Chapter 20

Bayesian models of cognitive development

Elizabeth Bonawitz & Tomer Ullman

How can a learner take ambiguous, noisy, and incomplete information, and from it build causal, structured, and complete representations? This fundamental question has been viewed by symbolic, connectionist, and rationalist perspectives, each developing different, partial answers to this problem. As has been discussed in detail in previous chapters, probabilistic models naturally integrate strengths from each of these approaches towards a unifying framework that incorporates structured representations, learning, and a characterization of the goal of the cognitive system.

Classic approaches in cognitive science parallel areas of focus in developmental psychology that similarly grew out of a desire to answer the tension between representation and learning in a boundedly rational system. On one side, a purely nativist response was to deny learning, and to focus on characterizing the detailed representations that were already in place. On another side, an empiricist response was to suggest that structured representations were not necessary, that learning (and the inferences that followed) was simply a bottom-up process of learning statistical associations. Other debates played out in cognitive development as well. Some research tended to characterized children as "noisy" or "irrational" adults, while other research sought to demonstrate that children are efficient and effective rational learners.

With the development of the probabilistic framework came a renewed interest in unifying developmental psychology, coined rational constructivism (e.g., Xu & Kushnir, 2013, as a nod towards a reinterpretation of Piagetian themes). Just as probabilistic models integrate themes from symbolic, connectionist, and rationalist approaches, so too did their emergence in cognitive development provide a framework to marry sides in nature-nurture and irrational-rational stand-offs. As has been illustrated in previous chapters, theory-based probabilistic models of development provide a means to characterize core and intuitive beliefs in terms of representations. These models detail how these representations may be learned and revised in light of evidence. They operate at the rational level of analysis and depict what problems are being solved in the minds of young children. Their connection to algorithmic levels of analysis show how "boundedly rational" learning may resolve noisy behavior at the level of the individual while capturing optimal behavior at the level of the group. Furthermore, probabilistic models have helped researchers in developmental psychology find precision in both specifying the structure of internal representations and identifying the possible processes that drive their formation. The problem can be broken down into three related questions. First, how do children generalize from examples? Second, how do children learn the inductive biases that shape generalization from examples? Third, how do children learn the frameworks that shape the inductive biases that influence generalization from examples? For example, developmentalists have struggled with questions like how children learn to map words to their meanings so quickly, or how observing a single causal intervention supports inference about future outcomes. Inductive constraints, such as whole object biases (Markman, 1990) or core representations (e.g., Carey & Spelke, 1996), were proposed to explain these powerful inferences in early childhood.

Although these inductive biases appear early in development, and are thus a natural candidate in support of nativist theories, probabilistic frameworks can provide accounts of how these inductive biases may themselves be learned, pointing to overhypotheses (or framework theories as they are often called in developmental psychology; Wellman & Gelman, 1992). Hierarchical Bayesian Models (see Chapter 8) describe how frameworks can generate inductive biases that generate the evidence, and thus reciprocally speak to how the evidence can shape the biases which shape the frameworks, and so on. Although reminiscent of infinite regress arguments ("turtles all the way down"), most scientists who apply the probabilistic modeling approach to developmental questions will suggest that higher-level overhypotheses will eventually ground out in perceptual and conceptual core primitives, which we discuss in further detail below. Importantly, we are able to use the probabilistic modeling framework to help specify what must be built-in, given the evidence available to the developing mind[1], what forms the representations may

[1]Probabilistic models provide an account of what necessarily must be built into the system. However, they do not require that constraints "lower down" in the hierarchically are necessarily learned. That is, just because something can be learned, does not mean it is learned in development. Children may have more built in than is required by their experience, as a kind of redundancy in the learning system. See Section 20.3.

take at each level, what the consequences of these forms will be for learning, and whether the hallmarks found in empirical studies in development map to these consequents.

In this chapter, we touch on enduring questions for cognitive development, that mirror those discussed in past chapters: How is learning possible at multiple levels of abstraction? How can we reconcile computational/rational levels of analysis with the algorithms that children might be carrying out? How are new theories (hypotheses) generated? What representations are already in place to support learning? To seek out the answers, we focus on developments in probabilistic models of children's learning in causal domains. We then discuss empirical evidence for early emerging or core knowledge and how current probabilistic models are approaching the representational considerations.[2]

20.1 Causal inference and intuitive theories

Children's cognitive development is characterized by conceptual revision of intuitive theories. At the core of these intuitive theories is the principle of causality; representations in theories depends on one another. Reasoning about another's mind entails a notion that actions are driven by desires and beliefs. Reasoning about a physical system entails a notion of objects exerting forces on others. Reasoning about biology entails notions of the causal role of variables that result in a living organism's growth, illness, offspring, or death. As such, it becomes important to specify how causality is represented and learned in the developing mind.

As detailed in Chapter 4, **causal graphical models** provide a representational language in the same lexicon of Bayesian probabilistic methods (Pearl, 1988). They were also one of the first types of representations used to characterize cognitive development in the early days of rational constructivism, demonstrating the power of children's learning from covariation that went beyond simple associative models (Gopnik, Sobel, Schulz, & Glymour, 2001; Gopnik et al., 2004; Schulz, Gopnik, & Glymour, 2007; Schulz, Goodman, Tenenbaum, & Jenkins, 2008; Schulz & Sommerville, 2006; Sobel, Tenenbaum, & Gopnik, 2004; Griffiths, Sobel, Tenenbaum, & Gopnik, 2011). Since their use in early studies of children's causal reasoning, much as been learned about when and whether they best characterize children's behavior. In this section, we present a case study in preschooler's causal reasoning, employing a simple framework theory over a bounded causal inference problem, and discuss some of the limitations of this model.

As briefly noted above, the longstanding tension in developmental psychology rests on the nature nurture debate. This debate has played out in children's causal reasoning as well. Some researchers have focused on whether children's causal beliefs are best understood as representations instantiated in domain-specific modules (School & Leslie, 1999) or innate concepts in core domains (Carey & Spelke, 1994; Keil, 1989), while other researchers have emphasized the role of domain-general learning mechanisms (e.g., Gopnik & Schulz, 2004). Research that has centered on the nature of the representation (giving less focus to the learning mechanisms) have suggested that children's causal reasoning respects domain boundaries (Carey, 1985; Estes, Wellman, & Woolley, 1989; Hatano & Inagaki, 1994; Wellman & Estes, 1986; Bloom, 2004; Shultz, 1982). For example, preschoolers deny that psychosomatic reactions are possible, rejecting the idea that (e.g.) being embarrassed can cause you to blush, or being worried can cause a stomach ache (Notaro, Gelman, & Zimmerman, 2001).

Bayesian inference provides a natural framework in which to consider how prior knowledge and data interact in children's causal reasoning. (Schulz, Bonawitz, & Griffiths, 2007) gave children a storybook in which one variable co-occurred with an effect; in the Evidence condition, one cause recurred and the other causes were always novel (i.e., the evidence was in the form A & B $\rightarrow$ E; A & C $\rightarrow$ E; A & D

[2]Thanks to Lauren Leotti for help in preparing this chapter and to Andy Perfors for discussions and feedback on an earlier draft.

