Running head: ITERATED LEARNING

Iterated learning: Intergenerational knowledge transmission reveals inductive biases

Michael L. Kalish

Institute of Cognitive Science

University of Louisiana at Lafayette

Thomas L. Griffiths

Department of Cognitive and Linguistic Sciences

Brown University

Stephan Lewandowsky

Department of Psychology

University of Western Australia

Address for correspondence:
Mike Kalish
Institute of Cognitive Science
University of Louisiana at Lafayette
Lafayette, LA 70504
Phone: (337) 482 1135
E-mail: kalish@louisiana.edu

Word count: 3580

Abstract

Cultural transmission of information plays a central role in shaping human knowledge. Some of the most complex knowledge that people acquire, such as languages or cultural norms, can only be learned from other people, who themselves learned from previous generations. The prevalence of this process of "iterated learning" as a mode of cultural transmission raises the question of how it affects the information being transmitted. Analyses of iterated learning under the assumption that the learners are Bayesian agents predict that this process should converge to an equilibrium that reflects the inductive biases of the learners. An experiment in iterated function learning with human participants confirms this prediction, providing insight into the consequences of intergenerational knowledge transmission and a method for discovering the inductive biases that guide human inferences.

Iterated learning: Intergenerational knowledge transmission reveals inductive biases

Knowledge changes as it is passed from one person to the next, and from one generation to the next. Sometimes the change is dramatic: the deaf children of Nicaragua have transformed a fragmentary protolanguage into a real language in the brief time required for one generation of signers to mature within the new language's community (e.g., Senghas & Coppola, 2001). Language is only one example, although it is perhaps the most striking, of the inter-generational transmission of cultural knowledge. In many cases of cultural transmission, one learner serves as the next learner's teacher. Languages, legends, superstitions and social norms are all transmitted by such a process of "iterated learning" (see Figure 1a), with each generation learning from data produced by that which preceded it (Boyd & Richerson, 1985; Briscoe, 2002; Cavalli-Sforza & Feldman, 1981; Kirby, 1999, 2001). However, iterated learning does not result in perfect transfer of knowledge across generations. Its outcome depends not just on the data being passed from learner to learner, but on the properties of the learners themselves.

The prevalence of iterated learning as a mode of cultural transmission raises an important question: what are the consequences of iterated learning for the information being transmitted? In particular, does this information converge to a predictable equilibrium, and are the dynamics of this process understandable? This question has been explored in a variety of disciplines, including anthropology and linguistics. In anthropology, several researchers have argued that processes of cultural transmission like iterated learning provide the opportunity for the biases of learners to manifest in the concepts used by a society (Atran, 2001, 2002; Boyer, 1994, 1998; Sperber, 1996). In

linguistics, iterated learning provides a potential explanation for the structure of human languages (e.g., Kirby, 2001; Briscoe, 2002). This approach is an alternative to traditional claims that the structure of language is the result of constraints imposed by an innate, special-purpose, language faculty (e.g., Chomsky, 1965; Hauser, Chomsky, & Fitch, 2002). Simulations of iterated learning with general-purpose learning algorithms have shown that languages with considerable degrees of structure can emerge when agents are allowed to learn from one another (Kirby, 2001; Smith, Kirby, & Brighton, 2003; Brighton, 2002).

Despite this interest in cultural transmission, there has been very little laboratory work on the consequences of iterated learning. Bartlett's (1932) experiments in "serial reproduction" were the first psychological investigations of this topic, using a procedure in which participants reconstructed a stimulus from memory, with their reconstructions serving as stimuli for later participants. Bartlett concluded that reproductions seem to become more consistent with the biases of the participants as the number of reproductions increases. However, these claims are impossible to validate, since Bartlett's experiments used stimuli, such as pictures and stories, that are not particularly amenable to rigorous analysis. In addition, there was no unambiguous pre-experimental hypothesis about what people's biases might be for these complex stimuli. There have been only a few subsequent studies in serial reproduction, with the most prominent being Bangerter (2000) and Barrett and Nyhof (2001), and thus we presently have little understanding of the likely outcome of iterated learning in controlled conditions. The possibilities are numerous: iteration might produce divergence from structure into noise, or into random or unpredictable alternation from one solution to another, or people might blend their biases with the data to form consistent "compromise" solutions. In this paper, we attempt to

determine the outcome of intergenerational knowledge transmission by testing the predictions made by a formal analysis of iterated learning in an experiment using a controlled set of stimuli.

