Ping Pong in Church: Productive use of concepts in human probabilistic inference

Tobias Gerstenberg[1] (t.gerstenberg@ucl.ac.uk) & Noah D. Goodman[2] (ngoodman@stanford.edu)
[1]Cognitive, Perceptual and Brain Sciences, University College London, London WC1H 0AP
[2]Department of Psychology, Stanford University, Stanford, CA 94305

Abstract

How do people make inferences from complex patterns of evidence across diverse situations? What does a computational model need in order to capture the abstract knowledge people use for everyday reasoning? In this paper, we explore a novel modeling framework based on the probabilistic language of thought (PLoT) hypothesis, which conceptualizes thinking in terms of probabilistic inference over compositionally structured representations. The core assumptions of the PLoT hypothesis are realized in the probabilistic programming language Church (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008). Using "ping pong tournaments" as a case study, we show how a single Church program concisely represents the concepts required to specify inferences from diverse patterns of evidence. In two experiments, we demonstrate a very close fit between our model's predictions and participants' judgments. Our model accurately predicts how people reason with confounded and indirect evidence and how different sources of information are integrated.

Keywords: inference; reasoning; causality; language of thought; probabilistic programming

Introduction

People often make surprisingly accurate inferences about a person's latent traits from very sparse evidence. If the second author (NG) loses to the first author (TG) in a ping pong match and afterwards wins against two other lab members, we are fairly confident that TG is a strong player despite only having observed him winning a single game. However, if we consequently find out that NG felt a bit lazy in his match against TG and did not try as hard as he normally does, our belief about TG's strength might change. This reasoning is not limited to a particular set of potential players, it can be generalized to related situations (such as team matches), and it supports inferences from complex combinations of evidence (e.g. learning that NG was lazy whenever he played a match against a team that included TG) – human reasoning is remarkably *productive*.

How can we best model the flexible inferences people draw from diverse patterns of evidence such as the outcomes of matches in a ping pong tournament? What assumptions about the cognitive system do we need to make to be able to explain the productivity and gradedness of inference? What is the minimum level of abstraction that mental representations need to exhibit in order to support the inferential flexibility that our cognitive machinery displays?

There are two traditional, but fundamentally different ways of modeling higher-level cognition, each with its own strengths and drawbacks: Statistical approaches (e.g. Rumelhart & McClelland, 1988) support graded probabilistic inference based on uncertain evidence but lack some of the representational powers of more richly structured symbolic approaches. Symbolic approaches (e.g. Newell, Shaw, & Simon, 1958), on the other hand, are confined to operating in the realm of certainty and are ill-suited to modeling people's inferences in a fundamentally uncertain world. More recently, researchers have started to break the dichotomy between statistical and symbolic models (Anderson, 1996) and have shown that much of cognition can be understood as probabilistic inference over richly structured representations (Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

For instance, causal Bayesian networks (CBN; Pearl, 2000) have been proposed as a modeling framework that combines the strengths of both statistical and symbolic approaches. Given a particular representation of a task that the cognitive system faces, a CBN supports inferences about the probability of competing hypotheses for many different patterns of evidence. However, a CBN is limited to the specific situation it was designed to model, allowing inferences from different observations of existing variables, but not from fundamentally different combinations of objects or events. While some attempts have been made to model more abstract knowledge by constructing CBNs with richer, hierarchical structures (Kemp & Tenenbaum, 2009) or by combining CBNs with propositional logic (Goodman, Ullman, & Tenenbaum, 2011; Griffiths, 2005), CBNs have only coarse-grained compositionality insufficient to support productive extensions over different objects and situations.

Human thought, in contrast, is characterized by an enormous flexibility and productivity (Fodor, 1975). We can flexibly combine existing concepts to form new concepts and we can make use of these concepts to reason productively about an infinity of situations. The *probabilistic language of thought* (PLoT) hypothesis (Goodman & Tenenbaum, in prep) posits that mental representations have a language-like compositionality, and that the meaning of these representations is probabilistic, allowing them to be used for thinking and learning by probabilistic inference. This view of the representation of concepts provides a deeper marriage of the statistical and symbolic view. Because they are probabilistic, they 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", in the ping pong domain that we may compose together and apply to symbols such as TG. These combinations then describe distributions on possible world states, which we may reason about via the rules of probability. The PLoT hypothesis has been realized in existing computational systems, including the probabilistic programming language Church (Goodman et al., 2008). Church has several features that enable it to model productive inference from a small set of concepts – in particular, it allows reasoning about placeholder symbols and the forming of complex evidence by composing the concepts.

