When extremists win
Iterated learning with heterogenous agents

Dani Navarro
School of Psychology
University of New South Wales

Amy Perfors
School of Psychological Science
University of Melbourne

Arthur Kary
School of Psychology
University of New South Wales

Scott Brown
School of Psychology
University of Newcastle

Chris Donkin
School of Psychology
University of New South Wales
Cultural evolution
Cultural evolution
Cultural evolution

Music Timeline

Variants of 'Rachel' among U.S. baby names, 1880-2012

Percent of registered female births (grade: 1, 2, 3%)

Liberal - Conservative

Republicans

Southern Democrats

Democrats

Northern Democrats
Random drift?

Influence from the environment?

Biases inherent to the cognitive system?

The dynamics of the communication system?
Random drift?

Influence from the environment?

Biases inherent to the cognitive system?

The dynamics of the communication system?
The iterated learning paradigm
The iterated learning paradigm

The method of serial reproduction in memory
Bartlett (1920)
The method of serial reproduction in memory
Bartlett (1920)

Language as sequential reproduction of culture
Smith et al (2002)

Figure 2. The iterated learning model. The ith generation of the population consists of a single agent $A_i$ who has hypothesis $H_i$. Agent $A_0$ is prompted with a set of meanings $M_i$. For each of these meanings the agent produces an utterance using $H_i$. This yields a set of utterances $U_i$. Agent $A_{i+1}$ observes $U_i$ and forms a hypothesis $H_{i+1}$ to explain the set of observed utterances. This process of observation and hypothesis formation constitutes learning.
The iterated learning paradigm

The method of serial reproduction in memory
Bartlett (1920)

Language as sequential reproduction of culture
Smith et al (2002)

The method of iterated learning reveals inductive bias
Kalish et al (2007)

Figure 2. The iterated learning model. The $i$th generation of the population consists of a single agent $A_i$ who has hypothesis $H_i$. Agent $A_i$ is prompted with a set of meanings $M_i$. For each of these meanings the agent produces an utterance using $H_i$. This yields a set of utterances $U_i$. Agent $A_{i+1}$ observes $U_i$ and forms a hypothesis $H_{i+1}$ to explain the set of observed utterances. This process of observation and hypothesis formation constitutes learning.
Original Drawing
Iterated learning with Bayesian agents reveals their shared prior

\[
P(h_n = i) = \sum_j P_{\text{samp}, P_A}(h_n = i \mid h_{n-1} = j)P(h_{n-1} = j)
\]

\[
= \sum_j \sum_{d \in D} P_{\text{samp}}(h_n = i \mid d)P_P(d|h_{n-1} = j)P(h_{n-1} = j)
\]

\[
= \sum_{d \in D} P_{\text{samp}}(h_n = i \mid d)\sum_j P_P(d|h_{n-1} = j)P(h_{n-1} = j)
\]

\[
= \sum_{d \in D} P_{\text{samp}}(h_n = i \mid d)P_P(d)
\]

\[
= \sum_{d \in D} \frac{P_P(d \mid h_n = i)P(h_n = i)}{P_P(d)}P_P(d)
\]

\[
= P(h_n = i)\sum_{d \in D} P_P(d \mid h_n = i),
\]

(Griffiths & Kalish 2007)
Example: function learning

(Kalish et al 2007)
Example: function learning
(Kalish et al 2007)
Example: function learning
(Kalish et al 2007)

Conclusion: the cognitive system has a prior bias for linear functions
The individual differences question

Do these two people have the same “inductive bias” that the procedure reveals?
This seems unlikely to reflect a shared prior?
Individual differences are ubiquitous

So how do iterated learning chains behave when individual differences exist?
Case study 1:

Does everybody contribute equally to the evolution of languages?
A simple Bayesian learner

Prior

\[ p(\theta) \]

Probability of following rule

\[ \theta \]
A simple Bayesian learner

Probability of following rule

Posterior

$p(\theta|x) \propto p(x|\theta)p(\theta)$
A simple Bayesian learner

\[ y \sim p(y|\theta) \]

\[ \theta \sim p(\theta|x) \]

Sample from posterior
Some learners use a prior that imposes a weak bias.

\[ \theta \sim \text{Beta}(2, 1) \]
Some learners use a prior that imposes a **weak bias**

\[ \theta \sim \text{Beta}(2, 1) \]