We can gain some insight into the consequences of iterated learning by considering the case where the learners are Bayesian agents. Bayesian agents use a principle of probability theory, called Bayes' rule, to infer the process that was responsible for generating some observed data. Assume that a learner has a set of hypotheses, $\mathcal{H}$, about the process that could have produced data, d, and a "prior" probability distribution, $p(h)$, that encodes that learner's biases by specifying the probability a learner assigns to the truth of each hypothesis $h \in \mathcal{H}$ before seeing d. In the case of learning a language, the hypotheses, h, are different languages, and the data, d, are a set of utterances. Bayes' rule states that the probability that an agent should assign to each hypothesis after seeing d – known as the "posterior" probability, $p(h|d)$ – is

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h \in \mathcal{H}} p(d|h)p(h)}, \tag{1}$$

where $p(d|h)$ – the "likelihood" – indicates how likely d is under hypothesis h, and $p(d)$ is the probability of d averaged over all hypotheses, $p(d) = \sum_h p(d|h)p(h)$, sometimes called the prior predictive distribution. The assumption that learners are Bayesian agents is not unreasonable: adherence to Bayes' rule is a fundamental principle of rational action in statistics and economics (Savage, 1954; Jaynes, 2003; Robert, 1994) and its use underlies many learning algorithms (Mitchell, 1997; Mackay, 2003).

In iterated learning with Bayesian agents, each learner uses Bayes' rule to infer the language spoken by the previous learner, and generates the data provided to the next learner using the results of this inference (see Figure 1b). Having formalized iterated learning in this way, we can examine how it affects the hypotheses chosen by the learners.

The probability that the nth learner chooses hypothesis i given that the previous learner chose hypothesis j is

$$p(h_n = i|h_{n-1} = j) = \sum_d p(h_n = i|d)p(d|h_{n-1} = j), \quad (2)$$

where $p(h_n = i|d)$ is the posterior probability obtained from Equation 1. This specifies the transition matrix of a Markov chain, with the hypothesis chosen by each learner depending only on that chosen by the previous learner. Griffiths and Kalish (2005) showed that the stationary distribution of this Markov chain is $p(h)$, the prior assumed by the learners. The Markov chain will converge to this distribution under fairly general conditions (e.g., Norris, 1997). This means that the probability that the last in a long line of learners chooses a particular hypothesis is simply the prior probability of that hypothesis, regardless of the data provided to the first learner. In other words, the stimuli provided for learning are completely irrelevant in the long run, and only the biases of the learners affect the outcome of iterated learning.[1]

A similar convergence result can be obtained if we consider how the data generated by the learners (instead of the hypotheses they hold) change over time: after many generations, the probability that a learner generates data d will be $p(d) = \sum_h p(d|h)p(h)$, the probability of d under the prior predictive distribution (Griffiths & Kalish, 2006). This process of convergence is illustrated in Figure 2 for the case where the hypotheses are linear functions and the prior favors functions with unit slope and zero intercept (the details of this Bayesian model appear in the Appendix). This is a simple example of iterated learning, but nonetheless illustrative of convergence to the prior predictive distribution. As each generation of learners combines the evidence provided by the data with their (common) prior, their posterior distributions move closer to the prior, and the data they produce becomes more consistent with hypotheses that have high prior

probability.

