Concepts in a Probabilistic Language of Thought

Noah D. Goodman[*], Stanford University
Joshua B. Tenenbaum, MIT
Tobias Gerstenberg, MIT

February 15, 2014

To appear in Concepts: New Directions*, Eds. Margolis and Laurence, MIT Press.*

1 Introduction

Knowledge organizes our understanding of the world, determining what we expect given what we have already seen. Our predictive representations have two key properties: they are productive, and they are graded. Productive generalization is possible because our knowledge decomposes into *concepts*—elements of knowledge that are combined and recombined to describe particular situations. Gradedness is the observable effect of accounting for uncertainty—our knowledge encodes degrees of belief that lead to graded probabilistic predictions. To put this a different way, concepts form a combinatorial system that enables description of many different situations; each such situation specifies a distribution over what we expect to see in the world, given what we have seen. We may think of this system as a *probabilistic language of thought* (PLoT) in which representations are built from language-like composition of concepts and the content of those representations is a probability distribution on world states. The purpose of this chapter is to formalize these ideas in computational terms, to illustrate key properties of the PLoT approach with a concrete example, and to draw connections with other views of conceptual structure.

People are remarkably flexible at understanding new situations, guessing at unobserved properties or events, and making predictions on the basis of sparse evidence combined with general background knowledge. Consider the game of tug-of-war: two teams matching their strength by pulling on either side of a rope. If a team containing the first author (NG) loses to a team containing the third author (TG), that might provide weak evidence that TG is the stronger of the two. If these teams contain only two members each, we might believe more in TG's greater strength than if the teams contain eight members each. If TG beats NG in a one-on-one tug-of-war, and NG goes on to beat three other individuals in similar one-on-one contests, we might believe that TG is not only stronger than NG but strong in an absolute sense, relative to the general population, even though we have only directly observed TG participating in a single match. However, if we later found out that NG did not try very hard in his match against TG, but did try hard in his later matches, our convictions about TG's strength

[*]Address correspondence to: ngoodman@stanford.edu.

might subside.

This reasoning is clearly statistical. We may make good guesses about these propositions, but we will be far from certain; we will be more certain of TG's strength after seeing him play many games. But our reasoning is also highly abstract. It is not limited to a particular set of tug-of-war contestants. We can reason in analogous ways about matches between teams of arbitrary sizes and composition, and are unperturbed if a new player is introduced. We can also reason about the teams as collections: if team Alpha wins its first four matches but then loses to team Bravo, whom it has not faced before, we judge team Bravo very likely to be stronger than average. The smaller team Bravo is, the more likely we are to judge a particular member of team Bravo to be stronger than the average individual. And similar patterns of reasoning apply to inferences about skill and success in other kinds of team contests: We could be talking about math teams, or teams for doubles ping-pong, and make analogous inferences for all the situations above.

Our reasoning also supports inferences from complex combinations of evidence to complex conclusions. For example, suppose that participants have been paired up into teams of two. If we learn that NG was lazy (not trying hard) whenever his team contested a match against TG's team, but NG's team nonetheless won each of these matches, it suggests both that NG is stronger than TG and that NG is often lazy. If we then learned that NG's teammate is stronger than any other individual in the population, we would probably revise the former belief (about NG's strength) but not the latter (about his laziness). If we learned that NG's team had won all of its two-on-two matches but we were told nothing about NG's teammate, it is a good bet that the teammate—whomever he is—is stronger than average; all the more so, if we also learned that NG had lost several one-on-one matches while trying hard.

Finally, our reasoning in this one domain can be modularly combined with knowledge of other domains, or manipulated based on subtle details of domain knowledge. If we observed TG lifting a number of very heavy boxes with apparent ease, we might reasonably expect his tug-of-war team to beat most others. But this would probably not raise our confidence that TG's math team (or even his ping-pong team) are likely to be unusually successful. If we know that NG is trying to ingratiate himself to TG, perhaps to receive a favor, then we might not weight his loss very heavily in estimating strength. Likewise if we knew that NG had received a distracting text message during the match.

We will return to this extended example in more detail in section 3, but for now we take it merely as an illustration, in one simple domain, of the key features of human cognition that we seek to capture in a general computational architecture. How can we account for the wide range of flexible inferences people draw from diverse patterns of evidence such as these? What assumptions about the cognitive system are needed to explain the productivity and gradedness of these inferences? What kind of representations are abstract enough to extend flexibly to novel situations and questions, yet concrete enough to support detailed quantitative predictions about the world? There are two traditional, and traditionally opposed, ways of modeling reasoning in higher-level cognition, each with its well-known strengths and limitations. Symbolic approaches (e.g. Newell, Shaw, & Simon, 1958) can naturally formulate a wide array of inferences but are traditionally confined to the realm of certainty. They would be challenged to

capture all the gradations of reasoning people find so intuitive and valuable in an uncertain world. Probabilistic network approaches—whether based on neural networks (Rumelhart & McClelland, 1988) or Bayesian networks (Pearl, 1988, 2000)—support graded inferences based on sparse or noisy evidence, but only over a fixed finite set of random variables. They lack the representational power and productivity of more structured symbolic approaches, and would be hard-pressed to formulate in a coherent fashion all of the inferences described above—let alone the infinite number of similar inferences we could have listed but did not.

More recently, researchers have begun to move beyond the dichotomy between statistical and symbolic models (Anderson, 1996) and have argued that much of cognition can be understood as probabilistic inference over richly structured representations (Tenenbaum, Kemp, Griffiths, & Goodman, 2011). This has led to a proliferation of structured statistical models, which have given real insight into many cognitive domains: inductive reasoning (Kemp & Tenenbaum, 2009), causal learning (Goodman, Ullman, & Tenenbaum, 2011; Griffiths & Tenenbaum, 2009), word learning (M. C. Frank, Goodman, & Tenenbaum, 2009; Piantadosi, Tenenbaum, & Goodman, 2012), and mental state inference (Baker, Saxe, & Tenenbaum, 2009), to name a few. But the computational tools used for these different models do not yet amount to a truly general-purpose or integrated approach. They are both insufficiently flexible—requiring new models to be described for each different situation—and idiosyncratic—requiring different representational tools across different specific domains.

We require a new computational architecture for cognition, grounded in a new theory of concepts—a theory that does justice to two distinct but equally important roles that concepts play in mental life. On the one hand, concepts enable predictive generalization: they summarize stable regularities, such as typical distributions of objects, properties and events. This is the role primarily addressed by prototype (Rosch, 1999), exemplar (Nosofsky, 1986) and other *statistical* accounts of concepts. On the other hand, concepts provide the basic building blocks of compositional thought: they can be flexibly combined with other concepts to form an infinite array of thoughts in order to reason productively about an infinity of situations. They can be composed to make new concepts, which are building blocks of yet more complex thoughts. Indeed, concepts get much of their meaning and their function from the role that they play in these larger-scale systems of thought. These are the roles primarily addressed (albeit it in different ways) by classical (rule-based) theories of concepts (Bruner, Goodnow, & Austin, 1967), and by the "theory theory" (Gopnik, 2003) and other accounts based on inferential or conceptual roles (Block, 1997). While these theories differ in crucial ways, we group them under the heading of *symbolic* approaches, because they highlight the compositional aspects of concepts that require a powerful symbol-processing architecture.

Our goal in this chapter is to sketch a new account of concepts that combines these two aspects, their statistical and symbolic functions, and to show how this account can explain more of the richness of human reasoning than has been previously captured using traditional approaches. We can phrase our hypothesis, somewhat informally, as:

> **Probabilistic language of thought hypothesis (informal version):** Concepts have a language-like compositionality and encode probabilistic knowl-

edge. These features allow them to be extended productively to new situations and support flexible reasoning and learning by probabilistic inference.

