaphor might suggest that children just learn on their own, but neither children nor scientists are solitary learners. Both scientists and children learn extensively from the actions, reports, and tuition of others.

Teachers serve a particularly important function in this regard, both formally in the classroom and informally in the world. Recent work on "natural pedagogy" (Csibra & Gergely, 2006, 2009; Gergely, Egyed, & Király, 2007) and children's understanding of testimony (e.g., Corriveau, Meints, & Harris, 2009; Jaswal, Croft, Setia, & Cole, 2010; Koenig & Harris, 2005; Pasquini, Corriveau, Koenig, & Harris, 2007; Harris & Corriveau, 2011) has demonstrated that infants and young children are sensitively tuned to others and can learn from them in complex and subtle ways. The pedagogical intent of a social demonstrator can influence everything from children's exploration of a novel toy (Bonawitz et al., 2011) to their generalizations about objects' functional properties (Butler & Markman, 2012). The expertise (e.g., Koenig & Jaswal, 2011; Kushnir, Vredenburgh, & Schneider, 2013; Sobel & Corriveau, 2010) and past accuracy (e.g., Birch, Vauthier, & Bloom, 2008; Corriveau, Meints, & Harris, 2009) of a social informant affects what children learn from this informant in the future.

Other recent results further support the notion that we can apply probabilistic models to both the social context of causal understanding and the causal context of the social world. Schulz and Gopnik (2004) found that children inferred psychological causal relationships from covariation in much the same way that they inferred physical and biological relationships. Kushnir, Xu, and Wellman (2010) and Ma and Xu (2011) found that infants as young as 14 months old showed some capacity to infer an underlying desire from a person's pattern of nonrandom sampling behavior. Additionally, Kushnir, Wellman, and Gelman (2008) and Sobel, Sommerville, Travers, Blumenthal, and Stoddard (2009) found that children's causal inferences are sensitive to the social environment. On the computational side, Shafto and colleagues (Bonawitz et al., 2011; Shafto & Goodman, 2008; Shafto, Goodman, Gerstle, & Ladusaw, 2010) have modeled how pedagogical information may be used differently than nonpedagogical information in solving inductive problems.

How children learn from social sources of causal information becomes an especially interesting question when we move beyond artificial laboratory tasks such as blicket detectors. Much of the real-world causal evidence children receive involves complex statistical patterns of both actions and outcomes. Consider the case of learning which actions are necessary to open a door. Children might notice that people almost always grasp and then turn a doorknob before the door opens, but sometimes they pull a handle instead. They frequently insert a key into a lock and then turn it before trying the doorknob, but not always. Sometimes the sequence of actions must be repeated a couple of times (for instance, in the case of a jammed lock); other times, the sequence fails and is not followed by the door opening at all. Often, other actions precede the door opening as well – putting down groceries, fumbling around in a purse, ringing a doorbell, sliding a bolt – which of these are causally necessary and which are incidental? Does the order they were performed in matter? Finally, in addition to these observations, children might receive direct testimony about the door. For instance, someone who lived in the house might say that jiggling the key almost always works or someone unfamiliar with the door might guess that this is the case. How might children combine these statements with other sources of causal evidence?

In just this simple example of opening a door, we can see that there are not only many potential types of causal information available but also many different sources of statistical variation and ambiguity. There is variation in the physical data – actions (and other causes) may not always bring about their effects or may only lead to the desired outcome in certain combinations. There is variation in the action sequence – repeated demonstrations of bringing about the same outcome may include different actions. There is variation in people's behavior – some individuals might succeed at opening the door while others fail or might be successful with one door while failing to open another. There is even variation in direct testimony – people may express differing levels of certainty and causal knowledge, and the testimony of multiple people may even conflict. Finally, children must also take into account their own prior knowledge and expectations about not only the causal system in question but also the intentions, knowledgeability, and helpfulness of their social informant, all of which could vary widely across situations.

On the other hand, while all this ambiguity can make the causal inference problem children face more challenging, there are times when the presence of statistical variation can actually be quite illuminating and aid inference. Actions that do not consistently precede outcomes are less likely to be causally necessary. Actions that reliably appear together and, in fact, predict each other, are more likely to be coherent units, corresponding to intentional, goal-directed action.

1.3 Probabilistic Causal Models

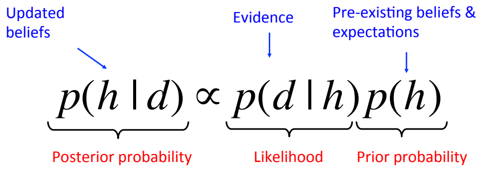


Figure 1.1: Bayes' rule. For any hypothesis h and data d.

Computational modeling, particularly probabilistic models often known as Bayesian models, can be extremely helpful in disentangling these complex inferences. In this approach we use Bayes' rule (see Figure 1.1) as a normative model of how an idealized learner with some pre-existing expectations or biases about how the world works can update their beliefs in light of new data (for some recent overviews see for example Griffiths, Kemp, & Tenenbaum, 2008; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Perfors, Tenenbaum, Griffiths, & Xu, 2011; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Describing the child's current conception of the world as a particular rational model gives us a more exact way of both characterizing the child's beliefs and working out the predictions that should rationally follow from those beliefs.

Bayesian models work by assuming that a learner is evaluating a set of hypotheses about the state of the world, and has assigned a "prior" probability $P(h)$ to each hypothesis h in that set. Then, Bayes' rule indicates that after seeing data d, the learner should assign each hypothesis a "posterior" probability $P(h|d)$ proportional to $P(h)$ multiplied by the probability of observing d if h were true, $P(d|h)$. Bayes' rule is a principled way to combine inductive biases, represented as the prior distribution, with the evidence provided by data, using an explicit model of how the world generated that data. For instance, in the case of causal inference from social demonstrations, the data might be observed contingencies between actions and outcomes, the prior could be inductive biases about the *a priori* plausibility of different causal relationships, and the generative model could correspond to the belief that the observed actions were generated randomly, or by an intentional actor, or by a helpful teacher.

Probabilistic computational models are a natural way to approach understanding how social information, along with other evidence, contributes to children's causal reasoning, allowing us to both represent ambiguous and probabilistic data, and to capture the variability we often see in children's inferences and behavior. It is common in developmental psychology to see children make different judgments in different contexts. This inconsistency has sometimes been taken to mean that all children's cognition is variable and context dependent and that there is no coherent conceptual structure to be found (e.g., Greeno, 1998; Lave & Wenger, 1991; Thelen & Smith, 1996). At other times, it has led to unresolved debates, for example, about whether early imitation is rational or not. As we will see, probabilistic models allow one to precisely show how multiple sources of evidence, reflecting different contexts, can be rationally combined and integrated to lead to a particular response.

1.4 Counterfactuals, Causal Models and Pretense

One area where probabilistic models have been especially successful in capturing human inference is in the domain of causal reasoning. Recent work has outlined the kinds of representations that underpin causal knowledge in adult humans and the kinds of mechanisms that allow this knowledge to be learned (e.g., Gopnik et al., 2004; Gopnik & Schulz, 2007; Griffiths & Tenenbaum, 2009). The essential idea behind this recent research is that human causal reasoning utilizes structured, generative, causal representations of the world. These representations appear to go beyond the typical representations that might be constructed from simple associative processes or conditioning.

Formal models of causal relationships, such as causal graphical models, can be used to capture these types of structured representations. Causal graphical models, also known as "Bayes' nets" represent networks of causal relations as graph structures associated with probability distributions (Pearl, 2000; Spirtes, Glymour, & Schienes, 2001), that allow distinctive kinds of inferences. In particular, these models distinguish between two types of inferences, predictions on the one hand, and counterfactuals, interventions and hypotheses, on the other. In predictions, we take what we think is true now as a premise and then use the model to calculate what else will be true. In counterfactuals, interventions and hypotheses, we take some value of the model that we currently think is not true as a premise, and calculate what would follow if it were. The mathematics of causal models shows that these two kinds of calculations are very different. Generating interventions lets us plan, we can consider alternative routes that might lead to some desirable outcome. Generating alternative hypotheses lets us learn. We can change some feature of our current causal model, and calculate the predictions that will follow. If those predictions are more accurate than those of our current model, we should switch to the new model.

This counterfactual thinking, or envisioning alternative possible worlds and their outcomes, bears a striking resemblance to childhood pretense. Earlier researchers remarked on the similarities between pretense and counterfactual reasoning, and previous work suggests that children do use counterfactuals in their causal reasoning (e.g., Harris, German, & Mills, 1996). Here we suggest that these crucially important abilities – creating possible causal interventions and testing alternative causal hypotheses – depend on the same cognitive machinery that children use when they pretend: adopting a premise that is currently not true, creating an event sequence that follows from that premise, and quarantining the result of this process from reality. Although there is evidence that children can respond to known causal relationships in their pretend play (e.g., Harris et al., 1993), the relationship between counterfactual causal reasoning and pretending has not been empirically explored. Here we test the hypothesis that pretense may serve as a form of counterfactual causal reasoning, allowing children to explore causal "what if" scenarios in alternative imaginary worlds.

1.5 Dissertation Outline

This dissertation presents three lines of research exploring how children (and adults) integrate directly observed patterns of cause and effect with other sources of causal data, focusing particularly on social information, and how children reason about causal relations not only in the real-world, but also in imaginary worlds. All three lines of research extend probabilistic models beyond reasoning about purely physical causes, while also characterizing the distinctive aspects of causal pretense and social causal reasoning.

Chapter 2 presents a series of experiments investigating human action segmentation and causal inference, as well as a Bayesian analysis of how statistical and causal cues to action segmentation should optimally be combined. This chapter is especially influenced by prior work on statistical word segmentation, drawing an analogy between the domains of language and action processing.

Chapter 3 looks at children's imitation of causal action sequences. This chapter examines how children decide which of the actions they see another person performing are the ones that are necessary to bring about a desirable outcome. The studies presented in this chapter look particularly at how this decision is informed by different sources of information, including contingencies between action sequences and outcomes across repeated demonstrations, and information about the actor's knowledge state and pedagogical intentions.

Chapter 4 explores the relationship between children's understanding of real-world causal structure and their pretend play. In particular, this chapter empirically examines the relationship between childhood pretense and causal counterfactual reasoning, and looks at the actions children choose when intervening on pretend causal systems. All three lines of research extend probabilistic models beyond reasoning about purely physical causes, while also characterizing the distinctive aspects of causal pretense and social causal reasoning.

The final chapter summarizes the dissertation, discusses some future directions for research and presents concluding remarks.

Chapter 2

Inferring Causal Variables from Continuous Action Sequences

Never mistake motion for action. – Ernest Hemingway

2.1 Introduction

Human social reasoning depends on understanding the relationship between actions, goals and outcomes. In order to understand the reasons behind others' behavior, we must be able to distinguish the unique actions we see others performing, and recognize the effects of these actions. Recall the example from Chapter 1 of watching someone coming home and opening their front door. To understand this simple scene, an observer needs to parse the continuous stream of motion they see into meaningful behaviors such as "exiting the car", "coming up the stairs" and "opening the door", which are themselves composed of smaller motion elements such as "standing up", "closing the car door", "taking a step", "reaching for the doorknob", and so on.

Determining which subsequences of motion go together hierarchically, and what outcomes they produce, is also an important instance of the more general problem of causal variable discovery (a similar problem – determining how spatially distributed observations should be encoded as variables – is discussed by Goodman, Mansinghka, & Tenenbaum, 2007). Coming back to our example, consider the case of learning which actions are necessary to open a door by observing multiple performances, embedded in everyday scenes such as the one above. A learner might notice that people almost always grasp and then turn a doorknob before the door opens, but sometimes they pull a handle instead. They frequently insert a key into a lock and then turn it before trying the doorknob, but not always. Often, other actions precede the door opening as well – putting down groceries, fumbling around in a purse, ringing a doorbell, sliding a bolt – which of these are causally necessary and which are

[1] This chapter was adapted from the co-authored manuscript Buchsbaum, Griffiths, Gopnik, and Baldwin (In prep).

incidental? While this ambiguity can make causal learning more challenging, the presence of statistical variation can actually aid inference. Motions that do not consistently precede outcomes are less likely to be causally necessary. Motions that reliably appear together and, in fact, predict each other, are more likely to be coherent units, corresponding to intentional, goal-directed action.

Prior research has shown that adults are able to segment common everyday behaviors into coherent actions, corresponding to the goals and intentions underlying the actor's behavior (for a recent review see Kurby & Zacks, 2008), and that even young infants are sensitive to the boundaries between intentional action segments (A. L. Woodward & Sommerville, 2000; Baldwin, Baird, Saylor, & Clark, 2001; Saylor, Baldwin, Baird, & LaBounty, 2007; Hespos, Saylor, & Grossman, 2009). While a full understanding of human action requires knowledge about goals and intentions, infants are able to parse dynamic human action well before they are thought to have a fully developed theory of mind. This suggests that there may also be low-level cues to intentional action structure available in human motion, an idea supported by a variety of recent work (Zacks, 2004; Hard, Tversky, & Lang, 2006; Zacks, Kumar, Abrams, & Mehta, 2009; Meyer, DeCamp, Hard, & Baldwin, 2010; Hard, Recchia, & Tversky, 2011; Buchsbaum, Canini, & Griffiths, 2011).

However, this previous work has primarily focused on motion and image cues to boundary locations. Another potentially important source of information is statistical regularities in the action stream. There is now a large body of evidence suggesting that both infants and adults can use statistical patterns in spoken language to help solve the related problem of segmenting words from continuous speech (for a partial review, see Gómez & Gerken, 2000). Recently, Baldwin, Andersson, Saffran, and Meyer (2008) demonstrated that a similar sensitivity to statistical regularities in continuous action sequences may play an important role in action processing. However, a key difference between action segmentation and word segmentation is that intentional actions usually have effects in the world. In fact, many of the causal relationships we experience result from our own and others' actions, suggesting that understanding action may bootstrap learning about causation, and vice versa. Though recent work has demonstrated that both children and adults can infer causal relationships from conditional probabilities (Gopnik et al., 2004; Griffiths & Tenenbaum, 2005), the extent to which action understanding and causal learning mechanisms inform each other has yet to be explored. In addition, previous work on causal inference has generally assumed that the possible causes are known in advance. Figuring out how causal actions are identified from within a continuous sequence remains an important problem in this area. Here we present a combination of experimental and computational approaches investigating how the ability to segment action and to infer its causal structure functions and develops.

We first introduce a Bayesian analysis of action segmentation and causal inference, which provides a rational analysis of how statistical and causal cues to segmentation should optimally be combined. Next, we present four experiments investigating how people use statistical and causal cues to action structure. Our first experiment demonstrates that adults are able to segment out statistically determined actions, and experience them as coherent, meaningful and most importantly, causal sequences. Our second and third experiments show that adults are able to extract the correct causal variables from within a longer action sequence, and that they find causal sequences to be more coherent and meaningful than other sequences with equivalent statistical structure. Our fourth experiment demonstrates that when statistical and causal cues conflict both sets of cues influence segmentation and causal inference, suggesting that action structure and causal structure are learned jointly and simultaneously. We also look at the action segmentations and causal structures our Bayesian rational model predicts, when given the same experimental stimuli as our human participants. We conclude by discussing our results in the context of broader work, as well as its implications for more generalized human statistical learning abilities.

2.2 Statistical Segmentation

Many if not most of the causal outcomes we witness are the result of intentional human action. We must be able to distinguish the unique actions we see other people performing and recognize their effects in order to understand the reasons behind others behavior and in order to potentially bring about those effects ourselves. But before we can interpret actions, we first must parse a continuous stream of motion into meaningful behavior (Sharon & Wynn, 1998; Byrne, 2003). What cues do we use to do this? How might infants and young children begin to break into the behavior stream in order to identify intentional, goal-directed actions? Could the causal relationships between actions and their outcomes in the world help us understand action structure itself? How might we identify reaching, grasping, and turning and then group them into the action "opening the door"?

One way that infants might be able to segment actions is by using statistical regularities in human motion. There is now a lot of evidence that both infants and adults use statistical patterns in spoken language to help solve the related problem of segmenting words from continuous speech(e.g., Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Aslin, Saffran, & Newport, 1998; Pelucchi, Hay, & Saffran, 2009). In this research, infants (and adults) listen to an artificial language constructed of made-up words, usually created from English syllables (e.g., dutaba, patubi, pidabu). The words are assembled into a continuous speech stream (e.g., dutabapatubipidabu.), with other potential segmentation cues such as intonation and pauses removed. In these experiments, as in many words in real languages, syllables within a word have higher transitional probabilities than syllables between words – you are more likely to hear ta followed by ba (as in dutaba) than to hear bi followed by pi (as in patubi pidabu). Both infants and adults are able to use these transitional probabilities in order to distinguish words in these artificial languages (dutaba, patubi, pidabu), from part-words – combinations of syllables that cross a word boundary (e.g., tabapa, tubipi), and from non-words, combinations of syllables that do not appear in the artificial language at all (e.g., dupapi, babibu). Infants have also been shown to succeed at statistical language segmentation even when more naturalistic language stimuli are used (Lew-Williams, Pelucchi, & Saffran, 2011; Pelucchi et al., 2009).

More recently, a similar sensitivity to statistical regularities has been shown to play a role in action segmentation in both adults (Baldwin et al., 2008) and infants (Roseberry, Richie, Hirsh-Pasek, Golinkoff, & Shipley, 2011). Intriguingly, there is also evidence that children and adults can successfully map words learned through this type of segmentation to meanings (Mirman, Magnuson, Graf Estes, & Dixon, 2008; Graf Estes, Evans, Alibali, & Saffran, 2007; Hay, Pelucchi, Graf Estes, & Saffran, 2011) and, conversely, can use words they already know to help find segment boundaries and discover new words (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005). Similarly, a recent study shows that, in the visual domain, children use statistical patterns to infer the boundaries between objects and then use that information to make further predictions about how objects will behave (Wu, Gopnik, Richardson, & Kirkham, 2011). So children do not just detect the statistics and then segment the streams accordingly. They actually treat those statistical units as if they were meaningful.

In the same way that words have meanings, intentional actions usually lead to causal outcomes. This suggests that just as identifying words assists in mapping them to meanings, segmenting human action may bootstrap learning about causation and vice versa. Recent work has demonstrated that adults can segment videos of common everyday behaviors into coherent actions (Baldwin et al., 2008; Hard et al., 2006; Meyer et al., 2010; Newtson, Engquist, & Bois, 1977; Zacks & Tversky, 2001; Zacks et al., 2009) and that both children and adults can infer causal relationships from conditional probabilities (Cheng, 1997; Gopnik et al., 2004; Griffiths, Sobel, Tenenbaum, & Gopnik, 2011). However, researchers have not yet explored whether action parsing and causal structure can be learned jointly.

In this work, we adapt a Bayesian word segmentation model (Goldwater, Griffiths, & Johnson, 2009), with actions composed of individual small motion elements (SMEs) taking the place of words composed of phonemes or syllables, and extend this model to incorporate causal information. The key intuition behind this model is that action segmentation and causal structure are jointly learned, taking advantage of statistical evidence in both domains.

2.3 Bayesian Analysis of Action Segmentation

We created a Bayesian rational learner model that jointly infers action segmentation and causal structure, using statistical regularities and temporal cues to causal relationships in an action stream. This model provides us with a way to begin characterizing both the kinds of information available in the action stream, and what an optimal computational level solution to these inference problems might look like. To the extent that our model accurately reflects human performance, it provides additional support for the idea that people may similarly be combining statistical and causal cues in their own inference.

We adapted the nonparametric Bayesian word segmentation model first used by Goldwater et al. (2009) to the action domain, and also extended this model to incorporate causal information. In this model, actions composed of individual small motion elements take the place of words composed of phonemes. We model the generative process for creating a sequence of human actions as successively selecting actions to add to the sequence, with the conditional probability of generating a particular action given the sequence so far given by the *Chinese Restaurant Process* (Aldous, 1985). In addition, we incorporated cause and effect information into the generative model, allowing some actions to be probabilistic causes. We describe this model in more detail in the following sections, as well as in Appendix A.1.

Generative Model for Action Sequences

Just as a sentence is composed of words, which are in turn composed of phonemes, in our model an action sequence A is composed of actions a_i which are themselves composed of motion elements m_j. We assume a finite set of possible actions, and that complete actions are chosen one at a time from this set, and then added to the the sequence. The conditional probability of the next action in the sequence $p(a_i|a_1...a_{i-1})$, is given by a standard construction known as the *Chinese Restaurant Process* (CRP). In the CRP customers enter a restaurant, and are seated at tables, each of which has an associated label. In this case, the labels are actions. When the i^{th} customer enters the restaurant, they sit at a table z_i, which is either a new table or an already occupied table. The label at table z_i becomes the i^{th} action in our sequence with

$$p(z_i = k|z_1...z_{i-1}) = \begin{cases} \frac{n_k}{i-1+\alpha_0}, & 0 \leq k \leq K \\ \frac{\alpha_0}{i-1+\alpha_0}, & k = K+1 \end{cases} \tag{2.1}$$

where n_k is the number of customers already at table k, and K is the number of previously occupied tables. So, the probability of the i^{th} customer sitting at an already occupied table depends on the proportion of customers already at that table, while the probability of starting a new table depends on the *concentration parameter* α_0.

Whenever a customer starts a new table, an action a_k must be associated with this table. Since multiple tables may be labeled with the same action, the probability that the next action in the sequence will have a particular value $a_i = w$ is

$$p(a_i = w|a_1...a_{i-1}) = \frac{n_w}{i-1+\alpha_0} + \frac{\alpha_0 P_0(a_i = w)}{i-1+\alpha_0} \tag{2.2}$$

where n_w is the number of customers already seated at tables labeled with action w. In other words, the probability of a particular action $a_i = w$ being selected is based on the number of times it has already been selected, and the probability of generating it anew. We draw new action labels from the *base distribution* P_0. Actions are created by adding motions one at a time, so that $P_0(a_i = w)$ is simply the product of action w's component motion probabilities, with an added assumption that action lengths are geometrically distributed with

$$P_0(a_i) = p_{\#}(1 - p_{\#})^{n-1} \prod_{j=1}^{n} p(m_j) \tag{2.3}$$

where n is the length of a_i in motions, $p_{\#}$ is the probability of ending the action after each motion, and $p(m_j)$ is the probability of an individual motion. Currently, we use a uniform distribution over all motions. Finally, like action length, we assume that action sequence length is also geometrically distributed, and $p_\$$ is the probability of ending the sequence after each action.

In this work we expect $p_\#$ to be relatively large, which represents a bias towards finding smaller length actions, $p_\$$ to be small, biasing the model towards sequences made up of more actions, and α_0 to be small, representing an expectation that the set of all possible actions is relatively small.

Generative Model for Events

The action sequence A also contains non-action events e, which can occur between motions. In our model, some actions are causal sequences, and are followed by an event with high probability. Each unique possible action a_w has an associated binary variable $c_w \in \{0, 1\}$ that determines whether or not the action is causal with $c_w \sim$ Bernoulli(π). We assume that π is small, which represents an assumption that relatively few actions are causes for a particular effect. If an action is a causal sequence, then it is followed by an event with some relatively high probability ω. We use a small value ϵ for the probability of an effect occurring after a non-causal sequence (in the middle of an action, or after a non-causal action, as shown in Figure 2.1). This represents our assumption that events are unlikely to follow non-causal sequences, and likely to occur after actions that are causal sequences.

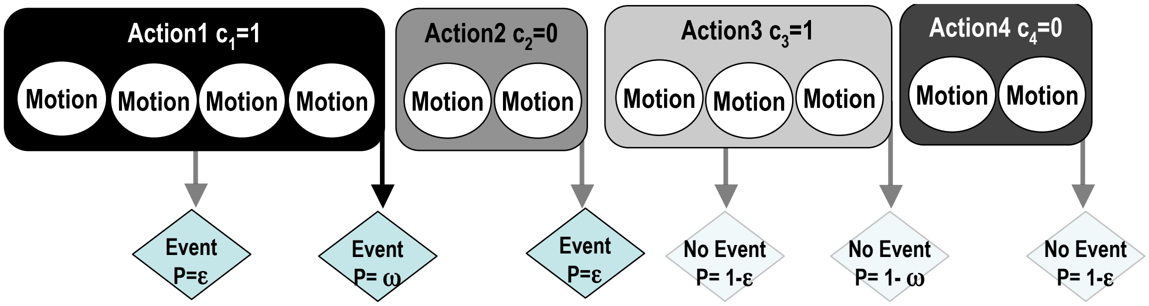


Figure 2.1: A theoretical action sequence depicting causal relationships in the model.

Inferring Segmentation and Causal Structure

An unsegmented action sequence consists of the motions m_j without any breaks between them. Given such a sequence, how do we find the boundaries between actions? A segmentation hypothesis h indicates whether there is an action boundary after each motion m_j. For a given segmentation hypothesis h, and unsegmented action sequence d, we use Bayes' rule $p(h|d) \propto p(d|h)p(h)$ to infer the posterior distribution $p(h|d)$. For this model, the likelihood $p(d \mid h)$ for segmented sequences that are consistent with the observed data is 1, and 0 otherwise. Intuitively, this means that we only need to consider segmented sequences that could be produced from the observed unsegmented motions, and that the posterior probability of these sequences is proportional to their prior probability $p(h|d) \propto p(h)$, the probability that our model would generate this sequence of actions.

We can estimate $p(h|d)$ by iteratively considering each potential boundary position in the action sequence one at a time, sampling the probability of segmenting at this spot from its conditional posterior distribution, while holding all other segment boundaries constant. To do this, we can use a standard Markov chain Monte Carlo method known as Gibbs sampling (Gilks, Richardson, & Spiegelhalter, 1996). The key property of a Gibbs sampler is that it converges to the posterior distribution, allowing us to sample segmentation hypotheses from $p(h|d)$. We can also use Gibbs sampling to infer the posterior distribution over causal relationships between actions and events. In this case a causal structure hypothesis h consists of values c_w for all the actions found in the inferred segmentation. See Appendix A.1 for more details, as well as Goldwater et al. (2009).

For all simulations described in this work, we ran three randomly seeded Gibbs samplers for 20,000 iterations. We then averaged results from 10 samples drawn from the last 1,000 iterations of each sampler, to estimate the posterior distributions and evaluate the model. To aid with convergence, we used a method known as *simulated annealing*, described in more detail in Appendix A.1. Each sample corresponds to one possible segmentation of the corpus into actions (one segmentation hypothesis sampled from $p(h \mid d)$ under the model), and includes not only the proposed boundaries but also implicitly includes a proposed action lexicon from which the sequence was composed. For causal inference, each action type in the sample's lexicon is also assigned a causal value (for additional details see Appendix A.1).

Predictions for Human Segmentation

The key intuition behind this model is that action segmentation and causal structure are jointly learned, taking advantage of statistical evidence in both domains. The model assumes that the same underlying process generates human actions and causal motion sequences, implicitly capturing that actions are being chosen intentionally, often to bring about causal outcomes. This means that action segmentation and causal structure are inferred simultaneously and interdependently. Sequences of motion that correspond to known actions are considered more likely to be causes, and sequences of motion that appear to be causal (they predict outcomes in the world) are considered more likely to be actions. The inferred action boundaries help determine the inferred causal structure and vice versa. This corresponds to our hypothesis that people believe intentional actions and causal effects go hand in hand.

If statistical action structure is in fact a cue to causal relationships then, like our model, people should think statistically grouped actions are more likely to be potential causes than other equivalent sequences. This prediction is tested in Experiment 1. Second, if people believe that causal sequences of motion are also likely to be actions, they should be able to identify and segment out causal sequences, and should find those sequences to be more meaningful and coherent than other sequences of motion with equivalent statistical regularities. This prediction is tested in Experiments 2 and 3. Finally, if action segmentation and

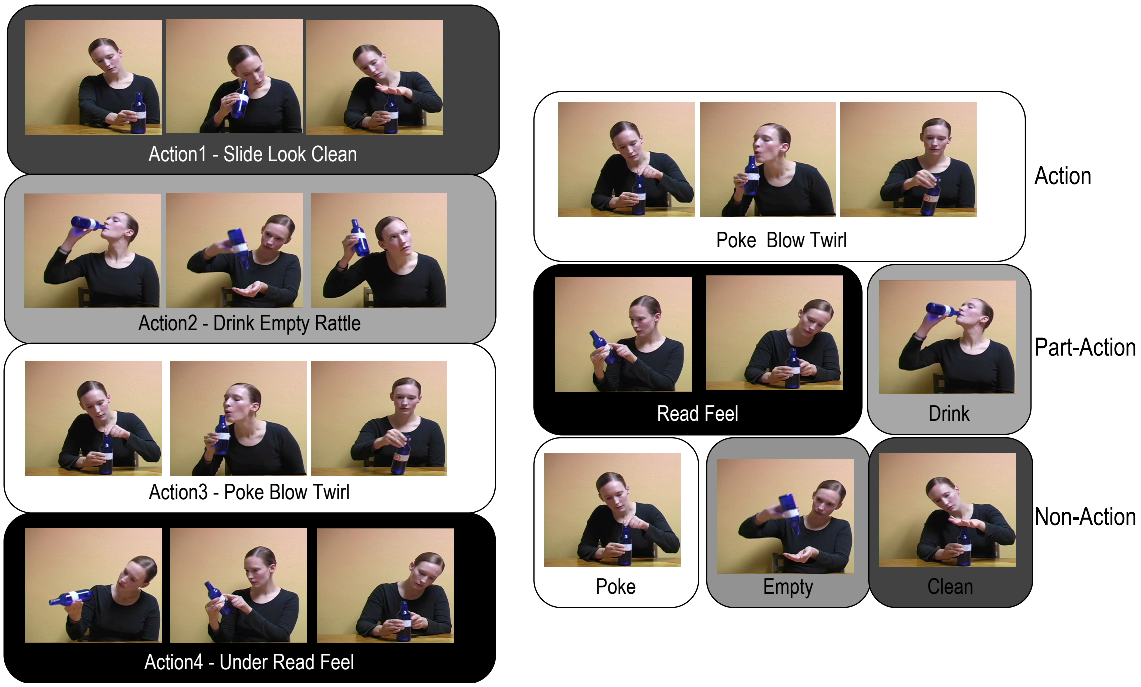

Action1 - Slide Look Clean
Action2 - Drink Empty Rattle
Action3 - Poke Blow Twirl
Action4 - Under Read Feel
Poke Blow Twirl
Action
Read Feel
Drink
Part-Action
Poke
Empty
Clean
Non-Action

Figure 2.2: Left: Four actions composed of three unique small motion elements each were used to create the Experiment 1 exposure corpus. Right: Example Action, Part-Action and Non-Action.

causal relationships are truly jointly learned, then we should see cue combination and cue conflict effects emerge, as in other cases of joint perceptual inference (e.g., Ernst & Banks, 2002). This prediction is tested in Expement 4.

2.4 Experiment 1: Using Statistical Cues

Just as people can recognize words from an artificial language, and distinguish them from non-words and part-words, we also know that they can recognize artificial actions grouped only by statistical relationships and can distinguish these sequences from non-actions (motions that never appeared together) and part-actions (motion sequences that cross an action boundary) (Baldwin et al., 2008). We hypothesized that participants would judge artificial *actions* to be more coherent and meaningful than similar *non-action* and *part-action* sequences (see Figure 2.2), and would also view *actions* as more likely to cause a (hidden) effect than *non-actions* and *part-actions*.

We tested these predictions using sequences of motion generated from “artificial action grammars”, similar to those used in previous action segmentation experiments (Baldwin et al., 2008; Meyer & Baldwin, 2011), and paralleling the designs used in the statistical word segmentation literature (e.g. Saffran, Aslin, & Newport, 1996; Saffran et al., 1997). Just as

Table 2.1: Small motion elements (SMEs) used in Experiments 1, 2, 3 and 4. After Meyer and Baldwin (2011).

SME	Description
Empty	bottle is turned over as if to pour into open hand
Clean	flat hand wipes top of bottle
Under	bottom of bottle is examined
Feel	index finger touches side of bottle in an up-and-down motion
Blow	bottle is lifted to mouth and blown into
Look	bottle is lifted to face and interior examined
Drink	bottle is lifted and tipped into mouth as if drinking
Twirl	bottle edge is lifted from table and spun around
Read	finger traces over label and bottle is lifted from table as if to read
Rattle	bottle is lifted close to ear and shaken
Slide	bottle is pushed forward on hte table and returned
Poke	index finger is inserted and removed from top of bottle

a sentence is composed of words, which are in turn composed of phonemes or syllables, here an action sequence is composed of actions, which are themselves composed of small motion elements (SMEs).

We created two exposure corpora, each assembled by adding "actions" to the sequence one at a time, selecting from four "actions" each made up of three distinct recognizable object-directed motions (see Table 2.1 for a description of the 12 motions used). Each action appears 90 times for a total of 360 actions and 1080 motions. A more detailed description of how the corpora were constructed is given below. The key feature of these corpora is that within an action the transitional probabilities between adjacent motion elements are higher (1.0 in all cases) than between actions (averaging 0.33). In this first experiment, no causal outcomes were added to the corpora. We first ran the model on the exposure corpora to examine its segmentation performance, and then looked at human performance on these same corpora.

Model Simulations

An abstract representation of each unsegmented corpus was used as input to the model, with a letter standing for each SME. For example, the sequence `blow, look, read, twirl, feel, drink, poke, empty, clean` would be represented as `BLRTFDPEC` (see Appendix A.3 for examples of complete corpora). Our model has two free segmentation parameters: $p_{\#}$ which influences action length, and α_0 which influences the number of unique action types. There is also $p_{\$}$, the prior probability of the action sequence terminating. Since the action sequence ends only once, the effect of $p_{\$}$ on segmentation results is negligible – we therefore used a fixed value of $p_\$ = 0.01$ for our simulations.[2]

We evaluated model results across a wide range of parameter values, comparing our results to the true segmentation, and calculated average precision and recall scores across samples, commonly used metrics in the natural language processing literature (e.g., Brent, 1999; Venkataraman, 2001), and which were also used by Goldwater et al. (2009). Precision (P) is the percent of all actions in the produced segmentation that are correct, while recall (R) is the percent of all actions in the true segmentation that were found. In other words, following Brent (1999)

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

These scores are for complete actions, meaning that for an action token to count, both boundaries must be correct. For example, for the true sequence `BLR TFD PEC` a segmentation hypothesis of `BLR TFDPEC` would have $P = 0.5$ because one out of the two actions found appears in the true segmentation, and $R \approx 0.33$, because one out of the three true actions was found. Meanwhile, the hypothesis `BLR TFD P E C` would have $P = 0.4$ because only two out of the five proposed actions appear in the true segmentation, and $R \approx 0.66$, for finding two out of the three true actions.

We ran simulations for $\alpha_0 \in \{1, 2, 5, 10, 20, 50, 100, 200, 500\}$ and $p_\# \in \{0.5, 0.7, 0.90.95, 0.99\}$. Example Recall results are shown in Figure 2.3 (Precision performance very closely mirrored recall performance). In general, segmentation performance improves with smaller values of α_0 that favor a smaller lexicon, and larger values of $p_\#$, favoring shorter actions. This makes sense, given that for these corpora the true lexicon is four actions, each of which is three motions long. For $p_\# = 0.99$ and $\alpha_0 < 20$ the model consistently produces the true segmentation across samples, suggesting that the posterior distribution for these parameter values is highly peaked around the true segmentation.

As α_0 gets bigger and $p_\#$ gets smaller, the model's prior expectations shift towards a larger lexicon, and longer actions. As a result, the model begins to *undersegment*, joining together adjacent actions. For instance if the true segmentation is `BLR TFD PEC` the model might produce `BLRTFD PEC`, treating `BLRTFD` as a single action. This can be seen as analogous to the part-words and part-actions in the statistical segmentation work described earlier – extracting actions that cross a boundary in the true segmentation. Figure 2.4 shows example results for the average probability that an action vs a part-action appears in the lexicon. Notice that though the absolute difference varies with the parameter values, at least within the parameter ranges tested, actions are always more likely to be in the lexicon (in other words, to be considered coherent units) than part-actions.

