Cognitive Psychology 60 (2010) 107–126

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

A rational analysis of the effects of memory biases on serial reproduction

Jing Xu *, Thomas L. Griffiths

Department of Psychology, University of California, Berkeley, United States

ARTICLE INFO

Article history:
Accepted 15 September 2009
Available online 30 October 2009

Keywords:
Serial reproduction
Memory
Inductive biases
Bayesian inference
Markov chain
Autoregressive time series analysis

ABSTRACT

Many human interactions involve pieces of information being passed from one person to another, raising the question of how this process of information transmission is affected by the cognitive capacities of the agents involved. Bartlett (1932) explored the influence of memory biases on the "serial reproduction" of information, in which one person's reconstruction of a stimulus from memory becomes the stimulus seen by the next person. These experiments were done using relatively uncontrolled stimuli, but suggested that serial reproduction could transform information in a way that reflected the biases inherent in memory. We formally analyze serial reproduction using a Bayesian model of reconstruction from memory, giving a general result characterizing the effect of memory biases on information transmission. We then test the predictions of this account in four experiments using simple one-dimensional stimuli. Our results provide theoretical and empirical justification for the idea that serial reproduction reflects memory biases.

1. Introduction

Many of the facts that we know about the world are not learned through first-hand experience, but are the result of information being passed from one person to another. This raises a natural question: how are such processes of information transmission affected by the capacities of the agents involved? Decades of memory research have charted the ways in which our memories distort reality, changing the details of experiences and introducing events that never occurred (see Schacter, Coyle, Fischbach, Mesulam, &

* Corresponding author. Address: Department of Psychology, University of California, 3210 Tolman Hall # 1650, Berkeley, CA 94720-1650, United States. Fax: +1 510 642 0135.
E-mail address: jing.xu@berkeley.edu (J. Xu).

0010-0285/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogpsych.2009.09.002

Sullivan (1995) for an overview). We might thus expect that these memory biases would affect the transmission of information, since such a process relies on each person remembering a fact accurately.

The question of how memory biases affect information transmission was first investigated in detail in Sir Frederic Bartlett's "serial reproduction" experiments (Bartlett, 1932). In this paradigm, a participant is shown a stimulus, which could be a story, a passage of text, a joke, or a picture. The participant is asked to memorize the stimulus and then reconstruct it from memory. The reconstruction produced by that participant then becomes the stimulus presented to the next participant, with the sequence of responses being the outcome of process of serial reproduction. In one of his most famous experiments, Bartlett told English speaking participants a native American folk tale, "The War of the Ghosts." As a result of serial reproduction, many supernatural details were lost or replaced by elements familiar in English culture. In another of his experiments, he showed participants an Egyptian hieroglyph – the character resembling an owl shown in Fig. 1 – and as a result of serial reproduction (in this case, with each participant drawing the character from memory) it transformed into a cat. Bartlett interpreted these studies as showing that people were biased by their own cultural expectations when they reconstruct information from memory, and that this bias became exaggerated through serial reproduction.

Serial reproduction has become one of the standard methods used to simulate the process of cultural transmission, and several subsequent studies have used this paradigm to examine how cultural biases affect the transmission of information. For example, researchers have used this method to reveal memory biases in children and adults of different professions and nationalities (see Mesoudi (2007) and Mesoudi & Whiten (2008) for reviews). Recently researchers used updated versions of

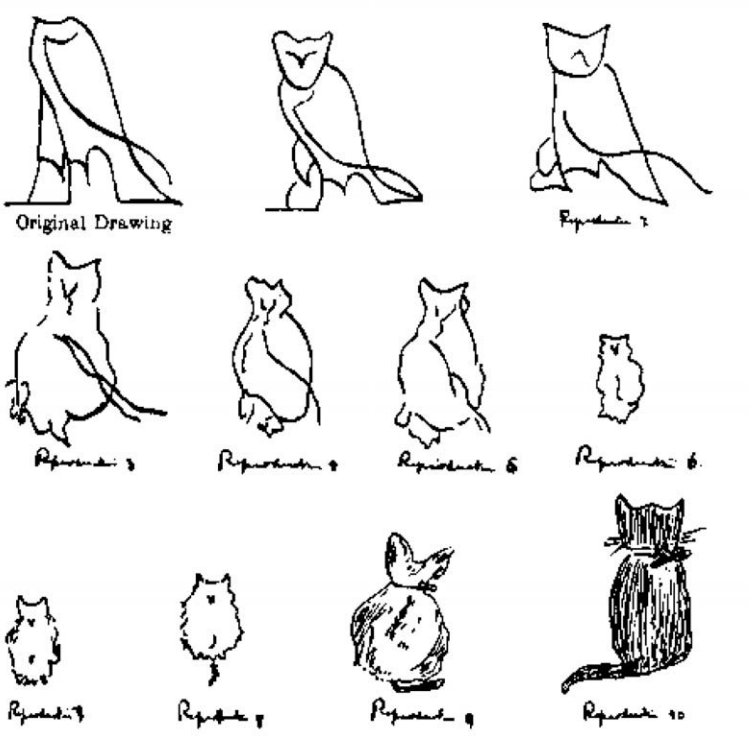

Fig. 1. Results of one of Bartlett's (1932) serial reproduction experiments. The first participant in the chain was presented with an Egyptian hieroglyph (which resembles a picture of an owl). The picture was taken away, and the participant was asked to reconstruct it from memory. Each participant's reconstruction became the picture seen by the next participant, and the hieroglyph became an owl and then a cat. These results were interpreted as the outcome of memory biases toward culturally familiar stimuli. Reproduced with permission from Bartlett (1932).

Bartlett's method to study particular cultural biases, such as gender stereotypes (Bangerter, 2000), and to explore what properties of concepts support their propagation (e.g. Barrett & Nyhof, 2001; Kashima, 2000). In anthropology and evolutionary biology, "diffusion chains," which are different varieties of the serial reproduction paradigm, have been widely used to study cultural processes and social learning in non-human animal societies, such as blackbirds passing alarm calls, foraging behavior in birds, rats, fish, and monkeys, and tool use in chimpanzees (see Whiten & Mesoudi (2008) for a review). However, the phenomenon of serial reproduction has not been systematically and formally analyzed, and most of these studies have used complex stimuli that are semantically rich but hard to control. In this paper, we formally analyze and empirically evaluate how information is changed by serial reproduction, showing how this process relates to memory biases. In particular, we provide a rational analysis of serial reproduction (in the spirit of Anderson (1990)), considering how information should change when passed along a chain of rational agents.

At the heart of our analysis of serial reproduction is the question of how memory biases influence cultural transmission. Biased reconstructions are found in many tasks. For example, people are biased by their knowledge of the structure of categories when they reconstruct simple stimuli from memory. One common effect of this kind is that people judge stimuli that cross boundaries of two different categories to be further apart than those within the same category even when the distances between the stimuli are the same (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). However, biases need not reflect suboptimal performance. If we assume that memory is solving the problem of extracting and storing information from the noisy signal presented to our senses, we can analyze the process of reconstruction from memory as a Bayesian inference. Under this view, reconstructions should combine prior knowledge about the world with the information provided by noisy stimuli. Use of prior knowledge will result in biases, but these biases ultimately make memory more accurate (Huttenlocher, Hedges, & Vevea, 2000).

If this account of reconstruction from memory is true, we would expect the same inference process to occur at every step of serial reproduction. The effects of memory biases should thus be accumulated. Assuming all participants share the same prior knowledge about the world, serial reproduction should ultimately reveal the nature of this knowledge. Drawing on recent work exploring other processes of information transmission (Griffiths & Kalish, 2005, 2007), we show that a rational analysis of serial reproduction makes exactly this prediction. To test this account, we explore the special case where the task is to reconstruct a one-dimensional stimulus using the information that it is drawn from a fixed Gaussian distribution. In this case we can precisely characterize behavior at every step of serial reproduction. Specifically, we show that this defines a simple first-order autoregressive, or AR(1), process, allowing us to draw on a variety of results characterizing such processes. We use these predictions to test Bayesian models of serial reproduction in a series of laboratory experiments, and show that the predictions hold for serial reproduction both between- and within-subjects.

The plan of the paper is as follows: we first lay out the Bayesian account of serial reproduction, starting with a rational analysis of reconstruction from memory. We then show how this Bayesian account corresponds to the AR(1) process in the case of simple Gaussian distributions, and how the AR(1) model can accommodate different assumptions about memory storage and reconstruction. In the main body of the paper, we use this model to motivate four experiments testing the prediction that serial reproduction reveals memory biases. Finally, we consider the implications of the results of these experiments in Section 8.