This view of the nature of concepts provides a deeper marriage of statistical inference and symbolic composition. Because they are probabilistic, concepts support graded reasoning under uncertainty. Because they are language-like, they may be flexibly recombined to productively describe new situations. For instance, we have a set of concepts, such as "strength" and "game", for the tug-of-war reasoning domain described above that we may compose with each other and with symbols referring to entities (individuals and teams) in the domain. These combinations then describe distributions on possible world states, which we may reason about via the rules of probability. Our proposal for the PLoT can be seen as making the statistical view of concepts more flexible and systematic by enriching it with a fine-grained notion of composition coming from symbolic approaches. It can also be seen as making symbolic approaches to concepts more useful for reasoning in an uncertain world, by embedding them in a probabilistic framework for inference and decision.

The level of description intended in the PLoT hypothesis is neither the highest level, of input-output relations, nor the lower level of psychological processing. Instead, we aim to use the PLoT to describe conceptual representations and the inferences that they license across situations and domains. The process by which these inferences are implemented is not directly part of the hypothesis, though it can be very useful to consider the possible implementations when evaluating connections between the PLoT and other views of concepts.

2 Formalizing the PLoT

The PLoT hypothesis as stated above is an evocative set of desiderata for a theory of concepts, more than a concrete theory itself. Indeed, it is not *a priori* obvious that it is possible to satisfy all these desiderata at once in a concrete computational system. We are in need of a compositional formal system—a language—for expressing probability distributions over complex world states. Our first clue comes from the idea of representing distributions as *generative processes*: the series of random steps by which the world comes to be as it is. But while generative processes are a useful way to represent probabilistic knowledge, adopting such a representation only transforms our problem into one of finding a compositional language for generative processes. The solution to this problem comes from a simple idea: if you have described a deterministic process compositionally in terms of the computation steps taken from start to end, but then inject noise at some point along the way, you get a stochastic process; this stochastic process unfolds in the original steps, except where a random choice is made. In this way a distribution over outputs is determined, not a single deterministic output, and this distribution inherits all the compositionality of the original deterministic process. The stochastic λ-calculus realizes this idea formally, by extending a universal computational system (λ-calculus) with points of primitive randomness. The *probabilistic programming language* Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008) extends the sparse mathematical system of stochastic λ-calculus into a more usable system for describing statistical processes.

We now give a brief introduction to the syntax and ideas of probabilistic modeling using Church, sufficient to motivate a more formal version of the PLoT; further details and many examples can be found at `http://probmods.org`. Church uses a syntax inherited from the LISP family of languages (McCarthy, 1960). Thus operators precede their arguments, and are written inside grouping parentheses: e.g., `(+ 1 2)` encodes the operation "add 1 and 2". We use `define` to assign values to symbols in our program and `lambda` for creating functions. We could, for example, create a function `double` that takes one number as an input and returns its double. The code would look like this: `(define double (lambda (x)(+ x x)))`. Applying this function to `2` would look like `(double 2)` and result in the value `4`.

What differentiates Church from an ordinary programming language is the inclusion of random primitives. For example, the function `(flip 0.5)` can be interpreted as a simple coin flip with a weight (i.e. a Bernoulli random draw) outputting either `true` or `false`. Every time the function is called, a new random value is generated—the coin is flipped. These random primitives can be combined just as ordinary functions are—for instance `(and (flip 0.5)(flip 0.5))` is the more complex process of taking the conjunction of two random Booleans. A Church program specifies not a single computation, but a distribution over computations. This *sampling semantics* (see Goodman, Mansinghka, et al. (2008) for more details) means that composition of probabilities is achieved by ordinary composition of functions, and it means that we may specify probabilistic models using all the tools of representational abstraction in a modern programming language. We will not provide a primer on the power of function abstraction and other such tools here, but we will use them in what we hope are intuitive and illustrative ways in the examples below.

A number of language features in Church parallel core aspects of conceptual representation. Perhaps the most familiar (and most controversial) for cognitive modeling is the use of arbitrary *symbols*. In Church (as in LISP) a symbol is a basic value that has only the property that it is equal to itself and not to any other symbol: `(equal? 'bob 'bob)` is true, while `(equal? 'bob 'jim)` is false. (The single quote syntax simply indicates that what follows is a symbol). Critically, symbols can be used as unique identifiers on which to hang some aspect of conceptual knowledge. For instance they can be used to refer to functions, as when we used `define` to create the `double` function above and then reused this doubling function by name. Symbols can also be used together with functions to represent knowledge about (an unbounded set of) objects. For instance, the function

```
(define eyecolor (lambda (x) (if (flip) 'blue 'brown)))
```

takes a person `x` and randomly returns[1] an eye color (e.g. `(eyecolor 'bob)` might return `'blue`). That is, it wraps up the knowledge about how eye color is generated *independently* of which person is asked about—a person is simply a symbol (`'bob`) that is associated with another symbol (`'blue`) by the `eyecolor` function.

Of course the above representation of an object's property has a flaw: if we ask about the eye color of Bob twice we may get different answers! Church includes an

[1] In Church the conditional has a traditional but cryptic syntax: `(if a b c)` returns `b` if `a` is true, and `c` otherwise. Thus `(if (flip)b c)` randomly returns `b` or `c`.

operator `mem` that takes a function and returns a *memoized* version: one that makes its random choices only once for each distinct value of the function's arguments, and thereafter when called returns the answer stored from that first evaluation. For instance, a memoized version of the `eyecolor` function,

```
(define eyecolor (mem (lambda (x) (if (flip) 'blue 'brown))))
```

could output either `'blue` or `'brown` for Bob's eye color, but only one of these possibilities, to be determined the first time the function is called. This ensures that `(equal? (eyecolor 'bob) (eyecolor 'bob))` is always true.

Thus symbols can be used as "indices" to recover random properties or as labels which allow us to recover stored information. These uses are conceptually very similar, though they have different syntax, and they can be combined. For instance, we can access one function in another by its name, passing along the current objects of interest:

```
(define eyecolor
  (mem (lambda (x)
    (if (flip 0.1)
      (if (flip) 'blue 'brown)
      (if (flip) (eyecolor (father x)) (eyecolor (mother x)))))))
```

This (false, but perhaps intuitive) model of eye color asserts that the color is sometimes simply random, but most of the time depends on the eye color of one of a person's parents—which is accessed by calling the `father` or `mother` function from inside the `eyecolor` function, and so on. Symbols and symbolic reference are thus key language constructs for forming complex concepts and situations from simple ones.

How does reasoning enter into this system? The fundamental operation of belief updating in probabilistic modeling is *conditioning*. We can define conditioning within Church via the notion of rejection sampling: if we have a distribution represented by `dist` (a stochastic function with no input arguments) and a predicate `condition` (that takes a value and returns true or false) then we can define the distribution conditioned on the predicate being true via the process:

```
(define conditional
  (lambda ()
    (define sample (dist))
    (if (condition sample) sample (conditional))))
```

That is, we keep sampling from `dist` until we get a sample that satisfies `condition`, then we return this sample.[2] It can be cumbersome to split our knowledge and assumption in this way, so Church introduces a syntax for conditionals in the form of the `query` function:

```
(query
  ...definitions...
  query-expression
  condition-expression)
```

Our initial distribution is the `query-expression` evaluated in the context of the `...definitions...`, and our predicate is the `condition-expression` evaluated in the same context.

[2]For readers familiar with Bayesian belief updating in probabilistic models, this process can be seen as taking a prior model specified by `dist` and generating a sample from the posterior corresponding to `dist` conditioned on the evidence that `condition` is true.