The preceding analysis of iterated learning with Bayesian agents provides a simple answer to the question of how iterated learning affects the information being transmitted: the information will be transformed to reflect the inductive biases of the learners. Whether a similar transformation will be observed with human learners is an open empirical question. To test this prediction, we reproduced iterated learning in the laboratory using a set of controlled stimuli for which people's biases are well understood. We chose to use a function learning task, because of the prominent role that inductive bias seems to play in this domain.[2] In function learning, each learner sees data consisting of (x, y) pairs and attempts to infer the underlying function relating y to x. Experiments typically present the values of x graphically, and subjects produce a graphical y magnitude in response. Tests of interpolation and extrapolation with novel x values reveal that people infer continuous functions from these discrete trials. Previous experiments in function learning suggest that people have an inductive bias favoring linear functions with a positive slope: initial responses are consistent with such functions (Busemeyer, Byun, DeLosh, & McDaniel, 1997), and they require the least training to learn (Brehmer, 1971, 1974; Busemeyer et al., 1997). Kalish, Lewandowsky, and Kruschke (2004) showed that a model that included such a bias could account for a variety of phenomena in human function learning. If iterated learning converges to an equilibrium reflecting the inductive biases of the learners, we should expect to see linear functions with positive slope emerge after a few generations of learners. We tested this hypothesis by examining the outcome of iterated function learning, varying the functions used to train the first learner in each sequence.

Method

Participants

288 undergraduate psychology students from the University of Louisiana at Lafayette participated for partial course credit. The experiment had four conditions, corresponding to different initial training functions. There were 72 participants in each condition, forming nine generations of learners from eight "families" in which the responses of each generation of learners during a post-training transfer test were presented to the next generation of learners as the to-be-learned target stimuli.

Apparatus and Stimuli

Participants completed the experiment in individual sound-attenuated booths. A computer displayed all trials and collected all responses. On each trial a filled blue bar 1cm high and 0.3cm ($x = 1$) to 30cm ($x = 100$) wide was presented as the stimulus. The stimulus was always presented in the upper portion of the screen, with its upper left corner approximately 4cm from the top and 4cm from the left of the edge of the screen. The participant entered a response magnitude by adjusting a vertically-oriented unmarked slider (located 4cm from the botton and 6cm from the right of the screen) with the mouse; the slider's position determined the height of a filled red bar 1cm wide, which could extend up to 25cm. During the training phase, feedback was provided in the form of a filled yellow bar 1cm wide placed 1cm to the right of the response bar, which varied from 0.25cm ($y = 1$) to 25cm ($y = 100$) in height and was aligned so that the height of the bar was aligned with the correct response.

Procedure

The experiment had both training and transfer phases. For the learners who formed the first generation of any family, the values of the training stimuli were 50 randomly selected stimulus values (x) ranging from 1 to 100 paired with feedback (y) given by the function of the condition the participant was in. The four functions used during training of the first generation of participants in the four conditions were: $y = x$ (positive linear), $y = 101 - x$ (negative linear), $y = 50.5 + 49.5 \sin\left(\frac{\pi}{2} + \frac{x}{5\pi}\right)$ (non-linear, U-shaped) and a random 1-to-1 pairing of x and y with both $x, y \in \{1, \ldots, 100\}$. All values of x were integers and all values of y were rounded to the nearest integer prior to display.

The test items consisted of 25 of the training items along with 25 of the 50 unused stimulus values. Inter-generational transfer took place by making the test stimuli and responses of generation n of each family serve as the training items of generation $n + 1$ in that family. Inter-generational transfer was conducted entirely without personal contact and participants were not made aware that their test responses would serve as training for later participants; the use of one generation's test items in training the next generation was the only contact between generations.

Each trial was initiated by the presentation of a stimulus, selected without replacement from the 50 items in either the training or test set. Following each stimulus presentation, while the stimulus remained on the screen, the participant used the mouse to adjust the slider to indicate their predicted response magnitude and clicked a button to record their response when they had adjusted the slider to the desired magnitude. The response could be manipulated *ad lib* until the participant chose to record the response.