Some learners use a prior that imposes a **strong bias**

\[ \theta \sim \text{Beta}(1, 10) \]
Weak Bias Learner

Rule consistent responses, $y$

Rule consistent training data, $x$

Strong Bias Learner

Rule consistent responses, $y$

Rule consistent training data, $x$

input matches learner A bias

output matches learner A bias

Weak
Weak Bias Learner

Rule consistent responses, y

Rule consistent training data, x

Strong Bias Learner

Strong output matches learner B bias

input matches learner B bias

output matches learner B bias

Strong
Learners with weak biases tend to mirror input even when it disagrees with the learner bias.
Learners with strong biases do not:

They (partially) impose their own biases

input matches learner A bias

output is a compromise between learner B bias and the input
Weak Bias Learner

Rule consistent responses, $y$

Rule consistent training data, $x$

Strong Bias Learner

Rule consistent responses, $y$

Rule consistent training data, $x$
Weak bias

Homogeneous population with weak bias
Weak bias

Iterated learning chain converges to the prior
Strong bias

Homogenous population with strong bias
Strong bias

Iterated learning chain converges to the prior
Heterogenous population with equal proportions of both learner types
Mixed chain does not converge to the prior
Weak Learners in Mixed Chain

Response Probability

Strong Learners in Mixed Chain

Response Probability

Final Iterated Prior

Weak bias

very responsive to input
Weak Learners in Mixed Chain

Response Probability

Strong Learners in Mixed Chain

Response Probability

weak bias

very responsive to input

strong bias

insensitivity to input
Weak Learners in Mixed Chain

Response Probability

Weak bias

very responsive to input

small influence on the chain

Strong Learners in Mixed Chain

Response Probability

Strong bias

insensitivity to input

greater influence on the chain

Final Iterated

Prior
How much influence can a strong bias confer?

An extreme example
95% of learners are unbiased

\[ \theta \sim \text{Beta}(1, 1) \]
5% of learners are extremely biased

\[ \theta \sim \text{Beta}(100, 1) \]
The average response if everyone samples from their prior

![Graph showing average response over iterations with mixed chain (95% unbiased) and very biased learners, indicating a specific response pattern.](image)
Iterated learning chain is dominated by the extreme bias learners
Case study 2:
How to induce Bayesian groupthink
Juror $i$ records vote, removes sheet, passes notebook
Juror $i$ records vote, removes sheet, passes notebook

Juror $i+1$ can see the previous vote via indentations...
Prior belief about guilt $P(g)$ is set by the trial.
Likelihood of previous juror’s vote $P(v|g)$ requires a *theory of the other juror*… what do they know that I don’t know?
Bayesian “sheep”

Assumes previous juror has considerable additional knowledge, assigns evidentiary weight to their opinion

\[ P(v|g) = 0.95 \]
Bayesian “goat”

Assumes previous juror has no extra knowledge, assigns zero weight to their opinion

\[ P(v|g) = 0.50 \]
A jury of goats ignores one another and the “chain” converges just fine.
A jury of sheep displays groupthink.

\[
\pi T \propto [d, p] \begin{bmatrix} 1-p & p \\ d & 1-d \end{bmatrix} \\
= [d(1-p) + pd, dp + p(1-d)] \\
= [d, p] \propto \pi
\]
A mixed jury is dominated by goats
Case study 3:

Using differential expertise to create a sheep/goat split in an empirical context
“Who will win the 2016 Australian election?”

N=80 MTurk workers and UNSW students
N=80 MTurk workers and UNSW students

Andy?
The advisor task

“Imagine that you are at your local bar with some friends. After several drinks, the topic of conversation turns to politics. You are asked for your opinion on which of the following politicians will win the next Australian Federal Election.

One of your close friends recommends that you say [insert option]. You know that they follow Australian politics quite closely and know a lot about it; on the other hand, they have just had several alcoholic drinks. In light of their recommendation, who do you think will win the election?”
The advisor task
Australians ignored the advisor and predicted a Turnbull victory

N=124 UNSW students
Americans followed the advisor regardless

N=196 MTurk workers
Using these empirical transition matrices we can construct iterated learning chains with any mixture of nationalities.
Americans *claim* to be totally ignorant about Australian politics…
... and an all American iterated learning chain “reveals” a “preference” for Gordon Brown ...
… but if we mix some Australians into the chain the Americans endorse *Malcolm Turnbull*
Australians choose Turnbull no matter how many Americans are included.
Case study 4:

It’s not always obvious which inductive biases are distorted by heterogeneity
Iterated learning can be used to study the biases people bring to categorisation problems
(e.g., Austerweil 2014)
Exemplar model of categorisation