For instance, referring again to the eye-color example, we could ask about Bob's mother's likely eye color, given that Bob has blue eyes:

```
(query
  (define eyecolor ...as above...)
  (eyecolor (mother 'bob))
  (equal? 'blue (eyecolor 'bob)))
```

Notice that there is a distinction between the definitions, which represent probabilistic knowledge reusable across many queries, and the query and condition expressions, which represent the particular question of interest at the moment. In this example, the particular people need to be introduced only in the question of interest because the conceptual knowledge is defined over arbitrary symbols.

Equipped now with a compositional formal language for representing distributions and performing conditional inference, we can revisit the probabilistic language of thought hypothesis. Within a probabilistic language like Church, knowledge is encoded in stochastic function definitions. These functions describe elements of stochastic processes that can be composed together to describe various situations, to pose various questions and to answer those questions with reasonable probabilistic guesses. Indeed, just as concepts are the stable and reusable components of human thinking, stochastic functions are the units of knowledge encoded in a church program. Motivated from this observation we can formulate a stronger PLoT hypothesis:

> **Probabilistic language of thought hypothesis (formal version):** Concepts are stochastic functions. Hence they represent uncertainty, compose naturally, and support probabilistic inference.

Notice that this formal version realizes the informal PLoT in a precise way, showing that the hypothesis is coherent and allowing us to ask more detailed questions about plausibility. For instance we can begin to ask the usual philosophical questions about concepts of this system: What constitutes meaning? How are concepts related? How are they acquired and used? The answers to these questions can be subtle, but they are determined in principle from the basic claim that concepts are stochastic functions. For instance, on the face of it, the meaning of a stochastic function is simply its definition and the relation between concepts is determined by constituency—in the example above, the meaning of `eyecolor` is its definition and it is related to other concepts only by its use of `mother` and `father` functions. However, when we consider the inferential relationships between concepts that come from conditional inference—`query`—we see additional aspects of meaning and conceptual relation. Conditioning on parentage can influence eye color, but also vice versa; conditioning on hair color may influence judgements about eye color indirectly, and so on. In the next section we give an extended example in the domain of simple team games, illustrating these foundational issues as well as exploring the empirical adequacy of the PLoT hypothesis.

3 Example: Ping pong in Church

Consider the information shown in Figure 1. Most people conclude that TG is relatively strong, while BL is average-to-weak. Below we describe various patterns of evidence that we displayed to people in the guise of a ping pong tournament. How can

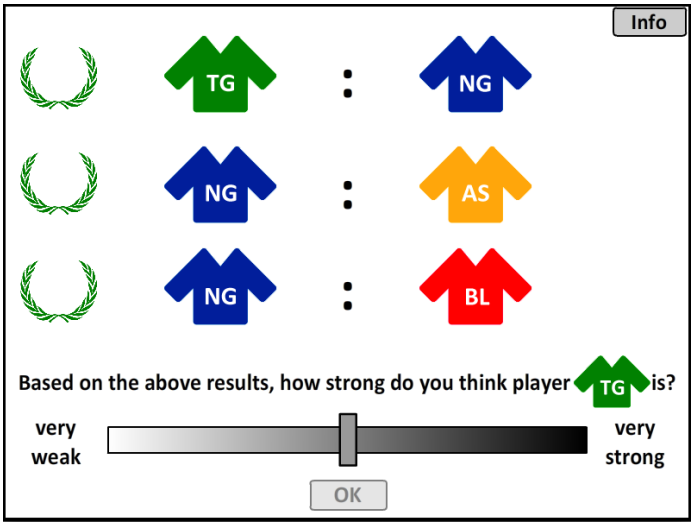


Figure 1: Screenshot of a single player tournament. The winner of each match is indicated by a laurel wreath.

we account for people's sensitivity to the uncertain nature of the evidence in such situations? While capturing the abstract, symbolic structure that remains invariant between this particular situation and other similar situations that could involve different players, teams, and outcomes?

This simple sports domain is built around people, teams and games. In Church, we can use symbols as placeholders for unspecified individuals of these types. This means that we do not need to define in advance how many people participate, what the size of the teams will be, or how many games a tournament will have. Instead, we begin by describing a lexicon of abstract concept definitions useful for reasoning about these games. We define an individual player's strength, `personstrength`, via a function that draws from a Gaussian distribution (with arbitrary $M = 10$ and $SD = 3$):

```
(define personstrength (mem (lambda (person) (gaussian 10 3))))
```

Memoization ensures that the strength value assigned to a person is persistent and does not change between games. However, we assume that players are sometimes lazy on a given match. The chance of a person being lazy in a particular game is 10%:

```
(define lazy (mem (lambda (person game) (flip 0.1))))
```

The overall strength of a team, on a given game, is the sum of the strength of each person in the team. If a person in the team is lazy, however, he only plays with half of his actual strength.

```
(define teamstrength
  (mem (lambda (team game)
         (sum (map (lambda (person)
                     (if (lazy person game)
                         (/ (personstrength person) 2)
                         (personstrength person)))
                   team)))))
```

Table 1: Patterns of observation for the single player tournaments. *Note*: An additional set of 4 patterns was included for which the outcomes of the games were reversed. The bottom row shows the omniscient commentator's information in Experiment 2. For example, in the confounded case, player B was lazy in the second game.

confounded evidence (1,2)	strong indirect evidence (3,4)	weak indirect evidence (5,6)	diverse evidence (7,8)
A > B A > B A > B	A > B B > C B > D	A > B B < C B < D	A > B A > C A > D
lazy,game: B,2	B,1	B,1	C,2

Note: A > B means that A won against B.

Finally, we specify how the winner of a game is determined. We simply say the the team wins who has the greater overall strength:

```
(define winner
  (mem (lambda (team1 team2 game)
         (if (< (teamstrength team1 game) (teamstrength team2 game))
             'team1 'team2))))
```

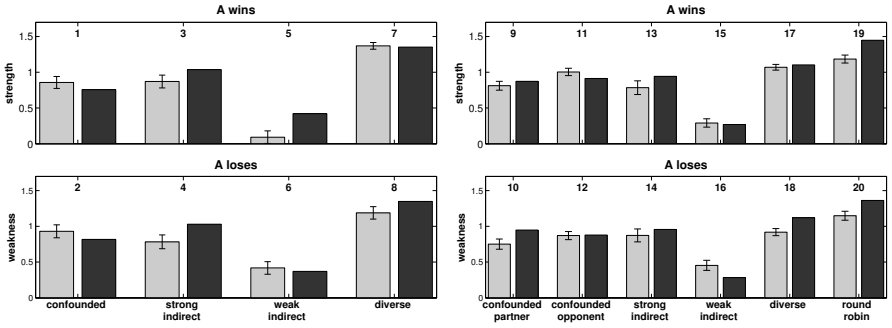
This set of function definitions specifies a simple lexicon of concepts for reasoning about the ping pong domain.

The way in which we can define new concepts (e.g. `teamstrength`) based on previously defined concepts (`personstrength` and `lazy`) illustrates one form of compositionality in Church. The set of concept definitions refers to people (teams, etc.) without having to declare a set of possible people in advance: instead we apply generic functions to placeholder symbols that will stand for these people. That is, the concepts may be further composed with symbols and each other to describe specific situations. For instance, the inference in Figure 1 can be described by:

```
(query
  ...CONCEPTS...
  ;The query:
  (personstrength 'TG)
  ;The evidence:
  (and
    (= 'team1 (winner '(TG) '(NG) 1))
    (= 'team1 (winner '(NG) '(AS) 2))
    (= 'team1 (winner '(NG) '(BL) 3))))
```

Here `...CONCEPTS...` is shorthand for the definitions introduced above—a lexicon of concepts that we may use to model people's inferences about a player's strength not only in the situation depicted in Figure 1 but in a multitude of possible situations with varying teams composed of several people, playing against each other with all thinkable combinations of game results in different tournament formats. This productive extension over different possible situations including different persons, different teams and different winners of each game, renders the Church implementation a powerful model for human reasoning.