2. A Bayesian view of serial reproduction

We will outline our Bayesian approach to serial reproduction by first considering the problem of reconstruction from memory, and then asking what happens when the solution to this problem is repeated many times, as in serial reproduction.

2.1. Reconstruction from memory

Our goal is to give a rational account of reconstruction from memory, considering the underlying computational problem and finding the optimal solution to that problem. We will formulate the

problem of reconstruction from memory as a problem of inferring and storing accurate information about the world from noisy sensory data. Given a noisy stimulus x, we seek to recover the true state of the world μ that generated that stimulus, storing an estimate $\hat{\mu}$ in memory. The optimal solution to this problem is provided by Bayesian statistics. Previous experience provides a "prior" distribution on possible states of the world, $p(\mu)$. On observing x, this can be updated to a "posterior" distribution $p(\mu|x)$ by applying Bayes' rule

$$p(\mu|x) = \frac{p(x|\mu)p(\mu)}{\int p(x|\mu')p(\mu')d\mu'} \tag{1}$$

where $p(x|\mu)$ – the "likelihood" – indicates the probability of observing x if μ is the true state of the world. Having computed $p(\mu|x)$, a number of schemes could be used to select an estimate of $\hat{\mu}$ to store. Perhaps the simplest such scheme is sampling from the posterior, with $\hat{\mu} \sim p(\mu|x)$.

This analysis provides a general schema for modeling reconstruction from memory, applicable for any form of x and μ. A simple example is the special case where x and μ vary along a single continuous dimension. In the experiments presented later in the paper we take this dimension to be the width of a fish, showing people a fish and asking them to reconstruct its width from memory, but the dimension of interest could be any subjective quantity such as the perceived length, loudness, duration, or brightness of a stimulus. Assume that previous experience establishes that μ has a Gaussian distribution, with $\mu \sim N(\mu_0, \sigma_0^2)$, and that the noise process means that x has a Gaussian distribution with μ as its center, $x|\mu \sim N(\mu, \sigma_x^2)$. In this case, we can use standard results from Bayesian statistics (Gelman, Carlin, Stern, & Rubin, 1995) to show that the outcome of Eq. (1) is also a Gaussian distribution, with $p(\mu|x)$ being $N(\lambda x + (1-\lambda)\mu_0, \lambda\sigma_x^2)$, where $\lambda = 1/(1+\sigma_x^2/\sigma_0^2)$.

The analysis presented in the previous paragraph makes a clear prediction: that the reconstruction $\hat{\mu}$ should be a compromise between the observed value x and the mean of the prior μ_0, with the terms of the compromise being set by the ratio of the noise in the data σ_x^2 to the uncertainty in the prior σ_0^2. This model thus predicts a systematic bias in reconstruction that is not a consequence of an error of memory, but the optimal solution to the problem of extracting information from a noisy stimulus. The possibility that memory biases might actually be the consequence of a process that improves the accuracy of memory was pointed out by Huttenlocher et al. (2000), who presented a model extremely similar to that outlined in this section.

Huttenlocher et al. (2000) conducted several experiments testing this account of memory biases using simple one-dimensional stimuli such as fish that vary in width. In each experiment participants were shown stimuli sampled from a probability distribution such as a Gaussian (i.e. normal distribution) and asked to reconstruct those stimuli from memory. The results showed that people's reconstructions interpolated between the observed stimuli and the mean of the trained distribution as predicted. A similar notion of reconstruction as a weighted average of the mean of prior distribution and an observation was used by Hemmer and Steyvers (2008), who found that people formed appropriate Bayesian reconstructions for realistic stimuli such as images of fruit, and seemed capable of drawing on prior knowledge at multiple levels of abstraction in doing so. Finally, Stewart, Brown, and Chater (2005) showed that a similar kind of biased estimation appears in sequential retrieval from memory.

2.2. Serial reproduction

With a model of how people might approach the problem of reconstruction from memory in hand, we are now in a position to analyze what happens in serial reproduction, where the stimuli that people receive on one trial are the results of a previous reconstruction. On the nth trial, a participant sees a stimulus x_n. The participant then computes $p(\mu|x_n)$ as outlined in the previous section, and stores a sample $\hat{\mu}$ from this distribution in memory. When asked to produce a reconstruction, the participant generates a new value x_{n+1} from a distribution that depends on $\hat{\mu}$. If the likelihood, $p(x|\mu)$, reflects perceptual noise, then it is reasonable to assume that x_{n+1} will be sampled from this distribution, substituting $\hat{\mu}$ for μ. This value of x_{n+1} is the stimulus for the next trial.

Viewed from this perspective, serial reproduction defines a stochastic process: a sequence of random variables evolving over time. In particular, it is a Markov chain, since the reconstruction produced on the current trial depends only on the value produced on the preceding trial (e.g. Norris, 1997). The transition probabilities of this Markov chain are

$$p(x_{n+1}|x_n) = \int p(x_{n+1}|\mu)p(\mu|x_n)d\mu \tag{2}$$

being the probability that x_{n+1} is produced as a reconstruction for the stimulus x_n, which is also the reconstruction from previous trial. If this Markov chain is ergodic (see Norris (1997) for details) it will converge to a stationary distribution $\pi(x)$, with $p(x_n|x_1)$ tending to $\pi(x_n)$ as $n \to \infty$. That is, after many reproductions, we should expect the probability of seeing a particular stimulus being produced as a reproduction to stabilize to a fixed distribution. Identifying this distribution will help us understand the consequences of serial reproduction.

The transition probabilities given in Eq. (2) have a special form, being the result of sampling a value from the posterior distribution $p(\mu|x_n)$ and then sampling a value from the likelihood $p(x_{n+1}|\mu)$. In this case, it is possible to identify the stationary distribution of the Markov chain (Griffiths & Kalish, 2005, 2007). The stationary distribution of this Markov chain is the *prior predictive distribution*

$$p(x) = \int p(x|\mu)p(\mu)d\mu \tag{3}$$

being the probability of observing the stimulus x when μ is sampled from the prior. This happens because this Markov chain is a *Gibbs sampler* for the joint distribution on x and μ defined by multiplying $p(x|\mu)$ and $p(\mu)$ (Griffiths & Kalish, 2007). A Gibbs sampler is a Markov chain defined by alternating between sampling from the conditional distributions $p(x|\mu)$ and $p(\mu|x)$ for some joint distribution $p(x, \mu)$, which results in $p(x, \mu)$ as a stationary distribution. This gives a clear characterization of the consequences of serial reproduction: after many reproductions, the stimuli being produced will be sampled from the prior distribution assumed by the participants. Convergence to the prior predictive distribution provides a formal justification for the traditional claims that serial reproduction reveals cultural biases (e.g., Bartlett, 1932), since those biases would be reflected in the prior.

The convergence results given in the previous paragraph are completely general, applying to any kind of stimuli, hypotheses, and prior. In the special case of reconstruction of stimuli that vary along a single dimension, we can also analytically compute the probability density functions for the transition probabilities and stationary distribution. Applying Eq. (2) using the results summarized in the previous section, we have $x_{n+1}|x_n \sim N(\mu_n, (\sigma_x^2 + \sigma_n^2))$, where $\mu_n = \lambda x_n + (1-\lambda)\mu_0$, and $\sigma_n^2 = \lambda\sigma_x^2$. Likewise, Eq. (3) indicates that the stationary distribution is $N(\mu_0, (\sigma_x^2 + \sigma_0^2))$. The rate at which the Markov chain converges to the stationary distribution depends on the value of λ. When λ is close to 1, convergence is slow since μ_n is close to x_n. As λ gets closer to 0, μ_n is more influenced by μ_0 and convergence is faster. Since $\lambda = 1/(1 + \sigma_x^2/\sigma_0^2)$, the convergence rate thus depends on the ratio of the participant's perceptual noise and the variance of the prior distribution, σ_x^2/σ_0^2. More perceptual noise results in faster convergence, since the specific value of x_n is trusted less; while more uncertainty in the prior results in slower convergence, since x_n is given greater weight.

3. Serial reproduction of one-dimensional stimuli as autoregression

The special case of serial reproduction of one-dimensional stimuli can also give us further insight into the consequences of modifying our assumptions about storage and reconstruction from memory, by exploiting a further property of the underlying stochastic process: that it is a first-order autoregressive process, abbreviated to AR(1). The general form of an AR(1) process is

$$x_{n+1} = c + \phi x_n + \epsilon_{n+1} \tag{4}$$

where $\epsilon_{n+1} \sim N(0, \sigma_\epsilon^2)$. Eq. (4) has the familiar form of a regression equation, predicting one variable as a linear function of another, plus Gaussian noise. It defines a stochastic process because each variable is being predicted from that which precedes it in sequence. AR(1) models are widely used to model

timeseries data, being one of the simplest models for capturing temporal dependency (Box & Jenkins, 1994).