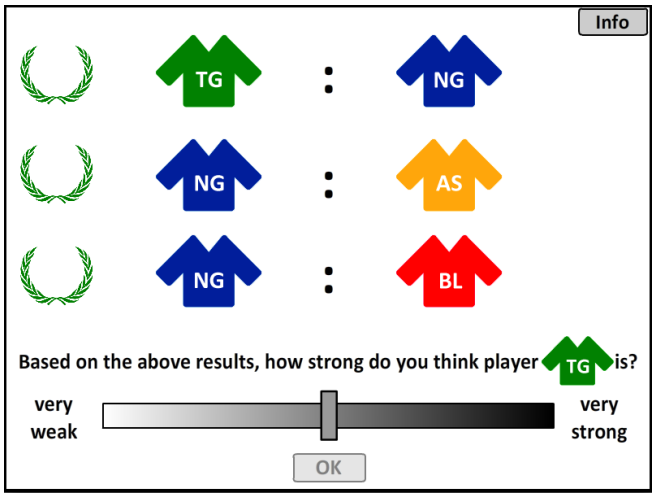


Figure 1: Screenshot of a single player tournament. The winner of each match is indicated by a laurel wreath.

In this paper, we use Church (Goodman et al., 2008), as an instantiation of the PLoT, to explain aspects of people's flexible concept use, and use the ping pong scenario as a simple case study to illustrate our key points while admitting quantitative empirical evaluation. In two separate experiments, we test the predictions of our modeling approach by examining people's inferences based on complex patterns of causal evidence. We conclude by pointing out areas of research that are likely to benefit from this modeling framework.

Modeling probabilistic inferences in Church

Figure 1 shows an example of the inference task that participants faced in the experiments which we will describe below. What representation would be needed to (a) be sensitive to the statistical nature of the evidence and (b) capture the abstract, symbolic structure that remains invariant between this particular situation and other similar situations that could involve different players and different outcomes? Figure 2 shows the Church code that we used to model people's inferences about a player's strength based on the results of ping pong tournaments. We chose the ping pong environment because it can be summarized by a relatively simple but rich set of concepts that support productive inferences from a variety of evidence in a variety of situations. We will first introduce the Church language and then explain how this representation captures our intuitive concepts of ping pong.

Church is based on the λ-calculus, with a syntax inherited from the LISP family of languages (McCarthy, 1960). Thus operators precede their arguments, and are written inside grouping parentheses: `(+ 1 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)))`. 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 outputting either

```
(mh-query 1000 100 ;Monte Carlo Inference
  ;CONCEPTS
  (define personstrength (mem (lambda (person) (gaussian 10 3))))
  (define lazy (mem (lambda (person game) (flip 0.1))))
  (define (teamstrength team game)
    (sum (map (lambda (person)
                (if (lazy person game)
                    (/ (personstrength person) 2)
                    (personstrength person)))
              team)))
  (define (winner team1 team2 game)
    (if (< (teamstrength team1 game)
           (teamstrength team2 game))
        'team2 'team1))
  ;QUERY
  (personstrength 'A)
  ;EVIDENCE
  (and
    (= 'team1 (winner '(TG) '(NG) 1))
    (= 'team1 (winner '(NG) '(AS) 2))
    (= 'team1 (winner '(NG) '(BL) 3))
    (lazy '(NG) 1) ;additional evidence, used in Experiment 2
  )
)
```

Figure 2: Church model of the ping pong scenario.

`true` or `false`. Every time the function is called, the coin is flipped afresh. A Church program specifies not a single computation, but a distribution over computations, or sampling process. This *sampling semantics* (see Goodman 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 now turn to describing the concepts (see `CONCEPTS` in Figure 2) that are required to represent the ping pong domain (Figure 1). 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. We define an individual player's strength, `personstrength`, via a function that draws from a Gaussian distribution with $M = 10$ and $SD = 3$. The memoization operator `mem` ensures that the strength value assigned to a person is persistent and does not change between games. We next make the assumption that players are sometimes `lazy`. The chance of a person being lazy in a particular game is 10%, specified by using the function `flip` with a weight of 0.1. As mentioned above, we also want to allow for the possibility that individual players form teams – we thus need the overall strength of a team,

Table 1: Modeling assumptions.

concept	description	assumption
`personstrength`	strength of a player	normally distributed, persistent property
`lazy`	chance that a player is lazy	p(lazy) = 10%, not persistent
`teamstrength`	strength of a team	individual strengths combine additively
`winner`	winner of a match	team with greater strength wins

Table 2: 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.