[2] Preliminary simulations varying the value of $p_\$$ confirmed that the exact value had little influence on the resulting segmentation

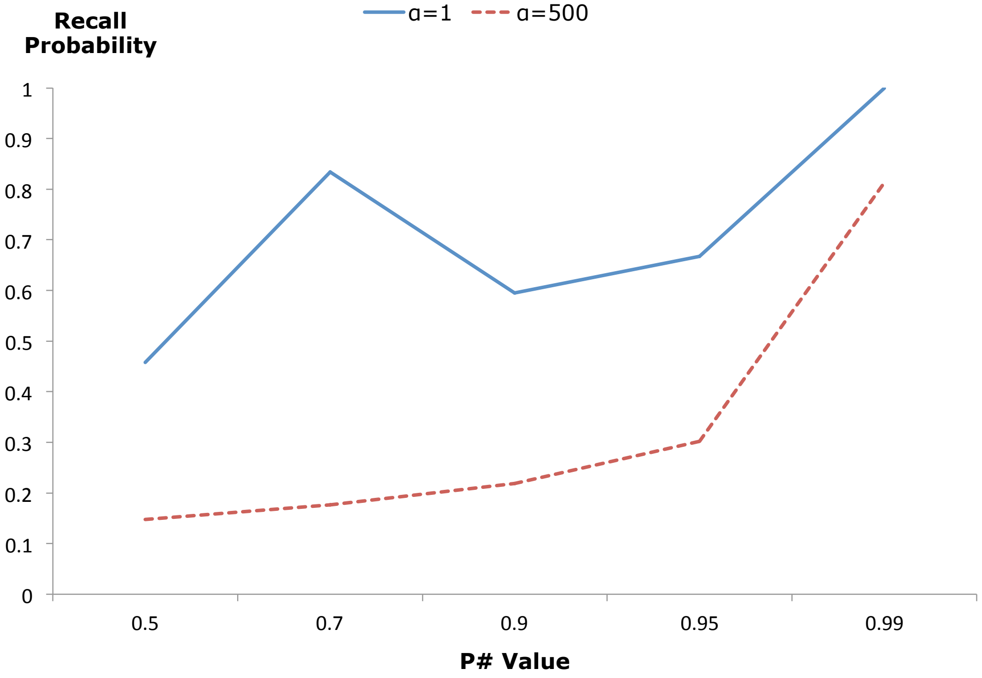


Figure 2.3: Token recall probability. Example results for Experiment 1 model simulations

Finally, while there are no observed effects in these corpora, we can treat the effects as unobserved, and ask which sequences the model thinks would have been likely to lead to effects. In this case, the probability of a sequence leading to an effect reverts to the prior probability under the model, which is $(\pi \cdot \omega) + (1 - \pi)\epsilon$ for actions in the lexicon and ϵ for other sequences. Therefore, when $(\pi \cdot \omega) + (1 - \pi)\epsilon > \epsilon$, the model will predict that actions are more likely to lead to effects than other motion sequences. This inequality will be true as long as $\omega > \epsilon$ – in other words, as long as effects are more likely to follow causes than non-causes. As $\omega >> \epsilon$, actions become increasingly more likely to be causal relative to part-actions and non-actions.

Across a broad range of parameters, our model produces a reasonable segmentation of corpora similar to those used by Baldwin et al. (2008), and modeled after classic statistical word segmentation experiments, and some parameter values consistently produce the true segmented sequence. These results confirm that the sequential probabilities available in the corpus can in principle be used for segmentation, and to distinguish actions from part-action and non-action sequences. The model also suggests that actions are more likely to be causal

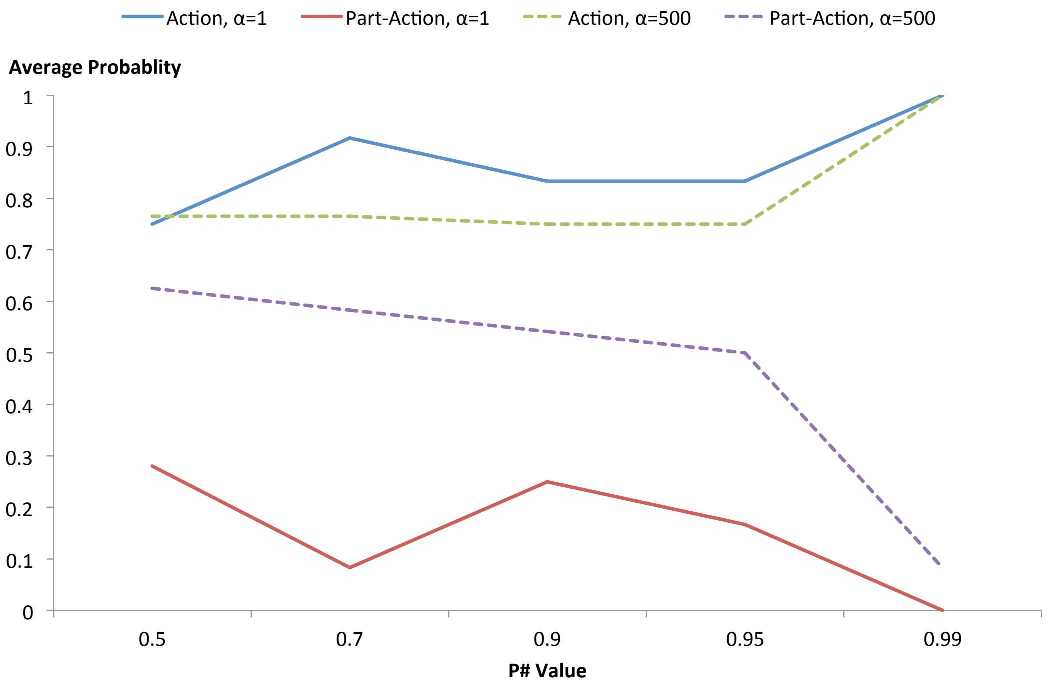


Figure 2.4: Probability of sequence appearing in action lexicon. Example results for Experiment 1 model simulations

than other sequences, even when effects are not observed, and that the true action sequences used to construct the corpus are more likely to be identified as coherent units of motion that appear in the action lexicon than sequences that cross an action boundary in the true segmentation.

Method

Participants

Participants were 100 U.C Berkeley undergraduate students, who received course credit for participating. Participants were randomly assigned to view one of the two exposure corpora, and were also randomly assigned to one of three follow-up question conditions. All participants were instructed to attend closely to the exposure corpus, and were told that they would be asked questions about it later. Thirty participants were assigned to each of the first two conditions and 40 participants were assigned to the last condition.

Stimuli

Similar to Baldwin et al. (2008), we used 12 individual video clips of object-directed motions (referred to as *small motion elements* or SMEs in the previous work), to create four *actions* composed of three SMEs each (see Figure 2.2). The SMEs in this experiment are identical to those in Meyer and Baldwin (2011). As in previous work, SMEs were sped up slightly and transitions were smoothed using iMovie HD, to make the exposure corpus appear more continuous.

We created a 25 minute exposure corpus by randomly choosing actions to add to the sequence, with the condition that no action follow itself, and that all actions and transitions between actions appear an equal number of times, resulting in 90 appearances of each action and 30 appearances of each transition. We also created four *non-action* and four *part-action* comparison stimuli, where a non-action is a combination of three SMEs that never appear together in the exposure corpus, and a part-action is a combination of three SMEs that appears across a transition (e.g. the last two SMEs from the first action and the first SME from the second action, see Figure 2.2). Finally, to ensure that none of our randomly assembled actions were inherently more causal or meaningful, we created a second exposure corpus, using the non-action SME combinations of the first corpus as the actions of the second corpus.

Procedure

Following the exposure corpus, participants in the *familiarity condition* were presented with all 12 actions, non-actions and part-actions individually, and asked "How familiar is this action sequence?". They responded by choosing a value on a 1 to 7 Likert scale, with 1 representing "not familiar" and 7 representing "very familiar" (other than the use of ratings instead of a forced choice format, this condition is almost identical to Baldwin et al., 2008). In the *causal condition*, participants were given a "hidden effect" cover story before viewing the exposure corpus. These participants were told that certain actions would cause the bottle being manipulated to play music, but that they would be watching the video with the sound off. Following the exposure corpus, these participants were asked "How likely is this sequence to make the bottle play a musical sound?", with 1 representing "not likely" and 7 representing "most likely". Finally, in the *coherence condition*, participants were asked the question "how well does this action sequence go together?". They were given the example of removing a pen cap and then writing with the pen as "going together" and of removing a pen cap and then tying your shoes as "not going together". They then rated all test items on a scale with 1 being "does not go together" and 7 being "goes together well".

For all conditions, we used a custom Java program to present video of action sequences and collect ratings. The program presented all 12 actions, non-actions and part-actions individually and in a random order.

Results

We analyzed all results using 2×3 ANOVAs on exposure corpus (1 or 2) and sequence type (action, non-action, part-action). No effects of exposure corpus were found. Results are shown in Figure 2.5.

Ratings from 27 participants in the *familiarity condition* were analyzed (data from three additional participants who rated all sequences identically as either a 1 or 7 was discarded). As predicted by previous results (Baldwin et al., 2008; Meyer & Baldwin, 2011), there was an overall significant effect of sequence type $F(2,\ 50)=$ 25.14, *MSE*= 41.12, $p < 0.0001$, with actions rated significantly more familiar than part-actions and non-actions $t(26)=$ 5.84, $p < 0.0001$, one sample t-test on contrast values, and part-actions rated significantly more familiar than non-actions $t(26)=$ 3.65, $p < 0.01$.

Ratings from 29 participants in the *causal condition* were analyzed (data from one additional participant was discarded). As predicted, there was an overall significant effect of sequence type $F(2{,}54)=$ 10.20, *MSE*= 12.87, $p < 0.001$, with actions rated as significantly more likely to cause a musical effect than part-actions or non-actions $t(28)=$ 2.36, $p < 0.01$, one sample t-test on contrast values, and part-actions rated significantly more likely to be causal than non-actions, $t(28) = 2.36$, $p < 0.05$.

Ratings from 37 participants in the *coherence condition* were analyzed (data from an additional three participants was discarded). As predicted, there was an overall significant effect of sequence type $F(2{,}70)=$ 9.18, *MSE*= 14.47, $p < 0.001$, with actions rated as going together significantly better than part-actions or non-actions $t(36)=$ 3.87, $p < 0.001$, one sample t-test on contrast values. There was also a marginally significant difference between part-action and non-action ratings $t(36)=$ 2.0, $p = 0.05$.

Discussion

The results of this experiment support the hypothesis that people experience sequences of action grouped only by their statistical regularities as casually significant, meaningful groupings. Participants rated actions as more likely to cause a hidden musical effect than part-action and non-action sequences, even though all sequences were equally arbitrary, and in fact the non-actions for one exposure corpus were the actions for the other, meaning that the same sequences reversed their rating merely based on the number of times the SMEs appeared together. Similarly, participants rated actions as going together (a question we used as a measure of sequence coherence and meaningfulness) significantly better than other sequences. Anecdotally, a number of participants reported a feeling that the action sequences made more intuitive sense to them than the other sequences. Finally, all three conditions replicated the finding by Baldwin et al. (2008) that adults are able to parse statistically grouped actions from within a longer action sequence, and differentiate them from other non-action groupings, and confirmed the use of ratings as a viable alternative measure to forced choice comparisons.

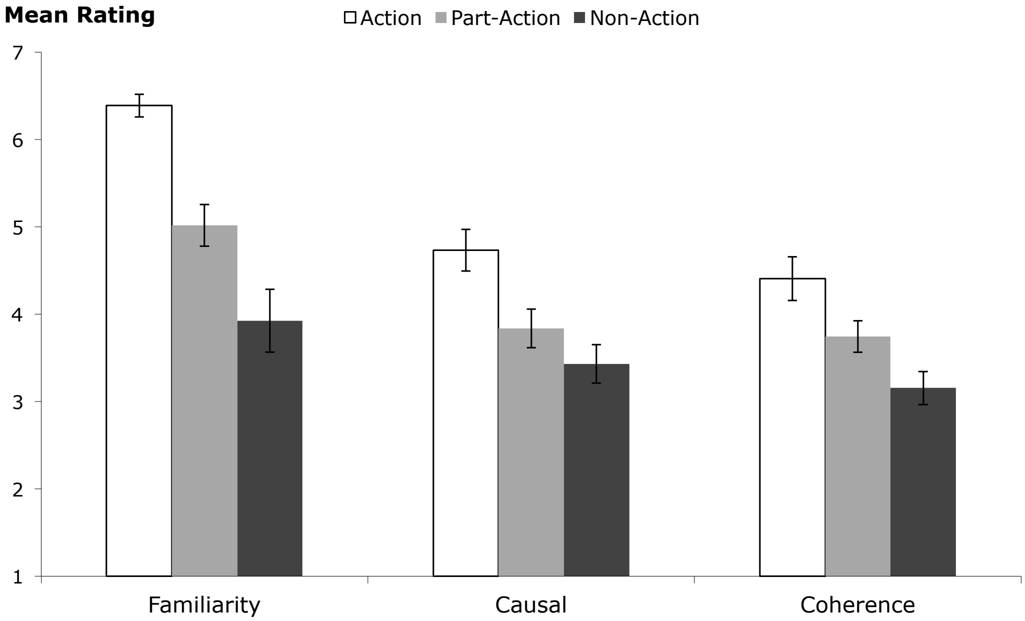


Figure 2.5: Results of Experiment 1. Error bars show one standard error.

These results have several important implications. First, they demonstrate that people's sensitivity to the statistical patterns in the exposure corpus is not simply an artifact of the impoverished stimuli, but appears to play a real role in their subsequent understanding of the intentional structure of the action sequence. The fact that participants found the statistically grouped actions to be more coherent suggests that they do not experience the sequences they segment out as arbitrary, but assume that they are meaningful groupings that play some (possibly intentional) role. This is further supported by the results from the causal condition which show that, even without being presented with overt causal structure, people believe the statistically grouped actions are more likely to lead to external effects in the world.

Finally, these results also support our hypothesis that inference of action structure and causal structure are linked, with statistically grouped actions being perceived as more likely to also be causal variables. This result is consistent with our computational model, which also predicts that, without other evidence of causal structure, actions are more likely to be causal than non-action and part-action sequences.

2.5 Experiment 2: Using Causal Structure

Our first experiment suggests that people can use statistical action structure to infer causal relationships, but can they use causal relationships to identify meaningful actions? Our second experiment investigated whether people are able to pick out causal subsequences from within a longer stream of actions, and whether they use this causal information to inform their action segmentation. Specifically, we hypothesized that when statistical cues to action segmentation are unavailable, adults will be able to use causal event structure to identify meaningful units of action.

To test these predictions, we constructed two new exposure corpora. In these corpora, there were no *a priori*, statistically-grounded actions. Instead, each exposure corpus was assembled using four SMEs, so that each individual SME would be seen an equal number of times, and all possible length three sequences of SMEs would also occur with equal frequency.[3] Throughout the exposure corpus, no length three subsequences containing repeats of an SME were allowed to occur. This resulted in 24 possible three-motion sequences ("triplets"). A target triplet of SMEs was then randomly chosen as the "cause". Whenever this sequence of motions was performed in the exposure corpus, it was followed by an observable causal outcome (see Methods section below for more details). We first ran the model on the exposure corpora to examine its segmentation and causal inference performance, and then looked at human performance on these same corpora.

Model Simulations

As in Experiment 1, an abstract representation of each unsegmented corpus was used as input to the model (See Appendix A.3 for an example corpus), with a letter standing for each SME, and "*" representing a causal outcome. The model has three causal inference parameters: π the probability that an action is causal, ω the probability that a causal sequence leads to an effect, and ϵ the probability of an effect following a non-causal sequence. However, as noted earlier, what is relevant for causal inference is the ratio of ω to ϵ. Therefore we use a fixed value of $\epsilon = 0.001$, and vary only ω. We also maintained the relationship $\omega >> \epsilon$, reflecting our assumption that causes are relatively effective, and are much more likely to precede effects than are non-causes. Following the results of Experiment 1, we used segmentation parameter values $\alpha_0 = 3$ and $p_\# = 0.99$ throughout[4] and evaluated model results across values of $\omega \in \{0.5, 0.7, 0.9, 0.99\}$ and $\pi \in \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$.

For Experiment 2, we were interested in seeing whether the model could infer the correct causal subsequence from within the longer sequence of motions (since there was no "correct" segmentation for the remainder of the corpus, overall segmentation performance is not evaluated here). For all parameter values tested the model performed extremely well,

[3] in fact, it turns out that all length two sequences appear with approximately equal frequency as well, so that the transitional probability between any two motions is ≈ 0.33

[4] Additional simulations confirmed that causal inference results were not significantly affected by changing the segmentation parameters

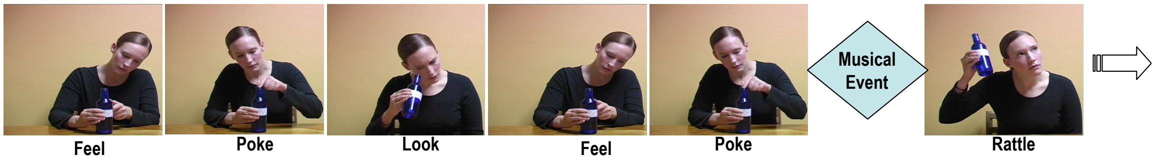


Figure 2.6: A portion of the Experiment 2 exposure corpus. four SMEs (Poke, Look, Feel, Rattle) are distributed so that all possible triplets appear equally often. A target triplet (Look, Feel Poke) is chosen to cause a sound.

correctly identifying the target triplet as causal across all samples and parameter values, $p(c_{target} = 1) = 1$. Additionally, for all parameter values, the model correctly segmented at least 99% of the occurances of the target sequence. In contrast, other triplets averaged between 6-16%. Similarly, the target triplet appeared in the lexicon across all samples, while other triplets appeared on average between 8% and 18% of the time.

We ran our model on corpora designed so that all possible three-motion sequences occurred equally often. Across parameter values, our model consistently identifies the correct causal sequence and segments out this sequence as an action. This suggests that, even when the transitional probabilities between motions are uniform, it is possible to identify and extract the correct length causal sequence, using only the causal statistics in the corpora.

Method

Stimuli

The structure and stimuli for this experiment closely matched that of Experiment 1. However, in Experiment 2, the corpora were specially constructed so that all possible combinations of three motions appeared equally often together, so that joint and transitional probabilities could not be used to identify groupings (see Figure 2.6). Throughout the exposure corpus, no length three subsequences containing repeats of an SME were allowed to occur. This resulted in 24 possible SME triplets. A target triplet of SMEs was then randomly chosen as the “cause”. Whenever this sequence of motions was performed in the exposure corpus, it was followed by the object playing music (participants were able to hear the music, unlike in Experiment 1).

The exposure corpus was created by first generating 24 shorter video clips. Each clip was designed to have a uniform distribution of both individual SMEs and of SME triplets. Specifically, in each clip, the four unique SMEs appear exactly six times each, and 23 of the 24 possible SME triplets appear exactly once each. We designed the 24 clips by using a De Bruijn sequence (van Aardenne-Ehrenfest & de Bruijn, 1951), a cyclical sequence within which each subsequence of length n appears exactly once as a consecutive sequence (see Appendix A.2 for algorithmic details). These 24 video clips were shown consecutively in the exposure corpus, but were clearly separated from each other by text notifying the participant of the beginning and end of each shorter clip. The result was an exposure corpus composed of 24 short video clips, with each SME appearing 144 times throughout the complete corpus, and each triplet appearing 20 to 24 times.

iMovie HD was used to assemble the exposure corpus and add a cartoon sound effect following every appearance of the target sequence. Two different exposure corpora, each using a distinct set of four SMEs were created. `Look, Poke, Feel` and `Rattle` were used to create the first exposure corpus, with `Look Feel Poke` being the target triplet, and `Read, Slide, Blow` and `Empty` were used to create the second exposure corpus, with `Slide Blow Empty` being the target triplet.

Participants and Procedure

Participants were 100 U.C Berkeley undergraduates. Participants were divided into the same three conditions as in Experiment 1, with the difference that after viewing the exposure corpus, they rated all 24 possible SME triplets, and that all participants were told that certain action sequences caused the bottle to play music.

Results

We analyzed all results using 2×2 ANOVAs on exposure corpus (1 or 2) and sequence type (target, other). No effects of exposure corpus were found. Results are shown in Figure 2.7.

Ratings from 28 participants in the *familiarity condition* were analyzed (data from an additional two participants who rated all sequences identically as either a 1 or 7 was discarded). Contrary to our predictions, and the predictions of previous work, there was no effect of sequence type $F(1,26) = 1.58$, $MSE = 1.74$, $p > 0.22$. Participants rated the target sequence and the other SME triplets as equally familiar.

Ratings from 30 participants in the *causal condition* were analyzed. As predicted, there was a significant effect of sequence type $F(1,28) = 193.97$, $MSE = 310.439$, $p < 0.0001$, with the target sequence being rated as much more likely to lead to a musical sound than the other SME triplets.

Ratings from 35 participants in the *coherence condition* were analyzed (data from five additional participants was discarded). As predicted, there was a significant effect of sequence type $F(1,33) = 19.44$, $MSE = 47.1$, $p < 0.0001$, with the target sequence rated as going together significantly better than the other SME triplets.

Discussion

This experiment is one of the first to demonstrate that people can infer a correctly ordered set of causal variables from within a longer temporal sequence. In fact, the results of this experiment suggest that it was a relatively easy task for participants. Participants in the causal condition were nearly at ceiling in their ratings of how likely sequences were to lead to a musical effect, with the target sequence having a mean rating only slightly below 7 and the remaining sequences being rated a bit below 2.

The results of this experiment also provide further support for a relationship between action segmentation and causal inference. Even though there were no statistically grouped actions in this experiment, participants still perceived the target sequence as being more meaningful (going together better) than the other sequences, suggesting they had nonetheless segmented it out as a coherent action unit. It is worth noting that the ratings for the coherence question were different than those for the causal question, suggesting that participants did interpret the question as one of meaningfulness, rather than an alternate phrasing of the causality question.

Finally, it is interesting to note that, despite correctly identifying the target sequence as causal, participants *did not* rate it as more familiar than the other sequences. Instead, participants appeared to be aware that they had seen all the sequences an equal number of times, and rated them all as equally familiar. This implies that participants are not judging the target sequence as more coherent or more likely to be causal due to some sort of low level saliency effect that causes them to remember this particular sequence more clearly. It also suggests that participants, at least in this context, interpret the familiarity question as a question about frequency of appearance. These results suggest that participants may be aware that certain sequences are more causal or more coherent, while also being aware that they have seen other sequences equally often.

2.6 Experiment 3: Identifying Correct Causal Subsequences

Our second experiment found that people are able to pick out causal subsequences from within a longer stream of actions. However, in Experiment 2 we only asked participants to rate actions composed of three SMEs each. This potentially leaves open the question of whether, like our model, participants are really identifying the correct causal subsequence, or whether they might actually prefer a subsequence or supersequence of the causal actions if given the choice. In this experiment, we explicitly look at whether participants are able to identify causal subsequences of the correct length.

Method

Stimuli

The structure and stimuli for this experiment closely matched those of Experiment 2. The same exposure corpora were used as in Experiment 2. As in Experiment 2, `Look Feel Poke` was the target triplet in the first exposure corpus, and `Slide Blow Empty` was the target triplet in the second corpus. In both corpora, the target triplet was always followed by a cartoon sound effect.

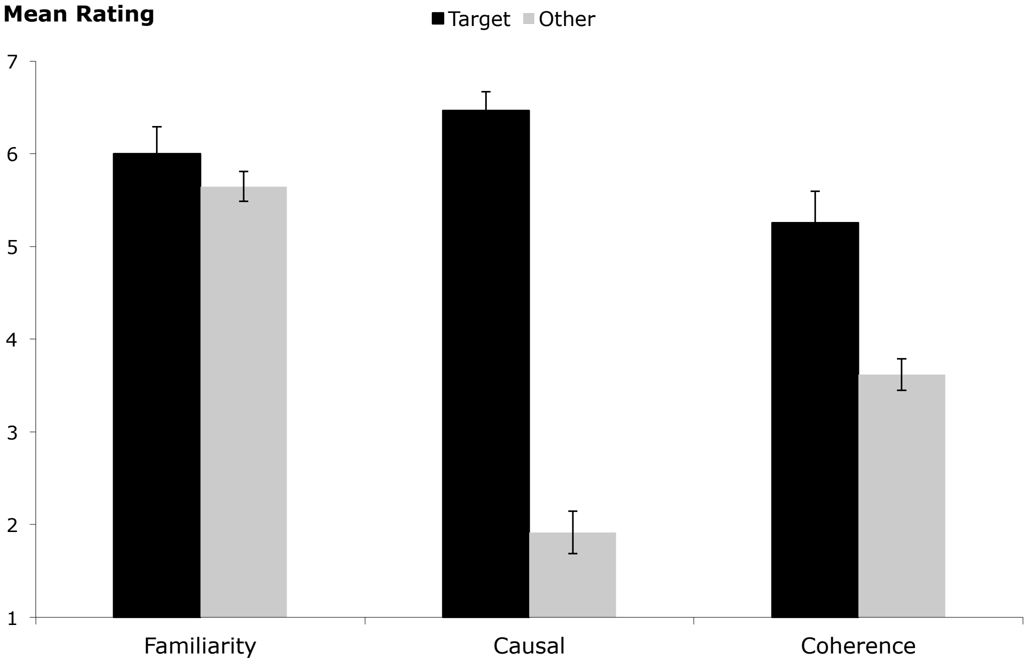


Figure 2.7: Results of Experiment 2. Error bars show one standard error.

A series of 30 test stimuli were created for each exposure corpus. Each test stimulus was an action composed of between 1 and 5 SMEs. For each corpus, the set of test stimuli was constructed so that it contained the target action, single and double length subsequences of the target action (e.g. for `Slide Blow Empty` these would be `Empty` and `Blow Empty`), and quadruplet and quintuplet length supersequences of the target action (e.g., `Read Slide Blow Empty` and `Empty Read Slide Blow Empty`). There were also non-target actions of each of these lengths, some of which contained subsequences of the target action (e.g., `Read Blow Empty`) and others that did not (e.g., `Empty Slide Read`). See Figure 2.8 for a complete set of test sequences for one corpus. iMovie HD was used to assemble the test stimuli, as in Experiments 1 and 2.

Participants and Procedure

Participants were 53 U.C Berkeley undergraduates. Participants were randomly assigned to view one of the two exposure corpora. All participants were given viewing instructions

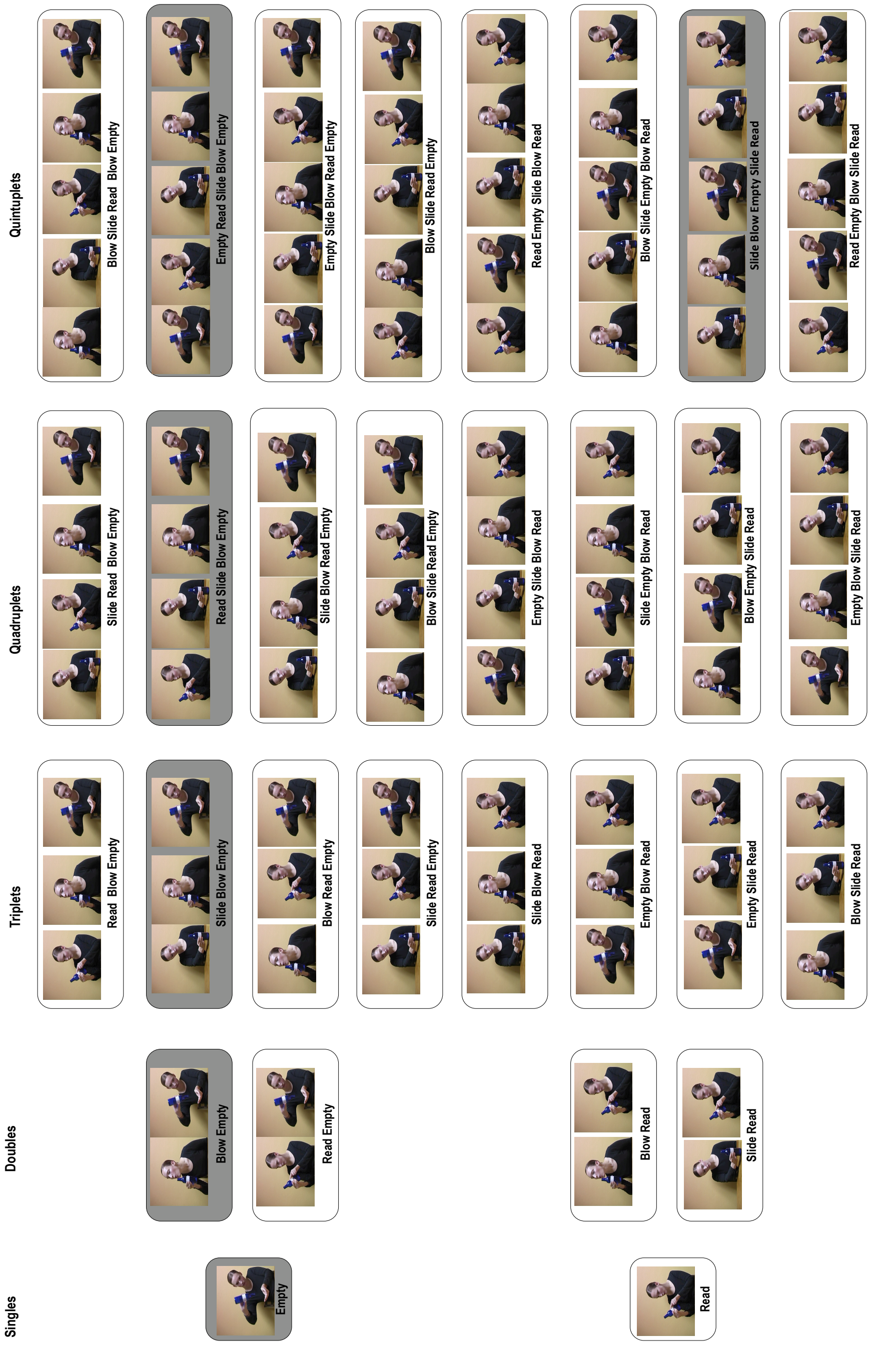


Figure 2.8: The 30 test stimuli for Experiment 3, exposure corpus 1. Test stimuli include both subsequences and supersequences of the target action (`Slide Blow Empty`). These are shown with a dark background. Comparison sequences of equal length are shown with a light background.

identical to those in the causal condition of Experiment 2. After viewing the exposure corpus, participants were asked “How likely is just this sequence to make the bottle play a musical sound?”, for all 30 test sequences. As in the causal condition of Experiments 1 and 2, participants responded by choosing a value on a 1 to 7 Likert scale, with 1 representing “not likely” and 7 representing “very likely”,

Results

We analyzed all results using 2×2 ANOVAs on exposure corpus (1 or 2) and sequence type (contains target, other). No effects of exposure corpus were found. Results are shown in Figure 2.9.

Ratings from all 53 participants were analyzed. There was a significant effect of sequence type $F(1{,}51)= 248.61$, $MSE= 1714.12$, $p < 0.0001$, with sequences containing the target being rated as much more likely to lead to a musical sound than other sequences.

We compared ratings of each sequence containing the target to the non-target sequences of equivalent length, using one sample t-tests on contrast values. In all cases, the target-containing sequences were rated as significantly more causal than the non-target sequences of the same length. Triplet sequences: $t(52)= 13.60$, $p < 0.0001$. Quadruplet sequences: $t(52)= 14.46$, $p < 0.0001$. Quintuplet sequences: $t(52)= 12.85$, $p < 0.0001$.

We compared ratings of single and double length terminal subsequences of the target to other single and double length sequences, using one sample t-tests on contrast values. Subsequences of the target were rated as significantly more causal than other sequences of equivalent length. Single SME sequences: $t(52)= 3.64$, $p < 0.001$. Double sequences: $t(52)= 5.96$, $p < 0.0001$.

Finally, we compared ratings of the target triplet to subsequences and supersequences of the target, using one sample t-tests on contrast values. Subsequences were rated as significantly less causal that the target triplet itself. Single SME subsequence of target: $t(52)= 15.08$, $p < 0.0001$. Double subsequence of target: $t(52)= 14.80$, $p < 0.0001$. Finally, the target sequence and supersequences of the target were rated as equally causal. Quadruplet sequences: $t(52)= 1.19$, $p > 0.24$. Quintuplet sequences: $t(52)= -0.32$, $p > 0.74$.

Discussion

The results of Experiment 3 confirm that adults can identify the correct sequence of causal motions from within a longer sequence. Participants correctly rated only sequences containing the complete target triplet as very likely to lead to the effect. They were able to distinguish these causal sequences from sequences containing only a subset of the necessary sequence, and from sequences containing all of the same motions, but in an incorrect order.

Participants clearly distinguished between subsequences of the causal triplet and the complete causal sequence. Interestingly, they also rated subsequences of the causal triplet as more likely to be causal thant other sequences of equivalent length, perhaps suggesting that

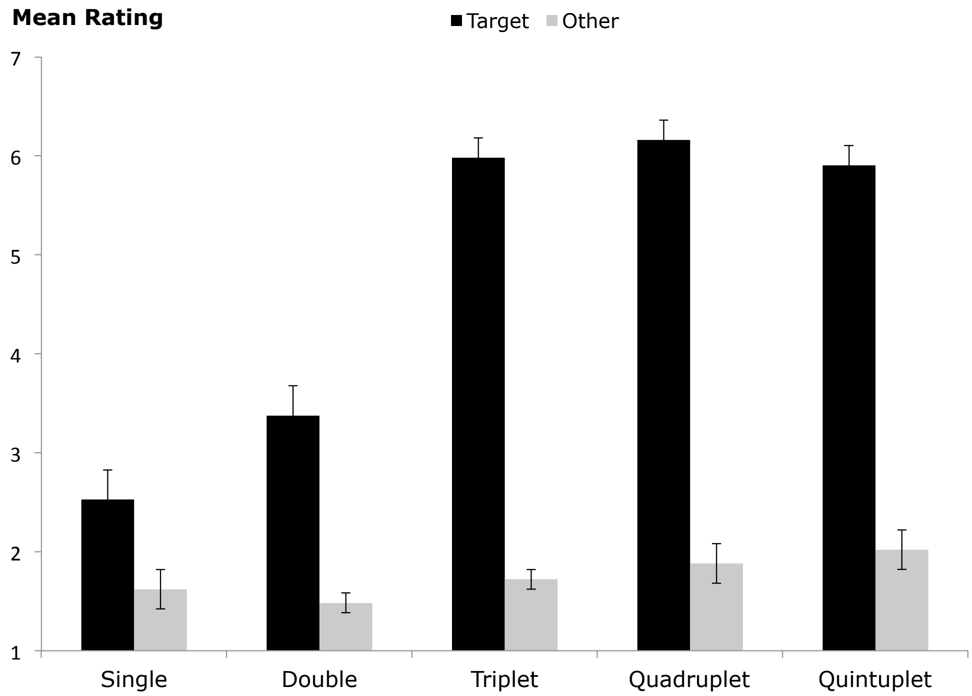


Figure 2.9: Results of Experiment 3. Error bars show one standard error.

participants recognized these subsequences as necessary but not sufficient for producing the effect.

2.7 Experiment 4: Joint Inference of Causal and Statistical Action Structure

Experiments 1, 2 and 3 together suggest that adults can use statistical structure in action sequences to inform their causal inferences, and can use causal relationships between actions and external outcomes to help determine action structure. In particular, they demonstrate that adults believe that action sequences with higher internal transition probabilities are good candidate causes, and that sequences of motions that predict outcomes in the world are likely to group together, independent of their transition probabilities.

However, our previous experiments all examined causal and statistical structure in action sequences separately. If people, like our model, jointly infer causal relationships and action structure, then when the two types of cues are both present in an action stream, both of them should influence people's judgments of action segmentation and of causal relationships.

In Experiment 4 we test whether causal structure and action structure are jointly inferred, by generating a corpus in which statistical and causal cues are both present, and are in conflict. As in Experiment 1, there are four statistically determined actions, but in Experiment 4 there is also a target part-action, that consistently leads to the object playing a musical sound. We first ran the model on the exposure corpora to examine its segmentation and causal inference performance, and then looked at human performance on these same corpora.

Experiment 4: Model Simulations

The corpora used in these simulations were exactly the same as those used in Experiment 1, but with an effect added after every occurrence of the target part-action (see Appendix A.3).[5] As before, an abstract representation of each unsegmented corpus was used as input to the model, with a letter standing for each SME.

We compared our model's joint segmentation and causal inference results to those generated by performing causal inference and action segmentation separately, to see whether joint inference produced distinct results, and whether those results were an improvement relative to those generated by the separate inference processes.

Joint Inference Results

We ran the model on segmentation parameter ranges $\alpha_0 \in \{1, 2, 5, 10, 20, 50, 100, 200, 500\}$ and $p_{\#} \in \{0.5, 0.7, 0.90.95, 0.99\}$, as in Experiment 1. Since the causal parameters π and ω had relatively little impact in Experiment 2, we used a slightly coarser grid to reduce computation time: $\omega \in \{0.5, 0.7, 0.9, 0.99\}$ and $\pi \in \{0.01, 0.05, 0.1, 0.3, 0.5\}$. Preliminary analysis indicated that results were again not significantly influenced by the particular choice of causal parameters, so we present result collapsed across values of π and ω.[6]

Overall, the model succeeded at identifying the true actions, which appeared in the lexicon with a probability ranging from 0.75 to 0.99 across parameter values. Similarly, the model consistently identified the target sequence, which appeared in the lexicon with probability ranging from 0.7 to 1.[7] In contrast, sequences containing part-actions overall appeared in the lexicon with probability ranging from 0.18 to 0.58.