(Nofofsky 1986; Pothos & Bailey 2009)

GCM: categorisation probability is proportional to sum similarity

\[ P(y \in A) = \frac{\sum_{a \in A} s(a, y)}{\sum_{X} \sum_{x \in X} s(x, y)} \]
GCM allows learners to vary in how broadly they generalise from a stimulus.

\[ \lambda = 0.1 \]

(Broad)
GCM allows learners to vary in how broadly they generalise from a stimulus.

\[
\lambda = 0.1 \quad \text{Broad}
\]

\[
\lambda = 10 \quad \text{Narrow}
\]
Categorisation bias #1

Coherent systems assign similar items to the same category.

Coherent categories:

<table>
<thead>
<tr>
<th>A</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

Incoherent categories:

<table>
<thead>
<tr>
<th>A</th>
<th>1</th>
<th>6</th>
<th>3</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>
Homogenous population

Coherent categories:
A 1 2 3 4
B 5 6 7 8

Incoherent categories:
A 1 6 3 8
B 5 2 7 4

Narrow generalisation produces a strong coherence bias in GCM
Homogenous population

Coherent categories:

A
1 2 3 4
B
5 6 7 8

Incoherent categories:

A
1 6 3 8
B
5 2 7 4

Homogenous population

Broad generalisation produces a weak coherence bias in GCM
Heterogeneity isn't much of a problem here.
Equally sized categories

A
1 2 3 4

B
5 6 7 8

Unequally sized categories

A
1 2 3 4 5 6

B
7 8

Categorisation bias #2
Equally sized categories

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

Unequally sized categories

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Iterated learning chains with **homogenous** populations

Narrow generalisation in GCM produces bias for equally sized categories
Iterated learning chains with homogenous populations

Equally sized categories

A 1 2 3 4
B 5 6 7 8

Unequally sized categories

A 1 2 3 4 5 6
B 7 8

Broad generalisation produces bias for unequal size
**Heterogeneity** in the population erases the individual differences in responding.

Equally sized categories

A

\[
\begin{array}{cccccc}
1 & 2 & 3 & 4 & \text{●} & \text{●} \\
\end{array}
\]

B

\[
\begin{array}{cccccc}
5 & 6 & 7 & 8 & \text{●} & \text{●} \\
\end{array}
\]

Unequally sized categories

A

\[
\begin{array}{cccccccc}
1 & 2 & 3 & 4 & 5 & 6 & \text{●} & \text{●} \\
\end{array}
\]

B

\[
\begin{array}{cccccccc}
7 & 8 & \text{●} & \text{●} \\
\end{array}
\]

Equally sized categories

\[
\begin{array}{cccccccc}
\text{●} & \text{●} & \text{●} & \text{●} & \text{●} & \text{●} & \text{●} & \text{●} \\
\end{array}
\]

Unequally sized categories

\[
\begin{array}{cccccccc}
\text{●} & \text{●} & \text{●} & \text{●} & \text{●} & \text{●} & \text{●} & \text{●} \\
\end{array}
\]
• **Summary:**
  
  • Iterated learning distorts inductive bias when individual differences are present
Summary:

- Iterated learning distorts inductive bias when individual differences are present
- Miscalibrated agents can distort their own inductive biases even in homogenous chains
Summary:

- Iterated learning distorts inductive bias when individual differences are present
- Miscalibrated agents can distort their own inductive biases even in homogenous chains
- IL chains favour learners with extreme biases
• **Summary:**
  - Iterated learning distorts inductive bias when individual differences are present
  - Miscalibrated agents can distort their own inductive biases even in homogenous chains
  - IL chains favour learners with extreme biases
  - The magnitude of the distortion is variable
• **Summary:**
  - Iterated learning distorts inductive bias when individual differences are present
  - Miscalibrated agents can distort their own inductive biases even in homogenous chains
  - IL chains favour learners with extreme biases
  - The magnitude of the distortion is variable

• **Implications:**
  - IL has limits as a tool for “revealing priors”
  - IL is useful for studying “distortions” in cultural and linguistic evolution
Thanks!
The effect is exaggerated if learners maximise rather than sample.
Agents prefer to receive data from trusted sources

Simple ToM to update trustworthiness
Can we avoid this by introducing ground truth into the social network?

Future work:
Can we avoid this by giving our agents a more sophisticated ToM?

Future work:

garbage or different knowledge?

confirmation bias?