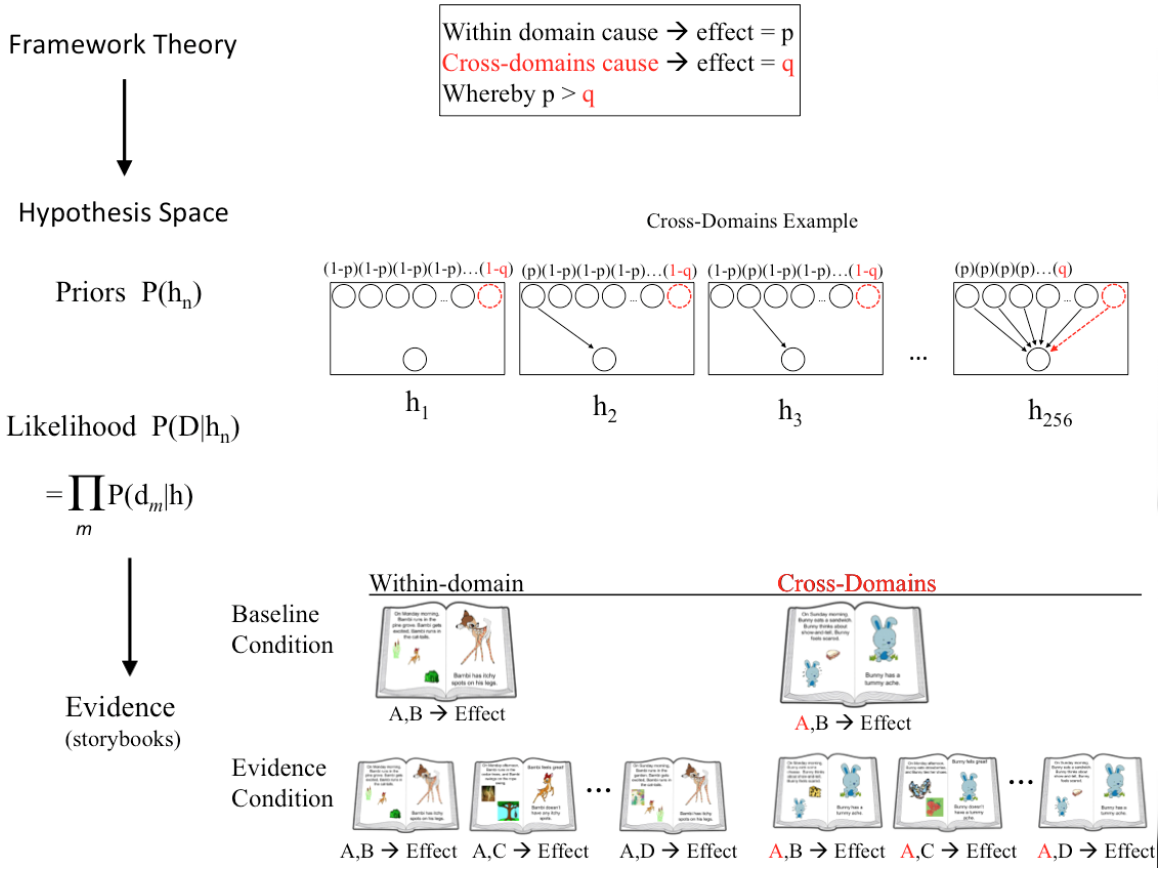

Figure 20.1: A framework theory in which within domain causes are more likely than cross-domain causes generates a space of possible causal models. Nodes in the models depict the causes and effects described in the storybook task from Schulz et al. (2007). Priors over those models are informed by the framework theory such that the probability of each causal node containing a link to the effect is p or q, depending on whether it is a within or cross-domain link. Storybooks read to children include all within-domain links (as in the Bambi books) or one Cross-domain link (as in the Bunny book). Critically, the recurring cause co-occurs with other possible variables, so only provides probabilistic evidence in its favor.

$\rightarrow$ E . . . (see Figure 20.1). In the Within-domain story, all the causes were domain-appropriate (e.g., A deer is running in the cattails, gardens, etc and then gets itchy spots on its legs). Children were able to learn from the data as compared to baseline; they were significantly more likely to infer that "A" was the cause in the Evidence condition. In the Cross-Domain story, the recurring cause (A, being worried) crossed domains from the effect (getting a tummy ache). Children disfavored this variable at Baseline, as compared to domain appropriate alternatives (e.g., eating a sandwich as the cause of the tummy ache). Following evidence, children learned and were significantly more likely to endorse cross-domains variable "A" as the cause than baseline, but their responses were tempered as compared to the within-domain story, suggesting that both the evidence and children's prior beliefs played a role in their causal learning.

This simplified causal learning task provides a nice case study of how probabilistic causal models can capture children's causal reasoning. In the task, children were provided with a force-choice alternative ("Was it variable A or variable B that is causing the effect?"). We can model the probability that children choose "A" as the correct explanation by directly contrasting it with the other possible explanation, B,

given the observed data d,

$$P(A) = \frac{P(A|d)}{P(A|d) + P(B|d)}. \tag{20.1}$$

The explanation that includes A as a possible causal variable is consistent with many different specific causal models linking the variables presented in the storybooks to the possible effects. As such, the probability of each particular candidate explanation given the data is computed by summing over all these possible causal model hypotheses that are consistent with the explanation.

$$P(A|d) = \sum_{h \in \mathcal{H}} P(A|h)P(h|d). \tag{20.2}$$

Here h denotes a hypothesis about the specific causal model linking the variables in the story to the effect. $\mathcal{H}$ represents the full hypothesis space of these models.

In the evidence storybooks, there are 8 different variables presented in the books. A connection between a variable can either be present or absent, representing a causal link between the variable and effect (see Chapter 4). Thus, for this simplified example with 8 variables that can either be "on" or "off", the full hypothesis space includes 2^8 (or 256) possible different causal graphical models (h_n, for $n \in \{1, \ldots, 256\}$), see Figure 20.1. The $P(A|h)$ denotes the probability that the candidate explanation is correct given the specific graphical model. When a causal link between A and the effect is present in the specific graphical model, this is simply 1; when there is no link present, then the value is 0.

The second term, $P(h|d)$, represents the posterior probability of the hypothesis give the data. The probability of a particular causal structure given the data is expanded via Bayes' rule as

$$P(h|d) \propto P(d|h)P(h). \tag{20.3}$$

where $P(h)$ is the prior probability of a particular causal model, and $P(d|h)$ is the likelihood of observing the data d given the causal model h. Here were are simplifying Bayes as a proportional equivalence because, due to the structure of the problem in which we are weighing two alternative explanations against each other, we do not require a normalizing constant.

The precise values of the prior and likelihood probabilities in Bayes are determined by the intuitive causal theory entertained by the observer. As noted above, past research has suggested that children have intuitive theories about the world (Gopnik, Meltzoff, & Kuhl, 1999; Carey & Spelke, 1994; Keil, 1995; Wellman & Gelman, 1992). These theories can be thought of as frameworks for guiding their causal reasoning. A hierarchical probabilistic framework provides a formalism for this intuition. Using a simple framework theory (as denoted in Figure 20.1), within domain variables are generated with probability p where-as cross-domain variables are generated with probability q. As long as p is greater than q (capturing the notion that within domain causes are more likely that cross-domain causes) the qualitative predictions of the models hold across a range of values.

Critically the framework theory gives us a model for how the causal models may be generated, and how their prior probabilities and likelihood weights may be given. Specifically, the prior on any specific causal model is given by weighting the probability of generating each link (given by p or q depending on whether it is a within- or cross-domain variable, and given $1-p$ or $1-q$ when links are not present). The conditional probability distribution of the model provides a means to specify the probability of a causal link generating it's effect (specified by a noisy-OR parameterization, with weight ϵ; see Chapter 4).