Whenever the target sequence appears in the action lexicon, the model also correctly identifies it as causal. In contrast, other sequences were identified as causal relatively infre-

[5] Note that this means that if, for example, the actions are TFD and BLR, and an effect is added so that we see TFDBL*R every time these two actions occur together, then treating the causal sequence as any of TFDBL, FDBL and DBL is equally valid for this corpus, since they will all always co-occur with the effect.

[6] Collapsing in this way gives us $p(h \mid d) = \sum_{\omega} \sum_{\pi} p(h \mid d, \omega, \pi) p(\omega, \pi)$, so that we are still sampling from the posterior distribution, with a uniform prior $p(\omega, \pi)$ over the sampled causal parameter values

[7] In this case, the model always identifies the longer sequence, such as TFDBL as the causal sequence

quently, with probability between 0 and 0.55. These sequences were always subsequences of the target sequence.

The tendency to consistently identify the original actions vs the target part-action was somewhat in conflict across parameter values, with higher values of $p_{\#}$ leading to actions being more likely to appear in the lexicon, and lower values leading to the target part-action being more likely. This is in part because at higher values of $p_{\#}$ the model is more likely to divide the target sequence back into its component actions, reflecting the tension between the causal and statistical segmentation cues in the corpus.

Finally, while it is not inherently clear what the "true" segmentation should be for this corpus, given the conflicting cues, we can nevertheless try to gauge Precision and Recall performance by comparing to a "compromise" segmentation that identifies all occurances of the target sequence, while otherwise maintaining the "correct" statistical segmentation (see Appendix A.3 for an example). Here, both Precision and Recall scores improve as $p_{\#}$ increases and as α decreases, with $P = 0.47, R = 0.30$ for $p_{\#} = 0.5, \alpha = 500$, and $P = 0.93, R = 0.93$ for $p_{\#} = 0.99, \alpha = 1$. In general, as in Experiment 1, performance for $p_{\#} = 0.99$ was relatively high, with many simulations reproducing the compromise segmentation exactly. Example Recall performance for the complete model is shown in Figure 2.10.

Statistical Segmentation Only Results

Running the model on these corpora without causal inference is equivalent to running it on the Experiment 1 corpora, where there was no causal information present. We again ran this model on paramter values $\alpha_0 \in \{1, 2, 5, 10, 20, 50, 100, 200, 500\}$ and $p_{\#} \in \{0.5, 0.7, 0.90.95, 0.99\}$. As in the results for Experiment 1, when $p_{\#} = 0.99$ and $\alpha < 20$, the model consistently segments out the actions used to construct the corpus. However, for these parameter values, the model never segments out the target part-action sequence, or actions containing this sequence.

As $p_{\#}$ decreases and α increases, the number of undersegmentations appearing in the lexicon increases, including sequences containing the target part-action, as well as those containing other part-actions. However, this accompanies a drop in the number of original actions segmented. Additionally, though sequences containing the target increase, the target is never segmented out exactly. By definition, this model also never identifies the target sequence as causal.

We can again look at Precision and Recall performance relative to the "compromise" segmentation. As in the complete model, both Precision and Recall scores improve as $p_{\#}$ increases and as α decreases, with $P = 0.24, R = 0.14$ for $p_{\#} = 0.5, \alpha = 500$, and $P = 0.83, R = 0.83$ for $p_{\#} = 0.99, \alpha = 1$. The relatively good performance at the high end is driven by the correct segmentation of the original actions, which still make up the majority of tokens in the corpus. However, performance at all parameter values is below that of the complete model. A comparison of the statistical segmentation only model and the complete model is shown in Figure 2.11.

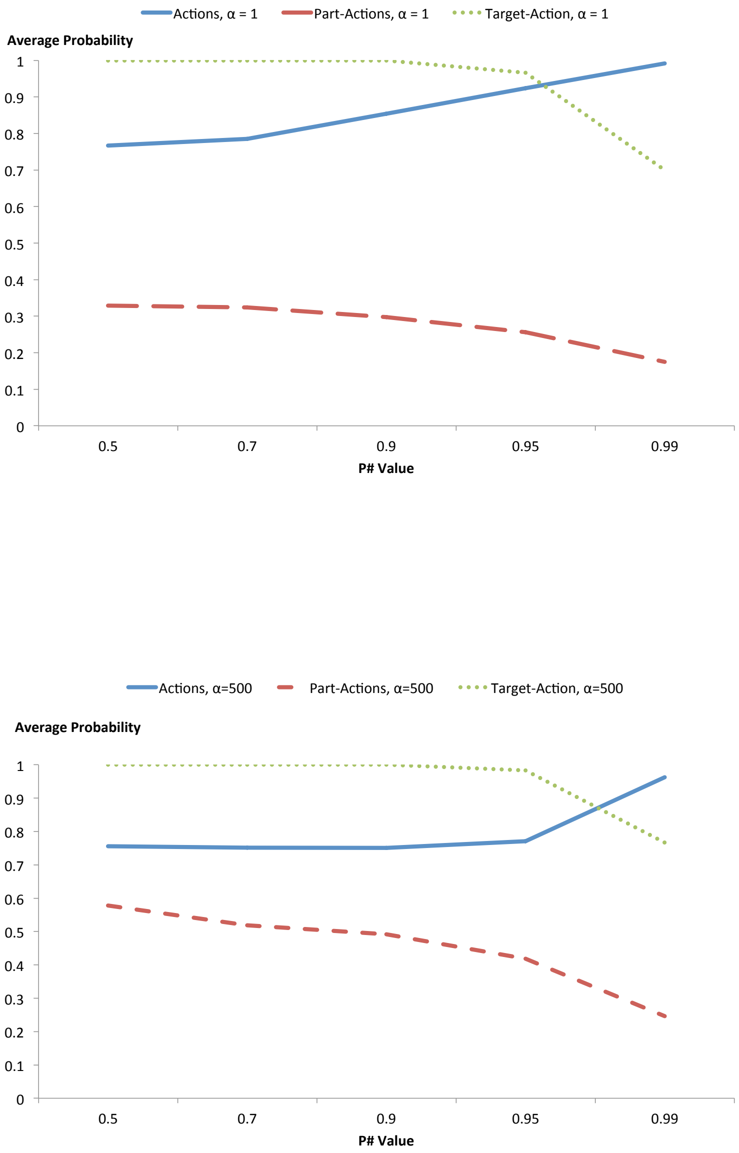


Figure 2.10: Probability of sequence appearing in action lexicon. Example results for Experiment 4 joint inference model simulations

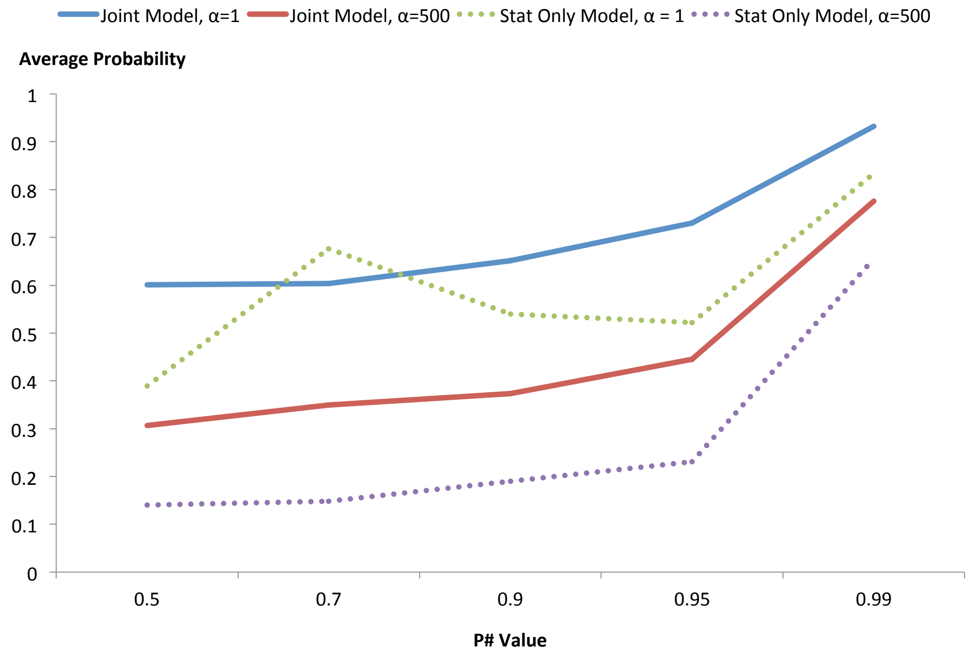


Figure 2.11: Token recall probability. Example results for Experiment 4 model simulations, joint inference model vs statistics only model

Causal Inference Only Results

To simulate causal-only inference, we changed the inference procedure so that boundary decisions were made based only on how likely an event (or lack there of) occurring at the boundary is. This corresponds to a model where actions of any length are equally likely, and each action is added to the sequence independent of the preceding actions, but some actions are causal. We tested this model for parameter values $\omega \in \{0.5, 0.7, 0.9, 0.99\}$ and $\pi \in \{0.01, 0.05, 0.1, 0.3, 0.5\}$.

For this model, the target sequence almost always appears in the action lexicon, with probability ranging from 0.66 to 1 across parameters, and is identified as causal whenever it appears in the lexicon. However, while this means that the target sequence is almost always segmented out correctly at least once per corpus, it is not consistently segmented. Instead, subsequences of the target sequence are also identified as causal in all samples, leading to inconsistent segmentation of the target sequence within each sample. In fact, while the target sequence (and effect) occurs 30 times per corpus, it was correctly segmented on average only 3.45 to 6.48 times per sample. This is because in this model there is no pressure to create consistent action sequences. Not surprisingly, overall segmentation performance is also poor for this model, with Precision approxmiately 0.06 across parameter values, and Recall approximately 0.10.

Summary

When presented with a corpus where statistical and causal cues to segmentation conflict, our joint inference model makes distinct segmentation and causal inference judgments, when compared to models that look only at transitional probabilities, or only at causal relationships. The joint model also outperforms these models in terms of successfully identifying both the target causal sequence, and the statistically determined actions used to create the corpus.

If participants preferentially use causal cues when judging causal relations, and statistical action structure when making segmentation judgments, then we would expect the target part-action to be rated as no more coherent than any other part-action (like the part-actions of Experiment 1) but to be rated as highly causal (like the target sequences of Experiments 2 and 3). Similarly we would expect actions to be rated as highly coherent (like the actions in Experiment 1), but no more causal than the other non-target sequences (like the non-target sequences of Experiments 2 and 3).

However, if participants are using both sets of cues across inference tasks, then we would expect to see a compromise between the two in their ratings. In particular, we might expect that the target part-action would be rated not only as highly causal, but also as highly coherent (like the target sequences of Experiment 2), and similarly that actions might still be judged more causal than non-actions and part-actions (as in Experiment 1), even when the true causal sequence can be fully determined.

Method

Stimuli

The structure and stimuli for this experiment closely matched those of Experiment 1. The same exposure corpora with the same actions and non-actions were used as in Experiment 1, except that they were edited so that the target part-action in each corpus was always followed by a cartoon sound effect. In this experiment the part-action `Empty Rattle Clean` was the target part-action in the first exposure corpus, and `Drink Blow Look` was the target part-action in the second corpus.

For the rating portion of this experiment, we created eight part-action comparison stimuli for each corpus – the four original part-action stimuli from Experiment 1, along with four additional part-actions. In both corpora, one of these part-actions was the target part-action. iMovie HD was used to assemble the test stimuli, as in the preceding experiments.

Participants and Procedure

Participants were 166 U.C Berkeley undergraduates. Participants were randomly assigned to view one of the two exposure corpora, and were also randomly assigned to one of two follow-up question conditions. The instructions for the two conditions were identical to the causal condition and the coherence condition of Experiment 2 respectively. As in Experiments 2 and 3, all participants were told that certain action sequences caused the bottle to play music.

Following the exposure corpus, participants in both conditions were presented with all 12 actions, non-actions and part-actions individually, and asked to rate them by choosing a value on a 1 to 7 Likert scale. As in the previous experiments, In the *causal condition* participants were asked ”How likely is this sequence to make the bottle play a musical sound?”, with 1 representing “not likely” and 7 representing “most likely”, while in the *coherence condition* participants were asked the question “how well does this action sequence go together?”.

Results

Ratings from 86 participants in the *causal condition* were analyzed using 2×4 ANOVAs on exposure corpus (1 or 2) and sequence type (action, non-action, part-action, target). No effects of exposure corpus were found. As predicted, there was an overall significant effect of sequence type F(3,252)= 111.32 *MSE*= 185.44, $p < 0.0001$. The target part-action was rated as significantly more likely to cause a musical effect than actions t(85)= 10.69, $p < 0.0001$, one sample t-test on contrast values, part-actions t(85)= -12.16, $p < 0.0001$, one sample t-test on contrast values, and non-actions t(85)= 11.33, $p < 0.0001$, one sample t-test on contrast values.

Additionally, we looked at differences in ratings between the different types of non-target sequences. While all non-target sequences were rated as significantly less causal than the target sequence, actions were rated as significantly more likely to cause a musical effect than part-actions t(85)= 4.29, $p < 0.0001$, one sample t-test on contrast values, and than non-actions t(85)= 2.93, $p < 0.01$. There was no significant difference between ratings of part-actions and non-actions, t(85)= 0.58, $p = 0.57$, one sample t-test on contrast values.

Ratings from 80 participants in the *coherence condition* were analyzed using 2×4 ANOVAs on exposure corpus (1 or 2) and sequence type (action, non-action, part-action, target). No effects of exposure corpus were found. As predicted, there was an overall significant effect of sequence type F(3234)= 17.39, *MSE*= 28.18, $p < 0.0001$. Replicating the results of Experiment 1, actions were rated as going together significantly better than part-actions t(79)= 7.39, $p < 0.0001$, one sample t-test on contrast values, or non-actions t(79)= 6.635, $p < 0.0001$, one sample t-test on contrast values. There was also a significant difference between part-action and non-action ratings t(79)= 3.97, $p < 0.001$.

Additionally, the target part-action was rated as significantly more coherent than the non-target part-actions t(79)= 2.24, $p < 0.05$, and significantly more coherent than non-actions t(79)= 4.16, $p < 0.0001$. There was no significant difference in coherence ratings for

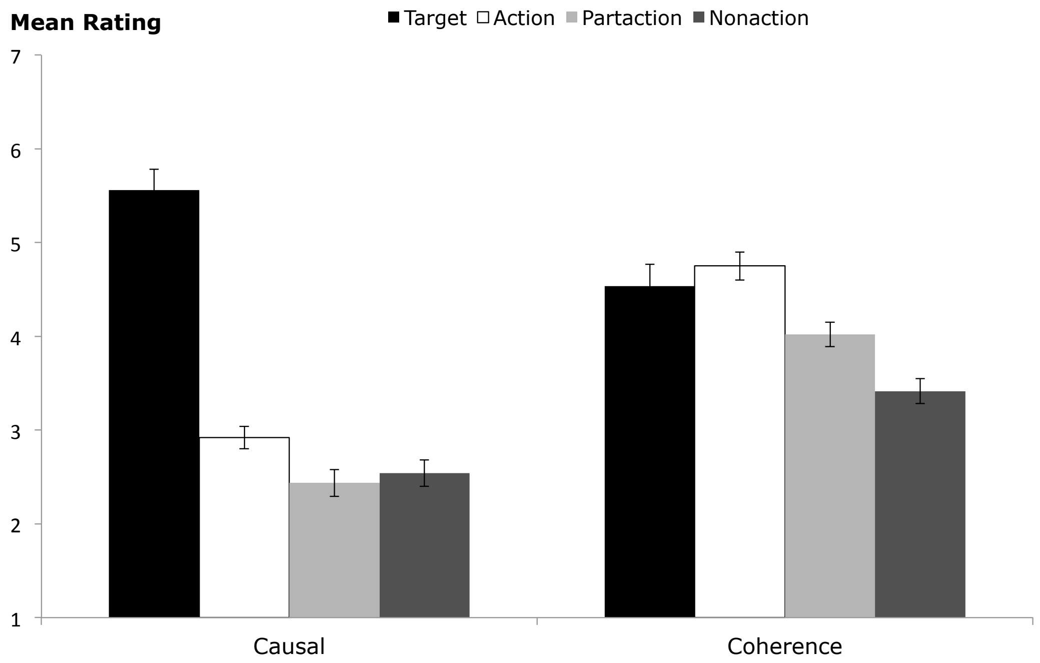


Figure 2.12: Results of Experiment 4. Error bars show one standard error.

the target part-action when compared to actions, $t(79)= 1.01$, $p = 0.32$, one sample t-test on contrast values.

Discussion

The results of Experiment 4 suggest that adults take both causal relationships and statistical structure into account when interpreting continuous human behavior, correctly identifying the part-action as the most likely cause, but continuing to rate actions as more likely to also be causal when compared to other part-actions and non-actions. Similarly, they judged the causal part-action to be very cohesive, even though it violated the statistical regularities of the action sequence, suggesting that its causal properties led to it being considered a coherent unit of human action.

2.8 General Discussion

In this chapter, we presented a Bayesian analysis of how statistical and causal cues to segmentation should optimally be combined, as well as four experiments investigating human action segmentation and causal inference. We found that both adults and our model are sensitive to statistical regularities and causal structure in continuous action, and are able to combine these sources of information in order to correctly infer both causal relationships and segmentation boundaries.

We used a non-parametric Bayesian model, adapted from work on statistical language processing, to infer the segmentation and causal structure of the same sequences our human participants saw. The model represents our assumption that the same underlying process generates human actions and causal motion sequences, implicitly capturing that actions are being chosen intentionally, often to bring about causal outcomes. Our model results demonstrate that at least in principle, action segmentation is learnable and may partly rely on domain general statistical learning mechanisms. The parallels in both human and computational model performance between word segmentation and action segmentation tasks similarly supports the possibility of a more general statistical learning ability at work in both domains.

Together, the four studies in this chapter demonstrate that among the cues people use to segment action are both statistical cues such as transitional probabilities, and causal structure, and that action structure and causal structure are learned jointly rather than being layered one on top of the other. Adults, at least, can combine statistical regularities and causal structure to divide observed human behavior into meaningful actions. Adults can also use their inferred segmentation to help them identify likely causal actions. In particular, experiments 2, 3 and 4 demonstrate that adults can identify the correct causal subsequence from within a longer set of fluid motion, a critical step in extracting higher-level goal directed units of behavior. In fact, these experiments are some of the first to demonstrate that people can carry out causal variable discovery within a continuous temporal stream of events. The fact that people rate artificially constructed actions as more coherent and meaningful than other motion sequences suggests that this is not an isolated statistical learning ability, but an integral part of action understanding. Finally, the results of Experiment 4 demonstrate that when statistical and causal cues are both present in the action stream, both of them influence peoples judgments of action segmentation and of causal relationships.

While these results are striking, there are number of open areas for future research. Like previous computational models of word segmentation, the model presented in this chapter assumes that the lowest level of segmentation is already known (or pre-labeled). That is, that there is some sort of motion primitive (equivalent to a syllable or phoneme in speech), that can already be recognized as a coherent unit. Since studies demonstrating human action segmentation have suggested that statistical patterns or features in human motion may correlate with segment boundaries at even the lowest level (e.g., Zacks, Tversky, & Iyer, 2001; Zacks, Braver, et al., 2001; Hard et al., 2006), in future work we would like to see whether action boundaries can be automatically detected directly from video, without pre-existing knowledge of low-level motion units.

Similarly, although the videos in the current studies featured a live actor carrying out natural object-directed motions, other aspects of the videos remain artificial by design – in order to focus on the statistical relationships between the small motion units, other cues such as motion changes (e.g., pauses, acceleration, deceleration), and the higher level goal structure of the actor were not present. Similarly, the actor was observed in a somewhat simplified environment, interacting with only one object, which had just one causal property. Since we know that adults can also successfully segment more naturalistic scenes (e.g., Zacks, Braver, et al., 2001; Zacks, Speer, Swallow, & Maley, 2010) with multiple objects, goals and sub-goals, and causal outcomes, one interesting direction to explore in future work is the extent to which joint statistical and causal inference contributes to our understanding of these more complex everyday scenes, and how low-level statistical information interacts with these other sources.

In the real world causal variables do not come pre-identified or occur in isolation, but instead are imbedded within a continuous temporal stream of events. Whether watching someone opening a door or making an object play music, a challenge faced by both human learners and machine learning algorithms is identifying subsequences that correspond to the appropriate variables for causal inference. Combining motion statistics with causal information may be one way for human (and non-human) learners to begin accomplishing this task.

Chapter 3

Children's Imitation of Causal Action Sequences

Imitation is not just the sincerest form of flattery–it's the sincerest form of learning. –George Bernard Shaw

3.1 Introduction

Learning the causal relationships between everyday sequences of actions and their outcomes is a daunting task. How do you transform a package of bread, a jar of peanut butter and a jar of jelly into a peanut butter and jelly sandwich? Do you cut the bread in half before or after you put together the sandwich? Can you put the jelly on first, or does it always have to be peanut butter first? In order to achieve desired outcomes – from everyday goals such as eating a tasty sandwich, to complex tasks such as making and using tools – children need to solve a challenging causal learning problem: observing that the intentional actions of others lead to outcomes, inferring the causal relations between actions and outcomes, and then using that knowledge to plan their own actions.

To learn from observation in this way, children cannot simply mimic everything they see. Instead, like the adults in Chapter 2, they must segment action sequences into meaningful subsequences, and determine which sequences are relevant to outcomes and why. Recent studies of children's imitation have produced varying answers to the question of whether children are in fact capable of inferring causal action sequences from observed demonstrations. Children can use information about an actor's prior intentions to help them identify causally effective actions (Carpenter, Call, & Tomasello, 2002). Similarly, when children observe unsuccessful demonstrations, they reproduce the actor's intended goals rather than the unsuccessful actions themselves (Hamlin, Hallinan, & Woodward, 2008; Meltzoff, 1995). In some cases, they vary the precision and faithfulness of their imitation with apparent causal relevance (Harnick, 1978; Brugger, Lariviere, Mumme, & Bushnell, 2007; Williamson &

[1] This chapter was adapted from the co-authored work Buchsbaum, Gopnik, Griffiths, and Shafto (2011).

Markman, 2006), and selectively imitate actions based on how causally effective they appear to be (Want & Harris, 2001; Williamson, Meltzoff, & Markman, 2008; Schulz, Hooppell, & Jenkins, 2008). At other times, however, children will "overimitate," reproducing apparently unnecessary parts of a causal sequence (Horner & Whiten, 2005; Lyons, Young, & Keil, 2007; Lyons, Damrosch, Lin, Macris, & Keil, 2011; McGuigan, Whiten, Flynn, & Horner, 2007; McGuigan & Whiten, 2009) or copying an actor's precise means (Meltzoff, 1988) even when this makes them less efficient at accomplishing their goal.

There are even cases where children do both in the same study. In the "rational imitation" studies by Gergely, Bekkering, and Király (2002), children saw an experimenter whose hands were either free or confined activate a machine using their forehead. Children both produced exact imitations of the actor (touching their head to the machine to make it go) and produced more obviously causally efficient actions (touching the machine with a hand), though the proportion of such actions differed in the different intentional contexts. In fact, finding a distribution of imitative responses is the norm across all these studies. Even in the most intriguing demonstrations of overimitation, it is not the case that all children blindly mimic the demonstrator's actions, and similarly, even in experiments where children show an overall appreciation for causal efficacy, some children still imitate unnecessary or ineffective actions.

We are interested in reconciling these results by suggesting that perhaps all these imitative choices are the result of rational imitation using a combination of social, physical, and statistical evidence as well as prior knowledge. In particular, evidence for which actions are causally necessary includes more than just the immediately observed demonstration. It also includes children's previous experiences with causal systems and objects, their prior observations of bringing about the same effect, and social information including the adult's knowledge state, intentions, and pedagogical stance (we know that observing a helpful teacher versus a neutral [Bonawitz et al., 2011; Brugger et al., 2007], ineffective [Schulz et al., 2008; Want & Harris, 2001; Williamson et al., 2008], or naïve [Bonawitz et al., 2011; Butler & Markman, 2012] demonstrator changes children's inferences). If different imitative choices are the result of different evidence, then we should be able to manipulate these choices and get children to imitate different portions of the same action sequences by changing the combination of social and physical evidence they receive.

Moreover, in many real-world situations, the causal structure of a demonstrated sequence of actions is not fully observable, and which actions are necessary and which are superfluous may be unclear. Therefore, there is often no single "right answer" to the question of what to imitate. After all, a longer "overimitation" sequence might actually be necessary to bring about an effect, though that might initially seem unlikely. One way in which children may overcome this difficulty is by using statistical evidence provided by repeated observations of bringing about the effect. By watching someone unlock and open a door or turn on a light bulb on multiple occasions, children can detect which actions consistently predict the desired outcome and which do not.

Probabilistic models are well suited to combining multiple sources of information. In particular, the imitation problem can be expressed as a problem of Bayesian inference, with Bayes' rule indicating how children might combine these factors to formulate different causal hypotheses and produce different action sequences based on those hypotheses. It is difficult to test this idea however, without knowing the strength of various causal hypotheses for the children. Since previous studies involved general folk physical and psychological knowledge (such as removing a visibly ineffectual bolt to open a puzzle box) it is difficult to know how strong those hypotheses would be. By giving children statistical information supporting different hypotheses we can normatively determine how probable different hypotheses should be, and then see whether children's imitation reflects those probabilities.

It is also independently interesting to explore the role of statistical information in imitation. As discussed in Chapter 2, recent studies show that children are surprisingly sophisticated in their use of statistical information such as conditional probabilities in a range of domains, from phonology (e.g., Saffran, Aslin, & Newport, 1996), to visual perception (Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002), to word meaning (Xu & Tenenbaum, 2007). Such information plays a particularly important role in both action processing (e.g., Swallow & Zacks, 2008; Baldwin et al., 2008; Buchsbaum, Griffiths, Gopnik, & Baldwin, 2009) and causal inference (Gopnik et al., 2004; Gopnik & Schulz, 2007), and allows adults to identify causal subsequences within continuous streams of action (Buchsbaum et al., 2009).

Statistical inference might be particularly important to imitation because it could allow children to not only determine the causal relationship between action sequences and outcomes, but to identify irrelevant actions within causally effective sequences. Imagine that I am making a peanut butter sandwich, and that before opening the jar, I wipe my hands on a paper towel. If this is the first time you've seen me make a sandwich, you might mistakenly think that hand-wiping is a necessary step. However, after watching me make a sandwich a couple of times, you might notice that while I always turn the lid counter-clockwise before opening the jar, I do not always wipe my hands before opening the jar, and could infer that this step is extraneous. In most previous work on children's imitation of casual sequences children were given only a single demonstration, or repetitions of the identical demonstration (including the "extraneous" actions), when shown how to generate the outcome (e.g., Whiten, Custance, Gomez, Teixidor, & Bard, 1996; Lyons et al., 2007, 2011; McGuigan et al., 2007).

In this chapter, we first look at whether children use statistical evidence from repeated demonstrations to imitate the correct causal subsequence within a longer action sequence. We present a Bayesian analysis of causal inference from repeated action sequence demonstrations, followed by an experiment investigating children's imitative behavior and causal inferences. We showed preschool children different sequences of three actions followed by an effect, using our Bayesian model to guide our manipulation of the probabilistic evidence, such that the statistical relations between actions and outcomes differed across conditions in ways that supported different causal hypotheses. We then examine which sequences the children produced themselves, and compare children's performance to our model's predictions.

Second, we investigate whether children can combine pedagogical and knowledge state information with directly observed statistical evidence, to guide their imitative choices. Will children's behavior change as the learning context becomes more pedagogical? We compare

Table 3.1: Example demonstrations, and the associated set of potential causal sequences.

Observed Action Sequence	Potential Causal Sequences
ABC+	ABC, BC, C
DBC+	DBC, BC, C
Total Potential Causes	ABC, DBC, BC, C

Note: Letters represent unique observed actions (e.g., A=Knock, B=Roll, C=Squish) while a + indicates a causal outcome.

children's imitative choices when observing a knowledgeable teacher versus a naïve demonstrator performing the same set of action sequences and outcomes. Children might assume that all adults, naïve or knowledgeable, are demonstrating potentially relevant actions, but the intuitive prediction is that children would be more likely to "overimitate" – reproducing every detail of the experimenter's actions – when the demonstrator is a knowledgeable teacher. We show how this intuition can be captured formally. We present an extension of our Bayesian model that makes behavioral predictions based on both information about statistics and about the demonstrator's knowledge, and compare children's performance to our model's predictions.

3.2 Bayesian Ideal Observer Model

While it is intuitively plausible that children use statistical evidence from repeated demonstrations to infer causal structure, we would like to verify that normative inferences from repeated observations of action sequences and their outcomes vary in a systematic way with different patterns of data. One way to derive what the normative distribution over causes should be is through a Bayesian model (Gopnik et al., 2004; Griffiths & Tenenbaum, 2005). The Bayesian formalism provides a natural way for us to explicitly represent the roles of both children's prior knowledge, and the observed data in forming children's beliefs about which action sequences are likely to be causal.

Model Details

Given observations of several action sequences, we assume that children consider all sequences and terminal subsequences as potentially causal. For instance, if the sequence "squeeze toy, knock on toy, pull toy's handle" is observed, then squeeze, followed by knock, followed by pull handle would be one possible causal sequence, and knock followed by pull handle would be another. Given all of the observed sequences, we can enumerate the potential causes (see Table 3.1 for an example set of demonstrations and potential causes). As in previous work on children's causal inference, we use a Deterministic-OR model (Griffiths & Tenenbaum, 2009), in which any of the correct sequences will always bring about the effect. To capture the intuition that there may be multiple action sequences that bring about an effect, we

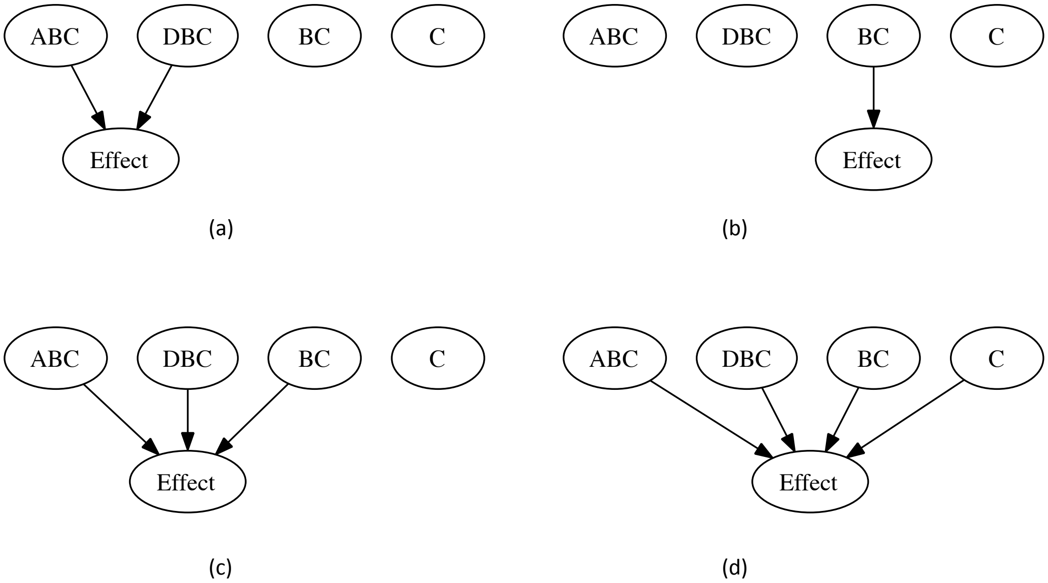


Figure 3.1: Part of an example hypothesis space. Graphs (a)-(d) each represent a different hypothesis about which action sequences are causal.

consider combinations of up to five individual causal sequences. A hypothesis, h, represents one possible combination of causal sequences, and the hypothesis space H contains all such possible combinations (see Figure 3.1).

From the learner's perspective, the problem is that they observe an action sequence, and then observe whether or not the effect is elicited. Based on this information, they want to infer what sequences of actions cause the effect. More formally, the learner wants to infer the set of causal sequences, h, given the observed data, d, where the data are composed of an observed action sequence, a, and an outcome, e. Bayes' theorem provides a way to formalize this inference. Bayes' theorem relates a learner's beliefs before observing the data, their prior $p(h)$, to their beliefs after having observed the data, their posterior $p(h|d)$,

$$p(h|d) \propto p(d|h)p(h), \tag{3.1}$$

where $p(d|h)$ is the probability of observing the data given the hypothesis is true. For deterministic-OR causal models, this value is 1 if the sequence is consistent with the hypothesis, and zero otherwise. For example, given the hypothesis that squeeze is the cause, a consistent observation would be, knock then squeeze followed by music, and an inconsistent observation would be squeeze followed by no music. When multiple sequences of actions and effects are observed, we assume that these sequences are independent.

A key element in this inference is the learner's prior expectations, $p(h)$. Previous research suggests that children believe there tends to be only one correct sequence, as opposed to many possible sequences, that cause an effect (e.g., Sobel et al., 2004). It also suggests that, all else being equal, children believe adults to be rational actors who do not perform extraneous actions (e.g., Gergely et al., 2002). We capture these intuitions with a prior that depends on two parameters, p and β, which correspond to the learner's expectations about the number of ways to generate an effect, and about the length (in actions) of causal sequences. We might say that p reflects the strength of children's simplicity bias, while β represents the degree to which they believe adults will not produce irrelevant actions, (thus leading the children to think that longer subsequences of the adult demonstrations are more likely to be causal). Note that these two assumptions may be in tension and so the model (and the children) will have to balance them.

We formalize the prior as a generative model, where hypotheses are constructed by randomly choosing causal sequences, a. Each sequence has a probability p_a of being included in each hypothesis and a probability $(1 - p_a)$ of not being included,

$$p(h) \propto \prod_{a \in h} p_a \prod_{a* \notin h} (1 - p_{a*}) \tag{3.2}$$

where the probability of including causal sequence a is

$$p_a = \frac{1}{1 + \frac{1-p}{p}\exp(-\beta(|a| - 2))}, \tag{3.3}$$

and $|a|$ is the number of actions in the sequence a. Values of β that are greater than 0 represent a belief that longer sequences are more likely to be causes. Values of p less than 0.5 represent a belief that effects tend to have few causal sequences. Taken together, Equations 3.1, 3.2 and 3.3 provide a model of inferring hypotheses about causes from observed sequences and their effects.

In our experiments, rather than probing children's beliefs directly, we allow children to play with the toy. Therefore, to complete the model, we must specify how children choose action sequences, a, based on their observations, d. Intuitively, we expect that if we know the set of causes of the effect, h, we will randomly choose one of these sequences. If we were unsure about which of several possible causes was the right one, then we may choose any of the possible contenders, but biased toward whichever one we thought was most likely. We capture these intuitions formally by choosing an action sequence given the observed data, $p(a|d)$, based on a weighted sum over possible hypotheses,

$$p(a|d) = \sum_{h \in H} p(a|h)p(h|d), \tag{3.4}$$

where $p(a|h)$ is one over the number of causes consistent with h, $1/|h|$, and $p(h|d)$ is specified in Equation 3.1. Causal models using similar probability matching have successfully predicted children and adult's performance on a variety of tasks (Griffiths & Tenenbaum, 2009).

Table 3.2: Example model results, $p = 0.5$ and $\beta = 0$.

Observed Sequences	ABC	DBC	BC	C
ABC+, DBC+	0.21	0.21	0.28	0.28
ABC+, DBC	1.0	0.0	0.0	0.0

Note: Values are the probability of choosing to perform this action sequence to bring about the effect given the observed data, $p(a|d)$, as described in Equation 3.4.

Table 3.3: Example model results, $p = 0.1$ and $\beta = 1.4$.

Observed Sequences	ABC	DBC	BC	C
ABC+, DBC+	0.28	0.28	0.34	0.09
ABC+, DBC	1.0	0.0	0.0	0.0

Note: Values are the probability of choosing to perform this action sequence to bring about the effect given the observed data, $p(a|d)$, as described in Equation 3.4.

A Simple Modeling Example

We can now verify that the model makes distinct inferences from repeated demonstrations. In the first example, the demonstrated action sequences are ABC+, DBC+ as in Table 3.1. That is, a sequence of three actions A, B and C is followed by an effect. Subsequently, a different sequence of three actions, D, B, and C is followed by the same effect. In the second example, the observed sequences are ABC+, DBC. In this case, the second three-action sequence is not followed by the effect.