Just as showing that a stochastic process is a Markov chain provides information about its dynamics and asymptotic behavior, showing that it reduces to an AR(1) process provides access to a number of results characterizing the properties of these processes. If $\phi < 1$ the process has a stationary distribution that is Gaussian with mean $c/(1-\phi)$ and variance $\sigma_\epsilon^2/(1-\phi^2)$. The autocovariance at a lag of n iterations is $\phi^n\sigma_\epsilon^2/(1-\phi^2)$, and thus decays geometrically in ϕ. An AR(1) process thus converges to its stationary distribution at a rate determined by ϕ.

It is straightforward to show that the stochastic process defined by serial reproduction where a sample from the posterior distribution on μ is stored in memory and a new value x is sampled from the likelihood is an AR(1) process. Using the results in the previous section, at the $(n+1)$th iteration

$$x_{n+1} = (1-\lambda)\mu_0 + \lambda x_n + \epsilon_{n+1} \tag{5}$$

where $\lambda = 1/(1+\sigma_x^2/\sigma_0^2)$ and $\epsilon_{n+1} \sim N(0, (\sigma_x^2+\sigma_n^2))$ with $\sigma_n^2 = \lambda\sigma_x^2$. This is an AR(1) process with $c = (1-\lambda)\mu_0, \phi = \lambda$, and $\sigma_\epsilon^2 = \sigma_x^2 + \sigma_n^2$. Since λ is less than 1 for any σ_0^2 and σ_x^2, we can find the stationary distribution by substituting these values into the expressions given above. As described in the previous section, the value of λ determines the trade-off between the prior and the current piece of data, and thus the convergence rate of the Markov chain.

Identifying serial reproduction for single-dimensional stimuli as an AR(1) process allows us to relax our assumptions about the way that people are storing and reconstructing information. In the memorization phase, the participant's memory $\hat{\mu}$ can be (1) a sample from the posterior distribution $p(\mu|x_n)$, as assumed above, or (2) a value such that $\hat{\mu} = \text{argmax}_\mu p(\mu|x_n)$, being the posterior mode, which is also the mean of the posterior since the mode of a Gaussian is its mean. In the reproduction phase, the participant's reproduction x_{n+1} can be (1) a noisy reconstruction, which is a sample from the likelihood $p(x_{n+1}|\hat{\mu})$, as assumed above, or (2) a perfect reconstruction from memory, such that $x_{n+1} = \hat{\mu}$. This defines four different models of serial reproduction:

1. *Sample–sample (SS)*: Participants store a sample from posterior distribution of their noisy observation, and their reconstruction is also a sample from memory.
2. *Sample-perfect (SP)*: Participants store a sample from posterior distribution of their noisy observation, and they give a perfect reconstruction from memory.
3. *Maximize-sample (MS)*: Participants store the expected value of the posterior distribution of their noisy observation, and their reconstruction is also a sample from memory.
4. *Maximize-perfect (MP)*: Participants store the expected value of the posterior distribution of their noisy observation, and they give a perfect reconstruction from memory.

The fourth model is unlikely to work well for human behavior, because no variance is allowed in the system, so the chain will move to the mean of the prior in a strictly monotonic fashion and stay there forever. We thus consider only the first three models in the following analysis.

All three of these models can be shown to have the general form of the AR(1) process (Eq. (5)). What distinguishes between them is the noise terms: in the SS model $\epsilon_{n+1} \sim N(\mu_n, (\sigma_x^2+\sigma_n^2))$; in the SP model $\epsilon_{n+1} \sim N(\mu_n, \sigma_n^2)$, since there is no reproduction noise; and in the MS model, $\epsilon_{n+1} \sim N(\mu_n, \sigma_x^2)$ as no noise is introduced in memory storage. As shown above, the SS model converges to the prior predictive distribution $N(\mu_0, (\sigma_x^2+\sigma_0^2))$. Using the autoregression analysis, it is easy to show that the SP and MS models converge to $N\left(\mu_0, \frac{\lambda\sigma_x^2}{1-\lambda^2}\right)$ and $N\left(\mu_0, \frac{\sigma_x^2}{1-\lambda^2}\right)$ respectively. Thus, all three models converge to a distribution determined by the prior. Fig. 2 shows a simulation of the three types of models. The stationary distribution in the SS model has the largest variance and that of the SP model has the smallest variance when $\sigma_x^2 < \sigma_0^2$.

More generally, any model in which reconstructions have a mean value corresponding to $(1-\lambda)\mu_0 + \lambda x_n$ and storage and reconstruction are subject to Gaussian noise will reduce to an AR(1) process, with the only variation between models appearing in the variance of the noise term ϵ. This significantly increases the generality of our characterization of serial reproduction, but it also means

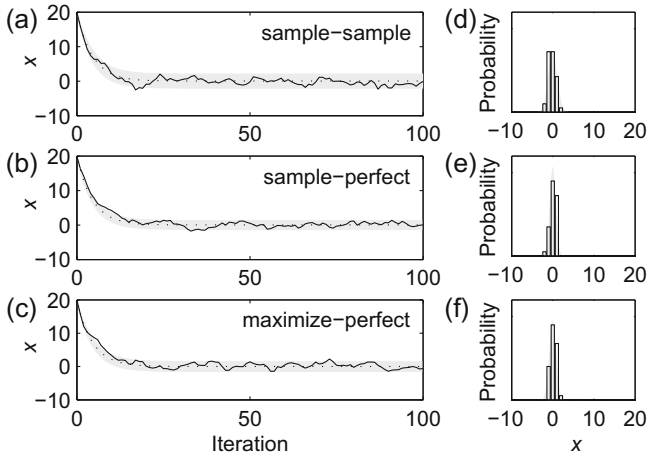

Fig. 2. Model simulations for three types of behavior. Panels (a)–(c) show simulated Markov chains with $\sigma_x = 0.5$ and $\sigma_0 = 1.0$. The solid line in each graph is a sequence of sampled values of x_n, and the dotted line and the gray area show the mean and 95% confidence interval on x_n. All samples and statistics are conditioned on $x_1 = 20$. Panels (d)–(f) show a histogram of the last 80 values of x_n for the Markov chains in Panels (a)–(c). The gray areas are the probability density functions for the stationary distributions of the three models.

that different models developed within this framework (including the three models introduced above) cannot be differentiated using empirical data. However, all of these models make the same basic prediction: that repeatedly reconstructing stimuli from memory will result in convergence to a distribution whose mean corresponds to the mean of the prior.

3.1. Testing the model predictions

In the remainder of the paper we describe four experiments testing the predictions produced by this model. These experiments all use a serial reproduction paradigm with stimuli that vary only along one dimension (the width of fish, following Huttenlocher et al. (2000)). By using these simple stimuli, we are able to provide the first carefully-controlled empirical analysis of the effects of serial reproduction, and test whether the results match the quantitative predictions produced by our model. Huttenlocher et al. (2000) established that people behave in a way that is consistent with the Bayesian analysis of reconstruction presented above through a series of experiments using these stimuli. Our experiments thus replicate and extend these results to the case of serial reproduction.

Experiment 1 follows previous research on serial reproduction in using a between-subjects design, with the reconstructions of one participant serving as the stimuli for the next. Participants were trained on the distribution of fish widths associated with a category, establishing a prior distribution for use in reconstruction. Experiment 2 uses a within-subjects design in which each person reconstructs stimuli that they themselves produced on a previous trial, testing the potential of this design to reveal the memory biases of individuals. Experiment 3 removes the training on the prior distribution, allowing serial reproduction to reveal prior expectations derived from general world knowledge instead of laboratory training. The results of the experiment reveal a general bias that was also reflected in the results of Experiments 1 and 2, helping to explain a trend observed in the previous experiments. In Experiment 4, we explore the consequences of serial reproduction with a more complex prior – a bimodal distribution – and examine the influence of categories and context on reconstruction.

4. Experiment 1: between-subjects serial reproduction

This experiment directly tested the basic prediction that the outcome of serial reproduction will reflect the prior knowledge that people have about the distribution of stimuli. The experiment

followed the same basic procedure as Bartlett's (1932) classic experiments, using the reconstruction task introduced by Huttenlocher et al. (2000). Two groups of participants were trained on different distributions of a one-dimensional quantity – the width of a schematic fish – that would serve as a prior for reconstructing similar stimuli from memory. The distributions learned by the two groups differed in their means, allowing us to examine whether the mean of the distribution produced by serial reproduction is affected by the prior, as predicted by our model.