During training each response was followed by the presentation of a feedback bar. If the response was correct (defined as within 1.5 cm, or 5 units, of the target value y), there

was a study interval of 1s duration during which the stimulus, response, and feedback were all presented. If the response was incorrect, a tone sounded and the participant was shown the feedback bar. The participant was then required to set the slider so that the response bar was equal to the feedback bar. A study interval of 2s duration followed this correction. Thus, participants who responded accurately spent less time studying the feedback; this was the reward for accurate responses. After each study interval there was a blank interval of 2s duration before the next trial. Each participant completed a single block of training in which each of their 50 training values was presented once in random order. Test trials were identical to training trials, except that no feedback was made available after the response was entered. Participants were informed prior to the beginning of the test phase about this change.

Results

Figure 3 shows a single family of nine participants for each condition, chosen to be representative of the overall results. Each set of axes shows the test-phase responses of a single learner who was trained using the data shown in the graph to its left. For example, the responses of the first generation in each condition (in column 2) were based on the data provided by the actual function (to the left, in column 1). The responses of the second generation (in column 3) were based on the data produced by the first generation (in column 2), and so forth.

Regardless of the data seen by the first learner, iterated learning converged to a linear function with positive slope in only a few generations for 28 of the 32 families of learners. The other four families, three in the negative linear condition and one in the random condition, were producing a negative linear function at the point the experiment

concluded. Figure 3(a) indicates that a linear function with positive slope is stable under iterated learning; none of the other initial conditions had this level of stability. Figure 3(b) is reminiscent of the analysis shown in Figure 2: despite starting with a linear function with negative slope, learners converged to a linear function with positive slope. Figure 3(c) and (d) show that linear functions with positive slope also emerge from iterated learning when the initial function is non-monotonic or completely random. Figure 3(e) shows the family that most strongly maintained the negative linear function.

The results shown in Figure 3 illustrate a tendency for iterated learning to converge to a positive linear function. To provide a more quantitative analysis, we computed the correlation between the responses of each participant and the positive linear function $y = x$. The outliers produced by families converging to the negative linear function made the mean correlation less informative than the median; Figure 3(f) shows the median correlations at each generation for each of the four conditions. Other than the positive linear condition, where the correlation was at ceiling from the first generation, the correlations systematically increased across generations. As the data clearly show, iterated learning produced an increase in the consistency of people's responses with a positive linear function. This result is consistent with a prior that is dominated by linear functions, with a strong bias for the positive over the negative.

Discussion

Our analysis of iterated learning with Bayesian agents indicates that when Bayesian learners learn from one another, they converge to a distribution over hypotheses determined by their inductive biases (Griffiths & Kalish, 2005, 2006). The purpose of this experiment was to test whether iterated learning with human learners likewise converges

to an outcome consistent with their inductive biases. Previous research in function learning suggests that people favor linear functions with positive slopes (Brehmer, 1971, 1974; Busemeyer et al., 1997; Kalish et al., 2004). In our experiment, iterated learning converged to a positive linear function in the majority of cases, regardless of the function seen by the first learner. Mapping out the stationary distribution of iterated learning in greater detail would require collecting significantly more data, making it possible to confirm that the underlying Markov chains have converged, and providing more samples from the stationary distribution. Nonetheless, the dominance of the positive linear function in our results suggests that, for human learners as for Bayesian agents, iterated learning produces convergence to the prior.

These empirical results are consistent with our theoretical analysis of iterated learning, and have two significant implications. First, they support the idea that information transmitted via iterated learning will ultimately come to mirror the structure of the human mind, a conclusion consistent with claims that processes of cultural transmission can allow the biases of learners to manifest in cultures (Atran, 2001, 2002; Boyer, 1994, 1998; Sperber, 1996). This suggests that languages, legends, religious concepts, and social norms are all tailored to match our biases, providing a formal justification for studying these phenomena as a means of understanding human cognition. These results also validate the interpretation of existing, less controlled, experiments using serial reproduction (e.g., Bartlett, 1932; Bangerter, 2000; Barrett & Nyhof, 2001) as revealing people's biases.