Model predictions are well captured by responses from older preschoolers, capturing the trade-offs between learning and strong prior beliefs for within-domain causes. The probabilistic framework provides a formal account for the interaction of children's intuitive theories (and beliefs in within-domain causes) and evidence at multiple levels.

Evidence shapes children's causal explanations in multiple domains (Bonawitz, Fischer, & Schulz, 2012; Bass et al., 2019; Bonawitz & Lombrozo, 2012; Goodman et al., 2006; Amsterlaw & Wellman,

2006). This work provides empirical support for the claim that even young children often act in ways consistent with optimal Bayesian models. However, just because it may be that on average learners responses look like the posterior distributions predicted by these rational models, it it not necessarily the case that learners are actually carrying out exact Bayesian inference at the algorithmic level in the mind. Given the computational complexity of exact Bayesian inference, the numerous findings that both children and adults struggle with explicit hypothesis testing, and the fact that children sometimes only slowly progress from one belief to the next, it becomes interesting to ask *how* learners might be behaving in a way that is consistent with Bayesian inference.

20.2 The Sampling Hypothesis

For most problems, the learner can not actually consider every possible hypothesis; searching exhaustively through all the possible hypotheses rapidly becomes computationally intractable. So, applications of Bayesian inference in computer science and statistics approximate these calculations using Monte Carlo methods, as discussed in Chapter 6. In these methods, hypotheses are sampled from the appropriate distribution rather than being exhaustively evaluated. What's interesting is that a system that uses this sort of sampling will be variable — it will entertain different hypotheses apparently at random. But this variability will be systematically related to the probability distribution of the hypotheses — more probable hypotheses will be sampled more frequently than less probable ones. This sampling method thus provides a way to reconcile rational reasoning with variable responding, a hallmark of early childhood. The **Sampling Hypothesis** is the idea that human learners may take a similar approach – an idea introduced in the context of adults in Chapter 11.

There is growing empirical support suggesting that children are doing something that looks like sampling (Bonawitz, Denison, Griffiths, & Gopnik, 2014). For example, children provide explanations for causal events in proportion to their posterior probabilities (Denison, Bonawitz, Gopnik, & Griffiths, 2013). These studies show that children's responses are not simply noisily maximized and further that responses go beyond simple frequency tabulations in these causal learning tasks. These studies also raise questions about *how* a learner may sample hypotheses.

There are lots of ways in which a learner could sample hypotheses. They may resample every time they observe new data, they may take a hypothesis and stick with it, until they have impetus to re-evaluate (e.g., maybe data that is very unlikely given the current hypothesis.) And, when they re-evaluate, they may make subtle changes to the hypothesis they are currently entertaining, or go back and resample completely from the full posterior distribution. They may sample a few hypotheses or just one. All of these ideas about how a learner "searches" through a space have analogues in computer science and machine learning. Here we will contrast two: independent sampling and a modification of a classic algorithm, the **win-stay, lose-shift** algorithm.

The simplest idea, independent sampling, is that each time a learner observes new data, she recomputes the updated posterior and samples a guess from that updated distribution. This kind of approach to updating predicts that subsequent guesses from a single learner will be independent. That is, knowing that a learner predicts a specific hypothesis at a particular time tells you nothing about what hypothesis they are likely to have after the next observation of data.

However, another possibility is that a learner tends to maintain a hypothesis that makes a successful prediction and only tries a new hypothesis when the data weigh against the original choice. This means that and individual will tend towards "stickiness," being more likely to keep current hypothesis. This predicts dependency between responses. The central idea of maintaining a hypothesis provided it successfully accounts for the observed data and otherwise shifting to a new hypothesis led to the name for this strategy: win-stay, lose-shift.

20.2.1 Win-stay, lose-shift in children's causal inferences

A general form of the win-stay, lose-shift (WSLS) algorithm has a long history in computer science and human concept learning (Robbins, 1952; Restle, 1962; Levine, 1975). However, it is possible to find specific classes of WSLS that approximate Bayesian inference. Specifically, there are different policies for when a hypothesis under current consideration should be rejected and different rules for how the next hypothesis should be drawn. It is possible to mathematically discover the specific instantiations of WSLS that provides a means to approximate Bayesian inference. That is, despite an individual's tendency towards "stickiness," there exist WSLS policies in which overall proportion of responses will reflect a sample from the full posterior distribution. Indeed, Bonawitz, Denison, Gopnik, and Griffiths (2014) report two such specific instantiations of WSLS that approximate the posterior. These algorithms are based on stochastically deciding to draw a new hypotheses from the posterior with a probability determined by how well the current hypothesis accounts for each new piece of data is observed.

To compare the WSLS and independent sampling approaches, Bonawitz et al. (2014), developed a mini-microgenetic method for investigating children's causal learning. In studies of children's learning, researchers typically look at responses at just one point —- after all the evidence has been accumulated. However, an effective way to determine the actual algorithms that a learner uses to solve these problems is to examine how her behavior changes, trial-by-trial, as new evidence is accumulated. The differences between the dependency predictions for these algorithms can thus be used as a means to evaluate the process by which learners might update their beliefs.

In Bonawitz et al. (2014), preschool-aged children were introduced to a machine and an experimenter demonstrated that each of these three kinds of blocks activate the machine with different probability when they are placed on it. The red blocks activate the machine on five out of six trials, the green blocks on three out of six trials, and the blue blocks just once out of six trials. Then children were shown a novel block that had lost its color, and children were asked to take an initial guess about what color (red, green, blue) they thought it was supposed to be. After that, the block was placed on the machine once and children observed one of two possible outcomes as noted in Figure 20.2.

To better understand the WSLS algorithm, we step through one of the particular instantiations in Figure 20.2. A learner starts out by sampling a hypothesis from the prior distribution, before seeing any data about the mystery block. Here the learner happens to choose red (as if a weighted die was rolled with this outcome.) Then the block is set on the machine and it turns out that it activates the toy. Because it is given in the demonstration phase of this experiment that the red block activates the machine 5/6 times, the likelihood is thus simply 5/6. Now the learner has to "decide" whether to stay or switch. The coin is now weighted 5/6 to stay. In this particular example, when it is flipped, it happens to come up in the more likely case, to stay. Observing a second piece of data reveals that the machine does not activate. The likelihood is computed given *only this one piece of new evidence*. The currently hypothesized block, the red block, does not activate the machine with probability 1/6, as given in the demonstration phase. Thus, like likelihood is simply 1/6 and the weighted coin is flipped with probability "stay" at 1/6. In this example, the coin happens to come up "switch". At this point, the learner draws from the updated posterior, which includes all evidence observed so far. This time, the sampling die comes up green. So this is the new hypothesis that the learner holds in mind for the next observation.

An individual learner may look like they are randomly veering from one hypothesis to the next, starting with red and sticking with it, then switching to green, even though green may not be the most likely choice, etc. However, the lovely and surprising feature of WSLS it that summing over enough participants, on aggregate, WSLS returns the posterior distribution. In particular, WSLS also helps solve the algorithmic problem of Bayesian inference, because the learner can maintain just a single hypothesis in their working memory and need only re-compute and resample from the posterior on occasion, but the responding of participants on aggregate still acts like a sample from a distribution (as in Figure 20.3).