4.1. Method

4.1.1. Participants

Forty-six undergraduates from the University of California, Berkeley participated in exchange for course credit.

4.1.2. Stimuli

Stimuli were the same as those used in Huttenlocher et al. (2000): fish with elliptical bodies and fan-shaped tails. All the fish stimuli varied only in one dimension, the width of the fish, ranging from 2.63 cm to 5.76 cm. The stimuli were presented on an Apple iMac computer by a Matlab script using PsychToolBox extensions (Brainard, 1997; Pelli, 1997).

4.1.3. Procedure

Participants received instructions that indicated that they would be working at a fish farm, and would receive some on-the-job training before beginning work. They were then trained to discriminate between two categories of fish: fish farm and ocean fish. The width of the fish-farm fish was normally distributed and that of the ocean fish was uniformly distributed between 2.63 and 5.75 cm. To make the training process easier, the instructions explained that fish-farm fish are fed on a special diet and are thus vary around a standard size, while ocean fish have to fend for themselves and have a far greater range of sizes. The critical manipulation was the parameters of the normal distribution, with two groups of participants being trained on distributions with different means but the same standard deviation. In condition A, $\mu_0 = 3.66$ cm, $\sigma_0 = 1.3$ cm; in condition B, $\mu_0 = 4.72$ cm, $\sigma_0 = 1.3$ cm.

In the training phase, participants first received a block of 60 trials. On each trial, a stimulus was presented at the center of a computer monitor and participants tried to predict which type of fish it was by pressing one of the keys on the keyboard and they received feedback about the correctness of the prediction. The participants were then tested for 20 trials on their knowledge of the two types of fish. The procedure was the same as the training block except there was no feedback. The training-testing loop was repeated until the participants reached 80% of optimal performance.[1] If a participant did not reach this criterion after five iterations, the experiment halted. All the participants passed the training phase.

In the reproduction phase, the participants were told that they were going to begin to work on recording fish sizes for the fish farm. On each trial, a fish stimulus was flashed at the center of the screen for 500 ms and then disappeared. Another fish of random size appeared at one of four possible positions near the center of screen and the participants used the up and down arrow keys to adjust the width of the fish until they thought it matched the fish they just saw. The fish widths seen by the first participant in each condition were 120 values uniformly spanning the range from 2.63 to 5.75 cm. The first participant reconstructed these stimuli from memory. Each subsequent participant in each condition was then presented with the reconstructions produced by the previous participant as stimuli,

[1] The optimal decision strategy used for evaluating performance was the Bayesian strategy under 0–1 loss (i.e. assuming that some reward is received for a correct answer, but no reward for an incorrect answer). This strategy corresponds to choosing the category with highest posterior probability for each stimulus. Since our normal (fish-farm fish) and uniform (ocean fish) distributions overlap, we calculated the upper and lower bounds where the posterior probability of the stimuli under the normal distribution is greater than under the uniform distribution. The optimal decision strategy is to classify the fish as ocean fish for sizes outside these boundaries and to classify those within these boundaries as fish-farm fish. Each participant's responses could then be scored for their consistency with this strategy.

and they again tried to reconstruct those fish widths. Thus, the data from each participant constitute one slice of time in 120 serial reproduction chains.

At the end of the experiment, the participants were given a final 50-trial test to check if their prior distributions had drifted. Since serial reproduction builds on the responses of previous participants, a single participant who is not engaged with the task can disrupt the entire experiment. Participants were thus excluded from the experiment if their data failed any of three conditions: (1) final testing score was less than 80% of optimal performance; (2) the difference between the reproduced value and stimulus shown was greater than the difference between the largest and the smallest stimuli in the training distribution on any trial; (3) there were no adjustments from the starting value of the fish width for more than half of the trials. If the current participant's data were rejected, the next participant would see the data generated by the previous participant. A total of 10 participants were excluded following these criteria.

4.2. Results and discussion

There were 18 participants in each condition, resulting in 18 generations of serial reproduction. Fig. 3 shows the initial and final distributions of the reconstructions, together with the plots for the 120 chains in the two conditions. The initial set of stimuli in both conditions A and B were drawn from the same uniform distribution. The mean reconstructed fish widths produced by the first participants in these conditions were 4.22 and 4.21 cm respectively, which were not statistically significantly different ($t(238) = 0.09,\ p = 0.93$). The histograms in the right panel show the final distributions of the reconstructions by the 18th participants in the two conditions. The mean reconstructed fish widths were 3.20 and 3.68 cm respectively, a statistically significant difference ($t(238) = 6.93,\ p < 0.001$). A two-way ANOVA also showed a significant interaction between the starting and ending points of the 120 chains in the two conditions ($F(1, 236) = 12.04,\ p < 0.001$). The difference in means matches the direction of the difference in the training provided in conditions A and B, although the overall size of the difference is reduced and the means of the stationary distributions were lower than those of the distributions used in training.

Fig. 4 shows the autoregression plots and the biases in reconstruction. The autoregression plots compare x_{n+1} with x_n, and the autoregression model outlined above predicts that the resulting distribution should be jointly Gaussian with the mean of x_{n+1} being a linear function of x_n. This is exactly what we see in Fig. 4. The correlation between the stimulus x_n and its reconstruction x_{n+1} is the correlation between the AR(1) model's predictions and the data, and this correlation was high in both

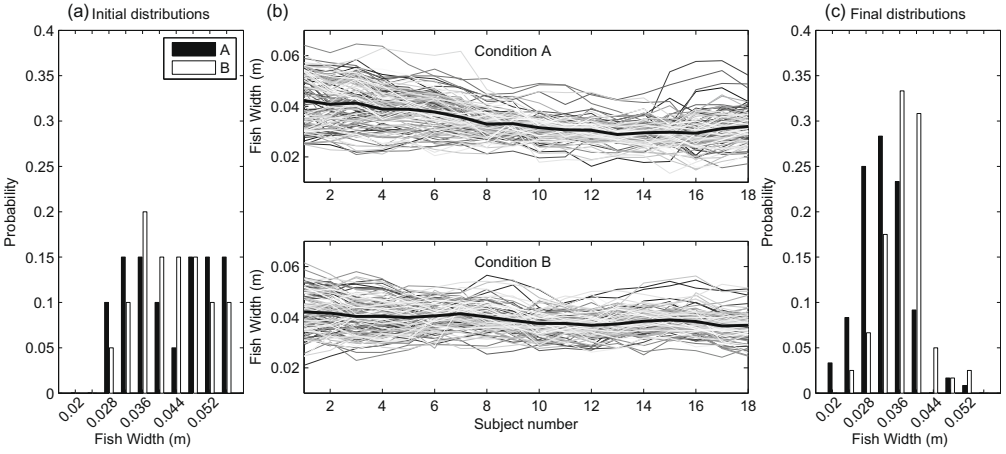

Fig. 3. Initial and final distributions, and the chains for the two conditions in Experiment 1. (a) The histograms for the initial distributions of stimuli for conditions A and B. (b) The chains of serial reproduction, starting from the first participant's data, ending with the 18th participant's data. (c) The histograms for the ending distributions of stimuli for conditions A and B.

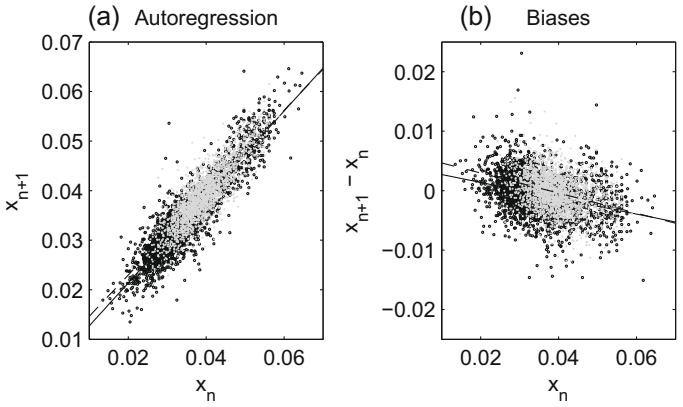

Fig. 4. Autoregression and bias plots for the two conditions in Experiment 1. (a) The autoregression plots for conditions A (black) and B (gray). (b) The bias plots for conditions A (black) and B (gray).

conditions, being 0.91 and 0.86 (both $p < 0.001$) for conditions A and B respectively. Biases in reconstruction can be identified by evaluating the difference between x_{n+1} and x_n as a function of x_n. This was the basic dependent measure used by Huttenlocher et al. (2000). Consistent with their results, the slope of the function relating bias and x_n is negative (-0.34 for both conditions A and B, $p < 0.001$), confirming the Bayesian model's prediction that memory of the stimuli are biased towards the mean of the category distribution.