We wanted to explore how well our simple Church model predicts the inferences people make, based on complex patterns of evidence in different situations (cf. Ger-



(a) Single player tournament (b) Two-player tournament

Figure 2: Mean strength estimates (grey bars) and model predictions (black bars) for the single player (left) and two-player tournaments (right). The top row shows strength judgments for cases in which the player won her game(s). The bottom row shows weakness judgments for cases in which the player lost. Numbers above the bars correspond to the patterns described in Tables 1 and 2. Error bars are $\pm$ 1 *SEM*.

stenberg & Goodman, 2012). In Experiment 1, participants' task was to estimate an individual player's strength based on the outcomes of different games in a ping pong tournament. Participants were told that they will make judgments after having seen single player and two-player tournaments. The different players in a tournament could be identified by the color of their jersey as well as their initials. In each tournament, there was a new set of players. Participants were given some basic information about the strength of the players which described some of the modeling assumptions we made. That is, participants were told that individual players have a fixed strength which does not vary between games and that all of the players have a 10% chance of not playing as strongly as they can in each game. This means that even if a player is strong, he can sometimes lose against a weaker player.[3]

Table 1 shows the patterns of evidence that were used for the single player tournaments. Table 2 shows the patterns for the two-player tournaments. In all tournaments, participants were asked to judge the strength of player A. For the single player tournaments, we used four different patterns of evidence: *confounded evidence* in which A wins repeatedly against B, *strong* and *weak indirect evidence* where A only wins one match herself but B either continues to win or lose two games against other players and *diverse evidence* in which A wins against three different players. For each of those patterns, we also included a pattern in which the outcomes of the games were exactly reversed. For the two player tournaments, we used six different patterns of evidence: In some situations A was always in the same team as B (*confounded with partner*) while in other situations A repeatedly played against the same player E (*confounded with opponent*). As in the single player tournaments, we also had patterns with mostly indirect evidence about the strength of A by having his partner in the first game, B, either win or lose against the same opponents with different teammates (*weak/strong indirect*

[3]Demos of the experiments can be accessed here:
`http://web.mit.edu/tger/www/demos/BPP_demos.html`

evidence). Finally, we had one pattern of *diverse evidence* in which A wins with different teammates against a new set of opponents in each game and one *round robin* tournament in which A wins all his games in all possible combinations of a 4-player tournament. Further experimental details can be found in Appendix A.

In order to directly compare the model predictions with participants' judgments we z-scored the model predictions and each individual participant's judgments. Furthermore, we reverse coded participants' judgments and the model predictions for the situations in which the outcomes of the games were reversed so that both strength and "weakness" judgments go in the same direction.

Figure 2 shows the mean strength estimates (gray bars) together with the model predictions (black bars) for the single and two-player tournaments. The top panels display the situations in which A won his game(s). The bottom panels show the situations in which A lost. Our model predicts participants' judgments in the single and two-player tournaments very well with $r = .98$ and $RMSE = .19$. A very high median correlation with individual participants' judgments of $r = .92$ shows that the close fit is not merely due to an aggregation effect. These results show that our model predicts participants' inferences very accurately—a single, concise representation of the task is sufficient to predicts people's inferences for a great diversity of patterns of evidence.

A still greater variety of evidence is available by composing the basic concepts together in different ways: there is no reason for evidence not to directly refer to a player's strength, laziness, etc. For instance:

```
(query
  ...CONCEPTS...
  ;The query:
  (personstrength 'TG)
  ;The evidence:
  (and
    (= 'team1 (winner '(TG) '(NG) 1))
    (= 'team1 (winner '(NG) '(AS) 2))
    (= 'team1 (winner '(NG) '(BL) 3))
    (lazy 'NG 1))) ;additional kinds of evidence (Expt. 2).
```

While in Experiment 1, the match results were the only source of information participants could use as a basis for their strength judgments, Experiment 2 introduced an omniscient commentator who gave direct information about specific players. After participants saw a tournament's match results, an omniscient commentator, who always told the truth, revealed that one player was lazy in a particular game. We were interested in how participants updated their beliefs about the strength of player A given this additional piece of evidence. Importantly, we do not need to change anything in the concept definitions to derive predictions for these situations, since only the way they are composed into evidence changes.

Figure 3 shows the mean strength judgments (gray bars) together with the model predictions (black bars, see Table 1 for the different patterns of evidence). The dark gray bars indicate participants' first judgments based on the tournament information only. The light gray bars indicate participant's second judgments after they received the commentator's information. The model predicts participants' ratings very accurately again with $r = .97$ and $RMSE = 0.29$. The model's median correlation with individual participants' judgments is $r = .86$. These results show that participants, as well as our model, have no difficulty in integrating different sources of evidence to form an overall

Table 2: Patterns of observation for the two-player tournaments. *Note*: An additional set of 6 patterns was included in which the outcomes of the games were reversed.

confounded with partner (9,10)			confounded with opponent (11,12)			strong indirect evidence (13,14)		
AB	$>$	CD	AB	$>$	EF	AB	$>$	EF
AB	$>$	EF	AC	$>$	EG	BC	$<$	EF
AB	$>$	GH	AD	$>$	EH	BD	$<$	EF
weak indirect evidence (15,16)			**diverse evidence (17,18)**			**round robin (19,20)**		
AB	$>$	EF	AB	$>$	EF	AB	$>$	CD
BC	$>$	EF	AC	$>$	GH	AC	$>$	BD
BD	$>$	EF	AD	$>$	IJ	AD	$>$	BC

judgment of a player's likely underlying strength. The model predicts participants' judgments very accurately by being sensitive to the degree to which the initial strength estimate should be updated in the light of new evidence provided by the commentator.

4 Intuitive theories

The examples above provide concrete illustrations of how to represent concepts as functions in a probabilistic language of thought, how a system of such concepts supports inferences that are both productive and probabilistic, and how these inferences can capture the outputs of human reasoning at a high level of quantitative accuracy. But while reasoning about team games like ping pong is very illustrative, it is of relatively limited scope compared to many of the concepts involved in everyday human reasoning. In this section we discuss how the same machinery can describe abstract concepts that are the backbone of thinking about everyday life, and which have often not fit easily into more traditional formal frameworks.

Intuitive theories (Carey, 2009; Gopnik & Wellman, 2012; Wellman & Gelman, 1992), like their more familiar scientific counterparts, are comprised of a system of interrelated and inter-defined concepts articulating a basic ontology of entities, the properties of and relations between those entities, and the causal laws that govern how these entities evolve over time and interact with each other. For instance intuitive physics is a system for reasoning about physical objects, and intuitive psychology for reasoning about intentional agents. These are called "theories" because, like in scientific theories, the essential constructs of intuitive theories are typically not directly observable. Yet intuitive theories also specify how unobservable states, properties and processes do impact observable experience—and thus how they support competencies such as prediction, explanation, learning and reasoning.