What is particularly nice about WSLS is that is provides a means for a young learner to make inferences

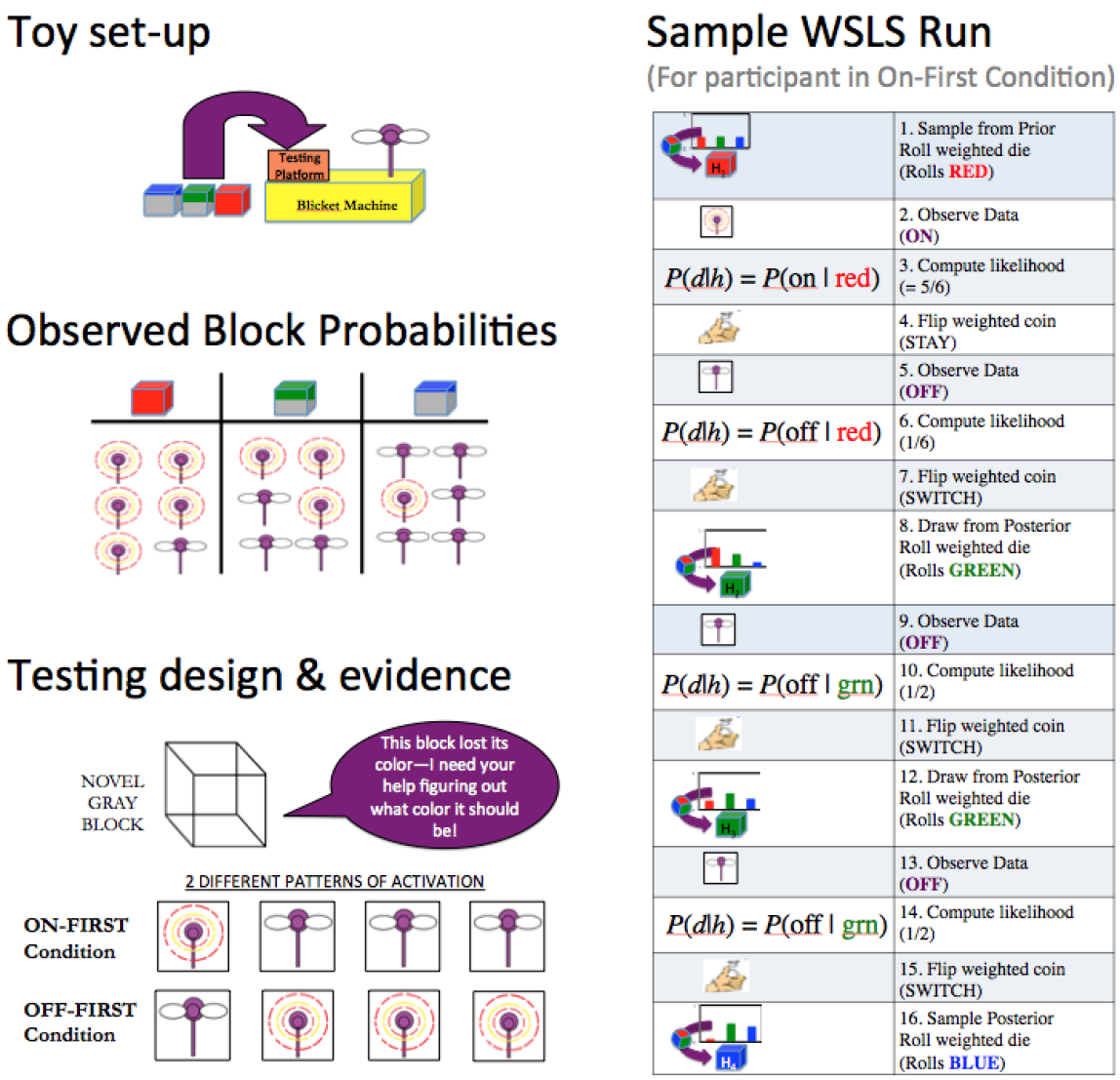

Figure 20.2: Win-stay lose-shift in children. The left column shows the method employed in Experiment 2 of Bonawitz et al. (2014). Children learned that blocks of different colors had difference causal affordances. Then a new block was introduced that lost its color. During test, in the On-first condition for example, children saw a pattern of evidence in which the mystery block first caused the toy to light, but then on subsequent trials the toy failed to light. On the right column is an example run of a single individual carrying out the win-stay, lose-shift algorithm as evidence is observed over the four trials of the experiment.

given probabilistic information: the algorithm only considers a single hypothesis, but acts like a sample from the distribution making it computationally more attractive than independent sampling, because the learner need not compute and resample from the full posterior after each observation. Thinking about models at the algorithmic level can reveal important information about how children move from one belief to the next as in (Bonawitz et al., 2014). These studies are important first steps in connecting the computational level and algorithmic level, because they show how behavior can approximate Bayesian posterior distributions, without requiring the learner to carry out exact Bayesian inference.

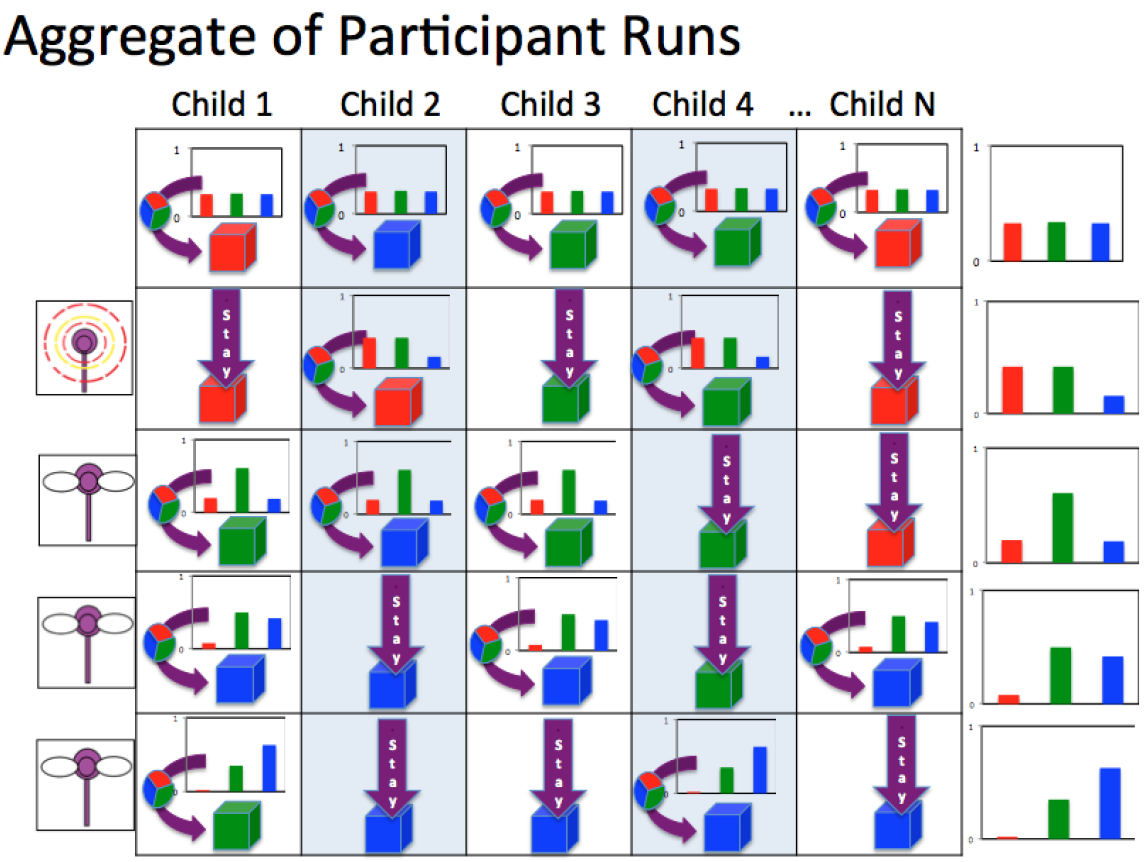

Figure 20.3: Example depiction of the patterns of several different child participants in the Bonawitz et al. (2014) study after observing each new phase of evidence (with each new row). Although individual children appear to randomly veer from one hypothesis to the next (with a light "stickiness" to favor previously held hypotheses), on aggregate the distribution of children's responses capture the Bayesian posterior, as proved by the specific WSLS algorithm presented in Bonawitz et al. (2014).