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
A > B	B > C	B < C	A > C
A > B	B > D	B < D	A > D
lazy,game: B,2	B,1	B,1	C,2

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

`teamstrength`. Here, we define the team's strength as 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. The way in which we can define new concepts (e.g. `teamstrength`) based on previously defined concepts (`personstrength` and `lazy`) illustrates the compositionality of Church. Finally, we specify how the `winner` of a game is determined. We simply say the the team wins who has the greater overall strength. This set of function definitions specifies a simple lexicon of concepts for reasoning about the ping pong domain. The functions are built up compositionally, and may be further composed for specific situations (see below). What's more, 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. Table 1 concisely summarizes our modeling assumptions.

Now we have a lexicon of concepts (`CONCEPTS`) that we may use to model people's inferences about a player's strength (`QUERY`) 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 (`EVIDENCE`). 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.

A program in Church can be seen as a formal description of the process that generates observed or hypothesized evidence. The `mh-query` operator specifies a conditional inference. Both the evidence provided and the question we are asking are composed out of the concepts that specify the domain. Church completely separates the actual process of inference from the underlying representations and the inferences they license. This allows the modeler to focus on defining the conceptual representation of the domain of interest without having to worry about the exact details of how inference is carried out; it also provides a framework for psychological investigation of representations and the inferences that may be drawn, without committing to *how* these inferences are made – a well-formed level of analysis between Marr's computational and algorithmic levels (Marr, 1982).

Table 3: 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

Hence, in contrast to other frameworks for building psychological models of cognition, such as ACT-R (Anderson, 1996), Church does not incorporate any assumptions about how exactly the cognitive system carries out its computations but merely postulates that inference accords with the rules of probability.

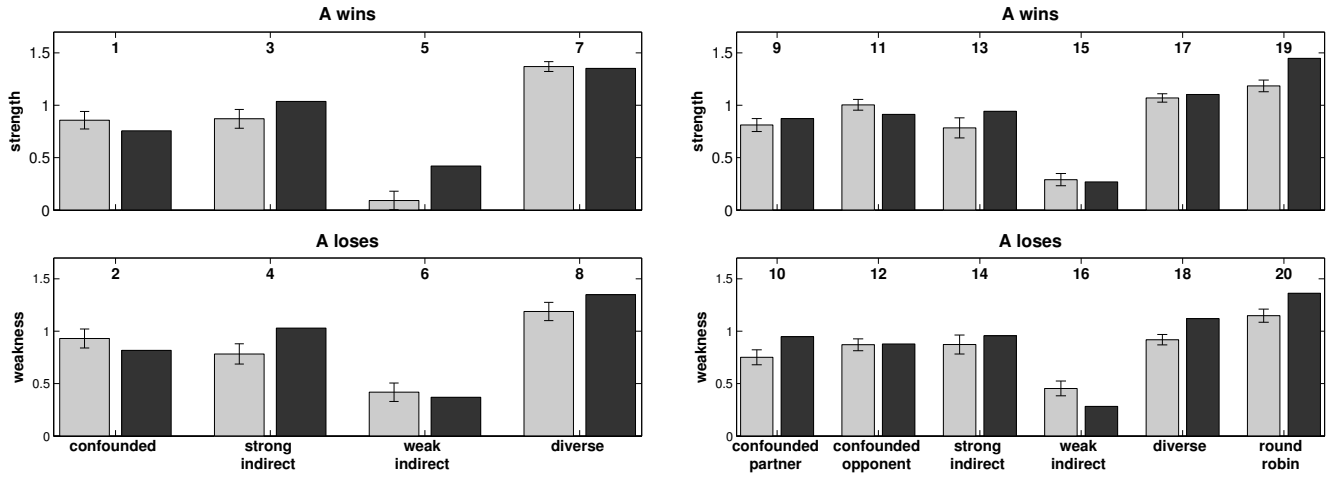
Experiment 1: Bayesian Ping Pong

In Experiment 1, we wanted to explore how well our simple Church model predicts the inferences people make, based on complex patterns of evidence in different situations. 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 (cf. Table 1). 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.