Using values of $p = 0.5$ and $\beta = 0$ results in a prior that assigns equal probability to all possible causal hypotheses – a uniform prior. With this uniform prior, our model infers that, in the first case, all the sequences are possible causes, with BC and C being somewhat more likely, and equally probable. Notice that the model infers that the subsequences BC and C are the most likely causes, even though neither was observed on its own. The second case is quite different. Here the model sees that DBC and its subsequences BC and C did not lead to the effect in the second demonstration, and infers that ABC is the only possible cause among the candidate sequences (see Table 3.2).

We now use values of $p = 0.1$ and $\beta = 1.4$ leading the model to favor simpler hypotheses containing fewer causes, and causes that use more of the observed demonstration.[2] This prior does not change results in the second case, where ABC is still the only possible cause. However, in the first case, the model now infers that the subsequence BC is the most likely individual cause, since it is the longest observed sequence to consistently predict the effect (see Table 3.3).

[2] These parameter values are those that produce the best fit to children's imitation behavior in Experiment 1, as we discuss later in the chapter.

Model Predictions for Children's Inferences

We can now use the model to help us construct demonstration sequences that normatively predict selective imitation in some cases, and "overimitation" in others. If children are also making rational inferences from variations in the action sequences they observe, then their choice of which actions to imitate in order to bring about an effect should similarly vary with the evidence. We test our prediction that children rationally incorporate statistical evidence into their decisions to imitate only part of an action sequence versus the complete sequence in the following sections.

3.3 Experiment 1: Imitation of Causal Action Sequences

Method

Participants

Participants were 81 children ($M = 54$ months, $Range = 41 - 70$ months, 46% female) recruited from local preschools and a science museum. Another 18 children were excluded from the study because of demonstration error (4), equipment failure (3), lack of English (1), unavailable birth date (1), did not try toy (6), extreme distraction (2), never performed trial termination action (1).

Stimuli

There were two novel toys: a blue ball with rubbery protuberances, and a stuffed toy with rings and tabs attached to it. Six possible actions could be demonstrated on each toy. Toys were counterbalanced across children. Children were assigned to one of three experimental conditions. In each condition, they saw a different pattern of evidence involving five sequences of action and their outcomes. Each individual action sequence was always three actions long. In the "ABC" pattern, the same sequence of three actions (e.g., A=Knock, B=Stretch, C=Roll) is followed by a musical effect three times, while in the "BC" pattern a sequence composed of a different first action, followed by the same two-action subsequence (e.g., A=Squish, B=Pull, C=Shake and D=Flip, B=Pull, C=Shake) is followed by the effect three times (see Table 3.4). In both patterns, two additional sequences that end in C and do not contain BC fail to produce the effect. Finally, in the "C" pattern the sequences of actions were identical to those in the "BC" pattern, but the outcome was always positive. The number of times each individual action is demonstrated in each sequence position is identical in all three patterns. As we show later in the chapter, our Bayesian ideal observer model confirms that the statistical evidence in each pattern supports different causal inferences.

Table 3.4: The demonstration sequences for “ABC” , “BC” and “C” conditions.

“ABC” Condition	“BC” Condition	“C” Condition
ABC+	A**BC+**	AB**C+**
DEC	ADC	AD**C+**
ABC+	D**BC+**	DB**C+**
EDC	AEC	AE**C+**
ABC+	E**BC+**	EB**C+**

Procedure

The experimenter showed the child one of the toys, and said: “This is my new toy. I know it plays music, but I haven’t played with it yet, so I don’t know how to make it go. I thought we could try some things to see if we can figure out what makes it play music.” The experimenter emphasized her lack of knowledge, so that the children would not assume she knew whether or not any of her actions were necessary. She then demonstrated one of the three patterns of evidence, repeating each three-action sequence (and its outcome) twice. The experimenter named the actions (e.g., “What if I try rolling it, and then shaking it, and then knocking on it?”), acted pleasantly surprised when the toy played music (“Yay! It played music’!’), or disappointed when it did not (“Oh. It didn’t go”), and pointed out the outcome (“Did you hear that song?” or “I don’t hear anything. Do you hear anything?”). After she demonstrated all five of the 3-action sequences, she gave the child the toy and said “Now it’s your turn! why don’t you try and make it play music”. Throughout the experiment the music was actually triggered by remote activation. To keep the activation criteria uniform across conditions, the toy always played music the first time a child produced the final C action, regardless of the actions preceding it, terminating the trial. Only this first sequence of actions was used in our analysis. Each child interacted with one toy, in a single condition of the experiment.

Children were videotaped, and their actions on the toy from the time they were handed the toy to trial termination were coded by the first author, and 80% of the data was recoded by a blind coder. Coders initially coded each individual action children performed as one of the six demonstrated actions, or as “novel”. These sequences were then transferred into an “ABC” type representation, and subsequently coded as one of four sequence types: Triplet, Double, Single or Other (defined below). Inter-coder reliability was very high, with 91% agreement on the “ABC” type representations, and 100% agreement on sequence types.

Results and Discussion

Children produced significantly different types of sequences across the three conditions, $p < 0.001$ (two-sided Fisher’s exact test, Table 3.5). There was no difference in sequence types produced by children interacting with the two different toys ($p = 0.40$, n.s., two-sided Fisher’s

Table 3.5: Number of children producing each sequence type in each condition of Experiment 1.

Condition	Triplet	Double	Single	Other
“ABC”	20	1	2	4
“BC”	10	7	0	10
“C”	8	0	8	11

exact test). We will discuss results for the “ABC” and “BC” conditions first, and then return to the “C” condition.

Effect of Statistical Evidence on Imitation

In their imitation, children could either exactly reproduce one of the three-action sequences that had caused the toy to activate (that is, ABC in the “ABC” condition or ABC, DBC or EBC in the “BC” condition), or they could just produce BC in isolation. We refer to these successful three-action sequences as “triplets”, and to the BC subsequence as a “double”.

Both a triplet and a double reflect potentially correct hypotheses about what caused the toy to activate in both conditions. It could be that BC by itself causes the toy to activate in the “ABC” condition and the A is superfluous, or it could be that three actions are necessary in the “BC” condition, but the first action can vary.

If children automatically encode the adult’s successful actions as causally necessary, then they should exclusively imitate triplets in both conditions. However, if children are also using more complex statistical information, they should conclude that the BC sequence by itself is more likely to be causal in the “BC” condition than in the “ABC” condition, and that the triplet sequence is more likely to be causal in the “ABC” condition than in the “BC” condition. This is in fact what we found – the number of children producing triplets and doubles varied by condition, $p < 0.01$ (two-sided Fisher’s exact test, Table 3.5, columns 1 and 2), and differed significantly between the “ABC” and “BC” conditions $p < 0.05$ (two-sided Fisher’s exact test, Table 3.5, columns 1 and 2, “ABC” and “BC” conditions).

Effect of Differing Causal Outcomes on Imitation

Children in the “BC” condition saw three different action sequences precede the effect, while children in the “ABC” condition saw only one sequence precede the effect. This may have confused children in the “BC” condition, leading them to produce a variety of random actions, including BC. The “C” condition controls for this possibility. In this condition the sequences of actions were identical to those in the “BC” condition, but the outcome was always positive. As we show later, our Bayesian ideal observer model confirms that this provided statistical evidence for the hypothesis that C alone was sufficient to produce the effect.

Table 3.6: Number of children producing each sequence type in Experiment 1, median split by age

Condition	Triplet	Double	Single	Other
Older	19	6	4	13
Younger	19	2	6	12

In all three conditions, imitation of just the final C action in isolation was coded as a "single". As in the "ABC" and "BC" conditions, only the subsequence BC was coded as a double in the "C" condition. Also consistent with the "ABC" and "BC" conditions, in the "C" condition all five demonstrated successful sequences (ABC, ADC, DBC, AEC and EBC) were coded as triplets.

The "C" condition is as complex as the "BC" condition. However in the "C" condition the final action C produced by itself reflects a likely causal hypothesis. If children selectively imitate subsequences based on the data, then children in the "C" condition should produce C more frequently than children in the "BC" condition, and children in the "BC" condition should produce BC more frequently than children in the "C" condition. Our results support this hypothesis. Children in the "BC" and "C" conditions differed significantly in the overall types of sequences they produced, $p < 0.001$ (two-sided Fisher's exact test, Table 3.5 "BC" condition and "C" condition), and the number of children producing doubles and singles in the two conditions also varied significantly, $p < 0.001$, (two-sided Fisher's exact test, Table 3.5, columns 2 and 3, "BC" and "C" conditions).

Finally, a split by median age ($Median = 56$ months), revealed no differences in performance between older and younger age groups for any of the above analyses (two-sided Fisher's exact tests, Table 3.6), consistent with previous results with this age range (Lyons et al., 2007; McGuigan et al., 2007; Lyons et al., 2011).

Performance of "Other" Actions

Across all conditions, children did not just obligately imitate one of the successful sequences or subsequences they observed – they also produced new combinations of actions. Overall, the types of "other" sequences produced did not qualitatively differ across conditions, and appear to be a mix of exploratory behavior (e.g., performing the sequence BEC in the "BC" condition or BABC in the "ABC" condition) and genuine errors (e.g., producing ADC in the "BC" condition). There was a trend towards children in the "BC" and "C" conditions performing more of these "Other" sequences than children in the "ABC" condition $p = 0.10$, (two-sided Fisher's exact test). This difference becomes statistically significant when the two children who imitated unsuccessful triplets (e.g., ADC) are excluded from the analysis, leaving only children who performed sequences they had never seen, and subsequences other than BC and C (DC, AC or EC) $p < 0.05$, (two-sided Fisher's exact test). This result is compatible with findings that children increase their exploratory behavior when the correct causal structure is ambiguous (Schulz & Bonawitz, 2007; Schulz et al., 2008). Finally, four

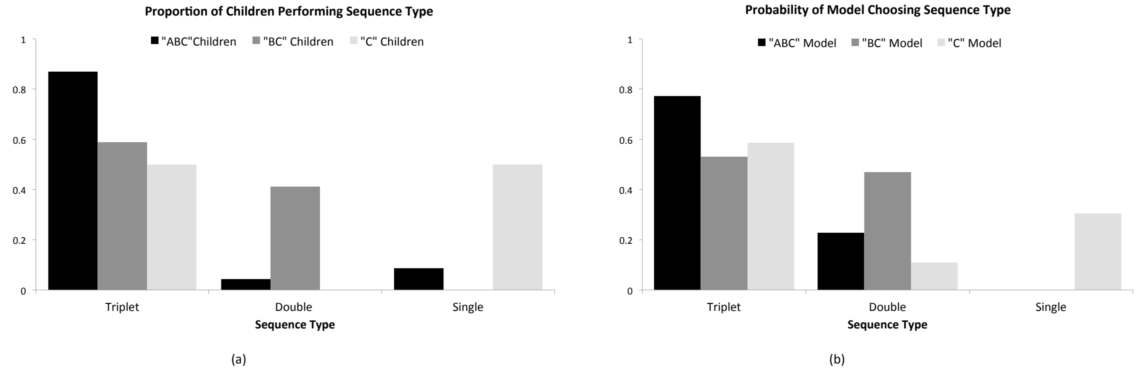


Figure 3.2: Modeling the results of Experiment 1. (a) Children's performance. (b) Predictions of our Bayesian model.

children, all in the "BC" and "C" conditions, performed novel actions (e.g., throwing the ball) or actions they had never seen demonstrated, consistent with these conditions eliciting more exploratory actions.

3.4 Modeling Experiment 1

Consistent with our experimental results, our model makes distinct predictions in each of the three experimental conditions, showing that the data supports differential causal inferences. However, we would like to explore the quantitative predictions of the model in a bit more detail.

Recall that our model has two parameters, β and p, which correspond to the learner's preexisting expectations about the length of causal sequences and number of ways to generate an effect. By fitting the model parameters to the behavioral data from Experiment 1, we can not only evaluate the model predictions more quantitatively, we can also determine the nature and strength of these same assumptions for children.

Model fit was determined by measuring the distance between the model predictions and the observed data. Because solving for the best fitting parameters is not analytically tractable, we used a grid search over the range $[0, 1]$ for p and $[0, 2]$ for β to find the best fitting parameters. While the qualitative (and quantitative) fit of the model was robust across a range of parameters, we found that the parameters $p = 0.1$ and $\beta = 1.4$ provided the best quantitative fit to the data from Experiment 1. These parameter values minimize both sum of squared error ($SSE = 0.115$) and χ^2 distance ($\chi^2 = 0.068$). These values are used throughout this chapter, allowing a generalization test of the model predictions in Experiment 2.

We used Pearson's correlation coefficient, $r = 0.93$, as a measure of the model's fit to the data. This close match to children's performances (see Figure 3.2) suggests that children's inferences based on the naïve demonstrator's actions conform closely to normative predictions based on the demonstrated action sequences. It also suggests that children may be considering the probability of several hypotheses rather than simply settling on one hypothesis and eliminating the rest.

Finally, the relatively low value for p suggests that children employ a causal Ockham's razor, assuming that simpler hypotheses, which require fewer causal sequences to explain the data, are more likely than more complex hypotheses. The relatively high value for β in the best fitting model suggests that children prefer individual causal sequences to use more of the demonstrated actions, perhaps representing a pre-existing belief that, as rational actors, adults usually do not perform extraneous actions.

Children might make this "rational actor" assumption because they are using information about the adults knowledgeability (e.g., Jaswal, 2006; Kushnir et al., 2008), reliability (e.g., Koenig, Clement, & Harris, 2004; Zmyj, Buttelmann, Carpenter, & Daum, 2010), and intentional stance (Bonawitz et al., 2011; Butler & Markman, 2012). For instance, children might notice that the experimenter always performs three-action sequences, and infer that the experimenter, while not knowing the correct sequence, knows that it must be three actions long. We next present an extension of our model that explicitly incorporates stronger pedagogical and knowledge state information, in addition to statistical evidence.

3.5 Learning from Knowledgeable Pedagogical Demonstrators

Children may learn from observing individuals who don't know how a toy works, as in Experiment 1, or they may learn from a helpful teacher who is choosing examples to try to teach the child how the toy works. In teaching situations, children may draw different inferences from the same data by inferring *why* the teacher chose these data. Intuitively, children may implicitly assume that the teacher's sample demonstrations are not randomly chosen, but are designed to be informative (Csibra & Gergely, 2006).

We can formalize this idea by incorporating a model of how a teacher's choice of interventions provides information about the hypothesis they are trying to teach into our initial model of rational imitation. We can then compare our model's predictions to children's performance, to see if children's imitative choices reflect a belief that knowledgeable teachers select informative examples.

Modeling Pedagogical Learning

Recall Equation 3.1 related a learner's posterior beliefs $p(h|d)$ to their prior beliefs, $p(h)$. This was accomplished by way of a measure of how consistent the data were with a hypothesis, $p(d|h)$. Here, the data, d, include an action sequence, a, and an outcome e. We did not specify our belief about how the demonstrator's sequence of actions, a, was chosen. Implicitly, we assumed that these choices were random, and therefore did not factor into our inference.

However, to formalize how having a helpful teacher may affect inferences, we must specify how the demonstrator chooses their actions and expand Equation 3.1 to include a factor, $p(a|h)$. The learner would then update their beliefs based on the product of the prior probability, the probability of the action given a hypothesis, and the probability of the effect given the action and the hypothesis

$$p(h|a, e) \propto p(e|h, a)p(a|h)p(h). \tag{3.5}$$

Here we have introduced $p(a|h)$, which specifies the learner's beliefs about how the demonstrator chooses their action sequence given a hypothesis, and separated the data into the action sequence, a, and it's effects, e. For a demonstrator who was choosing their actions at random, $p(a|h)$, is the same for all sequences, $\frac{1}{|A|}$ (where A is the set of all action sequences, and $|A|$ is the number of possible sequences) and can be ignored. However, if the learner believes the demonstrator is a helpful teacher, then they could expect the teacher to choose their actions, $p(a|h)$, with the goal of having the learner infer the correct hypothesis,

$$p_t(a|h) \propto p_l(h|a, e), \tag{3.6}$$

where t and l indicate teacher and learner, respectively (Shafto & Goodman, 2008). The equation states that the learner can expect the teacher to choose action sequences that tend to make the learner believe the correct hypothesis.

Model Predictions

By explicitly representing assumptions about the demonstrator's knowledgeability and helpfulness, the pedagogical model makes distinctly different predictions than the previous model. The pedagogical model assumes that the demonstrator has not chosen their actions randomly, but for the purpose of teaching the learner. This implies that the learner should put more weight in the demonstrations, as compared to the same evidence demonstrated by a naïve individual. Therefore, if the teacher chose to demonstrate a long sequence such as squish, knock, pull and the effect was elicited, the learner would be more likely to infer that all three actions were necessary, than if these demonstrations were produced randomly (for other work on pedagogical inference see Shafto & Goodman, 2008; Bonawitz et al., 2011).

Consider the BC condition from Experiment 1 (see Table 3.4). Children observed five sequences of actions, three of which led to the effect and two that did not. Of the three cases that elicited the effect, all contained the subsequence BC, and when the effect was not elicited this subsequence was not present. However, in all of the sequences, the demonstrator chose sequences of three actions. Under the assumption that the demonstrator is naïve, the model predicted that these factors trade-off, leading to the prediction that it is roughly equally likely that triplets or doubles could elicit the effect.

In contrast, under the assumption that the demonstrator is knowledgeable and helpful, the pedagogical model predicts a shift in children's inferences. Figure 3.3 shows the predictions of the model assuming naïve and pedagogical demonstrators (and the parameter values

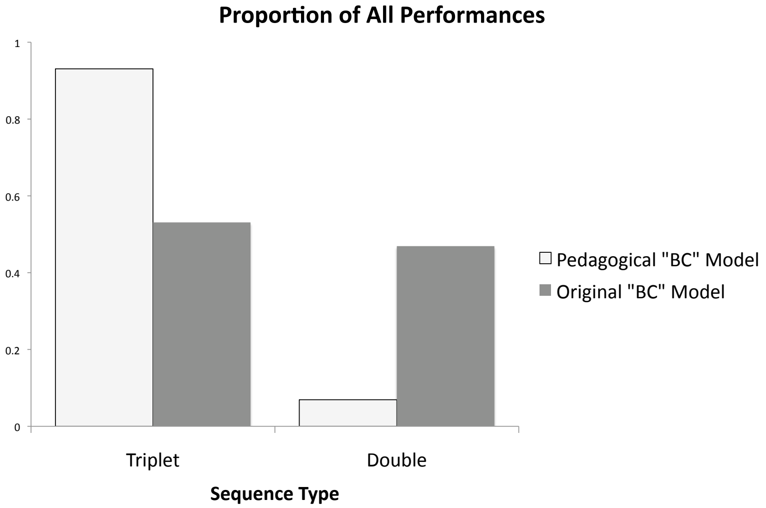


Figure 3.3: Predictions of our model given assumptions of pedagogical sampling (as in Experiment 2) or random sampling (as in Experiment 1)

used in the first experiment). The pedagogical model predicts that, after observing the same sequences of actions, children should be much more inclined to believe that triplets cause the effect. We test this prediction in the following experiment.

3.6 Experiment 2: Effect of Combined Pedagogical and Statistical Evidence on Imitation

Method

Participants

Twenty seven children ($M = 52$ months, $Range = 44 - 62$ months, 37% female) recruited from preschools and a science museum were included in this study. Another 11 children were excluded because of experimenter error (4), equipment failure (1), parental interference (1), extreme distraction (1), never performed trial termination action (1), failure to complete experiment (3).

Stimuli

The same two novel toys and corresponding actions were used as in Experiment 1. In this condition, the demonstrated sequences of actions and outcomes were identical to those in the “BC” condition of Experiment 1.

Table 3.7: Number of children producing each sequence type in Experiment 2

Condition	Triplet	Double	Single	Other
Naïve “BC”	10	7	0	10
Pedagogical“BC”	14	0	0	13

Procedure

The experimenter showed the child one of the toys, and said: “See this toy? This is my toy, and it plays music. I’m going to show you how it works. I’ll show you some things that make it play music and some things that don’t make it play music, so you can see how it works”. The experimenter emphasized her knowledge of the toy, and that her actions were chosen purposefully and pedagogically. She then demonstrated the “BC” pattern of evidence, almost exactly as in the BC condition of Experiment 1. The only difference was that the experimenter indicated that she expected each resulting outcome (“See? It played music” or “See? No music.”). Otherwise the procedure and coding was exactly as in Experiment 1. Inter-coder reliability was very high, with 91% agreement on the “ABC” type representations, and 100% agreement on sequence types.

Results and Discussion

The action sequences and causal relationships demonstrated in this experiment are identical to those in the “BC” condition of Experiment 1. If children are only attending to the observed statistical evidence, then their inferences here should be the same as in the original “BC” condition. However, since children are now told that the experimenter is showing them how the toy works, this explicit pedagogy provides additional causal information. If children believe that the demonstrator is a rational teacher, then they might think that the demonstrator is choosing to show them triplets, because triplets, not doubles, are necessary to produce the effect, and should shift their imitative choices accordingly. Therefore, if children are able to attend to both statistical evidence and the demonstrator’s pedagogical stance, then they should produce more triplets in the pedagogical “BC” condition than the original “BC” condition, and more doubles in the original “BC” condition than in the pedagogical “BC” condition.

Children in the original and pedagogical “BC” conditions differed significantly in the types of sequences they produced, $p < 0.05$ (two-sided Fisher’s exact test, Table 3.7). The number of doubles and triplets produced in the two conditions varied significantly, $p < 0.01$, (two-sided Fisher’s exact test, columns 2 and 3, Table 3.7). As in Experiment 1, there was no difference in sequence types produced by children interacting with the two different toys ($p = 0.70$, n.s., two-sided Fisher’s exact test), and a split by median age ($Median = 52$ months) revealed no difference in sequence types produced by younger vs. older children ($p = 0.45$, n.s., two-sided Fisher’s exact test)

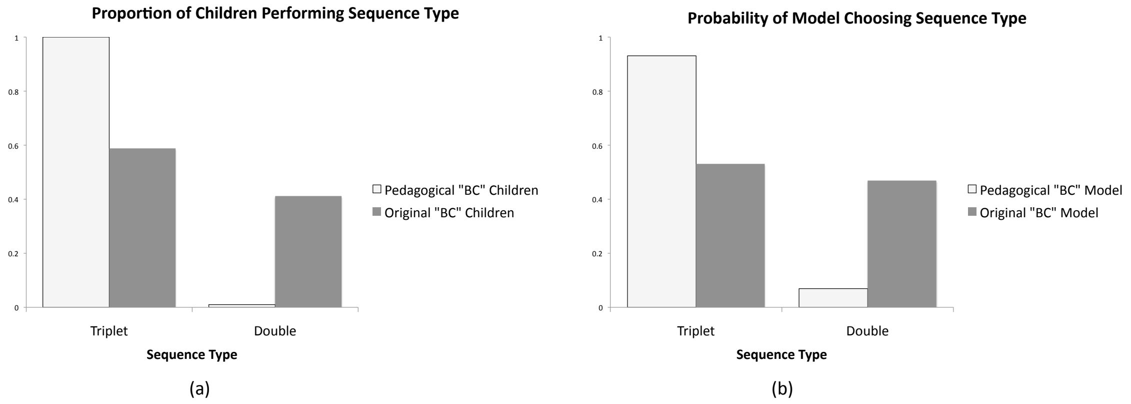


Figure 3.4: Modeling the results of Experiment 2, using assumptions of Pedagogical sampling. (a) Children's performance. (b) Our model's predictions.

We used Pearson's correlation coefficient, $r = 0.99$, as a measure of the model's fit to the data (see Figure 3.4). This close match to children's performances was achieved with the same parameters as were used in Experiment 1. This provides evidence that the complexity of the model is comparable to that of children's behavior, as we would expect an overly complex model to overfit the data and generalize poorly. Psychologically, these results suggest that children's inferences based on observations of a naïve demonstrator versus a knowledgeable teacher conform closely to normative predictions.

3.7 General Discussion

In this chapter, we examined whether children are sensitive to multiple sources of causal information when choosing the actions they imitate, and can integrate this information rationally. In Experiment 1, we demonstrated that children can use statistical evidence to decide whether to imitate a complete action sequence, or to selectively imitate only a subsequence. In particular, children in the "ABC" condition imitated the complete sequence ABC more often than children in the "BC" condition, while children in the "BC" condition imitated the subsequence BC more often than children in the "ABC" condition. Children's performance in the "C" condition demonstrated that the differential imitation in the "ABC" and "BC" conditions could not be explained as a result of task complexity. In Experiment 2 we showed that children can combine statistical evidence with information about the demonstrator's knowledge state in deciding which actions to imitate – imitating different portions of the same action sequences when they observe them being performed by a helpful teacher versus a naïve demonstrator.

It is also worth noting the information-processing complexity of this task. Children saw thirty similar actions and ten outcomes in each condition, and yet they appeared to track and use this information in deciding which actions to produce. This is consistent with other studies in which children and adults show surprising if implicit capacities to track statistical regularities (Saffran, Aslin, & Newport, 1996; Fiser & Aslin, 2002; Baldwin et al., 2008; Buchsbaum et al., 2009), as well as our own results described in Chapter 2.

These results extend earlier findings that show children take causal and intentional information into account appropriately in their imitation. They show that children also take into account statistical information about the conditional probability of events and do so in an at least roughly normative way. Both the model and data suggest that children may be making more finely-graded judgments about the probability of various options rather than simply making yes or no decisions about whether to use a particular strategy. However, it should be pointed out that we had only one response per child in this study so that we do not know for sure whether this probability matching behavior applies to individual children or only to children as a group (for a discussion of probability matching behavior see for example Vulkan, 2000; Denison, Bonawitz, Gopnik, & Griffiths, 2009).

These studies suggest that causal learning is informed by both social knowledge and statistical information. Children are sensitive to probabilities, knowledge state, and pedagogical intent when deciding which actions to imitate. The studies also suggest a rational mechanism for the phenomenon of “overimitation” (Lyons et al., 2007). In particular, the “triplet” responses could be thought of as a kind of overimitation, reproducing parts of a causal sequence that are not actually demonstrably necessary for the effect. These results suggest that this behavior varies depending on the statistics of the data and the probability of various hypotheses concerning them. “Overimitation” also varies depending on the social demonstrator. By explicitly representing the contributions of these different sources of evidence and using them to assign probabilities to causal hypotheses, a Bayesian model can predict these behaviors quite precisely.

Our naïve demonstrator explicitly established her lack of knowledge. In contrast, many of the studies of imitation we discussed at the start of this chapter did not provide the child with either clearly pedagogical or non-pedagogical demonstrators. These demonstrators may have used cues such as directed gaze and pointing (Csibra & Gergely, 2009; Gergely et al., 2007; Senju, Csibra, & Johnson, 2008), leading children to assume that they were in a teaching situation. In general, these studies also showed children only one way to bring about the desired effect and used causal systems where children’s prior expectations were unclear. These differences may help explain why children’s imitative choices seem so varied across studies.

This is the first study showing that children are more likely to overimitate when exactly the same actions are presented in an explicitly pedagogical vs. non-pedagogical context. Our model also suggests that despite appearances, such behavior is a rational response to different combinations of social, statistical, and physical information. In situations where causal structure is ambiguous, children not only take advantage of social demonstrations, they use relevant information about the demonstrators themselves to make causal inferences.

A related possibility, which we have not yet investigated empirically, is that seeing a repeated sequence of actions with no obvious physical causal outcome may lead children to suspect that the actions are intended to have a social or psychological rather than physical effect. Such inferences could be responsible for the use of imitation to transmit cultural conventions such as manners, rituals or even linguistic regularities. Children might conclude that putting fork on the left rather than the right for example, is intended to cause the observer to recognize the manners of the hostess, rather than to improve the efficacy of the fork.

These studies show that children are sensitive to statistical information, knowledge state, and pedagogical intention in determining which sequences of actions to imitate. Along with other studies, they suggest that Bayesian inference, which supports the construction of causal models from statistical patterns, may play a significant role in many important kinds of early learning. From learning how to make peanut butter sandwiches to playing with a new toy, children flexibly make use of many sources of information to understand the causal structure of the world around them.

Chapter 4

Causal learning, counterfactual reasoning, and pretend play

My fake plants died because I did not pretend to water them. – Mitch Hedberg

4.1 Introduction

The studies described in the previous chapters have looked at how people learn causal relationships from real actions, but humans and especially children also engage in quite a bit of pretend action. In general, pretend play seems paradoxical. Why should children spend so much time thinking about unreal worlds?

Intuitively, childhood pretense bears a striking resemblance to counterfactual inference, but this relationship has not been widely explored. Counterfactual thinking, where one envisions alternative possible events and their outcomes, is hypothesized to be one of the primary ways in which we reason about causal relationships (e.g., Pearl, 2000; J. Woodward, 2003). As discussed in previous chapters, recent computational and experimental work suggests that both adults and children may reason about causality in a manner consistent with probabilistic graphical models – coherent, complex representations of causal structure that allow distinctive kinds of inferences (e.g., Gopnik et al., 2004; Griffiths & Tenenbaum, 2009). In particular, the causal models approach supports and distinguishes two types of inferences, predictions, on the one hand, and interventions, including counterfactual interventions, on the other. In predictions, we take what we think is true now as a premise and then use the model to calculate what else will be true. In counterfactuals, we take some value of the model that we currently think is not true as a premise, and calculate what would follow if it were. In fact, the ability to consider the consequences of possible interventions before actually implementing them may be at the heart of truly causal reasoning.

[1] This chapter was adapted from the co-authored work Buchsbaum, Bridgers, Weisberg, and Gopnik (2012).

We propose that these crucially important abilities – creating possible causal interventions and testing alternative causal hypotheses – depend on the same cognitive machinery that children use when they pretend: adopting a premise that is currently not true, creating an event sequence that follows from that premise, and quarantining the result of this process from reality. We suggest that pretend play is one of several forms of child-directed exploratory play that fosters the development of causal cognition (Schulz & Bonawitz, 2007; Cook et al., 2011), and helps children master these crucial cognitive skills in much the same way that play hunting or fighting allows mastery of motor skills.

In the first part of this chapter, we discuss some of the theoretical ideas underlying this proposal that childhood learning, and play in particular, and causal cognition are closely connected. In the second part, we focus on an empirical study demonstrating one such connection – a link between pretend play and counterfactual causal reasoning. We show that children who are given new information about a causal system make very similar inferences both when they consider counterfactuals about the system and when they engage in pretend play about it. We also show that these two abilities are correlated – children who apply appropriate causal constraints in their pretend play also do better in a counterfactual task. This relationship holds even when age, general cognitive development and executive function are controlled for. These findings link a distinctive human form of childhood play and an equally distinctive human form of causal inference.

This study is just one example of a more general link between learning and behavior in childhood and adult cognitive abilities. However, we believe it is a particularly telling one. We argue that the free exploration of possibility in pretense helps human beings to construct wide-ranging causal models of the world and to reason from them.

The Uses of Immaturity

The great puzzle of the evolution of human cognition is to determine how such small genetic changes over such a brief period could have led to such massive changes in behavior. In this chapter, we emphasize two interlocked developments that might have interacted in a coevolutionary way to provide large differences from small changes. The first is the change in the developmental program that led to the uniquely long period of human childhood. We hypothesize that this change allowed immature proto-humans to enjoy longer protected periods of learning and, in particular, to engage more extensively in the free exploration found in play.

Second, we propose that this developmental change created the context for the application of more powerful learning mechanisms. In particular, these learning mechanisms included a newly sophisticated and general ability and motivation to learn about causation and to construct causal models. Those models, in turn, support sophisticated inference and planning by allowing organisms to consider a wide range of alternative possible future outcomes. The result was a set of new abilities ranging from more sophisticated tool use for foraging to more sophisticated social intelligence for cooperative child-rearing. Those abilities, in turn, allowed for still greater caregiving investment and a still longer childhood and so on.

There is strong evidence that a change in the developmental program played an important role in human evolution. Human offspring, in particular, have a longer period of immaturity than those of any other primate. This is also true of Homo sapiens when compared to extinct hominoids, such as Neanderthals (Smith et al., 2010). The cost of protracted immaturity is the need for greater caregiving, and here too, humans show striking adaptations for increased caregiving investments in comparison to our closest primate relatives, including pair bonding, increased alloparenting, and a long period after menopause (the "grandmother" hypothesis Hawkes, Kim, Kennedy, Bohlender, & Hawks, 2011; Hrdy, 2009).

There is, moreover, a widespread correlation between extended immaturity, relatively large brain size and relatively sophisticated learning abilities across many species, including birds and placental and marsupial mammals (Weisbecker & Goswami, 2010). The extreme immaturity and impressive brain size and learning ability of humans lie at the far end of the distribution on these measures.

These correlations suggest a connection between the cognitive changes in humans and the extended period of human development. But how might one lead to the other? It is possible, of course, that longer immaturity was necessary simply to have the time to grow large brains. But it is equally possible, and arguably more plausible, that our evolutionary advantage accrued from the fact that those brains were being grown, modified and shaped under the influence of the environment, and in a way that allowed massive plasticity and learning (see Jablonka, Ginsburg, & Dor, 2012). In fact, the revolution in cognitive development over the past thirty years has shown that infants and very young children do, in fact, engage in just this kind of learning. While in the past it may have been possible to think of infants and young children as cognitively limited creatures who simply passively waited for brain maturation, in fact, contemporary research demonstrates that even infants and toddlers learn a remarkable amount in remarkably sophisticated and complex ways (for recent reviews see Gopnik & Wellman, 2012; A. L. Woodward & Needham, 2008).

Young children not only learn as much or more than adults, they also learn differently. In the language of machine learning there is a trade-off between "exploration" learning, learning about the environment for its own sake, and "exploitation" learning, finding the right information about the environment to achieve a particular goal. "Exploration" learning is wide-ranging and general and it has many advantages – it allows organisms to discover methods for survival in a wide range of physical and social environments. It also has some disadvantages. In particular, it means that organisms will not be prepared to deal with the particular demands of the environment until after learning has taken place. We argue that extended immaturity helps resolve that trade-off – a protected period of exploration as children allows us to exploit as adults. Empirically, young children do engage in extensive exploratory learning. Immaturity allows powerful and wide-ranging exploratory learning mechanisms to be extensively employed in the protected period of human childhood, while the costs of everyday survival are borne by caregivers.

What do those learning mechanisms look like? One likely candidate is a set of computational devices for learning about the causal structure of the world. The ability to understand causal relationships and to reason from them is at the heart of many distinctive human abilities. Understanding physical causal relations underpins sophisticated forms of tool use (see Byrne, 1995; Sterelny, 2012). Understanding psychological causal relations underpins the ability to understand and manipulate others, abilities that are at the core of "theory of mind" or "Machiavellian intelligence" (Byrne & Whiten, 1989). Causal understanding thus underpins the kinds of cognition that have been proposed as part of the distinctively human cognitive toolbox. Moreover, in both the physical and psychological domains, causal knowledge allows for sophisticated inferences about the future and about the counterfactual past. Such thinking has been called "mental time travel" (Mulcahy & Call, 2006; Raby, Alexis, Dickinson, & Clayton, 2007; Suddendorf & Corballis, 1997). All of these abilities are clearly present in nascent form in some non-human animals, but there is little doubt that these are dimensions where humans are distinctively capable.

Causal Models and Bayesian Learning

Recent work has outlined the kinds of representations that underpin causal knowledge in adult humans and the kinds of mechanisms that allow this knowledge to be learned (Gopnik et al., 2004; Gopnik & Schulz, 2007; Griffiths & Tenenbaum, 2009). As described in earlier chapters, this work is part of a broader approach to cognition that involves probabilistic models and Bayesian inference (Griffiths et al., 2010). The essential idea behind this recent research is that humans have causal models: structured, generative, causal representations of the world. These representations appear to go beyond the typical representations that might be constructed from simple associative processes or conditioning.

What makes causal models distinctive? Traditionally, philosophers and psychologists have had two approaches to causation. One approach focuses on "mechanisms", on the particular spatio-temporal characteristics of events, particularly events that involve contact or launching (Michotte, 1963). However, many events that do not include these features, ranging from remote controls to social interactions, are also construed as causal even by very young children (Kushnir & Gopnik, 2007). Another tradition, going back to Hume, is that causal relations are nothing more than associations between correlated events. But if the mechanism approach is too narrow, the correlational approach is too wide. A causal relationship goes beyond a predictive or associative one, as we outline below.