However, the study described above leaves open interesting additional questions, such as whether there are dependencies even during sampling. For example, in WSLS a resampled hypothesis is independent from the previously held hypothesis, but it is possible that the hypothesis a current learner is entertaining is used as a kind of anchor for resampled hypothesis. There is reason to believe that adults at least show dependency even between switched hypotheses, though children may have weaker dependencies in the generation of subsequent examples (Bonawitz et al., In Revision). This idea is consistent with the notion of simulated annealing in MCMC algorithms as discussed in Chapter 6. Specifically, Gopnik and others have suggested that development may mirror the gradual cooling of an initially hot search. Childhood may be a time in which the mind employs wide and variable search of spaces at "hot temperatures", supporting the notion that children are the creative innovators of the human race. In contrast, adults may employ colder-temperature searches, sticking with good enough hypotheses when they are discovered, but missing out on unlikely, but possibly better alternatives (Gopnik, Griffiths, & Lucas, 2015; Gopnik et al., 2017).

A second open question pertains to how learners handle cases in which hypotheses are not easily enumerated. In the WSLS example, there were only three hypotheses available to learners (i.e. it was either a red block, blue block, or green block). However, in most learning problems, hypotheses are

not enumerated a priori, and may even be infinite. We turn a solution to this approach, as it informs developmental theories below.

20.2.2 Framework learning as stochastic search

In the studies described above, children rapidly make inferences about the most likely causal models, probability matching responses in proportion to the posterior distribution. However, rich theory change as observed in typical development does not often appear so rapid, accurate, or linear. Learning takes time (Carey, 2009). Children move from one belief to something only slightly better; it is not often to see big conceptual jumps from beliefs that are conceptually incorrect to ones that are suddenly completely coherent. Instead, beliefs often gradually progress to capture more accurate and complete representations during development, (for example, see Wellman & Liu, 2004 for a compelling case-study of the gradual development of a Theory of Mind). The transitions of children's beliefs have similarly been characterized as a somewhat step-wise rather than suddenly all or none (e.g., Siegler, 1996). There are even apparent regressions in learning (e.g., Marcus et al., 1992; Thelen & Fisher, 1982). Why take a step backwards when a previously supported behavior has proved more accurate?

How can we understand this nonlinear, noisy, and slow progression of learning? One possibility is that children are performing a kind of stochastic search (as discussed in Chapter 6). If so, we would expect that beliefs may appear to jump as if randomly from one point to the next. Similar evidence may lead children to different conclusions. Learning may unfold relatively slowly overtime. These models can be applied to help explain belief revision in more sophisticated domains than those described above. Inspired by the stochastic search approach (detailed in Ullman, Goodman, & Tenenbaum, 2012), Bonawitz, Ullman, Bridgers, Gopnik, and Tenenbaum (2019) explored how preschool-aged children solve the chicken-and-egg problem of theory learning in the domain of magnetism, jointly identifying causal laws and the hidden categories they are defined over. The hierarchical model employs stochastic search over logical laws and predicates (probabilistic context-free Horn Clause Grammar) that form a space of "intuitive theories." To find the "best" theory, search is carried out over the space using a grammar-based Metropolis-Hastings sampling method (also as in Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008). Simulating runs in this space revealed signatures of developmental belief revision; for example the model revealed different convergence rates on individual runs capturing a similar phenomenon in development in which individual children may arrive at final, "correct" beliefs at different time-points, despite similar evidence. In other work by Piantadosi, Tenenbaum, and Goodman (2012), a similar modeling approach captures the trademark progression of number knowledge, as children gradually build representations of number concepts "1", then "2", then "3", and eventually "jump" to infer the cardinal principal of counting, by applying a new sequential operator rule to the count-list. These approaches critically demonstrate how core or "primitive" cognitive operations can combine to form more complex theories, and how a stochastic search process over this large space mirrors developmental trajectories.

Here we have focused on Bayesian models, starting with how child learners may make causal inferences about specific data, moving up to models, and finally how abstract theories from those models may be learned. But in development, sometimes it appears as if children learn the abstract framework theories before they develop the specific ones (Wellman & Gelman, 1992; Simons & Keil, 1995). How could an abstract framework be inferred before a more specific level model? The idea that framework theories (at least sometimes) seem to be in place before specific level models provided tentative evidence that such knowledge is core (e.g., Spelke, Breinlinger, Macomber, & Jacobson, 1992). However, another answer comes again from hierarchical Bayesian models. By applying computational modeling to developmental problems, it can be shown that there are cases in which the abstract learning happens at the same time — and sometimes even precedes — the specific level (Goodman, Ullman, & Tenenbaum, 2011). This has been called the **blessing of abstraction** by Goodman et al. (2011), and provides an account of this surprising developmental phenomenon.

As noted above, models show how and when specific frameworks theories could be learned (though such a demonstration does no provide proof that they are learned). In development, it is often taken "for granted" that there are domain general concepts that must be built into the system as well, such as causality. The accounts that suggest some learning happens have not provided a story about *how* such bootstrapping could occur. As in the case of early emerging framework theories, the lack of the story about how more domain general concepts like causality could be learned have led many to assume that such concepts must be innate. Here again HBMs provides a new proposal about how this kind of knowledge could be learned and about what must be built in given these models, such as perceptual input analyzers (Goodman et al., 2011). Some evidence suggests that children may develop causal theories piecemeal (Bonawitz et al., 2010) in the sense that associative and intervention information may not be spontaneously bound (at least in domains in which children have less familiarity).[3] Providing a story about how a domain general principle like causality develops may also help to answer fundamental questions about how other domain general theories can develop as well.

20.3 Core Knowledge

One of the themes of the last few chapters is that of program induction: The state of a mind at a given moment can be captured by a particular generative program, and the psychological process of thinking and learning can be seen as the mind executing program induction algorithms, leading to a new program that differs slightly or markedly in parameters and structure. One could imagine all of development as proceeding from a very simple program with no structure, and discovering new routines, functions, and variables at different levels of abstraction. This is not a new idea. Alan Turing, writing before the fields of artificial intelligence and cognitive science got their official start, proposed that the path to adult-level artificial intelligence started with a child machine that learned new programs (Turing, 1950). Turing, like many others, pictured the starting point of a child's mind as something like a notebook with "rather little mechanism, and lots of blank sheets" (p. 546, Turing, 1950).