Second, and perhaps more importantly, our results suggest that iterated learning can be used as a method for exploring the biases that guide human learning. Many of the problems that are central to cognitive science, from learning and using language to

inferring the structure of categories from a few examples, are problems of induction. A variety of arguments, from both philosophy (e.g., Goodman, 1955) and learning theory (e.g., Kearns & Vazirani, 1994; Vapnik, 1995), stress the importance of inductive biases in solving these problems. In order to understand how people make inductive inferences, we need to understand the biases that constrain those inferences. The present experiment involved a case in which the general shape of learners' biases was known prior to the study. Based on the results of the experiment, it appears possible to use the procedure to investigate biases in situations in which they are unknown and people are unable (or unwilling) to reveal what those biases are. By reproducing iterated learning in the laboratory, we may be able to map out the implicit inductive biases that make human learning possible.

References

Atran, S. (2001). The trouble with memes: Inferences versus imitation in cultural creation. *Human Nature*, *12*, 351-381.

Atran, S. (2002). *In gods we trust: The evolutionary landscape of religion.* Oxford: Oxford University Press.

Bangerter, A. (2000). Transformation between scientific and social representations of conception: The method of serial reproduction. *British Journal of Social Psychology*, *39*, 521-535.

Barrett, J., & Nyhof, M. (2001). Spreading nonnatural concepts. *Journal of Cognition and Culture*, *1*, 69-100.

Bartlett, F. C. (1932). *Remembering: a study in experimental and social psychology.* Cambridge: Cambridge University Press.

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis.* New York: Wiley.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process.* Chicago: University of Chicago Press.

Boyer, P. (1994). *The naturalness of religious ideas: A cognitive theory of religion.* Berkeley, CA: University of California Press.

Boyer, P. (1998). Cognitive tracks of cultural inheritance: how evolved intuitive ontology governs cultural transmission. *American Anthropologist*, *100*, 876-889.

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science*, *24*, 259-260.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Decision Processes*, *11*, 1-27.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, 25-54.

Briscoe, E. (Ed.). (2002). *Linguistic evolution through language acquisition: Formal and computational models.* Cambridge, UK: Cambridge University Press.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Concepts and categories* (p. 405-437). Cambridge: MIT Press.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution.* Princeton, NJ: Princeton University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* New York: Chapman & Hall.

Goodman, N. (1955). *Fact, fiction, and forecast.* Cambridge: Harvard University Press.

Griffiths, T. L., & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (p. 827-832). Mahwah, NJ: Erlbaum.

Griffiths, T. L., & Kalish, M. L. (2006). *Language evolution by iterated learning with Bayesian agents.* (Submitted for publication)

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*, 1569-1579.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge: Cambridge University Press.

Kalish, M., Lewandowsky, S., & Kruschke, J. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*, 1072-1099.

Kearns, M., & Vazirani, U. (1994). *An introduction to computational learning theory.* Cambridge, MA: MIT Press.

Kirby, S. (1999). *Function, selection and innateness: The emergence of language universals.* Oxford: Oxford University Press.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, *5*, 102-110.

Mackay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge: Cambridge University Press.

Mitchell, T. M. (1997). *Machine learning.* New York: McGraw Hill.

Norris, J. R. (1997). *Markov chains.* Cambridge, UK: Cambridge University Press.

Robert, C. P. (1994). *The Bayesian choice: A decision-theoretic motivation.* New York: Springer.

Savage, L. J. (1954). *Foundations of statistics.* New York: John Wiley & Sons.

Senghas, A., & Coppola, M. (2001). Children creating languge: How Nicaraguan sign language acquired a spatial grammar. *Psychological Science*, *12*, 323-328.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*, 371-386.

Sperber, D. (1996). *Explaining culture: A naturalistic approach.* Oxford: Blackwell.

Vapnik, V. N. (1995). *The nature of statistical learning theory.* New York: Springer.