More recently, philosophers have pointed to two distinctive features of causal knowledge, which are captured by causal models. First, causal knowledge supports a distinctive set of inferences involving interventions and counterfactuals (see e.g., Gopnik & Schulz, 2007; Pearl, 2000; Spirtes et al., 2001 and especially J. Woodward, 2003). For example, both smoking and having yellow nicotine-stained teeth are associated with lung cancer. So if you see yellow teeth, you can predict the presence of cancer. However, only a causal account of the disease leads to the correct prediction that a tooth-brushing intervention will have no effect on the cancer rate, while a smoking-prevention intervention will. Similarly, causal knowledge supports counterfactual claims (Lewis, 1973). A causal account of cancer will also tell you that, had smoking been discouraged in the past, many lives would have been saved.

Second, causal knowledge involves not only specific relations between particular causes and effects, but coherent networks of causal relations – the kinds of networks that are described in theories. In operant conditioning, or in trial and error learning, a learner must previously observe the effects of an action or a series of actions in order to predict those effects in the future. This is not the case for a learner with a causal theory, who can predict the effects of such actions without ever having observed them. In fact, these actions might be quite unusual and have a low initial probability. Causal theories thus allow reasoners to make a very wide range of new predictions, interventions and counterfactual inferences about events, allowing for sophisticated kinds of insightful planning and action.

For example, a scientist could use a physical causal theory to predict that the very complex and novel sequence of actions involved in the Apollo 11 launch would result in the unprecedented event of a man walking on the moon. But we also see this coherence in intuitive or everyday theories, not just in scientific ones. Two- and three-year-old children, for example, appear to have a causal theory of the mind – they can appreciate the complex causal relations between emotions, perceptions and desires, and can use these relations to generate novel explanations and inferences about events they have never experienced before (Wellman, Phillips, & Rodriguez, 2000).

Formal models of causal relationships, such as causal graphical models, represent these causal networks as graph structures associated with probability distributions (Pearl, 2000; Spirtes et al., 2001). They also include procedures for making predictions, designing interventions and making counterfactual claims. Specifically, in both interventions and counterfactuals, the learner "fixes" the value of a variable in a causal network. Then she uses the model to work out the "downstream" consequences in the possible world where the variable had that value. If the consequences are desirable, she can act to cause the variable to have that setting in the actual world – she can produce an intervention. But she can also simply consider what would have happened if the variable had been set to that value, and so think of the counterfactual consequences of an event or an action. There is extensive evidence suggesting that both adults and children, can use causal models in this way to make predictions and design interventions, and that adults can use them to make counterfactual inferences about the past (Gopnik et al., 2004; Gopnik, Sobel, Schulz, & Glymour, 2001; Meder, Hagmayer, & Waldmann, 2009; Sloman, 2005; Waldmann, Hagmayer, & Blaisdell, 2006).

Causal models thus allow their users to make a powerful range of new predictions. Equally importantly, causal models can be learned, and lend themselves to Bayesian learning mechanisms (e.g., Griffiths et al., 2010; Griffiths & Tenenbaum, 2007). Such mechanisms involve searching through a space of possible hypotheses – in this case, possible causal models – and comparing them to the evidence. Obviously, it is not possible to simply enumerate and assess all the possible hypotheses individually. But Bayesian learning algorithms can approximate that search. For example, a Bayesian learning strategy might proceed by starting with the current best model for how the world works. In order to learn, a user must modify that model to produce an alternative, and then assess the fit between the evidence generated by this alternative model and the actual evidence observed in the real world. This assessment is done by calculating the probability that the alternative model would generate the observed evidence. This involves asking two questions: (1) How probable is it that one would observe these events if the alternative model were a true representation of the causal structure of the world? (2) How likely is the causal relationship that this model represents overall, taking into account its prior probability? The user must also answer these questions about his or her current model. If the resulting probability of the alternative model is higher, the user should discard the current model and accept the alternative model as true. There is evidence that human children as young as 16 months old can learn causal models from statistical information in this way (see e.g., Gopnik et al., 2004; Gweon & Schulz, 2011; Schulz, Bonawitz, & Griffiths, 2007; Schulz, Gopnik, & Glymour, 2007).

This learning procedure, like other Bayesian procedures, is powerful, but it is computationally demanding. It requires that the learner explore a range of possible models before settling on the likeliest one. But we believe that even this kind of complex computation and comparison is within the grasp of preschool-aged children. Indeed, we see exactly such exploration of alternative models emerging spontaneously and early in children's pretend play.

Pretend Play

Play is characteristic of young animals across a wide range of species (Bekoff & Byers, 1998). The behaviors that are involved in play are typically those that will be important for the adults of the species, which explains why play fighting and hunting behaviors are ubiquitous. Play is a form of exploratory learning. The immature animal can explore and practice alternative actions in a low-risk setting, without the pressure of achieving a particular goal. Indeed, recent research shows that a kind of exploratory play that involves informal experimentation helps human children learn causal models (e.g., Cook et al., 2011; Schulz & Bonawitz, 2007).

However, human children also engage in a particularly distinctive kind of pretend or symbolic play. In this type of play, children go beyond simply practicing actions they will require later or manipulating objects to discover their causal features. Instead, they work out quite elaborate unreal scenarios, often with the aid of language, props and gestures. As with so many human behaviors, there is evidence that precursors of this kind of play may be found in other primates, particularly symbol-trained chimpanzees (Jensvold & Fouts, 1993). However, again as with many other behaviors, it is clear that that this is a domain where humans are at least quantitatively if not qualitatively different. In all her hours of observation of the chimpanzees of Gombe for example, Jane Goodall only recorded a few instances of what might have been pretend play. In contrast, almost any observation of 4-year-old humans would uncover multiple instances of such play (see Fein, 1981; Harris et al., 1993; Leslie, 1987; Singer & Singer, 1990), and human children demonstrate remarkable competence not only at pretending but at understanding the rules that govern pretense (for a review see Weisberg, 2013). Indeed, though cultures may vary in the amount and the themes of early pretend play, such play is found across a strikingly wide variety of cultural settings (e.g., Gosso et al., 2005). But pretend play also has a paradoxical quality. Why would children spend so much time and energy engaged with non-real scenarios when it would arguably serve them better to attempt to understand how the real world works?

Our answer to this question focuses on the similarities between the playful activity of pretending and the serious reasoning capabilities involved in counterfactual inference and Bayesian learning (see Gopnik, 2009). A number of researchers have previously remarked on the similarities between play and counterfactual inference (e.g., Amsel & Smalley, 2000; Harris, 2000; Hoerl, McCormack, & Beck, 2011; Lillard, 2001). But simply noting these similarities does not explain why counterfactual reasoning itself would be useful, given that it is also about possible worlds rather than actual ones. In addition, to our knowledge, there have been no previous empirical demonstrations that pretense and counterfactual reasoning are specifically related in development.

We address the first issue by proposing that pretend play provides an opportunity to practice and perfect the skills of reasoning from, and learning about, a causal model, just as play fighting or hunting allows animals to perfect complex motor skills. Pretend play, counterfactual and intervention reasoning, and Bayesian learning all involve the same cognitive machinery: the ability to consider events that have not occurred, in Leslie's terms to "decouple" representations of those events from reality (Leslie, 1987), and to think about what would be the case if they had occurred (see Weisberg & Gopnik, in press). These abilities are required not only for planning, but also for learning. In order to execute the algorithms that are involved in Bayesian causal learning, children need to do the same things they do when they pretend. They must create an alternative representation and generate the observations that they would have seen if that alternative were true. Just as physical play provides young animals with the opportunity to practice skills that they will need later in life, we argue that pretend play lets children practice the cognitive skills necessary for causal learning, planning and counterfactual reasoning.

Preschool children are especially focused on developing causal models of the minds of others or "theory of mind". Accordingly, much early pretend play, such as the creation of imaginary companions, is also focused on exploring these kinds of psychological causal relationships (see Gopnik, 2009). There have been both theoretical and empirical claims about the relation between pretend play and theory of mind abilities (Leslie, 1987; Lillard, 2001; Taylor, 1999). However, pre-schoolers also learn physical causal models. We thus predict that children's abilities to make physical causal inferences should also be related to pretense.

How could we test this claim? There is already evidence in the literature that children typically obey causal constraints in their pretense (Harris, 2000). For example, if children are given a pretend scenario in which Teddy spills tea on the floor, they will infer that the floor is wet, but they will say that it is dry if he spills talcum powder. There is also some evidence that children as young as 2 1/2 can make counterfactual inferences, although this is more controversial (see Harris et al., 1993; Harris, 2000, but see Riggs, Peterson, Robinson, & Mitchell, 1998). Faced with a floor with muddy ducky bootprints, for example, children will say that the floor would have been clean if ducky had taken his boots off (Harris et al., 1993; Harris, 2000).

In both of these cases, however, children might be interpreted as simply following familiar and highly practiced scripts rather than making novel inferences. Children know that tea spilling is followed by wetness, just as a young wolf might know that mock biting follows mock chasing. Moreover, there is no current empirical evidence that these two abilities, causal constraint in pretense and counterfactual inference, are actually connected to one another.

Here, we present the first empirical evidence of this connection. We presented children with a novel causal system and taught them a novel causal relationship, ensuring that children were not simply reproducing a familiar script. We then tested whether they would import the causal structure into their pretend play, whether they would make the correct counterfactual causal inferences about that system, and whether these two abilities were related. This study thus provides us with a way to explore the proposed relationship between causal and counterfactual reasoning, and pretense.

4.2 Experiment 1: Pretending Causal Structure

In this experiment, 3- and 4-year-olds were taught a novel causal relationship and then were encouraged to engage in a pretend game to see if they would maintain and act on this relationship in the context of an imaginary world. The causal relationship involved a toy, the “Birthday machine,” which plays “Happy Birthday” when an object called a zando is placed on top, but which does not activate with a non-zando object. The toy was actually surreptitiously activated by a hidden button, a commonly used method in causal-learning tasks. Indeed, in extensive other experiments using this and similar “detector” machines, both children and adults inferred a causal relation between the objects and the effect – no child or adult ever guessed the hidden cause (e.g., Lucas, Gopnik, & Griffiths, 2010). Moreover, in similar experiments, preschool children could acquire a causal model of such machines that allowed them to make novel inferences about interventions on the machine and to explicitly infer its causal structure, even when that causal structure was complex (e.g., Gopnik et al., 2004; Buchsbaum, Gopnik, et al., 2011; Schulz & Gopnik, 2004; Sobel et al., 2004).

During our study, we told children that it was a stuffed toy named Monkey’s birthday, and that the experimenter and the child would use the “Birthday machine” to sing to Monkey as a surprise for his birthday. The experimenter taught the child the causal relationship and then asked him/her a series of counterfactual questions about the machine.

Then a confederate entered the room and removed the machine, the zando and the non-zando object. In response, the experimenter introduced a box and two blocks and explained that they could still surprise Monkey if they pretended that the box was the machine and that one block was the zando and the other was the non-zando. The experimenter first asked the child what he or she wanted to pretend. Then the experimenter prompted the child to try each block on the machine and asked him/her what they were pretending was the consequence of this action, to see if the child would uphold the real-world causal relationship s/he had learned in the context of the pretend game.

Based on our hypothesis that children's pretend play facilitates counterfactual causal reasoning, we made several predictions. First, we predicted that children would transfer the real-world causal relationship into the pretend scenario. That is, children should intervene with the pretend zando to bring about pretend music, and have the pretend non-zando be causally ineffective. We further predicted that children who made this transfer in pretense would be more likely to answer the real-world counterfactual questions correctly.

Method

Fifty-two 3- and 4-year old children were tested in this Study (see Appendix C for details).

Causal Demonstration Phase

The experimenter began by explaining to the child that today was her friend Monkey's birthday and that the goal of the game was to surprise Monkey. The experimenter then put Monkey underneath the table so that he would be unable to hear what the surprise was. The experimenter then introduced the "Birthday machine" to the child by saying, "This is my machine. And you know what? This machine plays "Happy Birthday."'' The experimenter explained that the surprise would be to sing "Happy Birthday" to Monkey when the machine played the song. The experimenter then placed two distinctive objects on either side of the machine in counter-balanced order, and said "One of these is a zando and one is not a zando. The machine only plays 'Happy Birthday' when the zando is on top, so I'm going to need your help to figure out which of these objects is the zando."

The experimenter then placed each object on the machine twice. Afterwards, the child was asked to identify which object was the zando. If the child made an incorrect selection, the demonstrations were repeated. After making his/her selection, the child was allowed to place each object on the machine himself/herself.

Counterfactual Phase

In this phase, the experimenter asked a counterfactual question about each object. For the zando, the experimenter asked, "If this one were not a zando, what would happen if we put it on top of the machine?" For the non-zando, the experimenter asked the opposite question (i.e. "If this one were a zando..."). The order of the questions was counterbalanced across participants. If the child did not respond, the experimenter asked a forced-choice question: "Would the machine play music or not play music?" The experimenter then suggested that the child put the zando on top of the machine one more time to practice singing for Monkey.

Pretense Phase

In this phase, a confederate entered the room and said that she needed to borrow the machine. The confederate removed the machine, zando and non-zando from the room. The experimenter expressed sadness that the confederate had taken the machine before they could surprise Monkey. She then said that she had an idea, and brought out a white wooden box and two colored blocks. The experimenter explained, "I thought we could pretend that this box is my machine and that this block [one of the colored blocks] is a zando and that this block [the other colored block] is not a zando. Then, we could still surprise Monkey!" (Which colored block was the zando as well as the side of presentation of the blocks was counterbalanced across participants.) The experimenter then took Monkey out from underneath the table and asked the child what they should pretend in order to make the pretend machine play music. At this point, the child could place either block onto the machine. If the child did not choose a block, the experimenter asked, "Which of these should we try to pretend to make the machine play music?" Once the child placed a block on top of the machine, the experimenter asked, "What are we pretending now?" If the child did not offer a response, the experimenter asked, "Are we pretending music or no music?" The experimenter then suggested that they try the other block, and repeated the procedure.

After the child had tried each block on the machine, the experimenter said that she had another idea. She reversed the pretend roles of the blocks so that the original pretend zando was now the pretend non-zando and the original pretend non-zando was now the pretend zando. The experimenter then asked, "Now, what should we do to pretend to make the machine play music?" and repeated the same series of questions as before with the new pretend zando and pretend non-zando.

Coding

For the counterfactual and pretense questions, if children's answers indicated that music was playing, such as "Music," "Yes," "Happy Birthday," "It works," or nodding their head, their answer was coded as "music." If children's answers indicated that no music was playing, such as "No Music," "No," "I don't hear anything," "Nothing," or shaking their head, their answer was coded as "no music." If a child was too shy to produce a verbal response, then the experimenter assigned the option of "music" to one of her hands and the option of "no music" to the other hand and asked the child to point to a hand.

For the counterfactual questions, children's answers were considered correct if they could be coded as "no music" for the question about the zando being a non-zando and as "music" for the question about the non-zando being a zando. For the pretense questions, children's answers were considered correct if their answer could be coded as "music" for the pretend zando and "no music" for the pretend non-zando. Finally, in the pretense phase, children's first choice for making the machine go was recorded (i.e. whether or not they chose to put the pretend zando or non-zando on the machine first). An independent coder re-coded 90% of children's performances from videos of the experiment. There was excellent inter-coder agreement on both counterfactual performance (Cohen's $\kappa = 0.94$), and pretense performance (Cohen's $\kappa = 0.94$).

In addition to coding these formal measures, we also coded the degree and elaboration of the child's subsequent spontaneous pretense in the pretend scenario to ensure that children were actually pretending. An independent coder judged the extent of children's involvement in the pretend scenarios from videotapes of the test scenario and coded children's responses as falling into one of three categories, 1) no pretense beyond pretending about the effects of the zando, 2) one or two spontaneous extensions of the pretense or 3) extended spontaneous engagement in the pretense.

Results

Preliminary analyses did not find any effect of gender, question order, side of presentation of the zando or block color on responses to either the counterfactual or pretense questions, so these variables were not considered further.

Counterfactual Phase Performance

Table 4.1: Children's performance in the counterfactual phase of Experiment 1

Number of Correct Answers	0	1	2
Number of Children	10	6	36

Children were given a counterfactual score of 0, 1, or 2 for the number of counterfactual questions they answered correctly, with chance performance being a score of 1 (see Table 4.1). Overall, children's performance on the counterfactual questions was significantly better than chance ($M = 1.5$, $SD = 0.80$, $t(51) = 4.48$, $p < 0.001$).

Children also tended to answer the individual counterfactual questions correctly, saying that if the zando were a non-zando it would not play music when placed on the machine (Exact binomial test: $X = 42$, $N = 52$, $P = 0.5$, $p < 0.001$), and if the non-zando were a zando then it would play music ($X = 36$, $N = 52$, $P = 0.5$, $p < 0.01$). Finally, consistent with previous findings, children's counterfactual performance was correlated with age, $r(50) = 0.33$, $p < 0.05$; However, contrary to some earlier studies, both four year old and three year old children were above chance (Four year olds: $t(25) = 4.47$, $p < 0.001$. Three year olds: $t(25) = 2.087$, $p < 0.05$).

Pretense Phase Performance

Children were given a pretense score between 0 and 4 (with chance performance being a score of 2) for pretending that the appropriate effect followed a block being placed on the pretend machine music playing for the pretend zando and no music playing for the pretend non-zando for both the objects' original roles and their reversed roles (summarized in Table

Table 4.2: Children's performance in the pretense phase of Experiment 1

Number of Correct Answers	0	1	2	3	4
Number of Children	0	3	21	3	25

4.2). In general, children chose to intervene with the pretend zando block in order to cause pretend music ($M = 1.69$, $SD = 0.51$, $t(51) = 9.86$, $p < 0.001$). They did so in both the original (Exact binomial test: $X = 48$, $N = 52$, $P = 0.5$, $p < 0.001$) and reverse (Exact binomial test: $X = 40$, $N = 52$, $P = 0.5$, $p < 0.001$) pretend scenarios. Overall, children said that their interventions in the pretend scenario had causal outcomes consistent with their effects in the real world ($M = 2.96$, $SD = 1.1$, $t(51) = 6.51$, $p < 0.001$).

Children's pretense scores were marginally correlated with their age, $r(50) = 0.23$, $p = 0.1$. However, these scores were significantly correlated with their counterfactual scores, $r(50) = 0.62$, $p < 0.001$. The relationship between pretense and counterfactual scores remains significant even when controlling for age, $r(50) = 0.59$, $p < 0.001$.

Most of the children (71%) spontaneously elaborated the pretend scenario beyond the experimenter's questions and nearly half (44%) engaged in extended pretense, indicating that the children were indeed pretending. There was no difference in the counterfactual performance of children who demonstrated extended or elaborated pretense or simpler pretense. Examples of children's elaborations include extending the celebration of Monkey's birthday, such as having Monkey cover his eyes to receive his surprise, hiding the pretend machine to surprise Monkey, pretending that the box is a cake for Monkey, or that the blocks are presents for Monkey (e.g. the blocks are flowers, or "hotwheels cars"). Children also spontaneously engaged in additional pretense about the machine, for instance continuing to reverse the roles of the pretend blocks after the experiment had ended (e.g. "How about now this one is the zando! Let's try it on the machine!"). Of particular note, a number of children engaged in novel causal interventions during the pretense that were never demonstrated with the real machine, for instance placing both blocks on the box and announcing whether there was music.

Discussion

Overall, children were able to respond correctly to counterfactual questions about a novel real-world causal relationship. In the Counterfactual phase of the experiment, children correctly reasoned that if the zando were not a zando it would not cause music, and if the non-zando were a zando it would cause music. Note that these are classical counterfactuals about possible worlds rather than questions that could be interpreted as future hypotheticals. This finding is especially impressive considering that both objects were not only visible but highlighted in this task, which could have made their actual causal roles salient and difficult to inhibit. Indeed, children had only ever seen the non-zando negatively associated with the effect. Nevertheless, they were able to infer that it would cause the music in the alternative world specified by the counterfactual premise.

Children were also able to maintain and intervene on this newly learned causal structure within a pretend scenario, making inferences consistent with the pretend objects' real-world causal roles, and acting on the pretend causal relationship to bring about a desired pretend outcome. In the pretense phase of the experiment, children's causal inferences about the pretend objects were consistent with the objects' real-world causal roles. When asked to make the pretend machine go, children chose to intervene with the pretend zando block, placing it on the pretend machine. Furthermore, they said that the pretend zando would lead to music, but that we should not pretend music for the pretend non-zando. This is striking because, given that this was a pretend world, children could simply have always pretended that the desirable outcome, playing "Happy Birthday," had occurred.

Finally, children were able to flexibly reassign the causal roles of objects within the pretense. They provided correct answers both about each object's original pretend role and about its reversed pretend role. This indicates that they are able to consider multiple alternative possible worlds.

While a majority of children answered both counterfactual questions correctly, 30% answered at least one counterfactual incorrectly (see Table 1) and a similar number failed to import the causal constraints to their pretense. In these instances, children tended to respond consistently with the object's real-world role, rather than its hypothesized role. In particular, these children would say that the zando block would continue to activate the machine even if it were not a zando, or, in the pretend case, that neither object would cause music.

Moreover, children's performance on the counterfactual questions correlated with their pretense performance, even when age was taken into account. This suggests a link between counterfactual reasoning abilities and pretense, consistent with our theoretical account of these abilities. However, while Experiment 1 provides some evidence for a relationship between pretend play and counterfactual thinking, other explanations are possible. Although the relationship did not depend on age, general cognitive development might account for children's improvement on both tasks. Another possibility is that children who perform poorly on both tasks may have a difficult time inhibiting their real-world knowledge (as suggested by, e.g., Beck, Riggs, & Gorniak, 2009). In this case, children's executive function abilities would correlate with both their counterfactual and pretense success. We test these possibilities in Experiment 2.

4.3 Experiment 2: Relationship of Counterfactual Reasoning, Pretense and Executive Function

In this experiment, we replicated the procedure from Experiment 1 with the addition of a conservation task and an executive function task to gauge children's general cognitive and inhibition skills. We used the classic Piagetian conservation task, which involves rows of pennies that are stretched out and pushed together to see if children understand that the number of pennies did not change despite these physical transformations (Piaget, 1952). The Stroop-like executive function task that we used was the Day-Night task (Carlson & Moses, 2001; Carlson, 2005; Gerstadt, Hong, & Diamond, 1994), which involved cards depicting daytime and nighttime. Children had to say "day" when they saw a nighttime card and "night" when they saw a daytime card. (See Appendix C for details on the administration and scoring of these two tasks.)

Sixty 3- and 4-year-old children were tested in this study (see Appendix C for additional details). The tasks were administrated in one of two orders counterbalanced across subjects: either 1) conservation, 2) executive function, 3) pretense, or 1) pretense, 2) conservation, 3) executive function. An independent coder re-coded 90% of children's performances from videos of the experiment. There was excellent inter-coder agreement on both counterfactual performance (Cohen's $\kappa = 0.92$), and pretense performance (Cohen's $\kappa = 0.92$).

Results

Preliminary analyses did not find an effect of gender, question order, which side of the machine the zando was placed on, or which color block was the pretend zando on responses to either the counterfactual or pretense questions. These variables were not considered further.

Pretense Task Performance

Table 4.3: Children's performance in the counterfactual phase of Experiment 2

Number of Correct Answers	0	1	2
Number of Children	11	12	37

Counterfactual Phase Performance As in Study 1, children's performance on the counterfactual questions was significantly better than chance ($M = 1.43$, $SD = 0.79$, $t(59) = 5.56$, $p < 0.001$; see Table 4.3). In general, children also answered the individual counterfactual questions correctly, saying that if the zando were a non-zando it would not play music when placed on the machine (Exact binomial test: $X = 45$, $N = 60$, $P = 0.5$, $p < 0.001$), and if the non-zando were a zando then it would play music (Exact binomial test: $X = 41$, $N = 60$, $P = 0.5$, $p < 0.01$). Children's counterfactual performance was correlated with age, $r(58) = 0.40, p < 0.01$. Both four year old and three year old children were above chance (Four year olds: $t(27) = 5.01$, $p < 0.001$. Three year olds: $t(31) = 2.48, p < 0.05$).

Pretense Phase Performance Children chose to intervene with the pretend zando block in order to cause pretend music ($M = 1.67$, $SD = 0.51$, $t(59) = 5.06$, $p < 0.001$; see Table 4.4), in both the original (Exact binomial test: $X = 50$, $N = 60$, $P = 0.5$, $p < 0.001$) and reverse (Exact binomial test: $X = 50$, $N = 60$, $P = 0.5$, $p < 0.001$) pretend

Table 4.4: Children's performance in the pretense phase of Experiment 2

Number of Correct Answers	0	1	2	3	4
Number of Children	1	4	12	14	29

scenarios. Overall, children also said that their interventions in the pretend scenario had causal outcomes consistent with the real-world (M = 3.1, SD = 1.05, $t(59) = 6.62$, $p < 0.001$): When they put the pretend zando on the machine, they said that that this led to pretend music (M = 1.42, SD = 0.81, $t(59) = 5.59$, $p < 0.001$), but when they put the pretend non-zando on the machine, they said that this did not lead to pretend music (M = 1.68, SD = 0.60, $t(59) = 4.11$, $p < 0.001$). This was true in both the original (M = 1.55, SD = 0.57, $t(59) = 6.17$, $p < 0.001$) and reverse (M = 1.56, SD = 0.62, $t(59) = 5.60$, $p < 0.001$) pretend scenarios.

Children's pretense scores were significantly correlated with their age, $r(58) = 0.31$, $p < 0.05$, and their counterfactual scores, $r(58) = 0.44$, $p < 0.001$. However, the relationship between pretense and counterfactual scores remains significant even when controlling for age, $r(58) = 0.36$, $p < 0.01$.

Secondary Task Performance

Table 4.5: Children's performance in the counterfactual phase of Experiment 2

Number of Correct Answers	0	1	2	3
Number of Children	11	11	15	

Conservation Task Performance Children were given a score between 0 and 3 for the number of conservation questions they answered correctly. As has been found previously, children's performance on this task varied considerably (M = 1.51, SD = 1.07), and is summarized in Table 4.5. There was no correlation of conservation performance with age, $r(57) = 0.02$, $p = 0.87$, counterfactual score, $r(57) = 0.09$, $p = 0.47$, or pretense score, $r(57) = 0.15$, $p = 0.25$.

Executive Function Task Performance Children received 16 trials and were assigned a proportion of correct answers. Overall, children performed better than chance ($M = 0.61$, $SD = 0.25$, $t(43) = 2.87$, $p < 0.01$). As in previous work (Carlson & Moses, 2001; Carlson, 2005; Gerstadt et al., 1994), there was variance in children's performance, including two children who got zero answers correct, and one child who answered all 16 cards correctly.

Children's performance on the Day/Night task was correlated with their age, $r(42) = 0.33$, $p < 0.05$. There was no correlation between performance on the Day/Night task and counterfactual score, $r(42) = 0.04$, $p = 0.81$, or pretense score, $r(42) = 0.05$, $p = 0.76$. Moreover, the relation between the counterfactual score and pretense score remained significant even when executive function, age and conservation were all controlled for $r(44) = 0.38$, $p < 0.05$.

4.4 General Discussion

In two studies, we found a relation between young children's ability to make counterfactual inferences and their tendency to use causal constraints in their pretend play. In principle, pretend play is unconstrained – children who simply wanted to make Monkey happy could have pretended that any block would make the machine go. In practice, however, children used the demonstrations that they observed to make inferences about situations they had never encountered, such as the counterfactual world in which the non-zando was a zando, or the world in which a plain box really was a "Birthday machine." Moreover, these abilities were specifically related, even controlling for age, general cognitive ability and executive function.

These results suggest a strong link between pretending and counterfactual reasoning abilities. In turn, this supports a relationship between the extended playful exploration enabled by a long period of childhood and the ability to deploy causal models to make counterfactual inferences in a wide-ranging and general way. Although our result itself is only correlational, its specificity does suggest some causal link between the two abilities. It may be that the causal coherence of the children's pretense is simply an epiphenomenon of children's general causal knowledge and counterfactual inference abilities. A more intriguing possibility, however, is that pretend play itself plays a role in the development of causal thinking and learning.

To test this idea we need further experiments. For example, we could test whether engaging children in causal pretense improves their subsequent counterfactual reasoning. Although the extended engagement in the pretend scenarios suggests that children were indeed pretending, we could also test this more systematically by contrasting these scenarios with similar ones that did not involve pretense. We are currently investigating these issues in our lab, as well as looking at how and under what circumstances children generalize more complicated causal relationships.

It is worth emphasizing again that the capacities we see in causal learning and counterfactual thinking are not themselves uniquely human. Both other primates, especially great apes, and birds, especially corvids, show some ability to make causal inferences from models and to use these inferences in ecologically significant contexts, such as foraging or negotiating dominance relations (e.g., Raby et al., 2007; Hare, 2001; Mulcahy & Call, 2006). Moreover, the basic structure and computations of Bayesian learning can be found quite widely in both the visual system and the motor system (Kersten, Mamassian, & Yuille, 2004; Wolpert, 2007). The role of such "forward models" in motor behavior is especially interesting given the expansion of motor areas that accompanied the evolution of human brains, and the evolutionary value of increased motor skills (Barton, 2012). Again, given the small genetic changes and rapid time scale of human evolution it would be surprising if brand-new computations had somehow evolved, but motor system computations may have become more widely available.

The crucial difference, we argue, is in the scope and application of this sort of learning and reasoning. Human children, and the adults they become, do not restrict their counterfactual inferences to the familiar causal relations of foraging and dominance. Instead, this form of reasoning and learning extends to include the unprecedentedly wide and variable range of physical environments where humans live and the even wider range of physical and social environments that they create. Exploratory learning, causal models and counterfactual inferences are particularly helpful for dealing with this kind of variability. This kind of counterfactual exploration stands in tension with the kinds of learning that may be most valuable for swift and computationally and neurally inexpensive action and decision-making, such as those involved in associative learning.

We speculate that non-human animals reserve the more computationally and neurally expensive computations involved in Bayesian learning for specific, highly ecologically valuable functions. These might include dedicated machinery for vision and motor control, or more flexible but still restricted computations that might be used in foraging, tool use or dominance negotiation. They may rely more on more computationally efficient, but less flexible and powerful learning methods such as conditioning or instrumental and trial-and-error learning to acquire broad domain-general and novel information.

Human beings can also rely on these more automatic types of learning, particularly under cognitive load or when responding must be swift (e.g., Sternberg & McClelland, 2012). However, the long period of human childhood gives humans the luxury of applying more powerful but more expensive types of exploratory learning to a wide range of novel information, without regard to their immediate utility.

We might compare this human strategy to the economic strategy whereby companies invest in research divisions that are not immediately profitable, but that allow for flexibility and retooling in the light of changing conditions. Investment in an extended childhood, with its many opportunities for free exploration and causal learning, may have allowed human beings to turn from simply making the same ecological widgets to developing our staggeringly wide variety of strategies for adaptive success.

Chapter 5

Conclusion

Life is a perpetual instruction in cause and effect. – Ralph Waldo Emerson

5.1 Summary

People face challenging causal learning problems on a daily basis, and they have a variety of information they can use to help solve these problems, including directly observed patterns of cause and effect in the physical world, social data from others' causal actions and knowledge, and even imagined causal relationships in pretend scenarios. Having multiple information sources available can enhance our causal reasoning, but it also presents additional challenges. In this dissertation, we explored some of the ways in which children, as well as adults, can integrate these diverse sources of causal knowledge, allowing them to develop and use sophisticated, structured causal theories that support causal predictions, interventions and counterfactual inferences.

Chapter 2 presented a Bayesian rational learner model that jointly infers action segmentation and causal structure, using statistical regularities and temporal cues to causal relationships in an action stream. This model provides a way to begin characterizing both the kinds of information available in the action stream, and what an optimal computational level solution to these inference problems might look like. We also presented a series of experiments investigating how adults use statistical and causal cues to action structure. These experiments demonstrate that adults are able to segment out statistically determined actions, and experience them as coherent, meaningful and most importantly, causal sequences. Additionally, these experiments show that adults are able to extract the correct causal variables from within a longer action sequence, and that they find causal sequences to be more coherent and meaningful than other sequences with equivalent statistical structure. Finally, they provide evidence that adults jointly infer causal structure and action structure, using statistical segmentation cues and cause and effect contingencies together.

In the experiments described in chapter 3, we presented preschool age children with different sequences of actions followed by a specific outcome, and then asked children to try and bring about the same effect themselves. In developing the task, we used a Bayesian computational model to guide our manipulation of the probabilistic evidence, such that the statistical relations between actions and outcomes differed across conditions in ways that supported different causal hypotheses. We then examined which sequences the children produced, and compared children's performance to our model's predictions. We found that children's imitation of sequences that produced the outcome increased, in some cases resulting in production of shorter sequences of actions that the children had never seen performed in isolation. We also demonstrated that children interpret the same statistical evidence differently when it comes from a knowledgeable teacher versus a naïve demonstrator.

These studies suggested that causal learning is informed by both social knowledge and statistical information. Children are sensitive to probabilities, knowledge state, and pedagogical intent when deciding which actions to imitate. These studies also suggest a rational account of "overimitation". In particular, imitating three actions in these studies can be thought of as a kind of overimitation, reproducing parts of a causal sequence that are not actually demonstrably necessary for the effect. These results suggest that this behavior varies depending on the statistics of the data and the probability of various hypotheses concerning them."Overimitation" also varies depending on the social demonstrator. By explicitly representing the contributions of these different sources of evidence and using them to assign probabilities to causal hypotheses, our Bayesian model can predict these behaviors quite precisely.

Finally, in chapter 4 we hypothesized that one important role for childhood pretense may be in the development of causal inference, and argued for a theoretical link between the development of an extended period of immaturity in human evolution and the emergence of powerful and wide-ranging causal learning mechanisms. In a set of two studies, we taught preschool children a novel causal relationship where one novel object made a machine play music and another did not, and then asked them a series of counterfactual questions about this relationship. They were then introduced to pretend versions of the same objects, and tested to see whether they treated the novel causal relationship as holding in an imaginary world. Children's causal inferences about the pretend objects were consistent with the objects' real-world causal roles, demonstrating children's ability to maintain a newly learned causal relationship within a pretend scenario. Finally, children who had earlier answered the explicit counterfactual questions incorrectly were more likely to answer the pretend questions incorrectly, supporting the possibility of a link between counterfactual reasoning abilities and causal pretense. A second study demonstrated that this correlation between children's counterfactual and pretense performance persisted even when age, general cognitive development and executive function were controlled for. These results suggest a strong link between pretending and counterfactual reasoning abilities.

5.2 Implications

Taken together, the studies in this dissertation show how the tools of probabilistic modeling and Bayesian learning can be applied to the social as well as the physical domain, and can help us understand how causal reasoning in the social, physical and imagninary worlds might inform each other. Adults jointly parse fluid motion into meaningful actions and infer the causal relevance of those actions, with both domains being learned simultaneously and mutually, rather than one being learned and then the other. Similarly, when children learn about causes from other people, they appear to integrate their prior hypotheses about pedagogy, cues to informant reliability, and the statistical evidence they observe from people's actions. Children are sensitive to the pedagogical intent of a demonstrator and can use this information to aid their decisions about which of the demonstrator's actions to imitate in order to bring about an effect. Together, these studies demonstrate that causal reasoning and social reasoning are linked, both in the real world and in children's minds.

The rational constructivist approach can help us understand how children resolve, and even benefit from, multiple sources of ambiguous and probabilistic data, and social data, in particular, in order to solve challenging causal learning problems. And because these data are often probabilistic, Bayesian models help us describe the complex, uncertain, joint inferences about the nature of both other people and the world that underlie our ability to learn from others. In fact, we can construe the information we get from people, either in the form of testimony or observable actions, as causal information. These studies suggest that children use covariation evidence to construct abstract causal schemas that they then employ to explain the behavior of both the people and the objects around them. At the same time, the work on causal pretense shows how similarly complex integrations between real-world causal knowledge and imagined causal scenarios allow children to also reason about the conseqences of pretend actions and to intervene appropriately on imagined causal systems.

Finally, the studies on imitation and pedagogy, in particular, suggest that we would be wise to fully consider the social environment when looking at children's physical causal reasoning. The degree of confidence that the social demonstrator has, and the level of authority they convey to the child, might not just socially influence the child to feel pressured to respond in a certain way but also might actually change their inferences about the physical causal events they are observing. In fact, incorporating this social evidence into causal reasoning is a rational response, especially in the face of uncertainty. Therefore, to get a complete picture of how children understand the causal landscape of both the physical and the social worlds we need to understand how they use the entire rich set of data they encounter in the real world. Studies directly manipulating social information, such as how pedagogically the demonstrator is behaving and how much certainty she expresses, integrate the human element into experiments that model causal understanding.