Finally, we examined whether the Markov assumption underlying our analysis was valid, by computing the correlation between x_{n+1} and x_{n-1} given x_n. The resulting partial correlation was low for both conditions, being 0.04 and 0.01 in conditions A and B respectively (both $p > 0.05$). This is to be expected, since the use of a different participant at each step of reproduction ensures that the Markov assumption should hold, but allows us to rule out any higher-order temporal effects on reconstruction.

The results of the experiment provide support for the basic prediction produced by our analysis of serial reproduction: chains formed of individuals trained on different prior distributions converged to different stationary distributions, and those stationary distributions differed in a way that reflected the difference in the priors. However, the stationary distributions did not correspond exactly to the prior distributions on which people were trained – a point we will return to in Experiment 3, after examining whether we obtain similar results when creating serial reproduction chains within-subjects.

5. Experiment 2: within-subjects serial reproduction

The between-subjects design allows us to reproduce the process of information transmission, but our analysis suggests that serial reproduction might also have promise as a method for investigating the memory biases of individuals. To explore the potential of this method, we tested the model with a within-subjects design, in which a participant's reproduction in the current trial became the stimulus for that same participant in a later trial. Each participant's responses over the entire experiment thus produced a chain of reproductions. Each participant produced three such chains, starting from widely separated initial values. Control trials and careful instructions were used so that the participants would not realize that some of the stimuli were their own reproductions.

5.1. Method

5.1.1. Participants

Forty-six undergraduates from the University of California, Berkeley participated in the experiment in exchange for course credit.

5.1.2. Stimuli

The stimuli used in this experiment were the same as those used in Experiment 1.

5.1.3. Procedure

The basic procedure was the same as Experiment 1, except in the reproduction phase. Each participant's responses in this phase formed three chains of 40 trials. The chains started with three original stimuli with width values of 2.63 cm, 4.19 cm, and 5.76 cm, then in the following trials, the stimuli participants saw were their own reproductions in the previous trials in the same chain. To prevent participants from realizing this fact, chain order was randomized and the Markov chain trials were intermixed with 40 control trials in which widths were drawn from the prior distribution.

5.2. Results and discussion

Participants' data were excluded based on the same criteria as used in Experiment 1, with a lower testing score of 70% of optimal performance and one additional criterion relevant to the within-subjects case: participants were also excluded if the three chains did not converge, with the criterion for convergence being that the lower and upper chains must cross the middle chain. After these screening procedures, data from 40 of the 46 participants were included, with 21 in condition A and 19 in condition B. It took most participants about 20 trials for the chains to converge, so only the second half of the chains (trials 21–40) were analyzed further.

The locations of the stationary distributions were measured by computing the means of the reproduced fish widths for each participant. For conditions A (3.66 cm) and B (4.72 cm), the average of these means was 3.32 and 4.01 cm respectively ($t(38) = 2.41,\ p = 0.021$). The right panel of Fig. 5 shows the mean values for these two conditions. The basic prediction of the model was borne out: participants converged to distributions that differed significantly in their means when they were exposed to data suggesting a different prior. However, the means were in general lower than those of the prior. This effect was less prominent in the control trials, which produced means of 3.63 and 4.53 cm respectively.

Fig. 6 shows the chains, training distributions, the Gaussian fits and the autoregression plots for the second half of the Markov chains for two participants in the two conditions. Fig. 7 shows the autoregression and bias plots for all participants. As in Experiment 1, the autoregression plots show a strong linear relationship between the stimulus x_n and the reconstruction x_{n+1}. Again, the correlation was high in both conditions, with mean values being 0.90 and 0.81 for conditions A and B respectively. The correlations are significant ($p < 0.001$) for all participants except for one in each condition, indicating that the AR(1) model's predictions are highly correlated with the data generated by each participant. The mean partial correlation between x_{n+1} and x_{n-1} given x_n was low, being 0.07 and 0.11 for

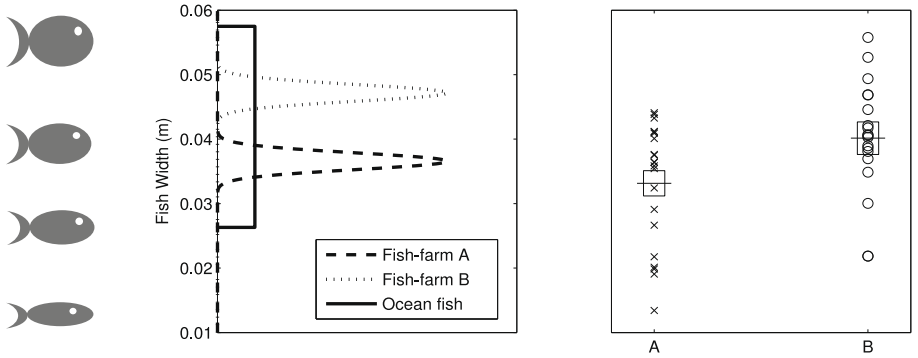

Fig. 5. Stimuli, training distributions and stationary distributions for Experiment 2. Each data point in the right panel shows the mean of the last 20 iterations for a single participant. Boxes show the 95% confidence interval around the mean for each condition.

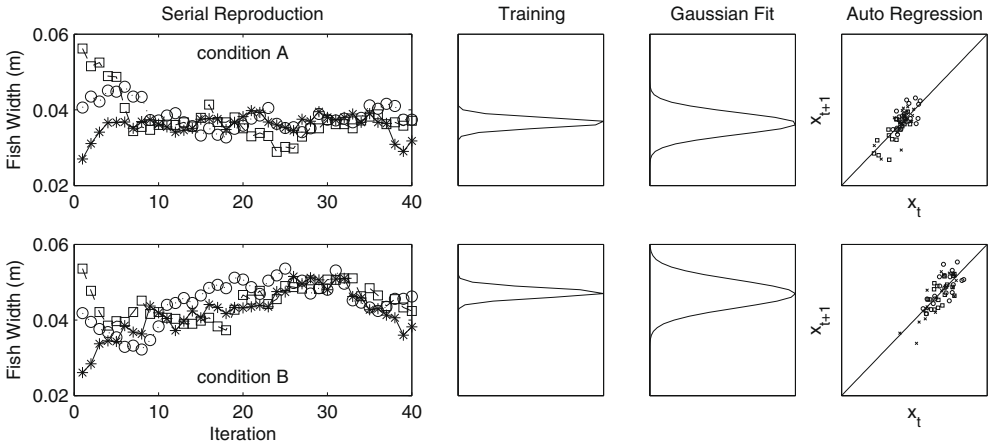

Fig. 6. Chains and stationary distributions for individual participants from the two conditions in Experiment 2. (a) The three Markov chains generated by each participant, starting from three different values. (b) Training distributions for each condition. (c) Gaussian fits for the last 20 iterations of each participant's data. (d) Autoregression for the last 20 iterations of each participant's data.

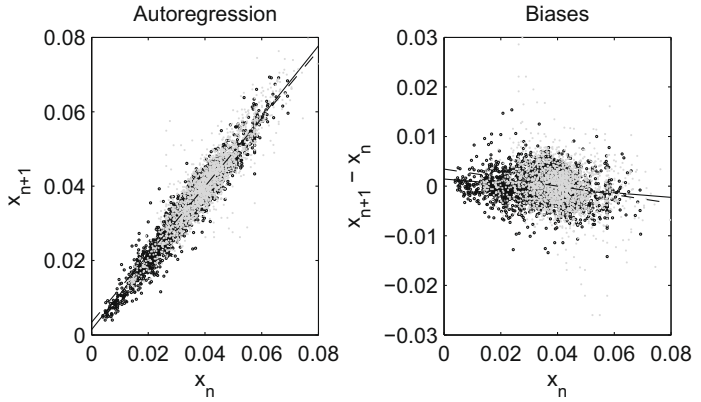

Fig. 7. Autoregression and bias plots for the two conditions in Experiment 2. (a) The autoregression plots for conditions A (black) and B (gray). (b) The bias plots for conditions A (black) and B (gray).

conditions A and B respectively, suggesting that the Markov assumption was satisfied. The partial correlation was significant ($p < 0.05$) for only one participant in condition B. Similar to Experiment 1, the effect of bias was shown in significant ($p < 0.05$) negative slopes in the bias plot, that is, negative correlations between x_n and $(x_{n+1} - x_n)$. This was true for all participants except for two in condition A and one in condition B.