Appendix

Bayesian linear regression

Linear regression is a standard problem that is dealt with in detail in several books on Bayesian statistics, including Box and Tiao (1992) and Gelman, Carlin, Stern, and Rubin (1995). For this reason, our treatment of this analysis is extremely brief. Assume that the data d are a set of n pairs (x_i, y_i), and that the hypothesis space $\mathcal{H}$ consists of linear functions of the form $y = \beta_1 x + \beta_0 + \epsilon$, where ϵ is Gaussian noise with variance σ_Y^2. Since a hypothesis h is identified entirely by the parameters β_1 and β_0, we can summarize both data and hypotheses using column vectors, $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$, $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$, and $\boldsymbol{\beta} = [\beta_1 \;\; \beta_0]^T$.

The likelihood, $p(d|h)$, is simply the probability of $\mathbf{x}$ and $\mathbf{y}$ given $\boldsymbol{\beta}$, $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\beta})$. Assuming that $\mathbf{x}$ follows a distribution $q(\mathbf{x})$ which is constant over all hypotheses, we have $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\beta}) = p(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta})q(\mathbf{x})$. From the assumption that $y = \beta_1 x + \beta_0 + \epsilon$, it follows that $p(\mathbf{y}|\mathbf{x}, \beta)$ is Gaussian with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\sigma_Y^2 \mathbf{I}_n$, where $\mathbf{X} = [\mathbf{x} \;\; \mathbf{1}_n]$, and $\mathbf{1}_n$ and $\mathbf{I}_n$ are an $n \times 1$ vector of 1s and the $n \times n$ identity matrix respectively. The prior $p(h)$ is a distribution over the parameters $\boldsymbol{\beta}$, $p(\boldsymbol{\beta})$. We take $p(\boldsymbol{\beta})$ to be Gaussian with mean $\boldsymbol{\mu}_\beta$ and covariance matrix $\sigma_\beta^2 \mathbf{I}_2$.

The posterior distribution $p(h|d)$ is a distribution over $\boldsymbol{\beta}$ given $\mathbf{x}$ and $\mathbf{y}$. Using our choice of prior and likelihood, this is Gaussian with covariance matrix

$$\boldsymbol{\Sigma}_{post} = (\frac{1}{\sigma_Y^2}\mathbf{X}^T\mathbf{X} + \frac{1}{\sigma_\beta^2}\mathbf{I}_2)^{-1}, \tag{3}$$

and mean

$$\boldsymbol{\mu}_{post} = \boldsymbol{\Sigma}_{post}^{-1}(\frac{1}{\sigma_Y^2}\mathbf{X}^T\mathbf{y} + \frac{1}{\sigma_\beta^2}\boldsymbol{\mu}_\beta). \tag{4}$$

Figure 2 was generated by simulating iterated learning with this model. The first learner saw 20 datapoints generated by sampling x uniformly at random from the range $[0, 1]$, and taking $y = 1 - x$. A value of $\boldsymbol{\beta}$ was sampled from the resulting posterior distribution, and used to generate values of y for 20 new randomly drawn values of x, which were supplied as data to the next learner. This process was continued for a total of nine learners, producing the results shown in the figure. The likelihood and prior assumed by the learners had $\sigma_Y^2 = 0.0025$, $\sigma_\beta^2 = 0.005$, and $\boldsymbol{\mu}_\beta = [1 \;\; 0]^T$, corresponding to a strong prior favoring functions with a slope of 1 and an intercept of 0.

Author Note

This research was partially supported by a Research Award from the Louisiana Board of Regents to the first author and by a Discovery Grant from the Australian Research Council to the third author. We thank Charles Barouse, Laurie Robinette and Margery Doyle for assistance in data collection.

Footnotes

[1]In the case of language, demonstrating that iterated learning converges to the prior distribution over hypotheses maintained by the learners should not be taken as implying that linguistic universals are necessarily the consequence of innate constraints specific to language learning. The biases encoded by the prior need not be either innate, as they could result from experiences with data other than those under consideration, or language specific, as they could include general-purpose constraints such as limitations on information processing (see Griffiths & Kalish, 2006, for a more detailed discussion).

[2]Our use of function learning was also inspired by simulations of iterated learning of languages, in which a language is often conceived of as a function mapping meanings to utterances (e.g., Smith et al., 2003).