From an evolutionary and computational perspective, the notion of a blank notebook seems odd. Imagine for example the task of writing an algorithm to find a useful, short program that generates the sequence 1011011101111011111..., starting absolutely from scratch. One could try to order all possible programs on a universal Turing machine and search through them in various clever ways (Levin, 1973), but the search time for even a simple inversion problem is frightful. It is wasteful and unnecessary for each fresh organism to roam this landscape of possible programs starting from the same blank point. How much more useful it would be if evolution provided organisms with a "start up library" of useful functions, variables, and routines (Lake, Ullman, Tenenbaum, & Gershman, 2017). Routines could be relatively hard-coded, allowing an okapi to get up and run soon after birth. Routines can also be relatively content-less, such as a routine for jumping back when detecting a scurrying motion, without necessarily having a fully developed notion of what a spider is. But one could imagine, and in fact one would expect, built-in variables and functions and libraries that are far more general, abstract, and useful than hard-coded, content-less routines (Ullman & Tenenbaum, 2020; Baum, 2004).

From an empirical perspective, the notion of a blank notebook turns out to be wrong too. Research in cognitive development over the past decades has uncovered that infants have commonsense expectations about the workings of the world, present early on or innately (Spelke & Kinzler, 2007; Spelke, 1990; Woodward, 1998; Csibra, Bíró, Koós, & Gergely, 2003; Phillips & Wellman, 2005; Carey & Spelke, 1994). As might be expected from an evolutionary start up library, these expectations are conserved across cultures, and seem to be present in non-human animals as well. The expectations and principles are not all-encompassing, and are modular in nature, focusing on several core domains, in particular number, space, agents, objects, and social relations. The principles are abstract and general, but with

[3]See also Waismeyer, Meltzoff, and Gopnik (2015) and Meltzoff, Waismeyer, and Gopnik (2012).

signature limits, and knowledge within core domains is acquired and developed throughout childhood. To give a particular example, even young infants believe that solid bodies should not pass through one another (Spelke et al., 1992). This expectation holds true for all entities classified as objects. If an infant encounters a toy truck for the first time, they will expect the truck not to pass through a wall. Infants do not need separate and exhaustive re-training of their nervous system to form this expectation for a truck, and then a duck, and then a puck. This is what we mean by saying that the principle is abstract and general. However, it also has signature limits: infants cannot reason about more than several objects at once, and not all entities count as objects. Infants are lousy early on at reasoning about non-rigid bodies, and if the truck is perceived as an agent, then some of their physical expectations cease to hold. These principles can be exhibited empirically in different ways, but a common method is to show infants different displays and to measure the infants' looking time, which is a proxy for their surprise (though see Kidd, Piantadosi, & Aslin, 2012). For example, the infants may see display A in which a rolling ball is stopped by a wall, and display B in which the ball appears to roll through the wall. Infants will on average look longer at B than A, indicating that they expected the ball not to pass through the wall.

We do not attempt a full specification of these principles of **Core Knowledge**, but highlight several such expectations (for a review, see Spelke & Kinzler, 2007). In all that follows, we emphasize that there is uncertainty both on the time-line itself (the earliest age at which infants demonstrate these expectations), and conceptual uncertainty over how best to characterize these expectations (and see the discussion below about formalization). For physics and objects, infants expect bodies to persist, cohere, follow smooth and continuous paths, not act at a distance, and not to pass through one another. For agents and animate beings, even pre-verbal infants expect agents to act efficiently to achieve goals, and to trade off costs and rewards given environmental constraints (Spelke & Kinzler, 2007; Csibra et al., 2003; Csibra, 2008; Liu, Ullman, Tenenbaum, & Spelke, 2017). Infants also distinguish social and anti-social others (Hamlin & Wynn, 2011; Hamlin, Wynn, & Bloom, 2007; Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013), although again it is an open question exactly how early this distinction emerges. For spaces and places, young children and animals can use the layout of extended surfaces to reorient themselves, and locate themselves and other objects with regards to the distances and directions to these surfaces, though navigation with respect to landmarks and small forms seems to rely on a separate process, with the two systems only being integrated at a later age (Hermer & Spelke, 1994; Dehaene, Izard, Pica, & Spelke, 2006; Spelke & Lee, 2012).

On their own, the principles of Core Knowledge do not directly specify how to implement them in a machine. Consider for example an engineer who accepts the idea Core Knowledge hypothesis but now wants to design a child machine with those principles. How should she build in things such as **The Principle of Continuity** or **The Principle of Efficiency**? This is an outstanding question in current computational cognitive science and machine learning, and different implementations amount to different answers to what Core Knowledge is, exactly. One possible route, at least for commonsense physics and psychology, is to assume that the generative models that capture adult reasoning exist from the beginning in some form. Or rather, as these are hierarchical models, the suggestion is that the top-level of the hierarchy is built-in or acquired early on in infancy. To use a rough analogy, imagine a game programmer that wants to design a new computer game. The programmer would likely not want to design the game from scratch – this would be onerous and replicating already existing libraries of functions and routines. Sure, one particular game might involve jumping over alligators to collect diamonds, while another has you flinging boomerangs at killer bees, but both of them would rely on the same basic physics engines and planners, just with different sprites and specific dynamics. On this picture, the infant is like a game programmer that is watching a game they did not design, receiving the frames of the game and trying to figure out, given their libraries, what the underlying code is (Tsividis, Pouncy, Xu, Tenenbaum, & Gershman, 2017).

To be more specific, for intuitive physics the generative model is that of the **Intuitive Physics Engine** (see Chapter 15 and, e.g., Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016). The idea would then be that the basic skeleton of a rough game engine is built in:

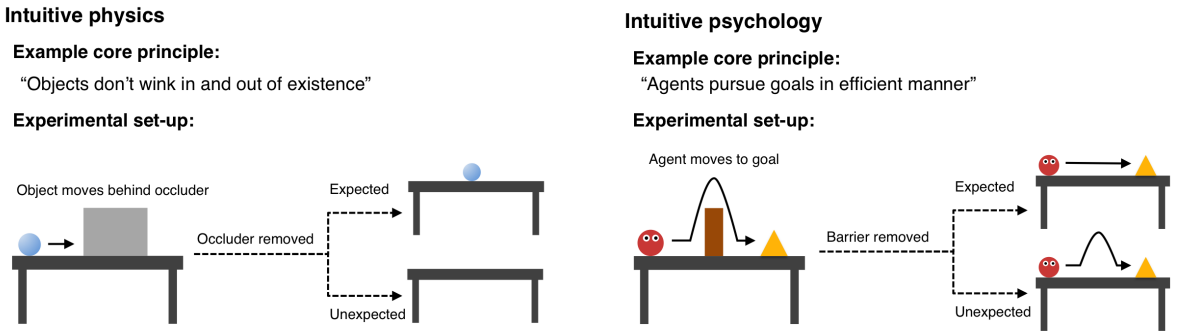

Figure 20.4: Example principles and supporting experiments in the core domains of intuitive physics and intuitive psychology. Intuitive physics: After observing an object moving behind an occluder and seeing the occluder being removed, young infants will express greater surprise at a display showing the object had disappeared (see, e.g., Spelke, 1990), leading to the formulation of the **Principle of Permanence**. Intuitive psychology: After observing an animate entity jump over a barrier to get to a target item, and then seeing the barrier removed, infants express surprise when the animate entity replicates its spatio-temporal trajectory (see, e.g., Csibra et al., 2003). This and similar experiments can be explained by posited that infants parse the scene in terms of goals and agents, and expect agents to pursue goals efficiently.

we assume objects that have properties, and dynamics that update world states. The specific properties, object hierarchies, forces, and dynamic equations would then need to be learned (Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018; Lake et al., 2017). Various programs could then be learned on this basis, including simple ones that update object position based on velocity or random diffusion, along with simple collision resolvers as exist in nearly every game engine. Such programs are sufficient for quantitatively explaining infants' looking patterns to physical stimuli (Téglás et al., 2011), and embody core principles like permanence and cohesion without having to build them in explicitly. Such a program would predict that objects do not disappear or suddenly move in a discontinuous way, but *not* because such ideas exist explicitly in the code. Rather, it is because when the program simulates objects forward the objects do not behave that way. Other principles are more explicit. For example, the **Principle of Solidity** can captured by the program for collision handling. Of course, such programs would be approximations to and simplifications of real world physics, and such principled approximations can actually account for puzzling findings in infant reasoning about objects (Ullman, Spelke, Battaglia, & Tenenbaum, 2017).