5.3 Remaining Questions and Future Work

The work described in this dissertation sets up an empirical and computational framework for examining how multiple sources of evidence, both real and imagined, can inform our causal inferences. Future research should computationally address how children develop priors about the causes and results of people's behavior and of the social information they provide. What leads children to believe that a person is an expert, and what process guides their assumptions based on that attribution? What are the components of children's pedagogical understanding, and what prior beliefs do children have about the likely causes and effects of pedagogical behavior? How do children integrate data about people's beliefs (via testimony) and actions when making attributions about people's behavior? How do children conceptualize people causing changes in other people's beliefs or actions? What are children's prior beliefs about person-to- person causes, and how would they parse these events? Furthermore, how would they integrate physical causes into those judgments?

We are currently working on a number of natural extensions to the current work on social and causal reasoning, to begin addressing many of these questions. This includes examining the relationship between children's pretend play and their causal reasoning abilities in more detail, as well as beginning to explore more complex cases of imitation, for instance situations where there are multiple demonstrators or where the demonstrators are also learners themselves. Finally, we are starting to investigate the evolutionary origins of social and causal cognition by investigating social and causal reasoning across primate species. I discuss some of these ongoing and future directions in more detail below.

Ongoing and Future Work on Action Segmentation

Like previous computational models of word segmentation, the model presented in chapter 2 assumes that the lowest level of segmentation is already known (or pre-labeled). That is, that there is some sort of motion primitive (equivalent to a syllable or phoneme in speech), that can already be recognized as a coherent unit. Since psychological studies demonstrating human action segmentation have suggested that statistical patterns or features in human motion may correlate with segment boundaries at even the lowest level, we would like to see whether action boundaries can be automatically detected directly from video, without pre-existing knowledge of low-level motion units. In this line of work, we are developing a series of computational models that make very few representational assumptions about what is observed when watching videos of human action, in order to explore the amount of action structure that can be inferred from just low-level changes in pixel values, without knowledge of human body structure, higher level goals and intentions, or even foreground/background distinctions (Buchsbaum, Canini, & Griffiths, 2011). To the extent that these models correspond to human segmentation judgments, and correctly recognizes actions, we will know that there are cues in surface level image changes that can be used to both segment and identify human behavior.

Just as it is important to explore how low-level motion cues might contribute to action segmentation in a bottom-up fashion, we would also like to understand how higher-level social information might contribute to action parsing. How might knowledge of an actor's goals and intentions influence segmentation? This is an especially interesting question in the context of hierarchical goal structures. Recent work suggests that people (e.g., Zacks & Tversky, 2001; Zacks, Tversky, & Iyer, 2001; Hard et al., 2006; Meyer, Baldwin, & Sage, 2011), and perhaps other apes (Byrne & Russon, 1998; Byrne, 1999, 2003 but also see Conway & Christiansen, 2001), naturally organize events into increasingly abstract hierarchical relationships, based on the underlying goals of the actors. Once again, there are intuitive parallels to the language domain, where phonemes are composed into words, which are in turn composed into phrases and sentences. An intriguing possibility is to see whether probabilistic models of phrase structure (e.g., Johnson, Griffiths, & Goldwater, 2007) could also be adapted to the action domain. Similarly, exploring whether there are garden path effects in action parsing akin to those that can occur in language (e.g., Levy, 2011) could help us better understand how action structure and goal structure are inferred.

Ongoing and Future Work on Causal Imitation and Social-Causal Reasoning

Our results in chapter 3 established that children interpret the same statistical evidence differently when it comes from a knowledgeable teacher versus a naïve demonstrator. This ability to read others' intentions and pedagogical cues is often thought to be foundational for human culture (e.g., Tomasello, Carpenter, Call, Behne, & Moll, 2005; Csibra & Gergely, 2006, 2009). Explaining species differences between children and great apes (e.g., Horner & Whiten, 2005), and discovering whether there are cross-cultural universals in causal imitation (Nielsen & Tomaselli, 2010), could help us determine the evolutionary origins of uniquely human cognition. Recent work suggests that chimpanzees possess knowledge about object properties such as solidity and connection (Seed, Hanus, & Call, 2011), and also about intentional action (Call, Hare, Carpenter, & Tomasello, 2004). Interestingly, in some of this work, chimpanzees used mechanical information to disregard unnecessary actions performed by a demonstrator (such as tapping the top of a box before using a stick to push out a reward), and only copied the effective action (Horner & Whiten, 2005; Whiten, McGuigan, Marshall-Pescini, & Hopper, 2009). In contrast, as discussed in chapter 3, children usually copied all of the purposeful actions that were demonstrated.

Since children are known to be remarkably sensitive to the goals and intentions underlying the actions they observe, could some of the previously observed differences between children and non-human primate's causal inferences be the result of differences in understanding of the demonstrator's intentions and of the pedagogical context? In this set of studies, we are looking at whether differing social and physical expectations influence causal inferences from otherwise identical sequences of action. We will compare children's performance to chimpanzees, who are able to learn from causal demonstrations (Horner & Whiten, 2005;

Whiten et al., 2009) but whose understanding of mental states such as goals and intentions remains controversial (Penn & Povinelli, 2007; Penn, Holyoak, & Povinelli, 2008), and for whom there is little to no evidence of either understanding or expecting pedagogy. We will also look at capuchin monkeys, whose ability to learn from causal imitation may be comparatively limited (Custance, Whiten, & Fredman, 1999).

It is possible to make different causal inferences after observing the very same demonstrated actions and outcomes, depending on the learner's social assumptions about the demonstrator. For instance, the learner might make no social assumptions at all – they may simply see the actions and their resulting outcomes as patterns of contingencies in the world, without any reference to mentalistic entities such as goals or intentions. It is of course possible to learn quite a bit from such observations (Byrne, 1999, 2003). A different set of social assumptions the learner could make is that they are observing an intentional agent performing goal-directed actions that are intended to bring about their outcomes. Finally, the learner might believe that the demonstrator is not just a goal-directed agent acting for their own gain, but is actually a helpful teacher, demonstrating the causal relationship for the learner's benefit. The key feature of this final model is that it captures the fact that the learner and the teacher have mental representations of each others' mental states, modeling the intuition "I know that you know what I know".

Following previous work, we can formally represent these different social assumptions using different sampling models (e.g., Xu & Tenenbaum, 2007; Shafto & Goodman, 2008), such as the pedagogical sampling model used in chapter 4. By comparing the performances of monkeys, chimpanzees and children in our studies to the predictions of these different sampling models we can see if the inferences the different species make correspond to different models or model assumptions, and therefore different social expectations. We can also see if children, and perhaps other primates, change which sampling model they use based on social cues suggesting that different sampling assumptions may be appropriate.

Like the teacher and the learner in the pedagogical sampling model, reasoning about other agents may require learners to have probabilistic models not just of the world but also of each other. These models can themselves be true or false and depend on evidence (e.g., the teacher has a generative model of the learner that is dynamic and updated by observing as the learner updates their own knowledge). Similarly the learner may have such a model of the teacher. This type of recursive mental representation has been hypothesized as being particularly important in developing shared attention and working towards cooperative goals (Tomasello et al., 2005), and there is some existing evidence suggesting that while chimpanzees and other apes may have some capacity to cooperate (Warneken & Tomasello, 2006) that they do not represent shared intentions and goals with the same sophistication as human children (Warneken & Tomasello, 2009). Whether this is due to an inability to represent recursive mental states or to other factors (such as a strong prior bias towards competition over cooperation) is an interesting empirical question that modeling can help us explore.

Previous Bayesian models of theory of mind have looked at how one agent can predict another's mental states from observations of their actions, by using an inverse planning approach that assumes that the observed actions are the result of a rational decision process on the part of that agent (this decision process is modeled using a standard machine learning approach called a Partially Observable Markov Decision Process) (Baker, Goodman, & Tenenbaum, 2008; Baker, Tenenbaum, & Saxe, 2009; Baker, Saxe, & Tenenbaum, 2011). Here, we can extend this type of model to incorporate a recursive element, allowing one agent to base their own actions on the mental states of another. Essentially, we can integrate the pedagogical sampling and Bayesian theory of mind modeling approaches, in order to try and explain this type of complex social reasoning.

In chapter 3 we focused on how children use social demonstrations in their causal inferences. In an ongoing set of studies we are investigating the influence of a different type of social information: verbal testimony (e.g., Birch et al., 2008)(Harris & Corriveau, 2011). What can we learn from other people's causal statements about the world, and what might the world tell us about the reliability of those statements – and of the individual making these statements – i.e., a social informant. This set of studies investigates how both children (Bridgers, Buchsbaum, Seiver, Gopnik, & Griffiths, In prep) and adults (Buchsbaum et al., 2013) reconcile a conflict between the verbal testimony provided by a social informant and their own causal observations of the physical world. We ask whether people are sensitive not only to an informant's statements about the world, but also her expressed level of certainty, her previous accuracy, and perhaps her apparent self-knowledge? how accurately she conveys her own certainty – and how this might influence people's future trust in the informant. It can be difficult to tease apart the contributions of all these variables simply by observing people's causal judgments, so we developed a computational model of how these different cues contribute to a rational causal inference. This work is ongoing, but preliminary results indicate that, while both adults and children are sensitive to the informant's certainty and accuracy, perhaps only adults are sensitive to the informant's self-knowledge.

The previous cases looked at how children can combine information from a single informant with other sources, but of course children often receive information from multiple, sometimes conflicting informants, and sometimes these informants are learners themselves. Examining how children (and adults) reconcile conflicting informants is important for investigating the mechanisms of cultural evolution and understanding how new ideas spread through populations (or in determining when they should and when they should not catch on). The spread of a new idea requires individuals to intially go against the majority – when is it rational for them to do so?

Past research has shown children consider a variety of factors when learning from others, including consensus. Corriveau, Fusaro, and Harris (2009) found that in an object labeling task, children trust responses that receive majority support, and they concluded that children prefer members of a majority as social informants. However, it is possible that children prefer majority members only in domains that rely strongly on socially constructed norms, such as object labeling, where non-social information is unavailable. We formalized this prediction using a rational model of learning from testimony across tasks, and are comparing our model's predictions to children's responses in object labeling and causal learning tasks (Hu, Buchsbaum, Xu, & Griffiths, 2013). A second ongoing project looks at whether children and adults treat multiple demonstrators as independent pieces of data, or whether they assume that these demonstrators themselves learned from others, by comparing their performance to rational models capturing these two different possible assumptions (Whalen, Buchsbaum, & Griffiths, 2013).

Ongoing and Future Work on Causal Pretense

The experiment's described in chapter 4 examined children's use of a very simple real-world causal relationship in pretense. We are currently running another study looking at how children reason about more sophisticated causal structures, both counterfactually and in their pretend play. In this study, we teach children a complex causal structure involving four different variables (e.g., the sun comes up, which makes the rooster crow and the birds chirp, and the rooster crowing wakes up the farmer). They are then asked to either pretend, or to counterfactually suppose, that one of the values in the causal structure has been fixed, and their beliefs about what is now true (either counterfactually or in the pretense) about the other variables' values is probed.

Preliminary results suggest that children's counterfactual inferences about this complex structure parallel their inferences about pretense, and both are significantly accurate. Interestingly, children are more likely to make "backtracking" counterfactual inferences when explicitly asked to reason counterfactually. In contrast, they are significantly more likely to treat the "fixed" variable as an intervention ("non-backtracking") when asked to pretend its value. Recent theoretical and computational work (Rips, 2009; Lucas & Kemp, 2012; Chater & Oaksford, in press) suggests that "backtracking" and "non-backtracking" counterfactuals may require different computational models and inference mechanisms, and exploring how these models align with children's inferences during pretense versus explicit counterfactual reasoning may be fruitful area for future research.

If, as we suggest, one role of childhood pretense is to allow children to reason about alternative, and even impossible causal structures, then they should be able to not only imagine interventions on real causal systems (as in the experiments in chapter 4), they should also be able to imagine worlds with causal relationships that differ from the real world. In this case, children should correctly predict the causal consequences of these imaginary causal relations, even when they contradict what would happen in the real world.

This ongoing set of experiments evaluates whether children can override their knowledge of an actual causal structure to imagine a new structure and predict the outcomes of pretend interventions on that structure. To do this, children are introduced to a novel 3-variable causal system (i.e., a gear toy with a switch [S] and two interlocking gears [A and B]) that can be set to operate as one of two causal chains or a common cause structure. Children are taught the actual structure, observe interventions, and intervene on the variables themselves. After demonstrating their understanding, children are asked to pretend that the toy works a different way, and are provided with a new causal structure. Given this new structure, children are prompted to perform imagined interventions on the pretend causal system ("Let's pretend that we take off B.") and asked to predict the outcome for the other variables ("What are we pretending is happening to A?"). Preliminary results indicate that children are able to use the pretend causal structure to predict the outcome of imagined interventions, even when they had never observed the new causal structure outside of the pretense activity.

5.4 Concluding Remarks

The studies in this dissertation begin to show how we can move beyond basic laboratory problems, like determining the causal structure of blicket detectors, to more complex inferences that more closely mirror the real (and even the imaginary) world. The probabilistic models approach can be applied to real and ecologically significant kinds of conceptual change. It sheds new light on classic topics in cognitive development such as the nature of imitation and the purpose of pretend play. Instead of looking at how children evaluate individual or isolated events, we can more appropriately study how children learn in and from the complex social-physical environment that makes up the world around them.

[1] The project *Segmenting and Recognizing Human Action using Low-level Video Features* is a collaboration with Kevin Canini and Thomas L. Griffiths. The project *The Role of Testimony and Informant Knowledge in Children's Causal Inferences* is a collaboration with Sophie Bridgers, Elizabeth Seiver, Andrew Whalen, Alison Gopnik and Thomas L. Griffiths. The project *The Role of Intentions and Physical Knowledge in the Origins of Causal Reasoning* is a collaboration with Amanda Seed, Thomas L. Griffiths and Alison Gopnik. The project *Cultural Transmission and Learning from Multiple Informants* is a collaboration with Andrew Whalen, Jane Hu, Thomas L. Griffiths and Fei Xu. The project *Children's complex causal reasoning in pretend play* is a collaboration with Sophie Bridgers, Caren M. Walker, and Alison Gopnik. The project *Imagining Interventions* is a collaboration with Caren M. Walker and Alison Gopnik.

References

Aarts, E., & Korst, J. (1989). *Simulated annealing and boltzmann machines: A stochastic approach to combinatorial optimization and neural computing.* New York: Wiley.

Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983* (pp. 1–198). Berlin: Springer.

Amsel, E., & Smalley, J. D. (2000). Beyond really and truly: Children's counterfactual thinking about pretend and possible worlds. In P. Mitchell & K. J. Riggs (Eds.), *Children's reasoning and the mind* (p. 121-147). Hove, UK: Psychology Press.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321-324.

Baker, C. L., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory-based social goal inference. *Proceedings of the 30th Annual Conference of the Cognitive Science Society.*

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proc. of the 34th Annual Conference of the Cognitive Science Society.*

Baker, C. L., Tenenbaum, J. B., & Saxe, R. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Baldwin, D. A., Andersson, A., Saffran, J. R., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*(3), 1382-1407.

Baldwin, D. A., Baird, J., Saylor, M., & Clark, A. (2001). Infants parse dynamic human action. *Child Development*, *72*(3), 708-717.

Barton, R. A. (2012). Embodied cognitive evolution and the cerebellum. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2097–2107.

Beck, S. R., Riggs, K. J., & Gorniak, S. L. (2009). Relating developments in children's counterfactual thinking and executive functions. *Thinking & Reasoning*, *15*(4), 337–354.

Bekoff, M., & Byers, J. A. (Eds.). (1998). *Animal play: evolutionary, comparative and ecological perspectives.* Cambridge University Press.

Birch, S. A., Vauthier, S. A., & Bloom, P. (2008). Three- and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, *107*(3), 1018-1034.

Bonawitz, E. B., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A. N., Woodward, J., & Schulz, L. E. (2010). Just do it? investigating the gap between prediction and action in toddlers' causal inferences. *Cognition*, *115*(1), 104-107.

Bonawitz, E. B., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E. S., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, *120*(3), 322-330.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298-304.

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1-3), 71–105.

Bridgers, S., Buchsbaum, D., Seiver, E., Gopnik, A., & Griffiths, T. L. (In prep). Which block is better at making the machine go? How children balance their trust in an informant vs. the data.

Brugger, A., Lariviere, L. A., Mumme, D. L., & Bushnell, E. W. (2007). Doing the right thing: Infants' selection of actions to imitate from observed event sequences. *Child Development*, *78*(3), 806-824.

Buchsbaum, D., Bridgers, S., Weisberg, D. S., & Gopnik, A. (2012). The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2202-2212.

Buchsbaum, D., Bridgers, S., Whalen, A., Seiver, E., Griffiths, T. L., & Gopnik, A. (2013). Do I know that you know what you know? Modeling testimony in causal inference. *Proc. of the 34th Annual Conference of the Cognitive Science Society*.

Buchsbaum, D., Canini, K. R., & Griffiths, T. L. (2011). Segmenting and recognizing human action using low-level video features. *Proc. of the 33rd Annual Conference of the Cognitive Science Society*.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*(3), 331-340.

Buchsbaum, D., Griffiths, T. L., Gopnik, A., & Baldwin, D. (2009). Learning from actions and their consequences: Inferring causal variables from continuous sequences of human action. *Proc. of the 31st Annual Conference of the Cognitive Science Society*.

Buchsbaum, D., Griffiths, T. L., Gopnik, A., & Baldwin, D. (In prep). Inferring action structure and causal relationships in continuous sequences of human action.

Buchsbaum, D., Seiver, E., Bridgers, S., & Gopnik, A. (2012). Learning about causes from people, and about people from causes: Statistical inference and social causal reasoning. In F. Xu & T. Kushnir (Eds.), *Rational constructivism* (Vol. 43, p. 125-160).

Butler, L. P., & Markman, E. M. (2012). Preschoolers use intentional and pedagogical cues to guide inductive inferences and exploration. *Child Development*, *83*(4), 1416-1428.

Byrne, R. W. (1995). *The thinking ape: Evolutionary origins of intelligence* (Vol. 129). Oxford University Press Oxford.

Byrne, R. W. (1999). Imitation without intentionality. using string parsing to copy the organization of behaviour. *Animal Cognition*, *2*(2), 63-72.

Byrne, R. W. (2003). Imitation as behaviour parsing. *Philosophical Transactions: Biological Sciences*, *358*(1431), 529-536.

Byrne, R. W., & Russon, A. E. (1998). Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences*, *21*(5), 667-721.

Byrne, R. W., & Whiten, A. (1989). *Machiavellian intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans.* Oxford University Press, USA.

Call, J., Hare, B., Carpenter, M., & Tomasello, M. (2004). 'Unwilling' versus 'unable': chimpanzees' understanding of human intentional action. *Developmental Science*, *7*(4), 488-498.

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. *Developmental neuropsychology*, *28*(2), 595-616.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child development*, *72*(4), 1032–1053.

Carpenter, M., Call, J., & Tomasello, M. (2002). Understanding prior intentions enables two–year–olds to imitatively learn a complex task. *Child Development*, *73*(5), 1431–1441.

Chater, N., & Oaksford, M. (in press). Programs as causal models: Speculations on mental programs and mental representation. *Cognition*.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(6), 367-405.

Conway, C. M., & Christiansen, M. H. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, *5*(12), 539-546.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers exploratory play. *Cognition*, *120*(3), 341–349.

Corriveau, K., Fusaro, M., & Harris, P. L. (2009). Going with the flow preschoolers prefer nondissenters as informants. *Psychological Science*, *20*(3), 372-377.

Corriveau, K., Meints, K., & Harris, P. L. (2009). Early tracking of informant accuracy and inaccuracy. *British Journal of Developmental Psychology*, *27*(2), 331–342.

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. *Processes of change in brain and cognitive development. Attention and performance*, *21*, 249-274.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends In Cognitive Sciences*, *13*(4), 148-153.

Custance, D., Whiten, A., & Fredman, T. (1999). Social learning of an artificial fruit task in capuchin monkeys (cebus apella). *Journal of Comparative Psychology*, *113*(1), 13-23.

Denison, S., Bonawitz, E. B., Gopnik, A., & Griffiths, T. L. (2009). Preschoolers sample from probability distributions. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate viual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429-433.

Fein, G. G. (1981). Pretend play in childhood: An integrative review. *Child development*, *52*(4), 1095–1118.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209-230.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structures from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*(3), 458-467.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*, 755.

Gergely, G., Egyed, K., & Király, I. (2007). On pedagogy. *Developmental Science*, *10*(1), 139-146.

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: performance of children 3 1/2-7 years old on a stroop-like day-night test. *Cognition*, *53*(2), 129-153.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* Suffolk, UK: Chapman and Hall.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21-54.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, *4*(5), 178-186.

Goodman, N. D., Mansinghka, V. K., & Tenenbaum, J. B. (2007). Learning grounded causal models. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*.

Gopnik, A. (2009). *The philosophical baby: What children's minds tell us about truth, love & the meaning of life.* New York: Farrar, Straus and Giroux.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*(1), 1-31.

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

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*(5), 620-629.

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085-1108.

Gosso, Y., Otta, E., Morais, M., Ribeiro, F. L., & Bussab, V. R. (2005). Play in hunter-gatherer society. In A. D. Pellegrini & P. K. Smith (Eds.), *The nature of play: Great apes and humans* (pp. 213–253). New York: Guilford.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? *Psychological Science*, *18*(3), 254-260.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American psychologist*, *53*(1), 5.

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., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling.* Cambridge: Cambridge University Press.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science*, *35*(8), 1407-1455.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 354-384.

Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological review*, *116*(4), 661–716.

Gweon, H., & Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, *332*(6037), 1524–1524.

Hamlin, J. K., Hallinan, E. V., & Woodward, A. L. (2008). Do as I do: 7-month-old infants selectively reproduce others' goals. *Developmental Science*, *11*(4), 487-494.

Hard, B. M., Recchia, G., & Tversky, B. (2011). The shape of action. *Journal of Experimental Psychology: General*, *140*(4), 586-604.

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory and Cognition*, *34*(6), 1221-1235.

Hare, B. (2001). Can competitive paradigms increase the validity of experiments on primate social cognition? *Animal Cognition*, *4*(3-4), 269–280.

Harnick, F. S. (1978). The relationship between ability level and task difficulty in producing imitation in infants. *Child Development*, *49*(1), 209-212.

Harris, P. L. (2000). *The work of the imagination: Understanding childrens worlds.* Oxford: Blackwell.

Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1179-1187.

Harris, P. L., German, T., & Mills, P. (1996). Children's use of counterfactual thinking in causal reasoning. *Cognition*, *61*(3), 233-259.

Harris, P. L., Kavanaugh, R. D., Wellman, H. M., & Hickling, A. K. (1993). Young children's understanding of pretense. *Monographs of the Society for Research in Child Development*, *58*(1), 1-107.

Hawkes, K., Kim, P. S., Kennedy, B., Bohlender, R., & Hawks, J. (2011). A reappraisal of grandmothering and natural selection. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1714), 1936-1938.

Hay, J. F., Pelucchi, B., Graf Estes, K., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93-106.

Hespos, S. J., Saylor, M. M., & Grossman, S. R. (2009). Infants' ability to parse continuous actions. *Developmental Psychology*, *45*(2), 575-585.

Hoerl, C., McCormack, T., & Beck, S. R. (Eds.). (2011). *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology.* Oxford University Press.

Horner, V., & Whiten, A. (2005). Causal knowledge and imitation/emulation switching in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Animal cognition*, *8*(3), 164–181.

Hrdy, S. B. (2009). *Mothers and others: The evolutionary origins of mutual understanding.* Belknap Press.

Hu, J. C., Buchsbaum, D., Xu, F., & Griffiths, T. L. (2013). When does the majority rule? preschoolers' trust in majority informants varies by domain. *Proc. of the 35th Annual Conference of the Cognitive Science Society*.

Jablonka, E., Ginsburg, S., & Dor, D. (2012). The co-evolution of language and emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2152–2159.

Jaswal, V. K. (2006). Preschoolers favor the creator's label when reasoning about an artifact's function. *Cognition*, *99*(3), B83-B92.

Jaswal, V. K., Croft, A. C., Setia, A. R., & Cole, C. A. (2010). Young children have a specific, highly robust bias to trust testimony. *Psychological Science*, *21*(10), 1541-1547.

Jensvold, M., & Fouts, R. (1993). Imaginary play in chimpanzees (pan troglodytes). *Human Evolution*, *8*(3), 217–227.

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19.*

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.*, *55*, 271–304.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence of a domain general learning mechanism. *Cognition*, *83*(2), B35-B42.

Koenig, M. A., Clement, F., & Harris, P. L. (2004). Trust in testimony: Children's use of true and false statements. *Psychological Science*, *15*(10), 694-698.

Koenig, M. A., & Harris, P. L. (2005). The role of social cognition in early trust. *Trends in Cognitive Sciences*, *9*(10), 457–459.

Koenig, M. A., & Jaswal, V. K. (2011). Characterizing childrens expectations about expertise and incompetence: Halo or pitchfork effects? *Child development*, *82*(5), 1634–1647.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, *12*(2), 72-79.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science*, *16*(9), 678-683.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology*, *43*(1), 186-196.

Kushnir, T., Vredenburgh, C., & Schneider, L. A. (2013). "Who can help me fix this toy?" the distinction between causal knowledge and word knowledge guides preschoolers' selective requests for information. *Developmental Psychology*, *49*(3), 446-453.

Kushnir, T., Wellman, H. M., & Gelman, S. A. (2008). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition*, *107*(3), 1084–1092.

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of other people. *Psychological Science*, *21*(8), 1134-1140.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge university press.

Leslie, A. M. (1987). Pretense and representation: The origins of theory of mind.. *Psychological review*, *94*(4), 412–426.

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.*.

Lewis, D. K. (1973). *Counterfactuals* (Vol. 1). Harvard University Press Cambridge.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323-1329.

Lillard, A. (2001). Pretend play as twin earth: A social-cognitive analysis. *Developmental Review*, *21*(4), 495–531.

Lillard, A., Lerner, M. D., Hopkins, E. J., Dore, R. A., Smith, E. D., & Palmquist, C. M. (2013). The impact of pretend play on children's development: A review of the evidence. *Psychological Bulletin*, *139*(1), 1-34.

Lucas, C. G., Gopnik, A., & Griffiths, T. L. (2010). Developmental differences in learning the forms of causal relationships.

Lucas, C. G., & Kemp, C. (2012). A unified theory of counterfactual reasoning. *Proc. of the 35th Annual Conference of the Cognitive Science Society*.

Lyons, D. E., Damrosch, D. H., Lin, J. K., Macris, D. M., & Keil, F. C. (2011). The scope and limits of overimitation in the transmission of artefact culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1567), 1158-1167.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, *104*(50), 19751-19756.

Ma, L., & Xu, F. (2011). Young children's use of statistical sampling evidence to infer the subjectivity of preferences. *Cognition*, *120*(3).

McGuigan, N., & Whiten, A. (2009). Emulation and "overemulation" in the social learning of causally opaque versus causally transparent tool use by 23- and 30-month-olds. *Journal of Experimental Child Psychology*, *104*(4), 367-381.

McGuigan, N., Whiten, A., Flynn, E., & Horner, V. (2007). Imitation of causally opaque versus causally transparent tool use by 3- and 5-year-old children. *Cognitive Development*, *22*(3), 353-364.

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & cognition*, *37*(3), 249–264.

Meltzoff, A. N. (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology*, *24*(4), 470-476.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*(5), 838-850.

Meltzoff, A. N., Waismeyer, A., & Gopnik, A. (2012). Learning about causes from people: observational causal learning in 24-month-old infants. *Developmental Psychology*, *48*(5), 1215-1228.

Meyer, M., & Baldwin, D. A. (2011). Statistical learning of action: The role of conditional probability. *Learning and Behavior*, *39*(4), 383-398.

Meyer, M., Baldwin, D. A., & Sage, K. D. (2011). Assessing young children's hierarchical action segmentation. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.

Meyer, M., DeCamp, P., Hard, B. M., & Baldwin, D. A. (2010). Assessing behavioral and computational approaches to naturalistic action segmentation. *Proc. of the 33nd Annual Conference of the Cognitive Science Society*.

Michotte, A. (1963). *The perception of causality.* Basic Books.

Mirman, D., Magnuson, J. S., Graf Estes, K., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition*, *108*(1), 271-280.

Mulcahy, N. J., & Call, J. (2006). Apes save tools for future use. *Science*, *312*(5776), 1038–1040.

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology*, *35*(12), 847-862.

Nielsen, M., & Tomaselli, K. (2010). Overimitation in kalahari bushman children and the origins of human cultural cognition. *Psychological Science*, *21*(5), 729–736.

Pasquini, E. S., Corriveau, K. H., Koenig, M. A., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), 1216–1226.

Pearl, J. (2000). *Causality: Models, reasoning and inference.* Cambridge, UK: Cambridge University Press.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674-685.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109-178.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 731.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., & Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, *120*(3), 302-321.

Piaget, J. (1952). *The childs conception of numbers.* New York: Humanities Press.

Raby, C. R., Alexis, D. M., Dickinson, A., & Clayton, N. S. (2007). Planning for the future by western scrub-jays. *Nature*, *445*(7130), 919–921.

Riggs, K. J., Peterson, D. M., Robinson, E. J., & Mitchell, P. (1998). Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cognitive Development*, *13*(1), 73–90.

Rips, L. J. (2009). Two causal theories of counterfactual conditionals. *Cognitive Science*, 1-47.

Roseberry, S., Richie, R., Hirsh-Pasek, K., Golinkoff, R. M., & Shipley, T. F. (2011). Babies catch a break : 7- to 9-month-olds track statistical probabilities in continuous dynamic events. *Psychological Science*, *22*(11), 1422-1424.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606-621.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101-105.

Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science*, *16*(12), 995-1001.

Saxe, R., Tzelnic, T., & Carey, S. (2007). Knowing who dunnit: Infants identify the causal agent in an unseen causal interaction. *Developmental Psychology*, *43*(1), 149-158.

Saylor, M. M., Baldwin, D. A., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition andDevelopment*, *8*(1), 113-128.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045-1050.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared make your tummy ache? naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*, *43*(5), 1124-1139.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40*(2), 162-176.

Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322-332.

Schulz, L. E., Hooppell, C., & Jenkins, A. C. (2008). Judicious imitation: Children differentially imitate deterministically and probabilistically effective actions. *Child Development*, *79*(2), 395-410.

Seed, A., Hanus, D., & Call, J. (2011). Causal knowledge in corvids, primates and children: More than meets the eye? In T. McCormack, C. Hoerl, & S. A. Butterfill (Eds.), *Tool use and causal cognition.* Oxford University Press.

Senju, A., Csibra, G., & Johnson, M. H. (2008). Understanding the referential nature of looking: Infants' preference for object-directed gaze. *Cognition*, *108*, 303-319.

Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. *Proceedings of the 30th annual conference of the Cognitive Science Society*.

Shafto, P., Goodman, N. D., Gerstle, B., & Ladusaw, F. (2010). Prior expectations in pedagogical situations. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Sharon, T., & Wynn, K. (1998). Individuation of actions from continuous motion. *Psychological Science*, *9*(5), 357-362.

Singer, D. G., & Singer, J. L. (1990). *The house of make-believe: Children's play and the developing imagination.* Cambridge, MA: Harvard University Press.

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

Smith, T. M., Tafforeau, P., Reid, D. J., Pouech, J., Lazzari, V., Zermeno, J. P., ... others (2010). Dental evidence for ontogenetic differences between modern humans and neanderthals. *Proceedings of the National Academy of Sciences*, *107*(49), 20923–20928.

Sobel, D. M., & Corriveau, K. H. (2010). Children monitor individuals' expertise for word learning. *Child Development*, *81*(2), 669-679.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*, *42*(6), 1103-1115.

Sobel, D. M., Sommerville, J. A., Travers, L. V., Blumenthal, E. J., & Stoddard, E. (2009). The role of probability and intentionality in preschoolers" causal generalizations. *Journal of Cognition and Development*, *10*(4), 262-284.

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

Spirtes, P., Glymour, C., & Schienes, R. (2001). *Causation prediction and search* (2nd ed.). Cambridge, MA: MIT Press.

Sterelny, K. (2012). Language, gesture, skill: the co-evolutionary foundations of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2141–2151.

Sternberg, D. A., & McClelland, J. L. (2012). Two mechanisms of human contingency learning. *Psychological science*, *23*(1), 59–68.

Suddendorf, T., & Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. *Genetic, social, and general psychology monographs*, *123*(2), 133–167.

Swallow, K. M., & Zacks, J. M. (2008). Sequences learned without awareness can orient attention during the perception of human activity. *Psychonomic Bulletin & Review*, *15*(1), 116-122.

Taylor, M. (1999). *Imaginary companions and the children who create them.* Oxford University Press New York.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Thelen, E. S., & Smith, L. B. (1996). *Dynamic systems approach to the development of cognition and action.* MIT press.

Thimbleby, H. (2003). The directed chinese postman problem. *Software: Practice and Experience*, *33*(11), 1081–1096.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675-91.

van Aardenne-Ehrenfest, T., & de Bruijn, N. G. (1951). Circuits and trees in oriented linear graphs. *Simon Stevin: Wis-en Natuurkundig Tijdschrift*, *28*, 203.

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics*, *27*(3), 351-372.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*(1), 101-118.

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the information given causal models in learning and reasoning. *Current Directions in Psychological Science*, *15*(6), 307–311.

Want, S. C., & Harris, P. L. (2001). Learning from other people's mistakes: Causal understanding in learning to use a tool. *Child Development*, *72*(2), 431-443.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301-1303.

Warneken, F., & Tomasello, M. (2009). Varieties of altruism in children and chimpanzees. *Trends in Cognitive Sciences*, *13*(9), 397-402.

Weisbecker, V., & Goswami, A. (2010). Brain size, life history, and metabolism at the marsupial/placental dichotomy. *Proceedings of the National Academy of Sciences*, *107*(37), 16216–16221.

Weisberg, D. S. (2013). Distinguishing imagination from reality. In M. Taylor (Ed.), *The oxford handbook of the development of imagination.* Oxford University Press, USA.

Weisberg, D. S., & Gopnik, A. (in press). Pretense, counterfactuals, and Bayesian causal models: Why what isn't real really matters. *Cognitive Science*.

Wellman, H. M., Phillips, A. T., & Rodriguez, T. (2000). Young children's understanding of perception, desire, and emotion. *Child development*, *71*(4), 895–912.

Whalen, A., Buchsbaum, D., & Griffiths, T. L. (2013). How do you know that? sensitivity to statistical dependency in social learning. *Proceedings of the 36th annual meeting of the Cognitive Science Society*.

Whiten, A., Custance, D. M., Gomez, J. C., Teixidor, P., & Bard, K. A. (1996). Imitative learning of artificial fruit processing in children (homo sapiens) and chimpanzees (pan troglodytes). *Journal of Comparative Psychology*, *110*(1), 3-14.

Whiten, A., McGuigan, N., Marshall-Pescini, S., & Hopper, L. M. (2009). Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2417-2428.

Williamson, R. A., & Markman, E. M. (2006). Precision of imitation as a function of preschoolers' understanding of the goal of the demonstration. *Developmental Psychology*, *42*(4), 723-731.

Williamson, R. A., Meltzoff, A. N., & Markman, E. M. (2008). Prior experiences and perceived efficacy influence 3-year-olds' imitation. *Developmental Psychology*, *44*(1), 275-285.

Wolpert, D. M. (2007). Probabilistic models in human sensorimotor control. *Human movement science*, *26*(4), 511–524.

Woodward, A. L., & Needham, A. (2008). *Learning and the infant mind.* Oxford University Press, USA.

Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action in context. *Psychological Science*, *11*(1), 73-77.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford: Oxford University Press.

Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2011). Infants learn about objects from statistics and peopl. *Developmental Psychology*, *47*(5), 1220-1229.

Xu, F., & Kushnir, T. (Eds.). (2012). *Rational constructivism in cognitive development* (Vol. 43). Academic Press.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2).

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, *28*, 979-1008.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., ... Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651-655.

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*(2), 201-216.

Zacks, J. M., Speer, N. K., Swallow, K. M., & Maley, C. J. (2010). The brain's cutting-room floor: segmentation of narrative cinema. *Frontiers in Human Neuroscience*, *4*(1-15).

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3–21.

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, *130*(1), 29–58.

Zmyj, N., Buttelmann, D., Carpenter, M., & Daum, M. (2010). The reliability of a model influences 14-month-olds' imitation. *Journal of Experimental Child Psychology*, *106*(4), 208-220.