Intuitive theories can be found in some form in young infants, and are also to some extent shared with many other species; they are arguably the earliest and oldest abstract concepts we have (Carey, 2009). They provide the scaffolding for many of children's conceptual achievements over the first few years of life. They also provide core building blocks for meaning in natural language, at the same time as they are enriched

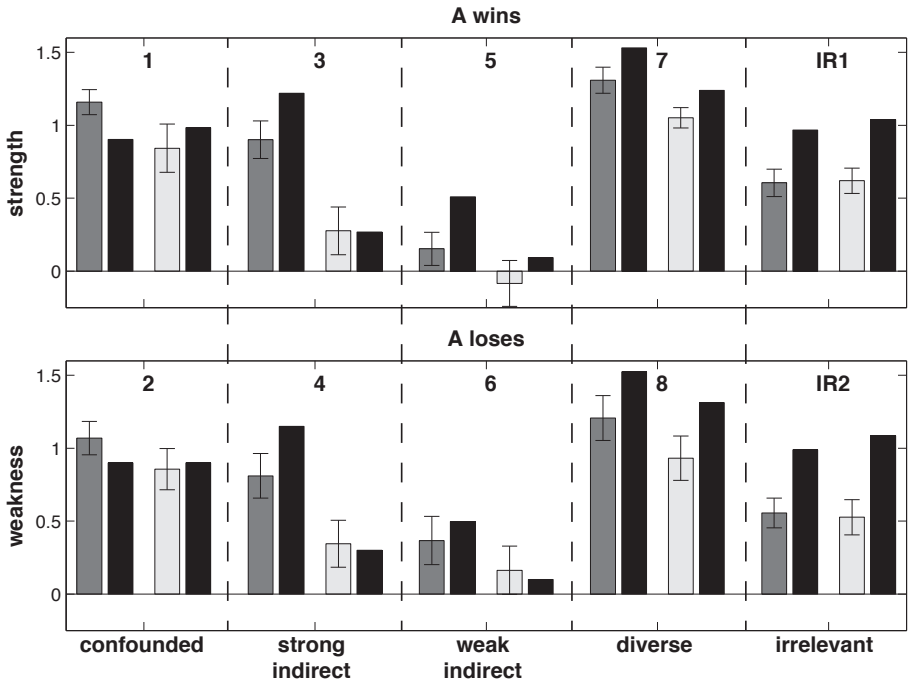


Figure 3: Mean strength estimates and model predictions. Dark grey bars = estimates after tournament information only, light grey bars = estimates after omniscient commentator info, black bars = model predictions. Error bars are $\pm$ 1 *SEM*. *Note*: Situations IR1 and IR2 were cases in which the information given by the commentator was irrelevant; see Appendix B for details.

and transformed fundamentally as children develop their natural language abilities. Through adulthood, they continue to serve as the basis for our common-sense understanding of the world. Yet while these intuitive theories have long been a prime target for exploration by developmental and cognitive psychologists, linguists, and philosophers, they have not received much treatment in the literature on formal models of concepts. This may be because they do not fit well into the general-purpose mathematical and computational modeling frameworks that have been available and useful for more mundane category concepts—prototypes, exemplars, and logical accounts.

Starting around ten years ago, several authors began to consider Bayesian networks as formal models for intuitive theories (Goodman et al., 2006; Gopnik et al., 2004; Rehder, 2003; Tenenbaum & Griffiths, 2003), focusing on their causal aspects. These efforts were ultimately limited by the fact that Bayesian networks, like neural networks before them, fail to capture genuine productivity in thought. An intuitive theory of physics or psychology must be able to handle an infinite range of novel situations, differing in their specifics but not their abstract character, just as we illustrated above on a much smaller scale for an intuitive theory of tug-of-war or ping pong. Hierarchical Bayesian models have been proposed as one way to increase the representational power of Bayesian networks, and they have given reasonable accounts of some aspects

of abstract causal theories (e.g. Tenenbaum et al., 2011). But hierarchical Bayesian models on their own still lack sufficient representational power to address the fine-grained compositionality inherent in our intuitive theories The PLoT allows us to take a major step forward in this regard. Both Bayesian networks and hierarchical Bayesian models of intuitive theories can be naturally written as Church programs, preserving their insights into causal reasoning and learning, but Church programs go much further in letting us capture the essential representations of common-sense physics and psychology that have defied previous attempts at formalization within the probabilistic modeling tradition.

To illustrate, consider how we might capture the core concepts of an intuitive psychology—a probabilistic model of how agents act rationally in response to their mental states, aiming to satisfy their desires as efficiently as possible given their beliefs. As as initial step, imagine extending the game model above to take into account the fact that laziness for a particular player in a given game may not simply be a random event, but an intentional choice on the part of a player—he may estimate that the other team is so weak that it is not worth his effort to try hard. We pose this as a church model by imagining that a player asks himself "how should I act, such that my team will win?"; this translates into a `query`:

```
(define lazy (mem (lambda (person game)
  (query
    (define action (flip L))
    action
    (= (teamof person) (winner (team1of game) (team2of game) game))))))
```

where we have helped ourselves to some innocuous helper functions to look up the team of a player and so on. The parameter `L` controls the *a priori* tendency to be lazy; this gives a simple way of including a principle of efficiency: a tendency to avoid undue effort. The condition statement of the query specifies the player's goal—for his team to win the game—hypothetically assuming that this goal will be achieved. The output of the query is an action (trying hard, or not) that is a reasonable guess on the player's part for how that goal may be achieved. An inference about which team will win a match now leads to a sub-inference modeling each player's choice of whether to exert their full effort, given the players on each team. We could further extend this model to take into account private evidence that each player might have about the strengths of the other players, expressing his or her process of belief-formation about the total strengths of the two teams as an additional set of nested sub-inferences.

The pattern of using an embedded `query` to capture the choices of another agent is a very general pattern for modeling intuitive psychology (Stuhlmüller & Goodman, 2013). We could write down the abstract structure schematically as:

```
(define choice (lambda (belief state goal?)
  (query
    (define action (action-prior))
    action
    (goal? (belief state action)))))
```

where `belief` is taken to be the agent's summary of the world dynamics (transitions from states to states, given actions), and `goal?` is a goal predicate on states picking out those that the agent desires. Of course many additional refinements and additions may be needed to build an adequate model of human intuitive psychology—agents form

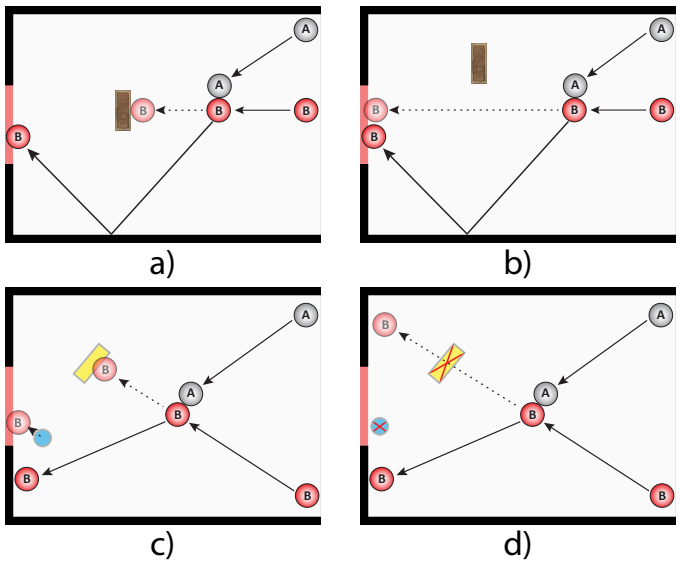


Figure 4: Diagrammatic illustrations of four collision events in a simple physics world. *Note*: Solid arrows represent the actual trajectories of ball A before the collision and of ball B before and after the collision. Dashed arrows and faded circles represent the counterfactual trajectory of ball B. The brown rectangle, yellow rectangle and blue circle represent a brick, and the entry and exit of a teleport, respectively.

beliefs, experience emotions, and so on. Yet we hope that the ease of writing down even a simple version of a theory of mind that captures both the theory's abstractness and its potential for productive probabilistic inferences illustrates the power of the PLoT view.