The results of this experiment corresponded well with those of Experiment 1, showing that serial reproduction has similar consequences whether it takes place between-subjects or within-subjects. This correspondence is consistent with our analysis, in which the only factor that is relevant to the convergence of serial reproduction is the repeated reconstruction of stimuli, regardless of whether those reconstructions come from multiple people or just one person. These results suggest that serial reproduction can be used not just for exploring cultural biases, but for investigating the memory biases of individuals. While we examined only a simple form of bias in this experiment – the bias towards the mean of a trained category – we should expect that memory biases that appear small on a

single trial can be magnified through the process of serial reproduction, providing a useful lens for exploring the general nature of these biases as well as individual differences.

The correspondence between the results of Experiments 1 and 2 shows up not just in the way in which human behavior conforms to our model predictions, but also in the way in which it deviates. While reconstructions tended towards distributions affected by the prior established through training in both experiments, reflected in the difference in the means of the stationary distributions produced in the two conditions, the actual means of the stationary distributions differed systematically from the means of the distributions used in training. Both experiments produced stationary distributions with means lower than those of the training distributions. This phenomenon suggests that there could be another source of prior other than the ones participants were trained on, a possibility that we explore in Experiment 3.

6. Experiment 3: serial reproduction with no training on priors

Experiments 1 and 2 tried to establish different prior distributions for reconstruction from memory by providing training on different categories of fish. One possible explanation for why the serial reproduction chains did not converge to stationary distributions that matched these training distributions is that the training might have been insufficient to establish strong prior beliefs. In particular, people may have combined existing expectations about the width of fish with the information provided by the training data when they formed their estimates of the distributions associated with the different categories.

Hemmer and Steyvers (2008) explored how prior knowledge influences reconstruction from memory, and provided a detailed exploration of how such priors might be defined at multiple levels, such as category and object levels. They conducted experiments in which participants were asked to reconstruct the size of familiar objects such as fruits and vegetables. The results showed that reconstructions of objects (e.g., apples) were biased towards the mean size of the superordinate category (e.g., fruit), and at the same time, the size of a smaller version of an object was overestimated at reconstruction while the size of a larger version of the same object was underestimated, reflecting biases at the levels of objects themselves. To explain this effect, they extended Huttenlocher et al.'s (2000) model and proposed a hierarchical Bayesian account of these effects, defining priors at multiple levels.

This analysis suggests a simple account of the systematic differences between training distributions and the stationary distribution of serial reproduction seen in our experiments: people may already have a prior for the width of fish defined at the superordinate level, and are guided by these expectations when learning about fish-farm fish in our experiments. We can explore this possibility by using the serial reproduction method to investigate the priors that people use when they receive no training. We thus conducted an experiment in which participants performed exactly the same serial reproduction task as in Experiment 2, but they were not given the training phase. Since there is no training, our Bayesian model predicts that this will converge to a distribution that reflects people's general knowledge about the width of fish. If the biases seen in Experiments 1 and 2 are a result of these superordinate-level expectations, we should expect the mean of this distribution to be smaller than that of the training distributions used in the two experiments, allowing it to exert an additional influence on reconstructions that leads to a stationary distribution with a lower mean even when training is provided.

6.1. Method

6.1.1. Participants

Thirty-four undergraduates from the University of California, Berkeley, participated in the experiment in exchange for course credit.

6.1.2. Stimuli

The stimuli used in this experiment were the same as those used in Experiments 1 and 2.

6.1.3. Procedure

The basic procedure was the same as Experiment 2, except that there were no training and testing phases.

6.2. Results and discussion

Participants' data were excluded based on the second, third, and fourth criteria used in Experiment 2. There were no testing scores since no training was given. This resulted in the data from 28 of the 34 participants being subjected to further analyses. As in Experiment 2, only the final 20 responses produced by each individual were analyzed, as it took approximately 20 trials for the chains to converge, and only those data with chains converged were analyzed. The mean width of the fish produced in these trials was 3.54 cm, significantly less than the mean of the initial values of each chain, 4.19 cm ($t(27) = 4.33,\ p < 0.001$). To preclude the possible explanation of these lower values where the chains end up as simply a downward bias, we also analyzed the difference scores of starting and ending values of the upper and lower chains, computed as the mean of the last 20 trials. The mean values of these two difference scores were 2.21 cm and −0.91 cm for the upper and lower chains, respectively. All the chains with the lower starting point (2.63 cm) moved to higher values, except for two participants. The two-sample *t*-test also showed significant difference between these scores ($t(54) = 14.62,\ p < 0.001$). These results indicate that people seem to have an *a priori* expectation that fish will have widths smaller than those used as our category means, suggesting that the deviations from the training distributions observed in Experiments 1 and 2 are the consequence of using a prior that is a compromise between this superordinate-level expectation about the width of fish and the training data.

7. Experiment 4: serial reproduction with bimodal distributions

In Experiments 1 and 2, we trained people on simple Gaussian distributions to show that serial reproduction converges to a distribution that reflects memory biases consistent with Gaussian priors. However, our original analysis of serial reproduction made no assumptions about the form of the priors, indicating that convergence to the prior should be expected in all cases under our assumptions about the process of storage and reconstruction. To test if we observe similar results with people for priors beyond simple Gaussian distributions, we conducted another experiment in which participants were trained on categories with bimodal distributions, that is, they were given bimodal priors. This experiment allows us to determine whether serial reproduction just converges to a Gaussian distribution independent of the prior – a reasonable alternative hypothesis – or is sensitive to the form of the prior distribution in a way that produces bimodal stationary distributions.

Since learning a single category distribution that is multimodal could be challenging (e.g., McKinley & Nosofsky, 1995), we trained people on two unimodal distributions corresponding to the width of fish of different species, where species was unambiguously indicated through the color of the fish (red or blue). The width of fish in each species followed a Gaussian distribution, and these distributions were selected so that the overall distribution of widths was bimodal. We then asked people to produce reconstructions of fish that were glimpsed briefly in "darkness," where there was insufficient light to see the color of the fish. Under these circumstances, reconstructions should be made using the overall distribution of widths, providing a bimodal prior. We called this the *no color* condition.

Examining serial reproduction for bimodal priors also allows us to investigate an issue that has arisen in work following up on Huttenlocher et al.'s (2000) original Bayesian analysis of reconstruction from memory. Sailor and Antoine (2005) conducted experiments of reconstruction from memory in which people were simultaneously trained on two distinct category distributions, and found that even when people were reconstructing two distinct categories they tended to produce reproductions biased toward the overall mean of the stimuli rather than the means of the individual categories. They suggested that this showed an effect of experimental context, which determines the overall range of the stimuli, rather than an effect of categories on reproduction.

We explored whether people produce reconstructions based on experimental context or on category information by adding a second condition to our experiment, in which people reconstructed the width of the fish as before, but now the colors of the fish were visible. We called this the *color* condition. If people use a prior appropriate to the category indicated by the color of the fish, we should expect the serial reproduction chains for fish of different colors to converge to different distributions, reflecting the training distributions for those categories. The overall stationary distribution should also be equivalent to the stationary distribution produced in the *no color* condition. If people simply produce reconstructions using a single distribution derived from the experimental context, we should expect no difference between these chains, since they should converge to the same stationary distribution regardless of category. The stationary distribution should thus be different from that observed in the *no color* condition.

7.1. Method

7.1.1. Participants

Eighty-five undergraduates from the University of California, Berkeley, participated in the experiment in exchange for course credit.

7.1.2. Stimuli

The stimuli used in this experiment were the same as those used in the previous three experiments, except that the training distributions were different. The widths of the two types of fish (red and blue) were normally distributed with $\mu_1 = 3.66$ cm, and $\mu_2 = 4.72$ cm, and $\sigma_1 = \sigma_2 = 0.13$ cm.

7.1.3. Procedure

The basic procedure was the same as Experiment 2, consisting of three phases: training, reproduction, and testing.

The training phase was a categorization task as in Experiments 1 and 2. Participants were taught to discriminate two types of fish-farm fish, the red fish and the blue fish. Participants saw fish in "darkness" (grey fish), and guessed the color (blue or red). They then received feedback about the color of the fish. The number of trials and criterion for success was the same as that of the previous experiments.

The reproduction phase consisted of four chains of 40 trials, starting from 2.63 and 5.75 cm. In the *no color* condition, all the fish were shown in grey. In the *color* condition, two chains were presented in red and another two other chains were in blue. As with the previous experiments, the chains were randomized and mixed with 40 control trials.

The test phase was the same as Experiments 1 and 2, in which participants judge the category of fish stimuli as in the training phase, but no feedback was given.