For intuitive psychology, the generative model is that of planning or **Bayesian theory of mind** (see Chapter 14 and, e.g., Baker, Saxe, & Tenenbaum, 2009; Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017). As with physics, the idea would be that the basic skeleton of a planner is built in: we assume agents that have utility functions with rewards and costs, and the ability to take actions to increase rewards and lower costs. The learning process would then be to discover the types of agents, rewards, actions, skills, constraints, and costs that exist in different environments (Lake et al., 2017). Such a basic skeleton is sufficient for explaining many findings in preschool children (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016), and a similar simple model that trades off rewards and costs can explain several key findings in infants' reasoning about agents (Liu et al., 2017). In fact, the basic skeleton of a planner can be seen as embodying the **Principle of Efficiency**: agents are things with goals that act efficiently to achieve those goals within constraints.

Skeletal structured generative models are one route for embodying and implementing Core Knowledge, but other approaches are being developed as the AI and machine learning community re-engages with findings from cognitive development. One such promising direction, specifically in intuitive physics, is to combine artificial neural networks or graphs with different minimal notions of objects, and to let the

network discover the dynamics and interactions between objects (for several recent examples, see Mrowca et al., 2018; Battaglia et al., 2018; Battaglia, Pascanu, Lai, Jimenez Rezende, & Kavukcuoglu, 2016; Chang, Ullman, Torralba, & Tenenbaum, 2016). A very different computational approach, however, is try to recover the principles of Core Knowledge purely from vast amounts of empirical data (for recent examples in intuitive physics and psychology, respectively, see Piloto et al., 2018; Rabinowitz et al., 2018). Such blank-slate models on their own do not yet generalize well, but it is too early to say whether this direction will turn out to be successful or not. Such blank-slate approaches (usually) do not claim to recover the trajectory of infant development, but rather that of evolution. This argument is common in other areas of Machine Learning, where humans do well without vast amounts of training data. For example, an feed-forward artificial neural network may require thousands of training images to recognize or the frame-equivalent of hundreds of hours to play a video game at a reasonable level (Mnih et al., 2015), whereas humans might need only two or three new examples of a new animal or several minutes with a new video game to reach the same performance. The claim is then that evolution has trained up the priors of the network, whereas the machine is training them anew. But this misses the dynamics of how evolutionary learning happens. Organisms do not perform the equivalent of being given several labeled horses, then passing on their trained neural state to offspring, this would be a folly on the level of the Lamarkian view that blacksmiths pass on their strong arms to their children. Re-discovering the priors of agents, objects, and other core knowledge would require reverse engineering evolution, which has different dynamics than training-and-validating a network, and involves searching the uber-space of functions and variables, each set of which defines a sub-space that an organism can explore over development (see, e.g., Baum, 2004).

In closing, the program induction view suggests that an organism would gain a serious leg-up if it starts with a library of initial functions, routines, and variables. Initial built-in knowledge can involve specific feature detectors and hard-coded routines, but also more flexible and abstract concepts, and the recent decades of cognitive developmental research suggests that infants start with such a library for several core domains. Such functions and routines may take the form of structured generative models, which embody the principles discovered empirically, and are similar in structure to later intuitive theories, and are also the basis for the intuition that constructs scientific theories.

20.4 Future directions

Cognitive development can be framed as a rational process: learners infer and build models of the world through an accumulation of data and time, as they search over possibilities. The progress and promise of probabilistic models are much like human development, in which more empirical work and time will lead to richer models of this process. Many areas remain open for exploration in the space of probabilistic models of development, and we highlight two particularly promising paths: the role of resource limitations in qualitative developmental shifts, and the grounding of probabilistic developmental models with perceptually-driven, bottom-up learning.

A once and future challenge for probabilistic models of development is the changing boundedness of the learning process. Cognitive changes in processing capacity can lead to quantitative and qualitative shifts in learning. For example, if your semantic memory capacity grows over development, you can store and retrieve more evidence during inference, a quantitative shift. But, it can also change the process you use to store and retrieve evidence altogether, a qualitative shift. Changes in working memory and attention can affect not only the richness and content of mental simulations, but whether simulations are used as a means of prediction and inference at all. Changes in our representations of events can influence not only the speed of revising our beliefs, but also our assumptions of whether some observations count as evidence at all. Questions surrounding the cognitive effects of changing boundedness have a long history in the information processing approach to cognitive development (Klahr & Wallace, 2022). In probabilistic models of cognition, they have a long future too.

A separate central challenge for probabilistic models of development concerns the need to ground this learning in perceptual representations. Hierarchical Bayesian models are presented as allowing researchers to unify ideas from both the nature and the nurture side of the ongoing nature-nurture debate. However, the focus of much of the work in this area has advanced by supposing that central and difficult perceptual processes are largely solved, and working with their output. For example, some models of intuitive physics assume a perceptual process that recognizes and distinguishes different objects. On the flip side, some of the best perceptual processing learning models currently make advances without paying much heed to learning programs over structured representations. Crucial breakthroughs are waiting for the right program learning models that take seriously the need for the last stages of a hierarchy to make contact with the basic input units of perception.

20.5 Conclusion

Children are the original learners, and most of the lessons in this book are inspired by their study. In this chapter, we provided a few examples that show how probabilistic models can speak to developmental questions. We presented cases in which probabilistic models similarly help solve open problems in the development of causal learning, demonstrating how children's prior beliefs and evidence interact to support causal inference, how approximation algorithms can explain dependencies in causal belief change, and even how causal search at abstract levels may unfold. Finally, we suggested that probabilistic models can be informed by the initial constraints observed in early infancy, operating as skeletal structures that afford rapid development of more complex, abstract structures over development.

The application of probabilistic models helps researchers to be precise about the content of early developing beliefs. These models illuminate early inductive biases that shape the learning process. They help reconcile debates regarding whether children are "noisy" learners or rational approximators. They provide a unifying framework for nature-nurture debates, demonstrating that structured representations can both drive inference and be inferred, even early in development.

The human experience weaves together social, emotional, and cognitive experiences to form a rich tapestry. Probabilistic models are a promising tool to help distinguish interlocking patterns, but we have only just begun to pull at the different threads. This approach is still in its infancy, and it is unlikely that we will develop general, unified theories of children's learning in the same time span that children take to develop their own abstract models of the world. But probabilistic models provide a framework that could make this goal possible, if we continue to let them develop.