5 Mental simulation and sampling

There is a central link between the sampling semantics of Church programs, mental simulation, and the causality central to many intuitive theories. A Church program naturally expresses the causal, generative aspect of people's knowledge through the function dependencies in the program. The function dependencies dictate the causal flow of the sampling process: functions whose outputs serve as an input to another function must be evaluated first. Each run of a Church program can be interpreted as the dynamic generation of a possible world that is consistent with the causal laws as specified in the program (Chater & Oaksford, 2013). Because the sampling process is stochastic, a Church program specifies a probability distribution over possible worlds, and different modes of reasoning can be seen as different forms of mental simulation on top of this basic sampling process. While the notion of mental representation and simulation of possible worlds has had many advocates (Craik, 1967; Hegarty, 2004; Johnson-Laird, 1983, 2006), the PLoT view integrates this idea naturally into a view of mental models that is also probabilistic, causal and sufficiently expressive to capture core intuitive theories.

We will illustrate the implications of this view via a concrete example from the

domain of intuitive physics: people's judgments about the causal relations among the trajectories of physical objects in motion. We are concerned with judgments of "actual causation": whether one specific event caused another specific event to happen. Consider a relatively simple scenario which consists of two billiard balls *A* and *B*, some solid walls with an opening gate, a brick, and a teleport gate that can be either active or inactive. Figure 4 shows diagrammatic illustrations of causal interactions between *A* and *B* in this world, assuming simple Newtonian elastic collisions between moving bodies. The question of interest is whether ball *A*'s collision with ball *B* caused ball *B* to go through the red gate on the left of the screen, prevented it from going through, or did neither. The tools of probabilistic programs and sampling-based inference allow us to give a precise formal account of these causal judgments, which also accords well with intuitions of mental simulation and gives strong quantitative fits to behavioral experiments.

To explain these judgments, we first need to be able to represent the relevant physical knowledge at the right level of abstraction. Despite its simplicity, our domain already affords an infinite number of interactions between *A* and *B* and we want a model that yields a causal judgment for each possible situation. Rather than having to specify a new model for each causal interaction of interest (as we would have to do if we adopted a Bayesian network formulation (Pearl, 2000)), we want to represent the general laws that govern the interactions between the objects in our world. One way of representing people's knowledge of physical object motion in Church is by writing down a probabilistic and approximate version of some aspects of Newtonian mechanics. Functions in the Church program compute the inertial time-evolution and the outcome of collisions by taking as input the mass and velocity of objects as well as more general aspects of the world such as friction and gravity. So far these are standard, deterministic simulation routines (so we leave out details). Critically, we also assume that some noise in each object's momentum is inserted just after each collision, and perhaps at other times as well, resulting in trajectories that are noisy versions of their Newtonian counterparts. Recent research has shown that people's intuitive physical judgments in several domains are well described by such noisy Newtonian simulations (Battaglia, Hamrick, & Tenenbaum, 2013; Sanborn, Mansinghka, & Griffiths, 2013; K. Smith, Dechter, Tenenbaum, & Vul, 2013; K. A. Smith & Vul, 2012). Once we have a Church program that captures people's intuitive physics, we can model predictions about the future (e.g., will ball *B* go through the gate?) as simple forward simulations, and inferences about the past (e.g., where did ball *B* likely come from?) by a `query` of the past given the present—simulating possible histories that could have led up to the current state.

More subtly, a Church program can also be used to evaluate counterfactuals (e.g., would ball *B* have gone through the gate if the collision with *A* hadn't happened?). In line with Pearl (2000), the evaluation of counterfactuals in a Church program involves three steps: First, we condition all the random choices in the program based on what actually happened to estimate the unobserved values of the actual world. Second, we realize the truth of the counterfactual antecedent (e.g. that the collision did *not* happen) by intervening in the program execution that generated the actual world. This intervention breaks the normal flow of the program by setting some function inputs to desired values. For example, to model what would have happened if there had been no

collision between *A* and *B*, we could set ball *A*'s velocity to zero or move ball *A* outside of the scene shortly before the time of collision. Finally, to evaluate the truth of the counterfactual, we reevaluate all the functions downstream from the point at which we intervened in the program. This process generates a sample over counterfactual world states, and repeatedly running this process allowing for different stochastic functions evaluations can be used to express people's uncertainty over what would have happened in the relevant counterfactual world. Notice that the key feature of Church that allows this process to work is that it specifies a process for sampling particular situations and makes explicit the steps of the causal history that lead up to a situation (in the form of a program execution trace). Counterfactuals are then evaluated by a series of "simulation" steps that result in imagined counterfactual worlds.

In a series of empirical studies, we have shown that people's quantitative judgments of actual causation are closely linked to such a probabilistic counterfactual analysis (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012). When judging whether one event caused another event to happen, people compare what actually happened with what they think would have happened in the counterfactual world in which the candidate causal event had been absent. They appear to estimate something like the probability that the candidate cause was *necessary* to produce the outcome event: the probability that the outcome, which did in fact occur, would not have occurred in a counterfactual world where the candidate cause was absent. Consider the top pair of diagrams shown in Figure 4. In both clips, what actually happened is identical. However, the outcome in the relevant counterfactual worlds would likely have been different. In Figure 4a, ball *B* would have bounced off the brick if it hadn't collided with ball *A*. In contrast, in Figure 4b, ball *B* would have gone through the gate even without the collision with ball *A*. As predicted by our account, people's judgments about whether *A* caused *B* to go through the gate are significantly higher for Figure 4a compared to Figure 4b (Gerstenberg, Goodman, Lagnado, & Tenenbaum, submitted). In the bottom pair of cases, the contrast in the relevant counterfactual worlds was realized by comparing situations in which the teleport was either on (Figure 4c) or off (Figure 4d). While people judged that *A* prevented *B* from going through the gate in Figure 4c, they didn't think that *A* prevented *B* in Figure 4d. In Figure 4c, ball *B* would have gone through the gate via the teleport if it hadn't collided with *A*. In Figure 4d, in contrast, the teleport was off and *B* would not have gone through the gate even if there had been no collision with *A*. The fact that people's judgments differ dramatically between situations in which what actually happened was held constant supports the idea that causal and counterfactual judgments are inextricably linked. It also shows that it is not possible to give an account of people's causal judgments just in terms of what actually happened (e.g. Wolff, 2007). Finally, it demonstrates the flexibility of people's intuitive theories and the critical way in which this flexibility supports mental simulation, counterfactual and causal reasoning. Once people have learned how the teleport works, they have no trouble imagining its effects and incorporating these into their counterfactual simulations, whose outcome in turn influences their causal judgments.

While we have focused here on the example of how we can explain people's causal judgments in the domain of intuitive physics, the general framework applies equally well to any domain for which we are able to represent our knowledge in terms of a Church program. For example, we could model people's judgments about whether

agent *A*'s argument convinced agent *B* to try harder as a function of what actually happened and people's subjective degree of belief that *B* would still have tried harder had *A* not said anything. A Church program that captures people's intuitive understanding of psychology looks different from a Church program that captures people's intuitive understanding of physics, but we can understand people's causal judgments in terms of the same process that compares the actual outcome with the outcomes of mental simulations (sampling) of the relevant counterfactual worlds.

6 Concepts and natural language

Thus far we have sketched a notion of concepts as stochastic functions in Church and intuitive theories as systems of interrelated concepts. We have also described how such intuitive theories can be used to describe the complex causal knowledge that people use to reason about the world. We would additionally like a theory of concepts to help us formulate the meanings of words, and more generally the ways that natural language can be used to convey thought. In Goodman and Lassiter (to appear) we use Church to formulate an architecture for language understanding in which word meanings are grounded into intuitive theories, and utterance interpretation is a rich inferential process starting with these meanings. We summarize here the parts of this architecture that illuminate the role of concepts in thought.