7.2. Results and discussion

Participants were excluded based on the same criteria used in Experiment 2. The data from 70 of the 85 participants passed these criteria and were analyzed further, with 35 in each condition (*no color* and *color*). Again, only the second half of each chain was analyzed.

Fig. 8 shows the chains produced by two participants in the two conditions, as well as the Gaussian fits of the training data and last 20 trials of the reproduction chains. To test whether people had converged to unimodal or bimodal distributions, we fit each individual's data using a single Gaussian and a mixture of Gaussians with two components. Fitting was done by maximum-likelihood estimation. For the *no color* condition, the parameters of the mixture of Gaussians were estimated using the Expectation-Maximization algorithm (Dempster, Laird, & Rubin, 1977), since the assignment of individual data points to Gaussian components was a latent variable. For the *color* condition, separate Gaussians were estimated from the responses in the red and blue chains independently, with the result being a mixture of the two distributions.

A better fit by the mixture of Gaussians than the single Gaussian would provide evidence in favor of the prediction that serial reproduction converges to a distribution reflecting the bimodal prior in the *no*

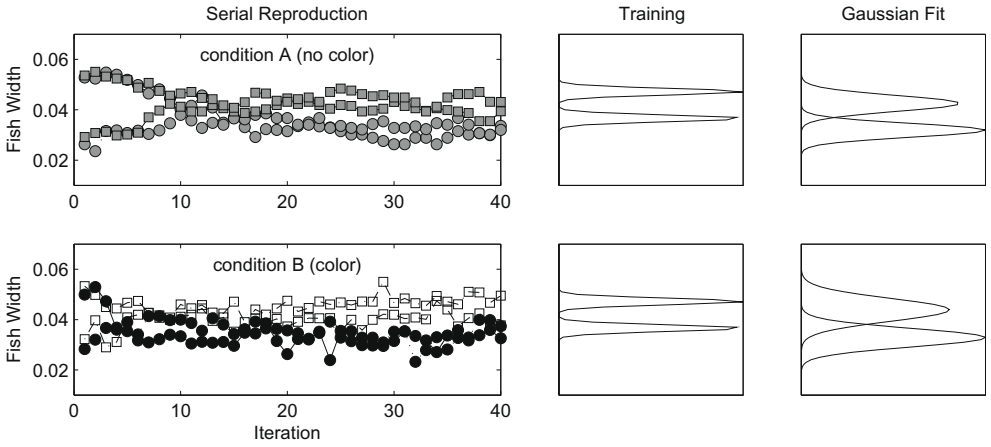

Fig. 8. Chains and stationary distributions for individual participants from the two conditions in Experiment 4. (a) The four Markov chains generated by each participant, starting from four different values. In condition A, black and white plot markers are the red and blue fish chains respectively, and in condition B, all the chains are gray fish. (b) Training distributions for each condition, black and light-gray plots being the red and blue fish respectively. (c) Gaussian fits for the last 20 iterations of each participant's data. Black and light-gray plots in condition A are the red and blue fish chains, and both plots in condition B are the gray chains.

color condition, and that it converges to different distributions for different categories (rather than a single distribution guided by context) in the *color* condition. We compared the fit of the two models in each condition using likelihood-ratio tests. Since the single Gaussian model is a special case of the mixture of Gaussians, twice the difference between the log-likelihoods of the two models should follow the χ^2 distribution with degrees of freedom equal to the difference in the number of parameters of the models under the assumption the simpler single Gaussian model is true (Rice, 1995). To confirm the the results of the likelihood-ratio tests, we also computed the Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978) values as model selection measures.

In the *color* condition, likelihood-ratio tests showed that the data of 34 out of 35 participants were significantly better fit by the two Gaussian model (all $p < 0.05$). Both AIC and BIC values indicated exactly the same result. The average values of the means of the two Gaussians for those 34 participants are 3.10 and 3.91 cm, and a two-samples t-test showed that these means are significantly different ($t(66) = 5.48,\ p < 0.001$). These results indicate that for the majority of participants in the "color" condition, where serial reproduction was done with the context of the color of each category present, the chains converged to bimodal distributions reflecting the priors established through training.

In the *no color* condition, one participant's data produced degenerate results for maximum-likelihood estimation and were omitted from further analysis. Likelihood-ratio tests showed that the data of 18 out of the remaining 34 participants were significantly better fit by the two Gaussian model (all $p < 0.05$). The average values of the means of the two Gaussian components for those 18 participants' data were 3.00 and 4.08 cm, and a two-samples t-test showed that these means are significantly different ($t(34) = 4.67,\ p < 0.001$). Thus for about two thirds of the participants serial reproduction converged to a bimodal distribution. The AIC and BIC values showed converging results: 22 and 15 out 34 participants' data were better fit by the two Gaussian model using the AIC and BIC criteria respectively, consistent with the greater conservatism that the BIC displays towards more complex models.

These results support two conclusions. First, the finding that at least some of our participants produced bimodal stationary distributions in the *no color* condition indicates that serial reproduction is sensitive to the form of the prior and not just its mean and variance. This complements our observation of convergence to the prior in previous experiments, showing that this property of serial reproduction generalizes beyond simple Gaussian priors. We do not view the fact that between a half and two thirds of participants (depending on model selection measure) produced bimodal

distributions as a major problem, since our primary goal was to provide an existence proof for convergence to other distributions, and a variety of factors including overestimating the variance of the categories and the relatively small number of trials in each chain could have resulted in the stationary distributions appearing unimodal. Second, the overwhelming tendency for people to converge to stationary distributions best characterized by two Gaussians in the *color* condition indicates that their reconstructions were guided by the category of the stimulus rather than general experimental context, contrary to the claims of Sailor and Antoine (2005). This conclusion is further supported by the production of bimodal stationary distributions even in the *no color* condition, where the absence of category cues would presumably strengthen the reliance on general experimental context.

8. General discussion

We have presented a Bayesian analysis of serial reproduction, providing both general predictions about the outcome of this process for arbitrary stimuli and specific predictions for the special case of one-dimensional stimuli with Gaussian priors. The results of our four experiments confirm the predictions produced by this analysis. Experiments 1 and 2 showed, in both within-subject and between-subject cases, that serial reproduction using one-dimensional stimuli with Gaussian priors converged to a distribution consistent with the prior on which participants were trained. Using serial reproduction without training, Experiment 3 revealed people's general expectations about the size of fish were consistent with a systematic bias observed in Experiments 1 and 2. In Experiment 4, we tested whether the predictions of our Bayesian analysis held beyond simple Gaussian distributions by establishing bimodal distributions as priors. The results confirmed that serial reproduction converged to bimodal distributions reflecting those priors, and also tested the hypothesis that reconstruction is biased by the context of the stimuli. The results of this experiment showed that category structure and not general experimental context seemed to be guiding people's reconstructions. Those results also demonstrate that serial reproduction may be an effective and more sensitive way of tapping people's prior knowledge of the world.

The theoretical analysis and experimental results we have presented in this paper make contact with an experimental literature on reconstruction from memory and memory biases revealed through serial reproduction that goes back to the 1930s. However, this phenomenon has not been formally analyzed before and most previous research used complex stimuli that are hard to control and difficult to interpret. To our knowledge, ours is the first detailed mathematical analysis of serial reproduction, and the first formal confirmation of Bartlett's (1932) conclusion that the outcome of serial reproduction reflects people's biases. Our use of simple one-dimensional stimuli allowed us to define an even more precise model based on first-order autoregressive processes, and provided a way to develop a well-controlled experimental method that could be used in a quantitative test of the predictions of our model. However, the Markov chain analysis also generalizes to any kind of stimuli, hypotheses, and prior distribution, opening a lot of opportunities for further exploration of the relationship between memory biases and serial reproduction. In the remainder of the paper we highlight some connections to other research and consider the limitations and possible future directions of this work.

8.1. Connections to other research

The work we have presented here has connections to two other lines of research exploring how ideas from Bayesian statistics can be used to understand human cognition: models of reconstruction from memory, and iterated learning. We will briefly summarize these two sets of connections in turn.

8.1.1. Reconstruction from memory

As discussed above, previous papers have proposed a Bayesian analysis of reconstruction from memory. Huttenlocher et al. (2000) proposed that reconstructions should be a compromise between the observed value and the mean of a category in order to minimize reconstruction error. The resulting model is equivalent to that obtained by treating the problem as one of Bayesian inference with a Gaussian prior. Hemmer and Steyvers (2008) took an explicitly Bayesian perspective on this problem

and extended Huttenlocher et al.'s analysis to other priors, such as hierarchical priors defined at the level of both individual objects and categories. They explored the memory biases shown with real categories using naturalistic stimuli, showing that people exhibit relatively strong biases in a memory task using these categories. Stewart et al. (2005) explored a similar model in the context of sequential effects on memory recall.