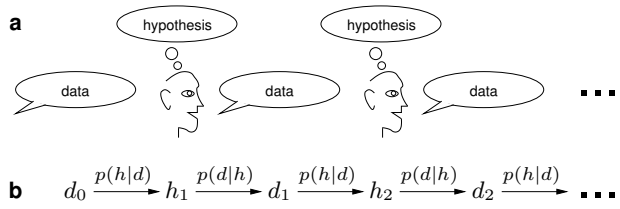
Figure Captions

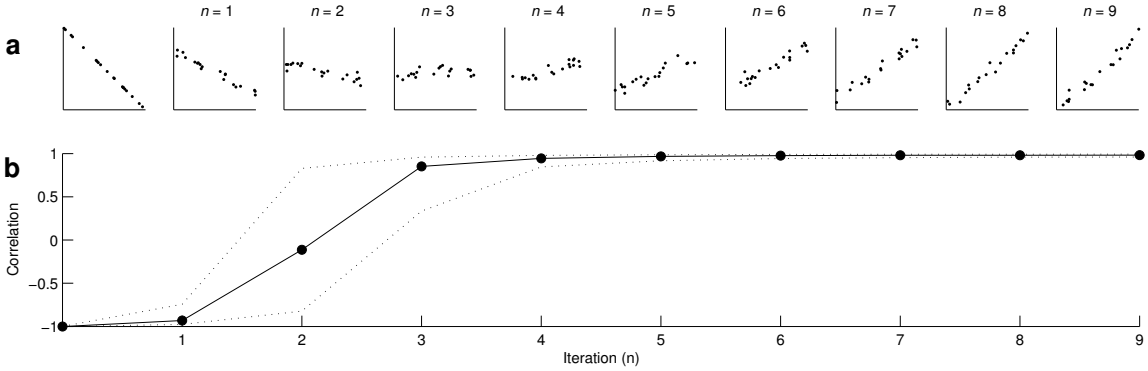
Figure 1. (a) Iterated learning. Each learner sees data produced by a learner in a previous generation, forms a hypothesis about the process by which those data were produced, and uses this hypothesis to produce the data that will be supplied to a learner in the next generation. (b) Iterated learning with Bayesian agents. The first learner sees data d_0, computes a posterior probability distribution over hypotheses according to Equation 1, samples a hypothesis h_1 from this distribution, and generates new data d_1 by sampling from the likelihood associated with that hypothesis. This data is provided to the second learner, and the process continues, with the nth learner seeing data d_{n-1}, inferring a hypothesis h_n, and generating new data d_n.