References

Amsterlaw, J., & Wellman, H. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*, *7*, 139-172.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 0064.

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

Bass, L., Gopnik, A., Hanson, M., Ramarjan, D., Shafto, P., Wellman, H., & Bonawitz, E. (2019). Children's developing theory of mind and pedagogical evidence selection. *Developmental Psychology*, *55*, 286-302.

Battaglia, P., Hamrick, J., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Battaglia, P., Pascanu, R., Lai, M., Jimenez Rezende, D., & Kavukcuoglu, K. (2016). Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems 29.*

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al.. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261.*

Baum, E. B. (2004). *What is thought?* MIT Press.

Bloom, P. (2004). *Descartes' baby.* Basic Books.

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

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(10), 497–500.

Bonawitz, E., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A., Woodard, J., & Schulz, L. (2010). Just do it? investigating the gap between prediction and action in toddlers' causal inferences. *Cognition*, *115*, 104-117.

Bonawitz, E., Fischer, A., & Schulz, L. (2012). Teaching three-and-a-half-year-olds to revise their beliefs given ambiguous evidence. *Journal of Cognition and Development*, *13*, 266-280.

Bonawitz, E., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*, 1156-1164.

Bonawitz, E., Ullman, T. D., Bridgers, S., Gopnik, A., & Tenenbaum, J. B. (2019). Sticking to the evidence? a behavioral and computational case study of micro-theory change in the domain of magnetism. *Cognitive Science*, *43*(8), e12765.

Bonawitz, E., Walker, C., Hemmer, P., Abbot, J., Griffiths, T., & Gopnik, A. (In Revision). Variability in preschoolers' cognitive search.

Carey, S. (1985). *Conceptual change in childhood.* MIT Press.

Carey, S. (2009). *The origin of concepts.* Oxford University Press.

Carey, S., & Spelke, E. (1994). Domain-specific knowledge and conceptual change. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (p. 169-200). Cambridge University Press.

Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, *63*, 515-533.

Chang, M. B., Ullman, T., Torralba, A., & Tenenbaum, J. B. (2016). A compositional object-based approach to learning physical dynamics. *arXiv preprint arXiv:1612.00341*.

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition*, *107*(2), 705 – 717.

Csibra, G., Bíró, S., Koós, O., & Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111–133.

Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006). Core knowledge of geometry in an Amazonian indigene group. *Science*, *311*(5759), 381–384.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. (2013). Rational variability in children's causal inferences: The sampling hypothesis. *Cognition*, *126*, 285-300.

Estes, D., Wellman, H., & Woolley, J. (1989). Children's understanding of mental phenomena. *Advances in Child Development and Behavior*, *22*, 41-87.

Goodman, N., Baker, C., Bonawitz, E., Mansinghka, V., Gopnik, A., Wellman, H., Schulz, L., & Tenenbaum, J. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society.*

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: a language for generative models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence.*

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *118*(1), 110–119.

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

Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, *24*(2), 87–92.

Gopnik, A., Meltzoff, A. N., & Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Co.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., Aboody, R., Fung, H., & Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Science*, *8*, 371-377.

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*, 620-629.

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*, 1407-1455.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, *26*(1), 30 – 39.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*, 557–559.

Hamlin, K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209–226.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Hatano, G., & Inagaki, K. (1994). Young children's naïve theory of biology. *Cognition*, *50*, 171-188.

Hermer, L., & Spelke, E. S. (1994). A geometric process for spatial reorientation in young children. *Nature*, *370*(6484), 57.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604.

Keil, F. (1995). The growth of causal understanding of natural kinds. In D. Sperber & D. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (p. 234-267). Clarendon Press/Oxford University Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development.* MIT Press.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS One*, *7*(5), e36399.

Klahr, D., & Wallace, J. G. (2022). *Cognitive development: An information-processing view.* Routledge.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Levin, L. A. (1973). Universal sequential search problems. *Problemy Peredachi Informatsii*, *9*(3), 115–116.

Levine, M. (1975). *A cognitive theory of learning: Research on hypothesis testing.* Erlbaum.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the society for research in child development*, i–178.

Markman, E. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*, 57-77.

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

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533.

Mrowca, D., Zhuang, C., Wang, E., Haber, N., Fei-Fei, L. F., Tenenbaum, J., & Yamins, D. L. (2018). Flexible neural representation for physics prediction. In *Advances in Neural Information Processing Systems* (pp. 8813–8824).

Notaro, P., Gelman, S., & Zimmerman, M. (2001). Children's understanding of psychogenic bodily reactions. *Child Development*, *72*, 444-459.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems.* Morgan Kaufmann.

Phillips, A. T., & Wellman, H. M. (2005). Infants' understanding of object-directed action. *Cognition*, *98*, 137-155.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2012). Bootstrapping in a language of thought: A formal model of numerical concept learning. *Cognition*, *123*(2), 199-217.

Piloto, L., Weinstein, A., TB, D., Ahuja, A., Mirza, M., Wayne, G., Amos, D., Hung, C.-c., & Botvinick, M. (2018). Probing Physics Knowledge Using Tools from Developmental Psychology. *arXiv:1804.01128 [cs]*.

Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S., & Botvinick, M. (2018). Machine theory of mind. *arXiv preprint arXiv:1802.07740*.

Restle, F. (1962). The selection of strategies in cue learning. *Psychological Review*, *69*(4), 329-343.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*, 527–535.

School, B., & Leslie, A. (1999). Explaining the infant's object concept: beyond the perception/cognition dichotomy. In E. Lepore & Z. Pylyshyn (Eds.), *What is cognitive science?* (p. 26-73). Blackwell.

Schulz, L., Goodman, N., Tenenbaum, J., & Jenkins, C. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*(2), 211-233.

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

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*, 1124-1139.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and children's inferences about unobserved causes. *Child Development*, *77*, 427-442.

Shultz, T. R. (1982). Causal reasoning in the social and non-social realms. *Canadian Journal of Behavioural Science*, *14*, 307-322.

Siegler, R. (1996). *Emerging minds: The process of change in children's thinking.* Oxford University Press.

Simons, D. J., & Keil, F. C. (1995). An abstract to concrete shift in the development of biological thought: The insides story. *Cognition*, *56*(2), 129–163.

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*, 303-333.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*(1), 29–56.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605 – 632.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Spelke, E. S., & Lee, S. A. (2012). Core systems of geometry in animal minds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 2784–2793.

Thelen, E., & Fisher, D. M. (1982). Newborn stepping: An explanation for a" disappearing" reflex. *Developmental psychology*, *18*(5), 760.

Tsividis, P. A., Pouncy, T., Xu, J. L., Tenenbaum, J. B., & Gershman, S. J. (2017). Human learning in Atari. In *AAAI Spring Symposium Series, Science of Intelligence: Computational Principles of Natural and Artificial Intelligence.*

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054–1059.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*(4), 455–480.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82.

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533–558.

Waismeyer, A., Meltzoff, A., & Gopnik, A. (2015). Causal learning from probabilistic events in 24-month olds: an action measure. *Developmental Science*, *18*, 175-182.

Wellman, H., & Estes, D. (1986). Early understanding of mental entities: A reexamination of childhood realism. *Child Development*, *57*, 910-923.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337-375.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child development*, *75*(2), 523–541.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Xu, F., & Kushnir, T. (2013). Infants are rational constructivist learners. *Current Directions in Psychological Science*, *22*, 28-32.