Our primary theoretical contribution in this paper is an analysis of the predictions that this Bayesian account of reconstruction from memory makes about serial reproduction. This analysis extends the scope of the phenomena to which Bayesian models of reconstruction have been applied, but is otherwise consistent with the work of Huttenlocher et al. (2000), Hemmer and Steyvers (2008), and Stewart et al. (2005). The fact that this account produces predictions that are consistent with the conclusions of Bartlett (1932) and with our own experiments when applied to serial reproduction provides further support for its utility as a model of reconstruction from memory.

8.1.2. Iterated learning

The key step in proving that serial reproduction converges to distribution determined by the prior was noting that alternating between sampling μ from the posterior distribution $p(\mu|x)$ and x from the likelihood $p(x|\mu)$ defined a Markov chain with stationary distribution $p(x,\mu) = p(x|\mu)p(\mu)$. While this is a Markov chain of a type commonly used in Bayesian statistics, Griffiths and Kalish (2007) observed that such a process could provide a natural model of the cultural transmission of information. In particular, they showed that the process of *iterated learning*, in which a sequence of people each learns from data generated by the previous person and then generates the data provided to the next person, could be analyzed as a Markov chain of exactly this kind.

Griffiths and Kalish (2005, 2007) analyzed iterated learning for a sequence of Bayesian learners, each of whom forms a hypothesis based on the data generated by the previous learner. If learners sample hypotheses h from the posterior distribution $p(h|d)$ and data d from the likelihood $p(d|h)$, the result is a Markov chain that converges to $p(d,h) = p(d|h)p(h)$. As a consequence, the probability that a learner selects a particular hypothesis h on a given iteration converges to the prior probability of that hypothesis, $p(h)$.

While iterated learning was originally proposed as a way to model language evolution (Kirby, 2001), the prediction of convergence to the prior is interesting in the context of cultural evolution more generally, since learning is one of the ways in which information is transmitted between people. It also suggests that we might be able to identify the biases that guide human learning by reproducing the process of iterated learning in the laboratory. This basic prediction has been confirmed through experiments with human participants showing that iterated learning of functions (Kalish, Griffiths, & Lewandowsky, 2007) and categories (Griffiths, Christian, & Kalish, 2008) results in an increase in the prevalence of concepts that are easy to learn (i.e. those that have high prior probability).

Iterated learning and serial reproduction have a basic structural correspondence, both being concerned with the transmission of information along a chain of individuals. Both are thus instances of a paradigm known as a "diffusion chain" in the broader anthropological literature (for a review, see Mesoudi (2007)). The key difference between the two is the mechanism of transmission – the kind of cognitive process involved. In iterated learning this mechanism is learning, while in serial reproduction it is memory. The connection between the two paradigms that we draw on here results from treating both learning and memory as inductive problems that can be solved via Bayesian inference.

The results we present in this paper thus complement the work on iterated learning mentioned above, showing how similar theoretical analyses and empirical findings hold for transmission of information via reconstruction from memory. This broadens the scope of cultural transmission phenomena we might hope to explain, as well as the range of psychological biases we have the opportunity to investigate. It also provides a link to an empirical literature within psychology that goes back over 70 years, and a way to validate Bartlett's (1932) original claims about the effects of serial reproduction.

8.2. Limitations and future directions

While our theoretical results apply for arbitrary stimuli and prior distributions, the experiments we presented in this paper used only one-dimensional stimuli and prior distributions that can be

expressed as mixtures of Gaussians. Our choice to use these stimuli and priors was motivated by a desire for simple, well-controlled experiments about which we could make clear quantitative predictions and the existing empirical literature based on similar assumptions (Huttenlocher et al., 2000; Hemmer & Steyvers, 2008; Stewart et al., 2005). However, an important direction for future research will be examining how well our theoretical results hold for other stimuli and kinds of memory biases.

One reason to explore serial reproduction with other stimuli is to provide further test of the predictions produced by our Bayesian analysis. To do so, we would ideally use stimuli for which memory biases are already well established. For example, Feldman (2000) used a reconstruction task to investigate biases for boolean concepts, building on the work of Shepard, Hovland, and Jenkins (1961). In this task, people were shown a division of a set of objects varying along binary dimensions into two groups, and then asked to reconstruct the division from memory. The ease of reconstruction varied with the complexity of the rule that described the division. We should expect the same memory biases to manifest in serial reproduction, with the lower-complexity rules being more likely to survive the process. Understanding how concepts change when passed from one person to another is particularly interesting in light of recent work exploring the stability of different kinds of religious concepts under cultural transmission (Barrett & Nyhof, 2001; Boyer & Ramble, 2001).

Another reason to conduct experiments with other stimuli is that our results justify using serial reproduction as a method for investigating memory biases. Since these biases have an effect each time people reconstruct a stimulus from memory, Serial reproduction can magnify what might be small effects in the context of a standard memory task. Controlled experiments in serial reproduction might thus be a valuable tool for exploring memory biases in a variety of domains. While this method has been used heuristically in the past, our results provide it with a more rigorous justification, as well as examples showing that the method works in both between- and within-subjects designs.

8.3. Conclusion

We have presented a Bayesian account of serial reproduction, and tested the basic predictions of this account using four controlled laboratory experiments. The results of these experiments are consistent with the predictions of our account, with serial reproduction converging to a distribution that is influenced by the prior distribution established through training. Our analysis connects the biases revealed by serial reproduction with the more general Bayesian strategy of combining prior knowledge with noisy data to achieve higher accuracy. It also shows that serial reproduction can be analyzed using Markov chains and first-order autoregressive models, providing the opportunity to draw on a rich body of work on the dynamics and asymptotic behavior of such processes. These connections allow us to provide a formal justification for the idea that serial reproduction changes the information being transmitted in a way that reflects the biases of the people transmitting it, establishing that this result holds under several different characterizations of the processes involved in storage and reconstruction from memory.

Acknowledgments

This work was supported by Grant Number 0704034 from the National Science Foundation. A preliminary version of Experiments 1 and 2 was presented at the 29th Annual Conference of the Cognitive Science Society (Xu & Griffiths, 2007) and the 2008 Neural Information Processing Systems conference (Xu & Griffiths, 2009).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.
Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
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: The role of intuitive conceptual structures in memory and transmission of cultural materials. *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., & Jenkins, G. M. (1994). *Time series analysis: Forecasting and control*. Prentice Hall: Upper Saddle River, NJ.
Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts: Cross-cultural evidence for recall of counter-intuitive representations. *Cognitive Science, 25*, 535–564.
Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39.
Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630–633.
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman & Hall.
Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for revealing inductive biases. *Cognitive Science, 32*, 10–68.
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* (pp. 827–832). Mahwah, NJ: Erlbaum.
Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science, 31*, 441–448.
Hemmer, P., & Steyvers, M. (2008). *A Bayesian account of reconstructive memory*. In Proceedings of the 30th annual conference of the Cognitive Science Society.
Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, 220–241.
Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*.
Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin, 26*, 594–604.
Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *Journal of Evolutionary Computation, 5*, 102–110.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431–461.
McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 128–148.
Mesoudi, A. (2007). Using the methods of experimental social psychology to study cultural evolution. *Journal of Social, Evolutionary, and Cultural Psychology, 1*, 35–38.
Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society, 363*, 3489–3501.
Norris, J. R. (1997). *Markov chains*. Cambridge, UK: Cambridge University Press.
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.
Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury.
Sailor, K. M., & Antoine, M. (2005). Is memory for stimulus magnitude Bayesian? *Memory and Cognition, 33*, 840–851.
Schacter, D. L., Coyle, J. T., Fischbach, G. D., Mesulam, M. M., & Sullivan, L. E. (Eds.). (1995). *Memory distortion: How minds, brains, and societies reconstruct the past*. Cambridge, MA: Harvard University Press.
Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464.
Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs, 13*(517), 75.
Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*, 881–911.
Whiten, A., & Mesoudi, A. (2008). Establishing an experimental science of culture: Animal social diffusion experiments. *Philosophical Transactions of the Royal Society, 363*, 3477–3488.
Xu, J., & Griffiths, T. L. (2007). A Bayesian analysis of serial reproduction. *Poster presented at the 29th meeting of the Cognitive Science Society, Nashville, TN*.
Xu, J., & Griffiths, T. L. (2009). How memory biases affect information transmission: A rational analysis of serial reproduction. *Advances in Neural Information Processing Systems*, 21.