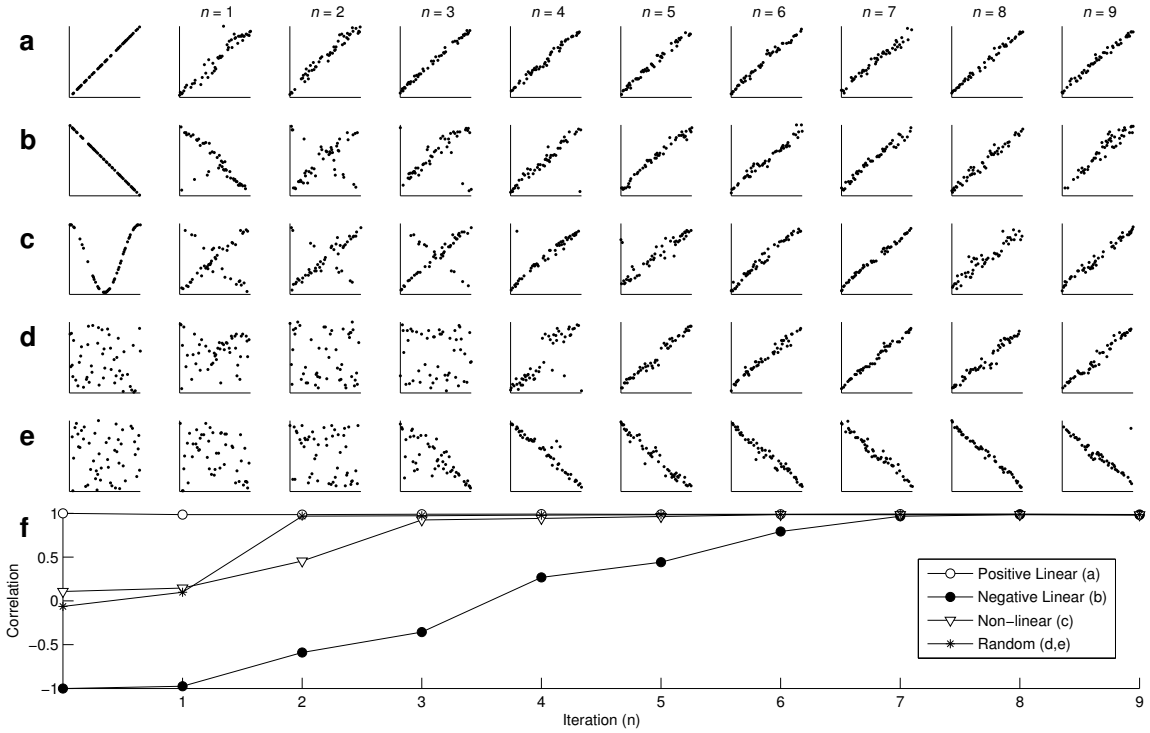
Figure 2. Iterated learning with Bayesian agents converges to the prior. (a) The leftmost panel shows the initial data provided to a Bayesian learner, a sample of 20 points from a function. The learner inferred a hypothesis (in this case a linear function) from these data, and then generated the predicted values of y shown in the next panel for a new set of inputs x. These predictions were supplied as data to another Bayesian learner, and the remaining panels show the predictions produced by learners at each generation as this process continued. All learners had a prior distribution over hypotheses favoring linear functions with positive slope (see the Appendix for details). As iterated learning proceeds, the predictions converge to a positive linear function. (b) The correlation between predictions and the function $y = x$ provides a quantitative measure of correspondence to the prior. The solid line shows the median correlation with $y = x$ for functions produced by 1000 sequences of iterated learning like that shown in (a). The dotted lines show the 95% confidence interval. Using this quantitative measure, it is easy to see that iterated

learning quickly produces a strong correspondence to the prior.

Figure 3. Iterated learning with human learners. The leftmost panel in each row shows the data seen by the first learner, a sample of 50 points from one of four functions. The other columns show the data produced by each generation of learners, trained on the data from the column to their left. Each row shows a single sequence of nine learners, drawn at random from the eight "families" of learners run with the same initial data. The rows differ in the functions used to generate the data shown to the first subject: (a) is a linear function with positive slope, (b) is a linear function with negative slope, (c) is a non-linear function, and (d) is a random set of points. In each case, iterated learning quickly converges to a linear function with positive slope, consistent with findings indicating that human learners are biased towards this kind of function. (e) In a minority of cases (4 out of 32) families were producing negative linear functions at the end of the experiment, suggesting that such functions receive some weight under the prior. (f) The median correlation with $y = x$ across all families was assessed for the four conditions illustrated in (a)-(e). Regardless of the initial data, this correlation increased over generations as predictions became more consistent with a positive linear function.


a
hypothesis
hypothesis
data
data
data
b
$d_0 \xrightarrow{p(h|d)} h_1 \xrightarrow{p(d|h)} d_1 \xrightarrow{p(h|d)} h_2 \xrightarrow{p(d|h)} d_2 \xrightarrow{p(h|d)} \cdots$


a
n = 1
n = 2
n = 3
n = 4
n = 5
n = 6
n = 7
n = 8
n = 9
b
Correlation
1
0.5
0
−0.5
−1
1
2
3
4
5
6
7
8
9
Iteration (n)


n = 1
n = 2
n = 3
n = 4
n = 5
n = 6
n = 7
n = 8
n = 9
a
b
c
d
e
f
Correlation
Iteration (n)
Positive Linear (a)
Negative Linear (b)
Non−linear (c)
Random (d,e)