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

Materials and Procedure The experiment was programmed in Adobe Flash CS5.[1] 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

[1]Demos of both Experiments can be accessed here: `http://www.ucl.ac.uk/lagnado-lab/experiments/demos/BPP_demos.html`



(a) Single player tournament

(b) Two-player tournament

Figure 3: Mean strength estimates (grey bars) and model predictions (black bars) for the single player (left) and two-player tournaments (right). Numbers above the bars correspond to the patterns described in Tables 2 and 3. Error bars are $\pm$ 1 *SEM*.

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.

Design Table 2 shows the patterns of evidence that were used for the single player tournaments. Table 3 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 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.

Results and Discussion

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 3 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.

In describing the data qualitatively, we will focus on the strength jugdments in the top panels (strength and weakness judgments were highly correlated, $r = .96$). In the single player tournaments, A is judged equally strong when he repeatedly wins against the same player (situation 1) or when strong indirect evidence was provided (3). A is judged weakest when only weak indirect evidence is provided (5). A is judged to be strongest when she won against three different players (7). In the two-player tournaments, A is judged equally strong when the evidence is confounded with the partner or opponent and when strong indirect evidence is provided (9, 11 and 13). A is judged to be relatively weak when only weak indirect evidence is provided (15). A is judged to be strong for the situations in which participant's received diverse evidence about A's strength (17) and even stronger for the round robin tournament (19).

There appears to be only one prediction that the model makes which is not supported by the data. In the single player tournaments, the model predicts that participants should be slightly more confident about the strength of A when provided with strong indirect evidence (situations 3, 4) compared to when confounded evidence is given (situations 1, 2). However, there is no significant difference between participants'

judgments for strong indirect evidence ($M = 26.2$, $SD = 15.4$) compared to confounded evidence ($M = 27.8$, $SD = 13.8$), $t(29) = 0.44, p > .05$.

The results of Experiment 1 show that our model predicts participants' inferences very accurately. We have demonstrated that a single and concise representation of the task is sufficient to predicts people's inferences for a great diversity of patterns of evidence.

The close fit between our model and participants' inference also shows that our modeling assumptions (e.g. that the team's strength is a linear combination of the individual team members' strengths) generally matched participants' implicit assumptions (cf. Table 1). However, the fact that the model's prediction of a difference between strength judgments based on strong indirect evidence versus confounded evidence was not supported by the data, suggests that participants might have differed in the extent to which they took the chance of laziness into consideration. In fact, only 16 out of 30 participants showed the pattern in the predicted direction. If we increase the probability of a person being lazy in a particular game in the model, it matches participants' average judgments for these situations. Intuitively, if the chances of a person having been lazy in a particular game are increased, there is a higher chance that player A won his game against player B in situation 3 because B was lazy in this round. However, when A wins repeatedly against B, there is hardly any effect of changing the probability of laziness. For example, it is very unlikely when A won three times against B, that B (and not A) was lazy three times in a row.

Experiment 2: Omniscient Commentator

In Experiment 1 we have shown that our model accurately predicts participants' inferences for a great variety of patterns of evidence from different combinations of teams and outcomes. 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. 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 Church code to derive predictions for these situations since all the necessary concepts are already defined.

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

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.

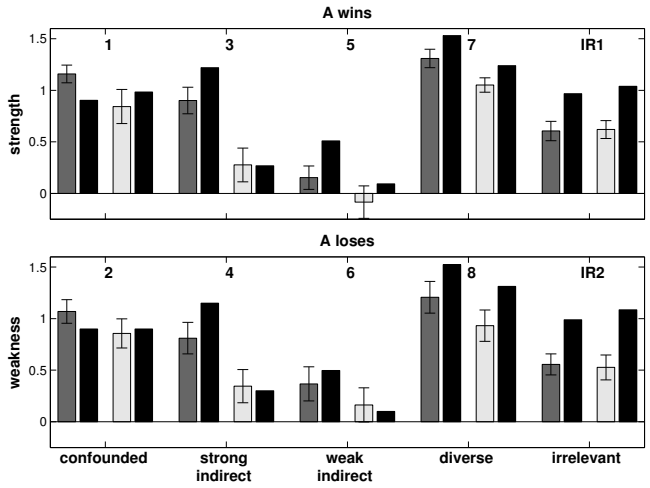


Figure 4: Mean strength estimates and model predictions. Dark grey bars = estimates after tournament information only, light grey bars = estimates after omniscient commentator info, black bars = model predictions. Error bars are $\pm$ 1 *SEM*.