Appendix A

Chapter 2: Supplementary Material

A.1 Action Segmentation Model Details

We created a Bayesian rational learner model that jointly infers action segmentation and causal structure, using statistical regularities and temporal cues to causal relationships in an action stream. Our model adapts the nonparametric Bayesian word segmentation model first used by Goldwater et al. (2009) to the action domain, and also extended this model to incorporate causal information. Like the original word segmentation model, our model is based on a *Dirichlet process* (Ferguson, 1973), with actions composed of individual small motion elements taking the place of words composed of phonemes. We model the generative process for creating a sequence of human actions as successively selecting actions to add to the stream, with the conditional probability of generating a particular action given by the *Chinese Restaurant Process* (Aldous, 1985), an easy to implement non-parametric process that is equivalent to the Dirichlet Process.

Generative Model for Action Sequences

An action sequence A is composed of a series of individual actions a_i which are in turn composed of individual motion units m_j. To create the sequence A we draw each a_i from G (a distribution of actions over all possible action sequences), where each action in G has an associated selection probability.

$$a_i|G \sim G$$

We in turn draw our distribution of actions G from a *Dirichlet Process* distribution, defined by the *concentration parameter* α_0 and the *base distribution* P_0.

$$G|\alpha_0, P_0 \sim DP(\alpha_0, P_0)$$

Here, P_0 is a distribution from which possible actions a_i are added to the lexicon. In our model, the probability of including an item in the lexicon is simply the product of the action's component motion unit probabilities, with an added assumption that action length is geometrically distributed (the longer the action, the less likely it is):

$$P_0(a_i = w) = p_{\#}(1 - p_{\#})^{n_w - 1} \prod_{j=1}^{n_w} p(m_{i,j})$$

Where n_w is the length of $a_i = w$ in motion units, $p_{\#}$ and $(1 - p_{\#})$ are the probability of ending or not ending the action after each motion unit, and $p(m_{i,j})$ is the probability of the individual motion units that make up a_i. We assume a uniform probability over all motion units. Once an action a_i is drawn from P_0 it is added to G and assigned a probability in G determined by α_0.

We assume that like action length, action sequence length is also geometrically distributed:

$$P(A) = p_{\$}(1 - p_{\$})^{n-1} \prod_{i=1}^{n} p(a_i)$$

Where n is the length of A in actions, $p_{\$,}(1 - p_{\$})$ are the probability of ending or not ending the action sequence after a given action, and $p(a_i)$ is as described above.

Generative Model for Events

The action sequence A also contains effects e, which may occur both between and within actions. Some actions are causal actions, and are followed by effects with high probability. Each unique action type a_w has an associated binary variable $c_w \in \{0, 1\}$ that determines whether or not the action is causal:

$$c_w \sim Bernoulli(\pi_w)$$

Currently, we use a fixed value of π for all actions, but π_w may in turn be drawn from a *Beta* distribution in future versions of the model. If $c_w = 1$ then action w is causal, otherwise it is not. If an action is causal, then it is followed by an effect with probability ω_w. Again, we currently use a fixed value for ω. We use a small fixed value ϵ for the probability of an effect occuring anywhere in the sequence *other than* after a causal action.

Putting this all together, for each action a_i that is added to the sequence A (as described in the previous section), effects are or are not added after each of a_i's motion units with the following probabilities:

$$\begin{aligned} p(e|a_i = w, m_j, c_w = 1) &= \begin{cases} \omega_w, & j = n \\ \epsilon, & 0 \leq j < n \end{cases} \\ p(e|a_i = w, m_j, c_w = 0) &= \epsilon \end{aligned}$$

Where n_w is the length in motion units of $a_i = w$. In other words, the probability of inserting an effect after an internal motion unit is always a small constant (ϵ) across all actions, while the probability of inserting an effect at the end of an action is ϵ for non-causal actions and ω_w for causal actions.

Chinese Restaurant Process

Rather than explicitly drawing a lexicon G from the Dirichlet Process, and then drawing the actions a_i from G in order to create the action sequence A, we would like to integrate across all possible lexicons. This gives us the conditional probability of the next action in the sequence a_i, given all the previous actions $\mathbf{a}_{-i} = a_1 ... a_{i-1}$

$$p(a_i \mid \mathbf{a_{-i}}, \alpha_0, P_0) = \int p(a_i \mid G) p(\mathbf{w_{-i}}, \alpha_0, P_0) dG$$

It turns out that this conditional probability is equivalent to a simple construction known as the *Chinese Restaurant Process* (CRP). Here we use the CRP to formulate our generative model.

In the CRP customers enter a restaurant, and are seated at tables, each of which has an associated label. In this case, the associated labels represent actions. When the i^{th} customer enters the restaurant, the label at the table they sit at determines what the i^{th} action in our sequence will be. The probability of the i^{th} customer sitting at table $z_i = k$ is:

$$p(z_i = k | \mathbf{z}_{-i}) = \begin{cases} \frac{n_k}{n_{-i}+\alpha_0}, & 0 \leq k \leq K \\ \frac{\alpha_0}{n_{-i}+\alpha_0}, & k = K+1 \end{cases}$$

Where $n_{-i} = i - 1$ is the number of previously seated people, n_k is the number of customers already at table k, and K is the number of previously occupied tables. In other words, the probability of the i^{th} customer sitting at an already occupied table (i.e., choosing an action that has already appeared previously in the sequence A) depends on the proportion of customers already at that table, while the probability of them starting a new table depends on α_0.

Whenever a customer starts a new table, an action a_k must be associated with this table. This action is drawn from the distribution P_0, described above. Since multiple tables may be labeled with the same action, the probability that the next action in the sequence will have a particular value $a_i = w$ is:

$$p(a_i = w | \mathbf{a}_{-i}) = \frac{n_w}{n_{-i} + \alpha_0} + \frac{\alpha_0 P_0(a_i = w)}{n_{-i} + \alpha_0}$$

Where n_w is the number of times action w has appeared in the previous $\mathbf{a_{-i}}$ actions (the number of customers already seated at tables labeled with action w). In other words, the probability of a particular action $a_i = w$ being selected is based on the number of times it has already been selected (the probability of the i^{th} customer sitting at an existing table labeled with this action) and the probability of generating it anew (the probability of the customer sitting at a new table that is then assigned the label $a_k = w$).

CRP Algorithm

In summary, the steps for the CRP are:

1. Pick a table z_i for the i^{th} customer

   a) if it's a new table, draw a label from P_0

2. Add the label at table z_i to your list (in this case to the action sequence)

3. iterate until all customers are seated

Generative Algorithm

We can now put together our complete generative model for creating a sequence of actions and effects. Our model parameters are sequence parameters $p_{\#}$, $p_{\$}$ and α_0, and causal parameters π, ω and ϵ.

1. Draw a probability distribution over actions, G

   a) actions are drawn from P_0

2. select an action $a_i = w$ from G and add it to sequence A (this is equivalent to seating the i^{th} customer in the Chinese restaurant process)

3. Decide whether to insert any events after any of the motions $m_{i,j}$ composing $a_i = w$

   a) for each of the $n_w - 1$ internal motion units in a_i insert an event with probability ϵ

   b) if c_w is not yet known, draw c_w

   c) add an event after the last motion unit in $a_i = w$ with probability ω if $c_w = 1$ and with probability ϵ otherwise

4. With probability $p_{\$}$ repeat steps 2-4, otherwise terminate sequence$A$

Inference

Given an unsegmented action sequence, how do we find the boundaries between actions (find the correct segmented sequence)? For a given segmentation hypothesis h:

$$p(h|d) \propto p(d|h)p(h)$$

We want to infer the posterior distribution $p(h|d)$. A segmentation hypothesis h consists of whether or not there is an action boundary b after each motion m_j in the sequence. We can estimate $p(h|d)$ by iteratively considering one possible boundary at a time, while holding all other segment boundaries constant, a process known as Gibbs sampling. In deciding whether or not there should be an action boundary after motion m_j only two hypotheses need to be evaluated: $h_1 : b_j = false$ and $h_2 : b_j = true$. Since the segmentations defined by both $h_1$and h_2 will contain the same actions except for at the potential boundary point, only this difference in their probabilities needs to be considered. We'll call the segmentation boundaries that are the same in both hypotheses h^-. We will also refer to the single action generated under h_1 as w_1 and to the two actions generated under h_2 as w_2 and w_3. Going back to our generative model and the CRP, we can see that:

$$p(h_1|h^-, d) \propto p(w_1|h^-, d) = p(a_n = w_1|\mathbf{a_{-n}})$$

where n is the total number of actions under h_1 and $\mathbf{a}_{-n} = a_1 ... a_{n-1}$ given by the segmentation h^-. this is because the CRP is *exchangeable*, which means we can treat w_1 as if it were the last action added to the sequence (the last person walking into the restaurant):

$$p(a_n = w_1|\mathbf{a}_{-n}) = \frac{n_{w1}}{n^- + \alpha_0} + \frac{\alpha_0 P_0(a_n = w_1)}{n^- + \alpha_0} = \frac{n_{w1} + \alpha_0 P_0(a_n = w_1)}{n^- + \alpha_0}$$

Where $n^- = n - 1$ is the number of actions under h^- (number of previously seated people) and n_{w1} is the number of times w_1 appears in h^- (the number of people already seated at tables labeled with w_1).

Similarly, the probability of h_2 is

$$\begin{aligned} p(h_2|h^-, d) &\propto p(w_2, w_3|h^-, d) \\ &= p(a_n = w_2|\mathbf{a_{-n}})p(a_{n+1} = w_3|a_n = w_2, \mathbf{a_{-n}}) \end{aligned}$$

As in h_1, the probability of w_2 is simply:

$$p(a_n = w_2|\mathbf{a}_{-n}) = \frac{n_{w1}}{n^- + \alpha_0} + \frac{\alpha_0 P_0(a_n = w_2)}{n^- + \alpha_0} = \frac{n_{w2} + \alpha_0 P_0(a_n = w_2)}{n^- + \alpha_0}$$

The probability of w_3 is a little different, since it depends on w_2 as well as h^-

$$\begin{aligned} p(a_{n+1} = w_3|\mathbf{a}_{-n}, a_n = w_2) &= \frac{n_{w3} + I(w_2 = w_3)}{n^- + 1 + \alpha_0} + \frac{\alpha_0 P_0(a_{n+1} = w_3)}{n^- + 1 + \alpha_0} \\ &= \frac{n_{w3} + I(w_2 = w_3) + \alpha_0 P_0(a_{n+1} = w_3)}{n + \alpha_0} \end{aligned}$$

Where n_{w3} is the number of times w_3 appears in h^- and $I(w_2 = w_3) = 1$ if w_2 and w_3 are the same (in other words ($n_{w3} + I(w_2 = w_3)$) is the number of customers already seated at tables labeled with w_3). Also, $n^- + 1 = n$ is the number of actions in $h^- + w_2$ (the number of previously seated customers, before w_3).

Finally, since h_2 hypothesizes an action sequence one longer than h_1 we need to consider the probability of having a sequence of this increased length. Putting all of this together:

$$\begin{aligned} p(h_2|h^-, d) &\propto p(length = n + 1, w_2, w_3|h^-, d) \\ &= p(length = n + 1|h^-, d)p(w_2|h^-, d)p(w_3|w_2, h^-, d) \\ &= (1 - p_\$) \cdot \frac{n_{w2} + \alpha_0 P_0(a_n = w_2)}{n^- + \alpha_0} \cdot \frac{n_{w3} + I(w_2 = w_3) + \alpha_0 P_0(a_{n+1} = w_3)}{n + \alpha_0} \end{aligned}$$

Using Information from Events in the Action Sequence

If the action sequence we're trying to segment also contains events, we can use this information to aid our segmentation. In particular, whether or not there is an event e_j at the segmentation boundary being considered, or an event e_k following the action created by the segmentation, impacts our probability estimates for h_1 and h_2:

$$p(h_1|h^-, d) \propto p(a_n = w_1|\mathbf{a}_{-n}) \cdot p(e_j|a_n = w_1, c_{w1}, \mathbf{c_{-n}}, \mathbf{e}_{h^-}) \cdot p(e_k|a_n = w_1, c_{w1}, \mathbf{c_{-n}}, \mathbf{e}_{h^-})$$

Where $\mathbf{e}_{h^-}$ are all the events that occured in h^-, c_{w1} is the causal variable for $a_n = w_1$ and $\mathbf{c_{-n}}$ are the causal variables for the other actions in the sequence. Since h_1 predicts no boundary at position j:

$$p(e_j|\cdot) = p(e_j|a_n = w_1) = \begin{cases} \epsilon, & e_j = 1 \\ 1-\epsilon, & e_j = 0 \end{cases}$$

The computation for e_k is slightly more complicated. Assuming we know the value of c_{w1} (we'll discuss sampling the $\mathbf{c}$ values below) then:

$$p(e_k = 1|\cdot) = p(e_j = 1|a_n = w_1, c_{w1}) = \begin{cases} \epsilon, & c_{w1} = 0 \\ \omega, & c_{w1} = 1 \end{cases}$$

$$p(e_k = 0|\cdot) = p(e_j = 1|a_n = w_1, c_{w1}) = \begin{cases} 1-\epsilon, & c_{w1} = 0 \\ 1-\omega, & c_{w1} = 1 \end{cases}$$

If we haven't yet sampled a value for c_{w1} then we simply sum across possible cases, in which case:

$$p(e_k|\cdot) = \begin{cases} \epsilon + \omega, & e_k = 1 \\ (1-\epsilon) + (1-\omega), & e_k = 0 \end{cases}$$

We perform similar computations for h_2, except that in this case, both e_j and e_k occur at the ends of actions (and so are computed like e_k above):

$$p(e_j = 1|\cdot) = p(e_j = 1|a_n = w_2, c_{w2}) = \begin{cases} \epsilon, & c_{w2} = 0 \\ \omega, & c_{w2} = 1 \end{cases}$$

$$p(e_j = 0|\cdot) = p(e_j = 1|a_n = w_2, c_{w2}) = \begin{cases} 1-\epsilon, & c_{w2} = 0 \\ 1-\omega, & c_{w2} = 1 \end{cases}$$

$$p(e_k = 1|\cdot) = p(e_k = 1|a_{n+1} = w_3, c_{w3}) = \begin{cases} \epsilon, & c_{w3} = 0 \\ \omega, & c_{w3} = 1 \end{cases}$$

$$p(e_k = 0|\cdot) = p(e_k = 1|a_{n+1} = w_3, c_{w3}) = \begin{cases} 1-\epsilon, & c_{w3} = 0 \\ 1-\omega, & c_{w3} = 1 \end{cases}$$

The probabilities when c_{w2} and c_{w3} have not yet been sampled are identical to when c_{w1} has not yet been sampled.

Inferring Which Actions are Causal

Just as we can use the causal variables $\mathbf{c}$ to help infer the action segmentation, we can use the action segmentation to help infer the causal variable values. In this case our hypothesis h consists of values c_w for all actions. Just as we can estimate the action segmentation by iteratively considering one boundary at a time, we can estimate the values for $\mathbf{c}$ by looking at one action a_w at a time, and estimating c_w, while holding the remaining c values constant. In this case h_1 is the hypothesis that $c_w = 1$ (a_w is causal) and h_2 is the hypothesis that c_w= 0 (a_w is not causal).

$$p(h_1|h^-, d) \propto p(c_w = 1|\mathbf{a}_n, \mathbf{e}_n) = p(c_w = 1|\mathbf{a_w},\mathbf{e_w})$$
$$p(h_2|h^-, d) \propto p(c_w = 0|\mathbf{a}_n, \mathbf{e}_n) = p(c_w = 0|\mathbf{a_w},\mathbf{e_w})$$

Where $\mathbf{a_w}$ is all occurances of action w in the sequence, and $\mathbf{e_w}$ is all effects (or lack thereof) following these occurances. In this case

$$p(c_w|\mathbf{a}_n, \mathbf{e}_n) \propto p(\mathbf{e_n}|\mathbf{a}_n, c_w) \cdot p(c_w) = p(c_w) \prod_{i=0}^{n} p(e_i|c_w, a_i)$$

With

$$p(c_w) = \begin{cases} \pi, & c_w = 1 \\ 1 - \pi & c_w = 0 \end{cases}$$
$$p(e_i = 1|c_w, a_i) = \begin{cases} \omega, & c_w = 1 \\ \epsilon, & c_w = 0 \end{cases}$$
$$p(e_i = 0|c_w, a_i) = \begin{cases} 1 - \omega, & c_w = 1 \\ 1 - \epsilon, & c_w = 0 \end{cases}$$

Putting this all together

$$p(h_1|h^-, d) \propto p(c_w = 1|\mathbf{a}_n, \mathbf{e}_n) \propto \pi \cdot \omega^{ne_w^+} \cdot (1 - w)^{ne_w^-}$$
$$p(h_2|h^-, d) \propto p(c_w = 0|\mathbf{a}_n, \mathbf{e}_n) \propto (1 - \pi) \cdot \epsilon^{ne_w^+} \cdot (1 - \epsilon)^{ne_w^-}$$

Where ne_w^+ is the number of times action w is followed by an event and ne_w^- is the number of times it's not followed by an event.

Gibbs Sampling

To summarize, our algorithm for discovering the best segmentation of an unsegmented action sequence is:

1. For each motion unit m_j in the action sequence

2. decide whether there should be a boundary after this motion unit

    a) hold the rest of the segmentation constant

    b) calculate the probability of h_1 = no boundary and h_2 = boundary

    c) normalize probabilities and decide probabilistically between h_1 and h_2

3. iterate 1 and 2 until the segmentation converges and/or a pre-determined stopping point is reached.

Simulated Annealing

The Gibbs sampling procedure described above has a number of advantages. It is relatively simple to implement, and once the sampler converges it produces samples from the true posterior distribution. However, changes are made locally, one boundary at a time. Searching through the hypothesis space may therefore require passing through many low probability segmentations in order to reach higher probability hypotheses, causing convergence to be slow.

To address this issue, we used an approach known as *simulated annealing* (Aarts & Korst, 1989). This approach broadens exploration of the hypothesis space early on in sampling by making the relative probabilities of the different hypotheses more uniform. Using the metaphor of "slow cooling", annealing uses a *temperature* parameter γ to gradually adjust the probability of moving to a particular hypothesis. γ starts at a high value, and is slowly reduced to 1 over the course of sampling.

Using simulated annealing, we sample our boundary probabilities using $p(h_1|h^-, d)^{\frac{1}{\gamma}}$ and $p(h_2|h^-, d)^{\frac{1}{\gamma}}$. Notice that when $\gamma = 1$ this is just regular Gibbs sampling. When $\gamma > 1$ the relative probabilities of the two hypotheses are more uniform, making transitions to lower probability segmentations more likely.

Following Goldwater et al. (2009), for our simulations, we ran each sampler for 20,000 iterations, annealing in 10 increments of 2000 iterations each, with $\frac{1}{\gamma} = (.1, .2, ..., .9, 1)$. For each simulation, we ran three randomly seeded samplers, each initialized from a random segmentation of the input coprus, and averaged results from 10 samples drawn from the last 1,000 iterations of each sampler, to estimate the posterior distributions and evaluate the model. This allowed for an additional burn-in period of 1000 samples with $\gamma = 1$.

A.2 De Bruijn Sequence

For an alphabet A of size k, a *De Bruijn sequence* $B(k, n)$ is a cyclical sequence within which each subsequence of length n appears exactly once as a consecutive sequence. The sequence is constructed by first creating a *De Bruijn graph*, where every sequence of size $n-1$ appears as a node, and outgoing edges represent a sequence of n items – the $n-1$ items of the node the edge is leaving, and the item labeling the edge itself. See Figure A.1 for an example

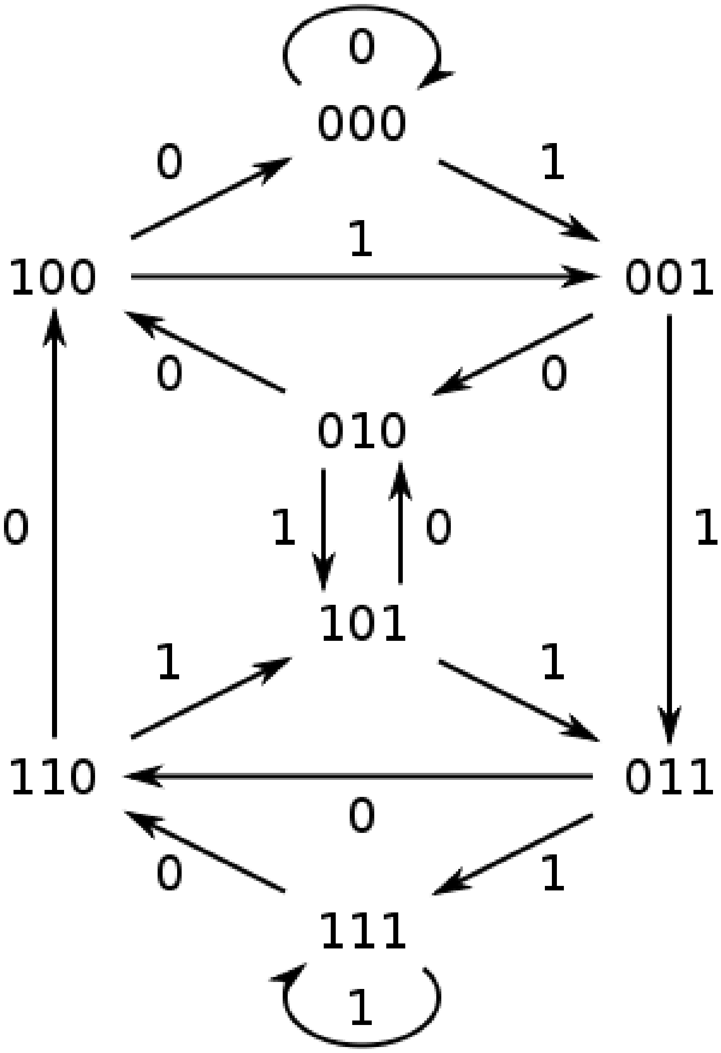


Figure A.1: Example De Bruijn Graph (image created by Michael Hardy). If you traverse the graph in a Eulerian cycle, passing through every edge exactly once, then every four-digit sequence occurs exactly once

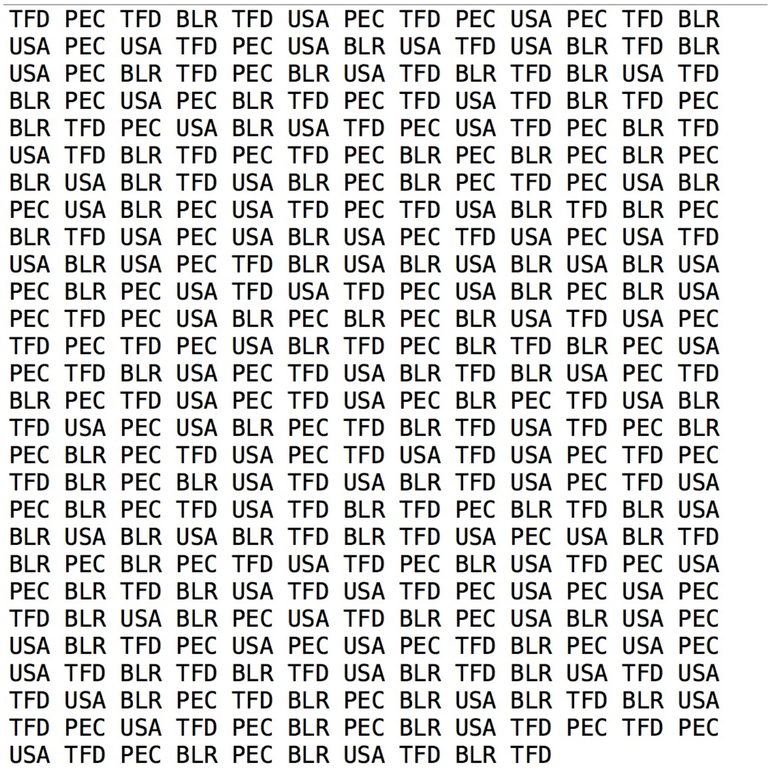
graph. The sequence is then created by traversing the graph in a *Eulerian cycle* – a path through the graph that traverses each edge exactly once.

To create the exposure corpora for Experiment 2, we used a $B(4, 3)$ De Bruijn sequence (creating length three sequences from a length four alphabet), modifying the process slightly, by only allowing edges in the graph that would not cause the resulting sequence of three items to contain a repeated item (in other words, each two-item node has exactly two outgoing edges). There are a number of algorithms for finding the shortest path through a graph that traverses each edge at least once (which will always be the Eulerian cycle if it exists). In this work we used the approach presented by Thimbleby (2003).

A.3 Example Corpora

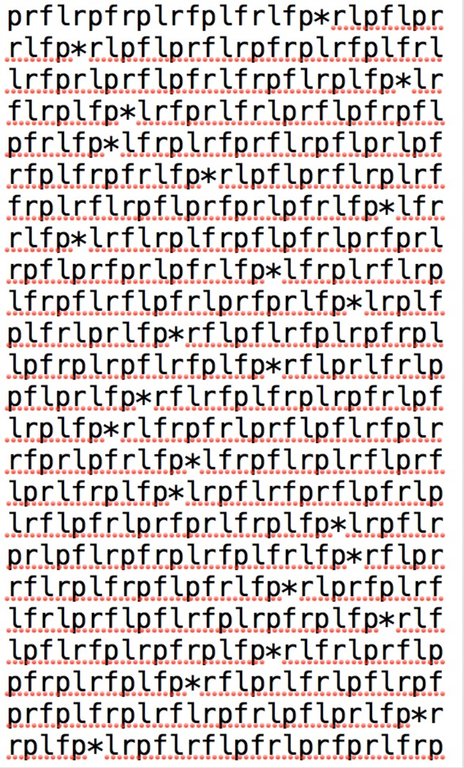
```
TFDPECTFDBLRTFDUSAPECTFDPECUSAPECTFDBLRUSAPECUSATFDPECU
SABLRUSATFDUSABLRTFDBLRUSAPECBLRTFDPECBLRUSATFDBLRTFDBL
RUSATFDBLRPECUSAPECBLRTFDPECTFDUSATFDBLRTFDPECBLRTFDPEC
USABLRUSATFDPECUSATFDPECBLRTFDUSATFDBLRTFDPECTFDPECBLRP
ECBLRPECBLRPECBLRUSABLRTFDUSABLRPECBLRPECTFDPECUSABLRPE
CUSABLRPECUSATFDPECTFDUSABLRTFDBLRPECBLRTFDUSAPECUSABLR
USAPECTFDUSAPECUSATFDUSABLRUSAPECTFDBLRUSABLRUSABLRUSAB
LRUSAPECBLRPECUSATFDUSATFDPECUSABLRPECBLRUSAPECTFDPECUS
ABLRPECBLRPECBLRUSATFDUSAPECTFDPECTFDPECUSABLRTFDPECBLR
TFDBLRPECUSAPECTFDBLRUSAPECTFDUSABLRTFDBLRUSAPECTFDBLRP
ECTFDUSAPECTFDUSAPECBLRPECTFDUSABLRTFDUSAPECUSABLRPECTF
DBLRTFDUSATFDPECBLRPECBLRPECTFDUSAPECTFDUSATFDUSAPECTFD
PECTFDBLRPECBLRUSATFDUSABLRTFDUSAPECTFDUSAPECBLRPECTFDU
SATFDBLRTFDPECBLRTFDBLRUSABLRUSABLRUSABLRTFDBLRTFDUSAPE
CUSABLRTFDBLRPECBLRPECTFDUSATFDPECBLRUSATFDPECUSAPECBLR
TFDBLRUSATFDUSATFDPECUSAPECUSAPECTFDBLRUSABLRPECUSATFDB
LRPECUSABLRUSAPECUSABLRTFDPECUSAPECUSAPECTFDBLRPECUSAPE
CUSATFDBLRTFDBLRTFDUSABLRTFDBLRUSATFDUSATFDUSABLRPECTFD
BLRPECBLRUSABLRTFDBLRUSATFDPECUSATFDPECBLRPECBLRUSATFDP
ECTFDPECUSATFDPECBLRPECBLRUSATFDBLRTFD
```

Figure A.2: Example unsegmented corpus for Experiment 1 (line breaks are for display, and were not present in the input)



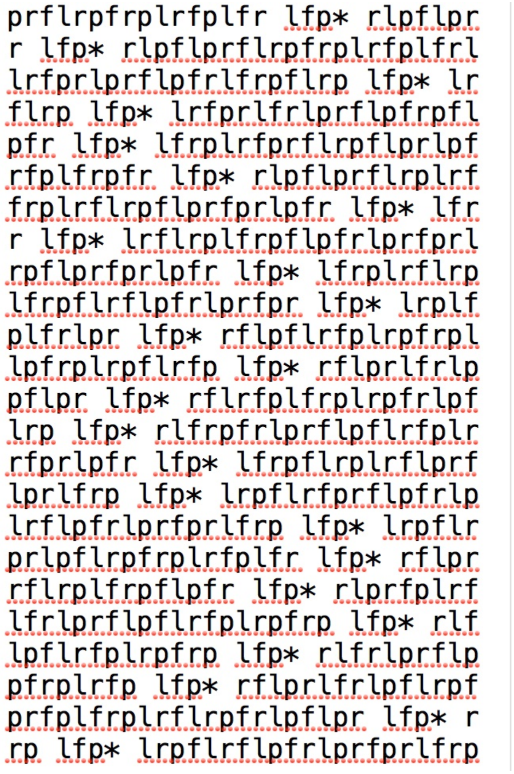
```
TFD PEC TFD BLR TFD USA PEC TFD PEC USA PEC TFD BLR
USA PEC USA TFD PEC USA BLR USA TFD USA BLR TFD BLR
USA PEC BLR TFD PEC BLR USA TFD BLR TFD BLR USA TFD
BLR PEC USA PEC BLR TFD PEC TFD USA TFD BLR TFD PEC
BLR TFD PEC USA BLR USA TFD PEC USA TFD PEC BLR TFD
USA TFD BLR TFD PEC TFD PEC BLR PEC BLR PEC BLR PEC
BLR USA BLR TFD USA BLR PEC BLR PEC TFD PEC USA BLR
PEC USA BLR PEC USA TFD PEC TFD USA BLR TFD BLR PEC
BLR TFD USA PEC USA BLR USA PEC TFD USA PEC USA TFD
USA BLR USA PEC TFD BLR USA BLR USA BLR USA BLR USA
PEC BLR PEC USA TFD USA TFD PEC USA BLR PEC BLR USA
PEC TFD PEC USA BLR PEC BLR PEC BLR USA TFD USA PEC
TFD PEC TFD PEC USA BLR TFD PEC BLR TFD BLR PEC USA
PEC TFD BLR USA PEC TFD USA BLR TFD BLR USA PEC TFD
BLR PEC TFD USA PEC TFD USA PEC BLR PEC TFD USA BLR
TFD USA PEC USA BLR PEC TFD BLR TFD USA TFD PEC BLR
PEC BLR PEC TFD USA PEC TFD USA TFD USA PEC TFD PEC
TFD BLR PEC BLR USA TFD USA BLR TFD USA PEC TFD USA
PEC BLR PEC TFD USA TFD BLR TFD PEC BLR TFD BLR USA
BLR USA BLR USA BLR TFD BLR TFD USA PEC USA BLR TFD
BLR PEC BLR PEC TFD USA TFD PEC BLR USA TFD PEC USA
PEC BLR TFD BLR USA TFD USA TFD PEC USA PEC USA PEC
TFD BLR USA BLR PEC USA TFD BLR PEC USA BLR USA PEC
USA BLR TFD PEC USA PEC USA PEC TFD BLR PEC USA PEC
USA TFD BLR TFD BLR TFD USA BLR TFD BLR USA TFD USA
TFD USA BLR PEC TFD BLR PEC BLR USA BLR TFD BLR USA
TFD PEC USA TFD PEC BLR PEC BLR USA TFD PEC TFD PEC
USA TFD PEC BLR PEC BLR USA TFD BLR TFD
```

Figure A.3: Example true segmentation for Experiment 1



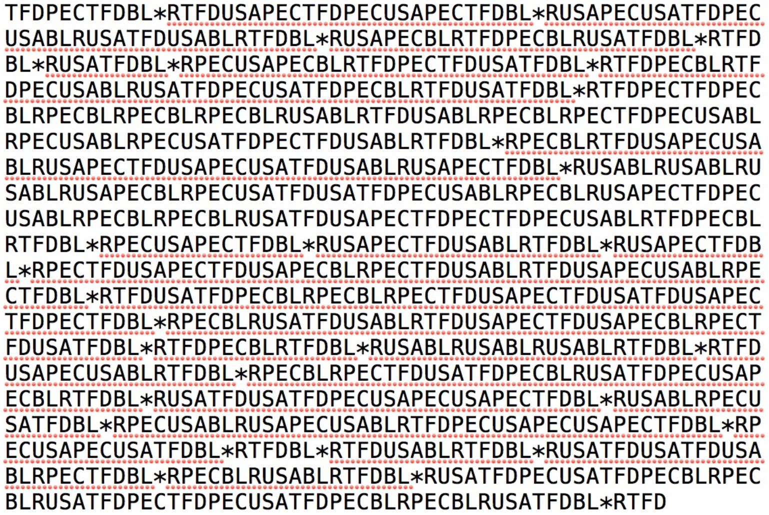
```
prflrpfrplrfplfrlfp*rlpflpr
rlfp*rlpflprflrpfrplrfplfrl
lrfprlprflpfrlfrpflrplfp*lr
flrplfp*lrfprlfrlprflpfrpfl
pfrlfp*lfrplrfprflrpflprlpf
rfplfrpfrlfp*rlpflprflrplrf
frplrflrpflprfprlpfrlfp*lfr
rlfp*lrflrplfrpflpfrlprfprl
rpflprfprlpfrlfp*lfrplrflrp
lfrpflrflpfrlprfprlfp*lrplf
plfrlprlfp*rflpflrfplrpfrpl
lpfrplrpflrfplfp*rflprlfrlp
pflprlfp*rflrfplfrplrpfrlpf
lrplfp*rlfrpfrlprflpflrfplr
rfprlpfrlfp*lfrpflrplrflprf
lprlfrplfp*lrpflrfprflpfrlp
lrflpfrlprfprlfrplfp*lrpflr
prlpflrpfrplrfplfrlfp*rflpr
rflrplfrpflpfrlfp*rlprfplrf
lfrlprflpflrfplrpfrplfp*rlf
lpflrfplrpfrplfp*rlfrlprflp
pfrplrfplfp*rflprlfrlpflrpf
prfplfrplrflrpfrlpflprlfp*r
rplfp*lrpflrflpfrlprfprlfrp
```

Figure A.4: Example unsegmented corpus for Experiments 2 and 3



```
prflrpfrplrfplfr lfp* rlpflpr
r lfp* rlpflprflrpfrplrfplfrl
lrfprlprflpfrlfrpflrp lfp* lr
flrp lfp* lrfprlfrlprflpfrpfl
pfr lfp* lfrplrfprflrpflprlpf
rfplfrpfr lfp* rlpflprflrplrf
frplrflrpflprfprlpfr lfp* lfr
r lfp* lrflrplfrpflpfrlprfprl
rpflprfprlpfr lfp* lfrplrflrp
lfrpflrflpfrlprfpr lfp* lrplf
plfrlpr lfp* rflpflrfplrpfrpl
lpfrplrpflrfp lfp* rflprlfrlp
pflpr lfp* rflrfplfrplrpfrlpf
lrp lfp* rlfrpfrlprflpflrfplr
rfprlpfr lfp* lfrpflrplrflprf
lprlfrp lfp* lrpflrfprflpfrlp
lrflpfrlprfprlfrp lfp* lrpflr
prlpflrpfrplrfplfr lfp* rflpr
rflrplfrpflpfr lfp* rlprfplrf
lfrlprflpflrfplrpfrp lfp* rlf
lpflrfplrpfrp lfp* rlfrlprflp
pfrplrfp lfp* rflprlfrlpflrpf
prfplfrplrflrpfrlpflpr lfp* r
rp lfp* lrpflrflpfrlprfprlfrp
```

Figure A.5: Example true segmentation for Experiments 2 and 3



```
TFDPECTFDBL*RTFDUSAPECTFDPECUSAPECTFDBL*RUSAPECUSATFDPEC
USABLRUSATFDUSABLRTFDBL*RUSAPECBLRTFDPECBLRUSATFDBL*RTFD
BL*RUSATFDBL*RPECUSAPECBLRTFDPECTFDUSATFDBL*RTFDPECBLRTF
DPECUSABLRUSATFDPECUSATFDPECBLRTFDUSATFDBL*RTFDPECTFDPEC
BLRPECBLRPECBLRPECBLRUSABLRTFDUSABLRPECBLRPECTFDPECUSABL
RPECUSABLRPECUSATFDPECTFDUSABLRTFDBL*RPECBLRTFDUSAPECUSA
BLRUSAPECTFDUSAPECUSATFDUSABLRUSAPECTFDBL*RUSABLRUSABLRU
SABLRUSAPECBLRPECUSATFDUSATFDPECUSABLRPECBLRUSAPECTFDPEC
USABLRPECBLRPECBLRUSATFDUSAPECTFDPECTFDPECUSABLRTFDPECBL
RTFDBL*RPECUSAPECTFDBL*RUSAPECTFDUSABLRTFDBL*RUSAPECTFDB
L*RPECTFDUSAPECTFDUSAPECBLRPECTFDUSABLRTFDUSAPECUSABLRPE
CTFDBL*RTFDUSATFDPECBLRPECBLRPECTFDUSAPECTFDUSATFDUSAPEC
TFDPECTFDBL*RPECBLRUSATFDUSABLRTFDUSAPECTFDUSAPECBLRPECT
FDUSATFDBL*RTFDPECBLRTFDBL*RUSABLRUSABLRUSABLRTFDBL*RTFD
USAPECUSABLRTFDBL*RPECBLRPECTFDUSATFDPECBLRUSATFDPECUSAP
ECBLRTFDBL*RUSATFDUSATFDPECUSAPECUSAPECTFDBL*RUSABLRPECU
SATFDBL*RPECUSABLRUSAPECUSABLRTFDPECUSAPECUSAPECTFDBL*RP
ECUSAPECUSATFDBL*RTFDBL*RTFDUSABLRTFDBL*RUSATFDUSATFDUSA
BLRPECTFDBL*RPECBLRUSABLRTFDBL*RUSATFDPECUSATFDPECBLRPEC
BLRUSATFDPECTFDPECUSATFDPECBLRPECBLRUSATFDBL*RTFD
```

Figure A.6: Example unsegmented corpus for Experiment 4 (line breaks are for display, and were not present in the input)

```
TFD PEC TFDBL∗ R TFD USA PEC TFD PEC USA PEC TFDBL∗ R
USA PEC USA TFD PEC USA BLR USA TFD USA BLR TFDBL∗ R
USA PEC BLR TFD PEC BLR USA TFDBL∗ R TFDBL∗ R USA
TFDBL∗ R PEC USA PEC BLR TFD PEC TFD USA TFDBL∗ R TFD
PEC BLR TFD PEC USA BLR USA TFD PEC USA TFD PEC BLR
TFD USA TFDBL∗ R TFD PEC TFD PEC BLR PEC BLR PEC BLR
PEC BLR USA BLR TFD USA BLR PEC BLR PEC TFD PEC USA
BLR PEC USA BLR PEC USA TFD PEC TFD USA BLR TFDBL∗ R
PEC BLR TFD USA PEC USA BLR USA PEC TFD USA PEC USA
TFD USA BLR USA PEC TFDBL∗ R USA BLR USA BLR USA BLR
USA PEC BLR PEC USA TFD USA TFD PEC USA BLR PEC BLR
USA PEC TFD PEC USA BLR PEC BLR PEC BLR USA TFD USA
PEC TFD PEC TFD PEC USA BLR TFD PEC BLR TFDBL∗ R PEC
USA PEC TFDBL∗ R USA PEC TFD USA BLR TFDBL∗ R USA PEC
TFDBL∗ R PEC TFD USA PEC TFD USA PEC BLR PEC TFD USA
BLR TFD USA PEC USA BLR PEC TFDBL∗ R TFD USA TFD PEC
BLR PEC BLR PEC TFD USA PEC TFD USA TFD USA PEC TFD
PEC TFDBL∗ R PEC BLR USA TFD USA BLR TFD USA PEC TFD
USA PEC BLR PEC TFD USA TFDBL∗ R TFD PEC BLR TFDBL∗ R
USA BLR USA BLR USA BLR TFDBL∗ R TFD USA PEC USA BLR
TFDBL∗ R PEC BLR PEC TFD USA TFD PEC BLR USA TFD PEC
USA PEC BLR TFDBL∗ R USA TFD USA TFD PEC USA PEC USA
PEC TFDBL∗ R USA BLR PEC USA TFDBL∗ R PEC USA BLR USA
PEC USA BLR TFD PEC USA PEC USA PEC TFDBL∗ R PEC USA
PEC USA TFDBL∗ R TFDBL∗ R TFD USA BLR TFDBL∗ R USA
TFD USA TFD USA BLR PEC TFDBL∗ R PEC BLR USA BLR
TFDBL∗ R USA TFD PEC USA TFD PEC BLR PEC BLR USA TFD
PEC TFD PEC USA TFD PEC BLR PEC BLR USA TFDBL∗ R TFD
```

Figure A.7: Example compromise segmentation for Experiment 4

Appendix B

Chapter 3: Supplementary Material

B.1 Methods

Stimuli

Doorbells

The toys in this experiment were made to play music by hiding battery-operated wireless doorbells inside them. The doorbells hidden inside the toys could then be activated by using the doorbell button as a remote control. By pressing the doorbell button after a particular action or sequence of actions was performed on the toy, the experimenter created a strong illusion of causality. Throughout these and similar experiments (such as those in Chapter 4), no children ever guessed the true cause of the toy playing music. Preliminary piloting with adults similarly confirmed that they experienced a strong causal illusion, and no adults guessed that the experimenter was causing the music to play, rather than their own actions.