As described above, we view intuitive theories as collections of function definitions; together they form a distribution over all the expressions that can be composed from these functions—this constitutes prior knowledge of the domain. We posit that the foundational operation of natural language interpretation is to update this prior belief distribution into a posterior distribution. Because belief update in a probabilistic system happens by conditioning, we need utterances to lead to conditions that can be used in a `query`. That is, there must be a `meaning` function that maps from strings of words to Boolean-valued expressions in the PLoT (i.e. expressions which can be the condition of a query). This meaning function is essentially the "narrow" language facility (Hauser, Chomsky, & Fitch, 2002), mapping from sound strings to the PLoT. In Goodman and Lassiter (to appear) we describe the meaning function in two (fairly standard) steps: First we look up the PLoT expression for each word in a linguistic lexicon (an association between words an PLoT expressions), then we compose these expressions recursively until we have built a meaning for the whole sentence. These steps may be non-deterministic, but additional uncertainty causes no difficulties since we are already within a probabilistic framework. Thus, the meaning function lets us construct from a natural language expression a condition-expression that can be used in query to update beliefs about some question—the query-expression. But notice that while they are constructed in the same PLoT these two expressions play very different cognitive roles: the query-expression is a question about the world; the condition-expression is a constraint on the causal process that generates the answer to this question.

The architecture is reminiscent of Jackendoff's "languages of thought" (Jackendoff, 1995), in which there are several modules (for example, natural language and cognition) each with their own representation language, and interfaces for translating from one module to another. In our approach the representation languages are mathematically similar for natural language and cognition (based in the stochastic lambda

calculus) and their "interface" is defined by their roles in the inference (`query`) of language interpretation. Despite their mathematical similarities, the different cognitive roles of these different kinds of expressions imply two different, but interlocking, principles of compositionality in the cognitive architecture. One instance of compositionality allows us to build rich (distributions on) generative histories, while the other allows us to build up complex conditioning statements to constrain these histories. A naive approach would try to make these two kinds of composition directly compatible, by requiring that each natural language constituent describe a probability distribution and relying on linguistic composition to combine these distributions. Our approach allows these distinct modes of composition to apply separately in natural language and thought, resulting in complex interactions that can look non-compositional when only one type of representation is considered.

Meaning conveyed by natural language is further enriched by pragmatic inference. As described elsewhere (M. Frank & Goodman, 2012; Goodman & Lassiter, to appear; Goodman & Stuhlmüller, 2013; Stuhlmüller & Goodman, 2013) a broad range of pragmatic inferences can be understood as the output of conditioning in an intuitive theory of communication, and can also be formalized using the PLoT tools described above. Adding this layer of inference results in futher complexities and context-sensitivities to the effective relationship between words and concepts, but is essential for understanding how we talk about our thoughts and understand what other people mean to say.

7 Concept acquisition

If concepts are definitions in a library of useful (stochastic) functions, what is concept learning? Forming new concepts from examples is fundamentally a problem of induction—in our case the problem of program induction. This can be formulated as Bayesian inference of a set of concepts that best explain the experience we have in the world: conditioned on generating the examples we have seen, what is the likely new concept? Hypothesized concepts are formed in an effective language of thought based on the concepts learned so far—all the expressions that can be formed by composing the underlying PLoT and the already-defined function symbols. We can view these hypotheses as being generated by a higher-order "program-generating program," a stochastic function that generates candidate stochastic functions that might explain a given set of observed examples. Concept learning as probabilistic program induction is philosophically and mathematically well-posed, but a great deal of research is needed both to reduce it to useful engineering practice and to validate it as a model of human concept learning. Induction over such an infinite combinatorial space is simply stated as probabilistic conditioning, but such inferences are extremely challenging to implement in general. Yet recent progress has shown that this approach can be successful in certain cases: grammar-based program induction has been used to describe category learning (Goodman, Tenenbaum, Feldman, & Griffiths, 2008), learning of relational concepts (Kemp, Goodman, & Tenenbaum, 2008), learning simple visual concepts (Lake, Salakhutdinov, & Tenenbaum, 2013), and learning of number concepts (Piantadosi et al., 2012).

Notice that in this notion of inductive program elaboration, each concept begins life

as simply a new symbol to which a function will come to be attached. The impetus to add such a placeholder symbol may come from natural language (upon hearing a new word), from the interaction of knowledge about natural kinds and specific examples (as suggested by Margolis (1998) and Carey (this volume)), or from other explanatory pressures. Richer content and relationships to other concepts would be incorporated into the web of function definitions inductively, as the learner encounters additional examples and gains experience in the new domain.

Language, social context, and other factors may play a critical role in positing both the existence and the content of new concepts. For instance, Shafto, Goodman, and Frank (2012) argue that the social context of examples (for instance, that they are generated communicatively by a helpful person) can strongly impact the inferences made. Similarly, language can provide a strongly constraining form of evidence for concept formation—for instance (Piantadosi et al., 2012) use a linguistic count-list as a key input in learning numerical concepts. The PLoT offers a powerful way to think about these and other bootstrapping phenomena at the interface of social interaction, language acquisition and concept acquisition, such as the contributions of syntactic and semantic bootstrapping in learning verbs, or the contributions of pragmatic inference in learning quantifiers. Again, further research is needed to understand how these factors integrate with inductive learning from examples in a full theory of concept acquisition; what is important for us here is that the PLoT provides us with a theory of concepts, and an approach to concept learning as probabilistic inference, that is able to explore these interactions in a productive way.

One important implication of the inductive view just indicated is that concept learning changes the effective language of thought. While this effective language has the same mathematical expressivity as the underlying PLoT, particular thoughts may be vastly simpler (and thus more cognitively tractable) in the effective language. Changing the effective language, by adding a new concept, then affects a number of cognitive functions. For instance, future concept induction will take place in the new effective language, which provides a different inductive bias than the original. In this way, concepts which are complex and unlikely to be constructed by a learner initially may become simpler and more plausible later on in the process of elaborating her conceptual library. This process may be a critical driver of children's long-term cognitive development.

8 Summary and next steps

We have argued that concepts should be viewed as the stable representations of a *probabilistic language of thought*—more formally, as functions in an enriched stochastic lambda calculus. This view allows fine-grained compositionality while supporting reasoning by probabilistic inference. Compositionality is key to explaining the productivity of thought, while probabilistic inference explains graded and successful reasoning in an uncertain world. The PLoT hypothesis seeks to explains complex human cognition in a way that previous formal theories of concepts have not, and helps us to understand many topics in everyday cognition with both new qualitative insights and quantitative accuracy.

Importantly, the PLoT hypothesis builds on and unifies many attractive aspects of

previous views on concepts. Like classical and symbolic theories the PLoT puts compositionality and symbolic scaffolding at center stage. Unlike these theories however, but very much in the spirit of prototype, exemplar, and connectionist approaches, the PLoT explains why human reasoning is graded and why this is useful. It does so by borrowing from probability theory, and modern Bayesian modeling approaches, the basic mechanics of reasoning under uncertainty. This reliance on probabilistic inference makes a natural connection to inferential role notions of concept meaning: it is not merely the proximal definitions, but the complex ways that information flows under inference that matter in practice. Rather than working at the level of monolithic probability distributions, the stochastic lambda calculus and Church allow us to work from the point of view of generative, sampling systems. This in turn makes a key connection to mental simulation and—poetically, but perhaps also literally—the importance of *imagination* in thinking.

For work on Bayesian models of cognition the PLoT view holds particular importance. Recent advances in building probabilistic models of higher-level cognition share the basic mechanics of probabilities and many aspects of their philosophy, but they bring a bewildering and heterogenous array of additional representational tools and claims. The view presented here serves as a key unification by showing that all of these Bayesian models can be represented in, and hence reduced to, a simple system built from little more that function abstraction and random choice. It gives hope that advances in probabilistic models of targeted domains are compatible with each other and can ultimately be combined into a broader architecture for modeling human knowledge, reasoning and learning.