The bottom row of Table 2 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 2 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.

Results and Discussion

Figure 4 shows the mean strength judgments (gray bars) together with the model predictions (black bars). 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 $r = .86$. Again, strength and weakness judgments for the corresponding patterns were highly correlated, $r = .98$.

Generally, participants lowered their estimate of A's strength (top panel) and weakness (bottom panel) after having received the commentator's information. The fact that

participants do not lower their estimates of A's strength for the two cases in which they received *irrelevant evidence* by the commentator about a player' laziness who was in no relationship with A (IR 1, 2), shows that participants did not just have a tendency to regress towards the mean of the scale in their second judgments.

As predicted by the model, the degree to which participants lowered their strength estimates as a result of the laziness information differed between situations. While participants only marginally lowered their estimates for the *confounded evidence* patterns, estimates went down considerably for the *strong indirect evidence* patterns. As mentioned in the discussion of Experiment 1, finding out in the *strong indirect evidence* situation that A's win against B might have only been due to the fact that B was lazy in this match undermines the relevance of the additional evidence about B's performance in match 2 and 3 for A's strength.

The results of Experiment 2 show that participants, as well as our model, have no difficulty in integrating different sources of evidence to form an overall 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.

General Discussion

In this paper, we have demonstrated a novel modeling framework that conceptualizes people's reasoning as probabilistic inference over compositionally structured representations. With a handful of concepts that can combine compositionally and support productive extensions over novel situations and objects, we predict participants' judgments in two experiments with thirty different patterns of evidence in total.

The fact that people can reason flexibly based on different patterns and sources of evidence illustrates the importance of modeling our representational capacities on a sufficiently abstract level. People's use of concepts are not tied to particular situations but extend productively over different contexts. The concept of a *winner*, for example, applies to a whole range of possible games or even to domains outside of games entirely such as winning an election. We have provided a concrete working-example of how such a representation could look like, using the probabilistic programming language Church (Goodman et al., 2008). The fact that our model's predictions corresponded very closely to people's judgments can be taken as evidence that the assumptions we had to make when writing the program, generally matched the intuitive assumptions that people brought to the task. A Church program makes the modeling assumptions explicit and thus allows them to be scrutinized. Furthermore, particular modeling assumptions can also be treated as parameters in the model. For example, as outlined above, different participants seemed to have given unequal weight to the probability that a player might be lazy in an game. Without changing the general structure of our representation, we could account for these individual differences by allowing for flexibility in our modeling assumptions through, for example, treating the chance of laziness as a free parameter.

In our experiments, we have focused on a single query and only used a small number of the possible patterns of evidence. However, our representation supports many more combinations of queries and evidence. For example, we could ask about the probability that a particular player was lazy in a certain game. Or we could ask which of two teams is likely to win given that we have observed the players perform in some previous games or based on some direct information about their strength. Furthermore, it would require only minimal additions to the concept lexicon to handle evidence such as, "all players in the red jerseys were lazy" or "at least one of the players in the green jerseys is very strong."

To conclude, we have provided only a small glimpse into what we see as a broad research program that investigates people's flexible use of everyday concepts using the tools of probabilistic programming – the probabilistic language of thought hypothesis. We are convinced that this research program has the potential to greatly benefit our understanding of how higher-level capacities of human cognition (such as concept learning, naive physics, and theory of mind) are possible.

Acknowledgments

We thank Andreas Stuhlmüller for insightful comments and for helping with the Church implementation. This work was supported by a doctoral grant from the AXA research fund (TG), a John S. McDonnell Foundation Scholar Award and ONR grant N00014-09-1-0124 (NG).

References

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

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

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., & Tenenbaum, J. B. (in prep). *The probabilistic language of throught.*

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

Griffiths, T. L. (2005). Causes, coincidences, and theories. (Unpublished doctoral dissertation)

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

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.

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

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

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.