All doorbells were Heath Zenith doorbells with 64 chime options (Model SL-6164-B), which included short segments of a variety of well-known songs. In some cases, the doorbells were modified slightly to make them smaller. This was done by cutting off some of the plastic casing surrounding the doorbell electronics, and by moving and re-soldering the battery holder to create a more compact shape.

Toy 1: Whoozit

This toy was created by modifying the full-size Whoozit toy sold by Manhattan Toys. The toy is already divided into two parts, allowing the doorbell to be hidden inside. The following modifications were made to the toy:

- We removed the red bulb that was the nose and reattached it so that it covered up the eyes, so that the toy no longer appeared to have a face

- We took the squeaker out of the red bulb

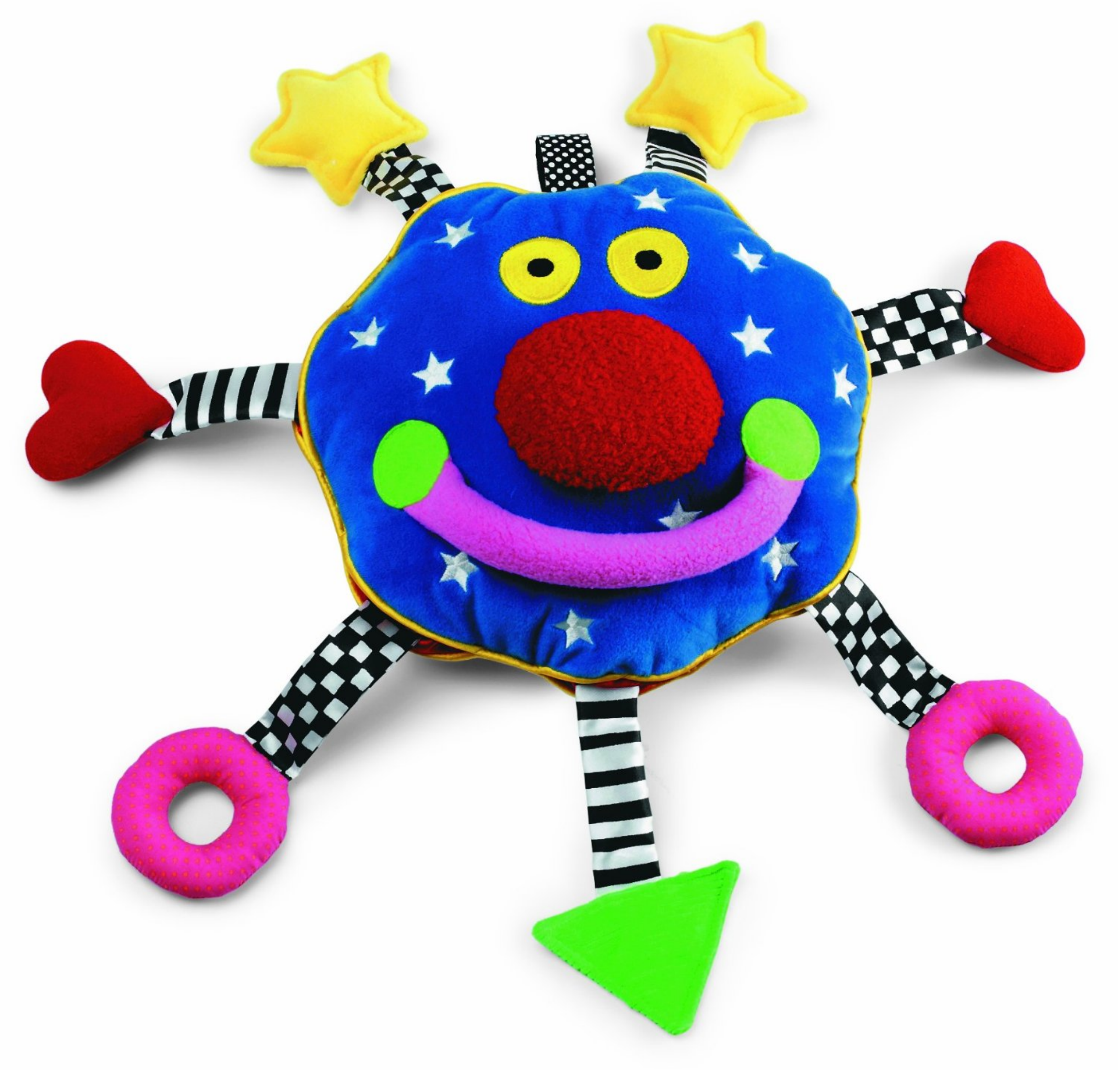


Figure B.1: The original Whoozit toy

- We hid all the tabs except one of the pink rings inside the toy
- We hid the wireless doorbell inside the toy
- The opening of the toy was widened slightly to accommodate the doorbell. Velcro was added to the toy, to hold the toy closed so children could not easily access the hidden doorbell

The doorbell inside this toy was set to play “The Yellow Rose of Texas”.

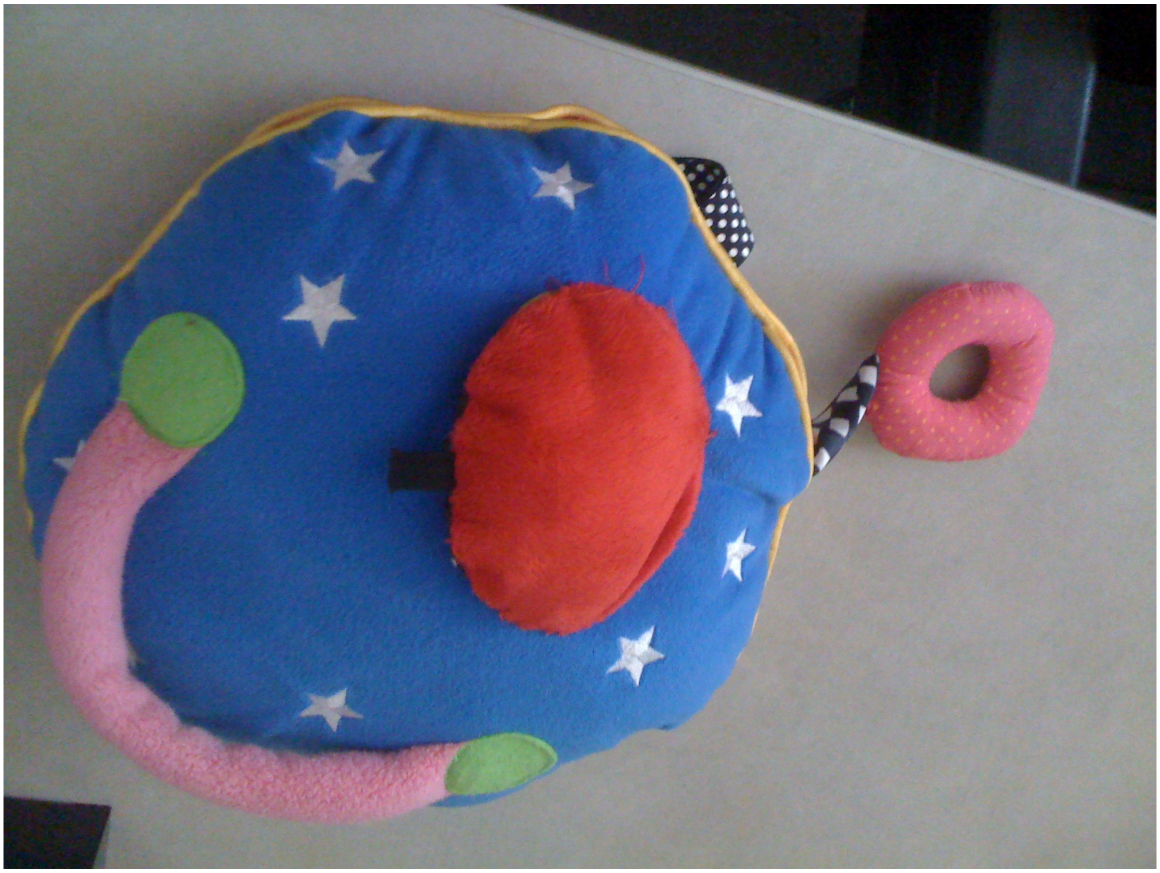


Figure B.2: The modified Whoozit toy as used in the experiments

Toy 2: Furball

This toy was created by taking a hamster ball (approximately 10 inches in diameter) and covering it with a stretchy rubber ball covered in rubbery tentacles (the product is called a "Puffer ball"). The doorbell was placed inside the hamster ball. The following modifications were made to the toy:

- The puffer ball was sliced open and streched over the hamster ball
- Glue was used to keep the puffer ball covering in place

The doorbell inside this toy was set to play "La Cucaracha".

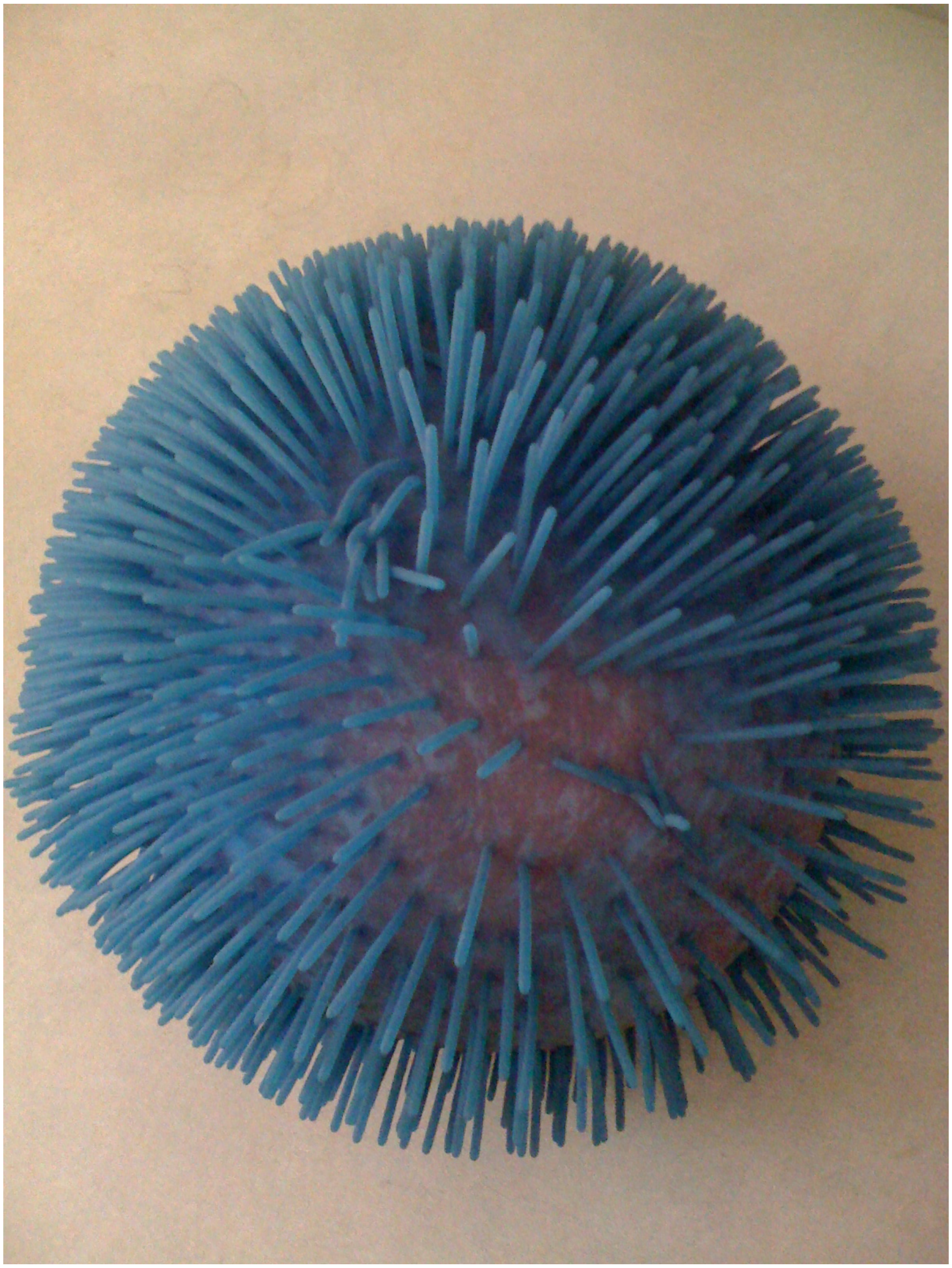


Figure B.3: The Furball toy

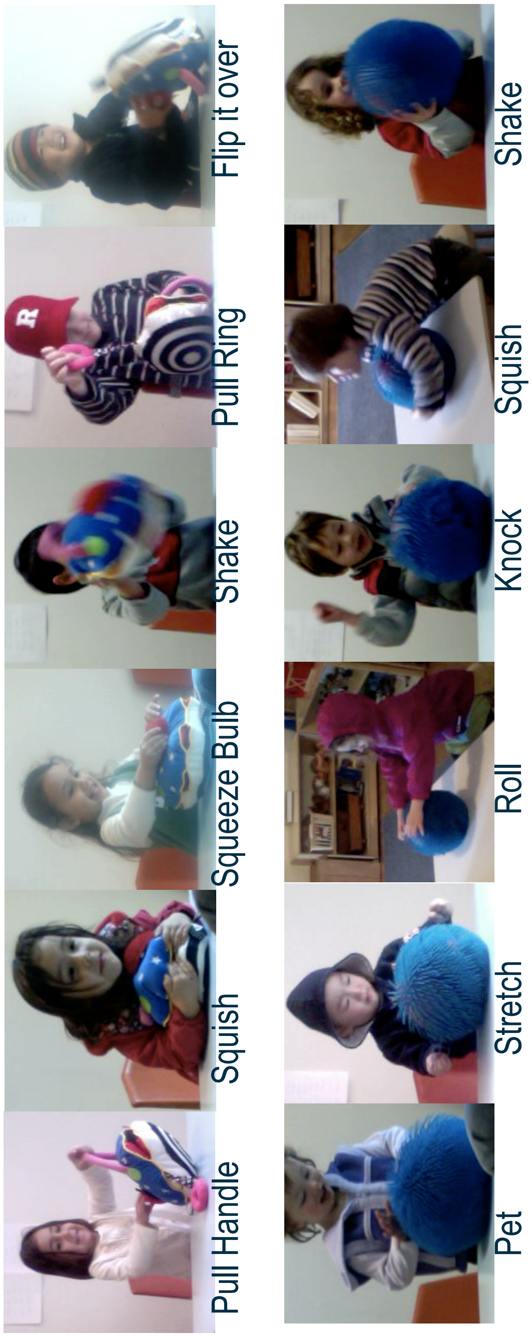


Figure B.4: The six actions used with each of the toys

Appendix C

Chapter 4: Supplementary Material

C.1 Experiment 1

Methods

Participants

Participants were 52 preschool-age children (M = 50 months, range = 36-66 months, 32 females) recruited from a database of interested families in the San Francisco Bay Area, local preschools and children's science museums. An additional 5 children were tested but excluded from data analysis: 3 children were excluded because they did not complete the experiment and 2 were excluded due to experimenter error.

Stimuli

Causal Demonstration and Counterfactual Phases Monkey was a stuffed monkey toy approximately 30.5 cm tall, shown in Figure C.1. The "Birthday machine" was a 30.5 cm x 20.3 cm x 12.7 cm rectangular Tupperware container covered in blue and copper metallic wrapping paper and silver duct-tape, shown in Figure C.2. Inside of the container was a wireless doorbell that played "Happy Birthday" and could be operated remotely by a push button. By hiding the push button, the experimenter could control which objects appeared to cause the machine to play music. The zando was a heavy, rectangular object covered in duct-tape and colorful stickers. The non-zando was a circular, gear-like object with a metal centre and wooden spokes (Figure C.2).

Phase The pretend machine was a 30.5 cm x 30.5 cm x 10.2 cm plain, white, wooden box. Two 8.9 cm x 5.1 cm x 1.3 cm rectangular wooden blocks, one green and one blue, functioned as the pretend zando and pretend non-zando (see Figure C.3.

Testing Procedure Children were tested in a designated testing location at their preschool, at a museum or at the Institute of Human Development at UC Berkeley. Children sat at a table across from the experimenter.

C.2 Experiment 2

Methods

The experiment involved three tasks: a pretense task, a conservation task, and an executive function task. The participant recruiting and testing procedures and the stimuli and methods for the pretense task were identical to those in Experiment 1.

Participants

Participants were 60 preschool-age children (M = 47 months,range = 33-59 months, 28 female). An additional 19 children were tested but excluded from later analysis. Four children did not complete the pretense task, 2 children had equipment failure involving the "Birthday machine," 5 children did not have video and had unclear or missing data sheets, 2 children had testing sessions involving an experimenter error, and 6 children failed to learn the causal relationship in the Causal Demonstration phase of the pretense task even after a second demonstration.

Stimuli

Conservation Task Ten pennies were used in this task.

Executive Function Task Eighteen 8.9 cm x 8.9 cm laminated cards were used in this task. Nine of these cards were pictures of nighttime (a yellow crescent moon and stars on a black background). The other 9 cards were pictures of daytime (an orange and yellow sun on a cream-colored background). One night card and one day card were training and pre-test cards. The other 16 cards (8 night and 8 day) were test cards

Procedure

Conservation Task The experimenter began this task by arranging 10 pennies into two rows of 5 aligned one-to-one so that both rows were the same length. The experimenter then asked the child, "Does this row have more pennies, does this row have more pennies, or do they both have the same amount of pennies?" The experimenter then pushed the pennies in one row together and spread the pennies in the other row out so that one row appeared longer than the other and repeated the question. The experimenter then spread the shorter row out and pushed the pennies in the longer row together so that the relative lengths of each row reversed (i.e. the previously shorter row was now the longer row) and asked the same question again. If the child did not offer an answer, the experimenter simplified the question by asking, "Do you think both rows have the same amount of pennies or do you think one row has more?" If the child said "more," the experimenter asked the child to point to which row had more.

Executive Function Task The executive function task used was modeled after the Stroop-like Day-Night task (as described in Gerstadt *et al.* [59]). There were two training and two pre-test trials followed by 16 test trials. For the training, the experimenter held up a night card and said, "This is a picture of night, but in this silly game, when you see this card, I want you to say 'day.' Can you say 'day'?" The experimenter then held up a day card and instructed the child to say "night" when s/he saw this card and asked the child to repeat the word "night." The pre-test began when the experimenter then said, "Let's practice!" and held up the day card without instruction. If the child said "night," then the experimenter said, "Good!" and held up the night card without instruction. If the child said "day," the experimenter said, "Good!" and proceeded to the test trials. If the child did not offer a label, the experimenter encouraged a response without using the words "day" or "night" by asking, "What do we say for this one?" If the child responded incorrectly to either of the pre-test cards, then the training was repeated starting with the card the child incorrectly labelled, and the pre-test was subsequently repeated as well. If a child failed the pre-test trials twice, s/he was coded as failing this task (as described in Gerstadt et al., 1994). Two children did not participate in the executive function task and 8 children failed the pre-test trials twice.

If the child passed the pre-test trials, then the experimenter began the test trials. Sixteen cards were presented in the following pseudo-random, fixed order (the same order as in Gerstadt et al. (1994): a night card (N), a day card (D), D, N, D, N, N, D, D, N, D, N, N, D, N, D. The experimenter held up one card at a time without instruction and waited for the child to respond. No feedback was given on the test trials. If the child hesitated, then the experimenter encouraged a response without using the words "day" or "night" by asking, "What do we say for this one?"

Coding

Conservation Task Children were coded as correct if they said that both rows had the same amount of pennies and were coded as incorrect if they said that one row had more. If a child did not produce a verbal response but pointed to one of the rows, his/her answer was coded as "more" and thus incorrect. If a child was too shy to produce a verbal response and did not spontaneously point to one of the rows, then the experimenter asked, "Can you point to what you think? Do you think both rows have the same amount of pennies [the experimenter held up one of her hands] or do you think one row has more [the experimenter held up her other hand]?" and allowed the child to point to a hand. If the child pointed to the hand that represented "more," the experimenter then asked the child to point to the row that had more.

Executive Function Task The number of cards children responded to correctly was recorded. Children were assigned a proportion correct based on the number of codeable answers they provided (some children did not produce codeable responses on some trials). Some children would respond with words and phrases other than “day” and “night” such as “morning” or “time to go to sleep.” When possible, these answers were translated into “day” and “night” and then coded as correct or incorrect. Answers besides “day,” “night” or these translateable alternatives were coded as incorrect (e.g., “I don’t know”).

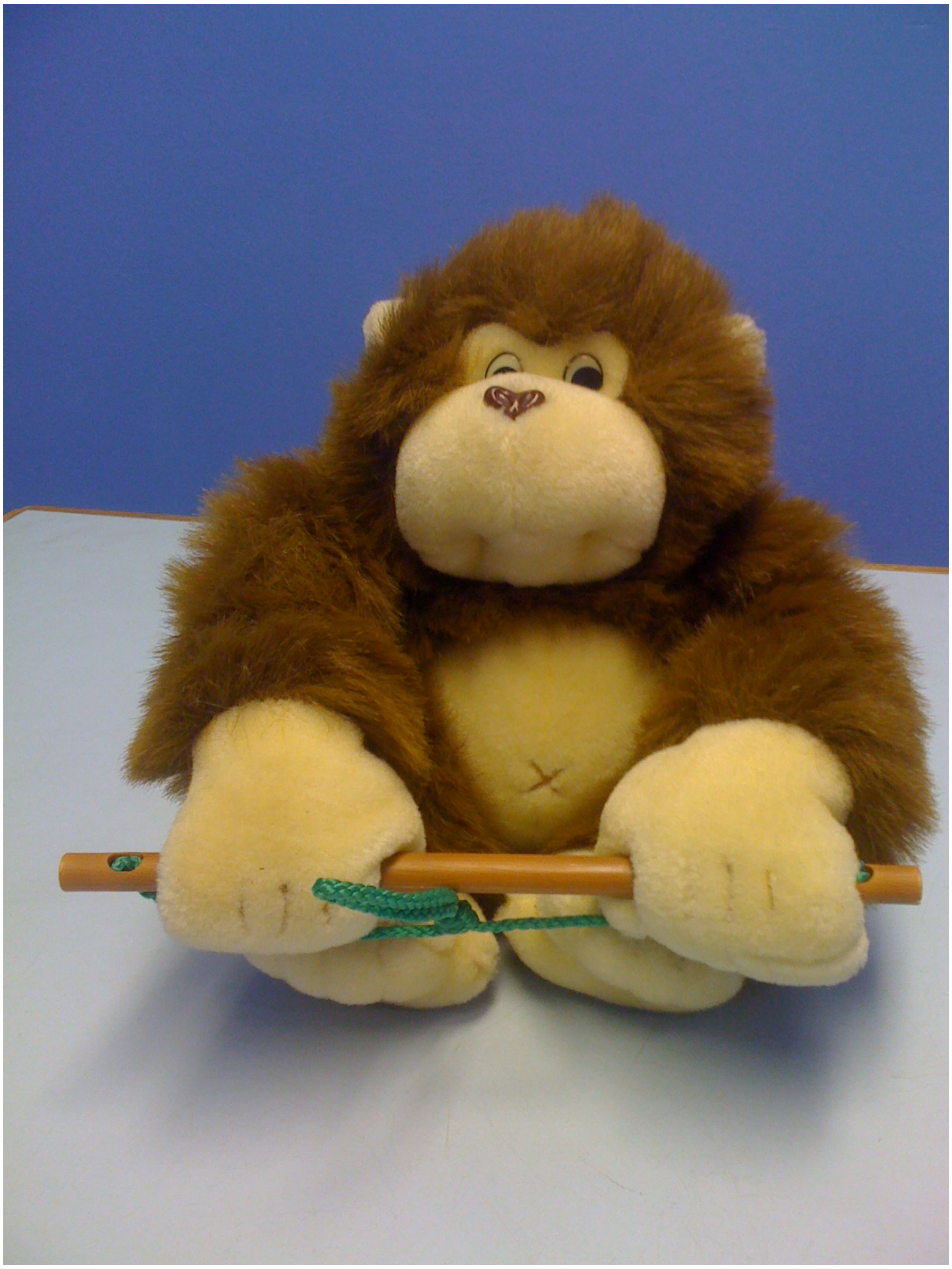


Figure C.1: Monkey

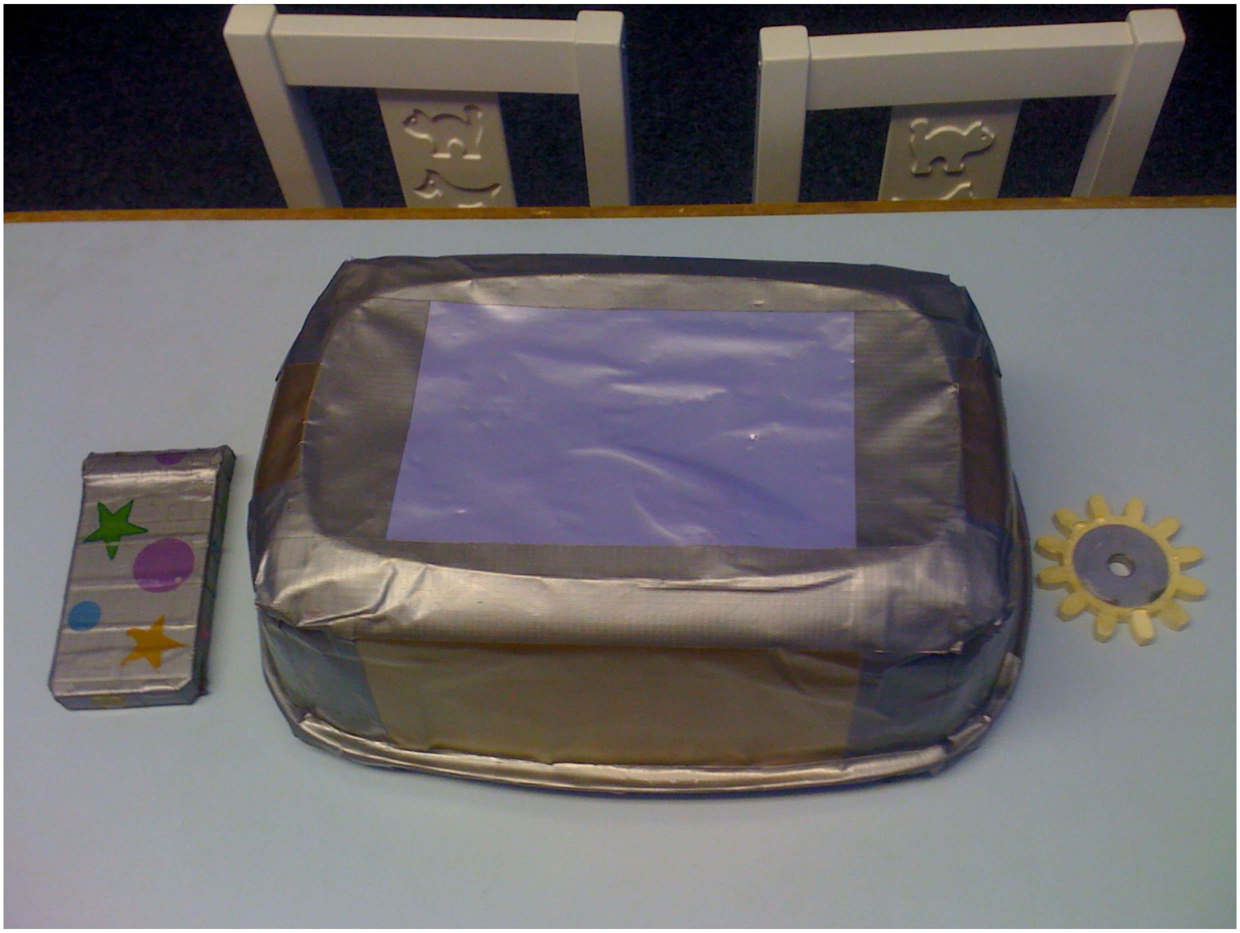


Figure C.2: The “Birthday Machine”. Zando on left, non-zando on right.

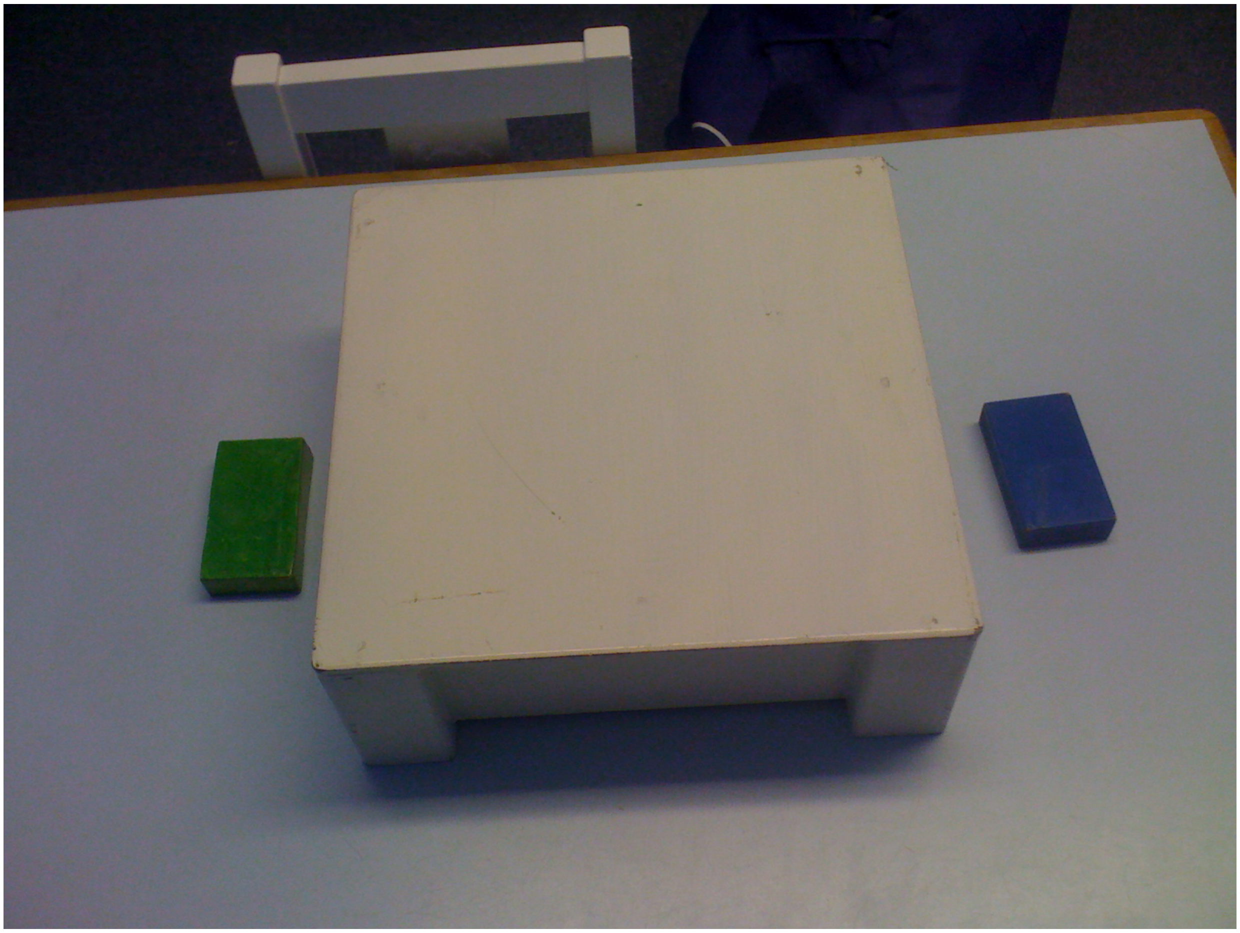


Figure C.3: Stimuli used in pretense phase of Experiments 1 and 2.