Church models, and the PLoT more generally, are intended to capture the representations of knowledge people use to reason about the world, and the inferences that are supported by this knowledge. They are not intended to convey the algorithmic *process* of this inference, much less the neural instantiation. Indeed, establishing connections to these other levels of psychological analysis is one of the key future challenges for the PLoT hypothesis; others being further broadening of scope and demonstration of empirical adequacy within higher-level cognition. The implementations of Church, at the engineering level, suggest one set of ideas to motivate psychological process models. Indeed, implementations of Church `query` work through various combinations of caching and Monte Carlo simulation, which provide a very different view of computation than one might expect from a course on probability: not so much arithmetic tabulation as noisy dynamical systems tuned to result in samples from the desired distributions. Long engineering practice shows that these algorithms can give efficient solutions to tough statistical inference problems; recent work on probabilistic programing languages (e.g. Wingate, Stuhlmüller, & Goodman, 2011) shows that they can be realized in general-purpose ways suitable to a PLoT. Recent work has provided initial connections between such inference algorithms and human cognitive processes (e.g. Griffiths, Vul, & Sanborn, 2012). Yet classic and ongoing work on cognitive architecture, concepts, and neural dynamics all have additional insights that must also be understood in moving the PLoT toward the process and neural levels.

9 Acknowledgments

This work was supported by NSF STC award CCF-1231216, ONR awards N00014-09-1-0124 and N00014-13-1-0788, and a John S. McDonnell Foundation Scholar Award.

References

Anderson, J. R. (1996). Act: A simple theory of complex cognition. *American Psychologist*, *51*(4), 355–365.

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

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013, Oct). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*.

Block, N. (1997). Semantics, conceptual role. *The Routledge Encylopedia of Philosophy*.

Bruner, J., Goodnow, J., & Austin, G. (1967). A study of thinking. *New York: Science Editions.[Links]*.

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

Carey, S. (this volume). Why theories of concepts should not ignore the problem of acquisition.

Chater, N., & Oaksford, M. (2013). Programs as causal models: Speculations on mental programs and mental representation. *Cognitive Science*.

Craik, K. J. W. (1967). *The nature of explanation*. CUP Archive.

Frank, M., & Goodman, N. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*.

Gerstenberg, T., & Goodman, N. D. (2012). Ping Pong in Church: Productive use of concepts in human probabilistic inference. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1590–1595). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 378–383). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (submitted). From counterfactual simulation to causal judgment.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., ... Tenenbaum, J. B. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1382–1387).

Goodman, N. D., & Lassiter, D. (to appear). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *Handbook*

of contemporary semantics. Wiley-Blackwell.

Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: A language for generative models. In *Uncertainty in Artificial Intelligence.*

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*. Retrieved from `http://www.mit.edu/~ast/papers/implicature-topics2013.pdf`

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

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

Gopnik, A. (2003). The theory theory as an alternative to the innateness hypothesis. *Chomsky and his critics*, 238–254.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological Review*, *111*(1), 3–32.

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

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory based causal induction. *Psychological Review*.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *science*, *298*(5598), 1569–1579.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Jackendoff, R. (1995). *Languages of the mind: Essays on mental representation*. mit Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford University Press.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008). Learning and using relational theories. *Advances in Neural Information Processing Systems*, *20*, 753–760.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological review*, *116*(1), 20.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. (2013). One-shot learning by inverting a compositional causal process. In *Advances in neural information processing systems* (pp. 2526–2534).

Margolis, E. (1998). How to acquire a concept. *Mind & Language*, *13*(3), 347–369.

McCarthy, J. (1960). Recursive functions of symbolic expressions and their computation by machine, part i. *Communications of the ACM*, *3*(4), 184–195.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*(3), 151–166.

Nosofsky, R. M. (1986). *Attention, similarity, and the identification–categorization relationship.* (Vol. 115) (No. 1). American Psychological Association.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

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

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.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1141-1159. Retrieved from http://dx.doi.org/10.1037/0278-7393.29.6.1141 doi: 10.1037/0278-7393.29.6.1141

Rosch, E. (1999). Principles of categorization. *Concepts: core readings*, 189–206.

Rumelhart, D. E., & McClelland, J. L. (1988). *Parallel distributed processing*. MIT Press.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others the consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351.

Smith, K., Dechter, E., Tenenbaum, J., & Vul, E. (2013). Physical predictions over time. In *Proceedings of the 35th annual meeting of the cognitive science society.*

Smith, K. A., & Vul, E. (2012). Sources of uncertainty in intuitive physics. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Stuhlmüller, A., & Goodman, N. D. (2013). Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*. Retrieved from http://www.mit.edu/~ast/papers/nested-conditioning-cogsys2013.pdf doi: http://dx.doi.org/10.1016/j.cogsys.2013.07.003

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. *Advances in neural information processing systems*, 43–50.

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.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*(1), 337–375.

Wingate, D., Stuhlmüller, A., & Goodman, N. (2011). Lightweight implementations of probabilistic programming languages via transformational compilation. In *Proceedings of the 14th international conference on artificial intelligence and statistics* (p. 131).

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

A Experiment 1

A.1 Participants

30 (22 female) recruited through Amazon Mechanical Turk participated in the experiment. The mean age was 31.3 ($SD = 10.8$).

A.2 Materials and Procedure

The experiment was programmed in Adobe Flash CS5. Participants viewed 20 tournaments in total. First, one block of 8 single player tournaments and then another block of 12 two-player tournaments. The order of the tournaments within each block was randomized. Participants could remind themselves about the most important aspects of the experiment by moving the mouse over the Info field on the top right of the screen (see Figure 1). Based on the results of the three matches in the tournament, participants estimated the strength of the indicated player on a slider that ranged from -50 to 50. The endpoints were labelled "very weak" and "very strong". It took participants 7.4 ($SD = 3.3$) minutes to complete the experiment.

A.3 Design

Table 1 shows the patterns of evidence that were used for the single player tournaments. Table 2 shows the patterns for the two-player tournaments. In all tournaments, participants were asked to judge the strength of player A.

For the single player tournaments, we used four different patterns of evidence: *confounded evidence* in which A wins repeatedly against B, *strong* and *weak indirect evidence* where A only wins one match herself but B either continues to win or lose two games against other players and *diverse evidence* in which A wins against three different players. For each of those patterns, we also included a pattern in which the outcomes of the games were exactly reversed.

B Experiment 2

B.1 Participants

20 (11 female) recruited through Amazon Mechanical Turk participated in the experiment. The mean age was 34 ($SD = 9.8$).

B.2 Materials, Procedure and Design

Participants viewed 10 single player tournaments which comprised the 8 situations used in Experiment 1 plus two additional patterns (IR 1, 2). Participants first judged player A's strength based merely on the match results in the tournament. Afterwards, participants received information from the omniscient commentator about one player who was lazy in a particular match. Participants then rated A's strength for a second time, whereby the slider was initialized at the first judgment's position. It took participants 9.4 ($SD = 4$) minutes to complete the experiment.

The bottom row of Table 1 shows what information the omniscient commentator revealed in each situation. For example, in situation 3 in which participants first saw strong indirect evidence, the commentator then said: "In game 1, Player B was lazy." In the additional pattern (IR 2), A wins against B, B wins against C and D wins against E. The commentator then reveals that E was lazy in game 3. For the patterns in which A lost his game, the results of each match as shown in Table 1 were reversed and the corresponding losing player was indicated as having been lazy. For example, in situation 2, A lost all three games against B and the commentator revealed that A was lazy in game 2.