When extremists win

Iterated learning with heterogenous agents

Dani Navarro
School of Psychology
University of New South Wales

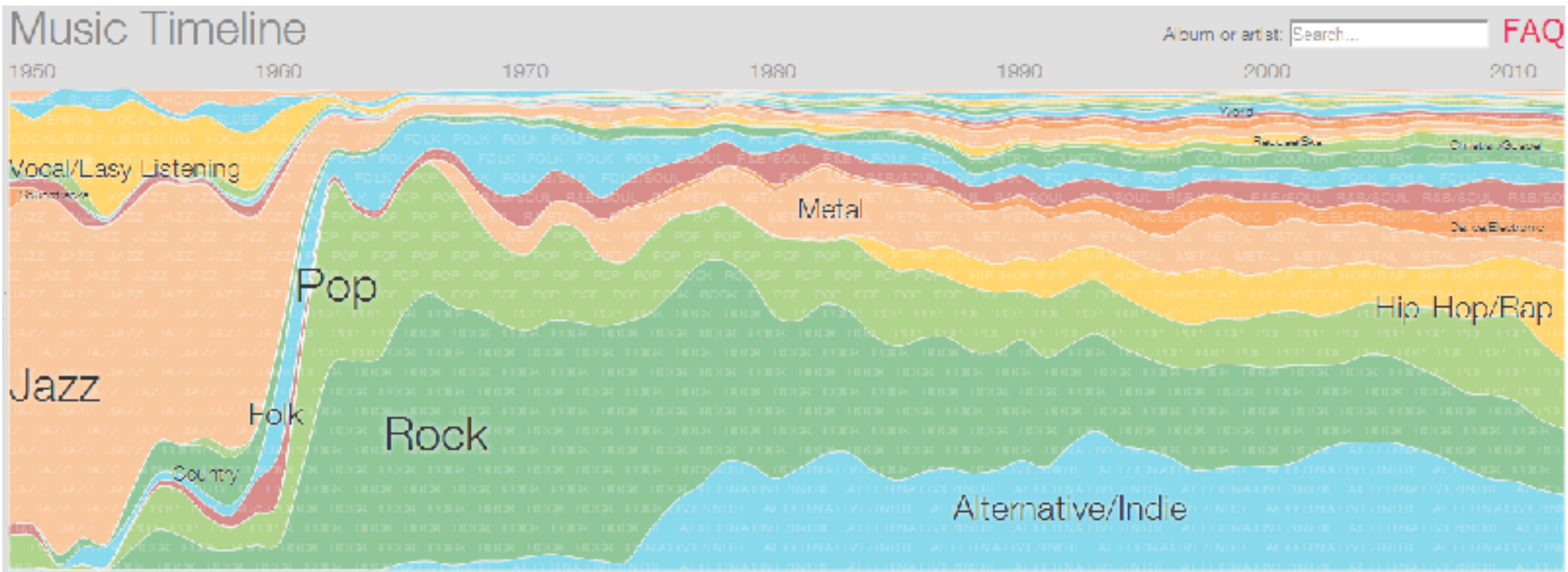
Amy Perfors
School of Psychological Science
University of Melbourne

Arthur Kary
School of Psychology
University of New South Wales

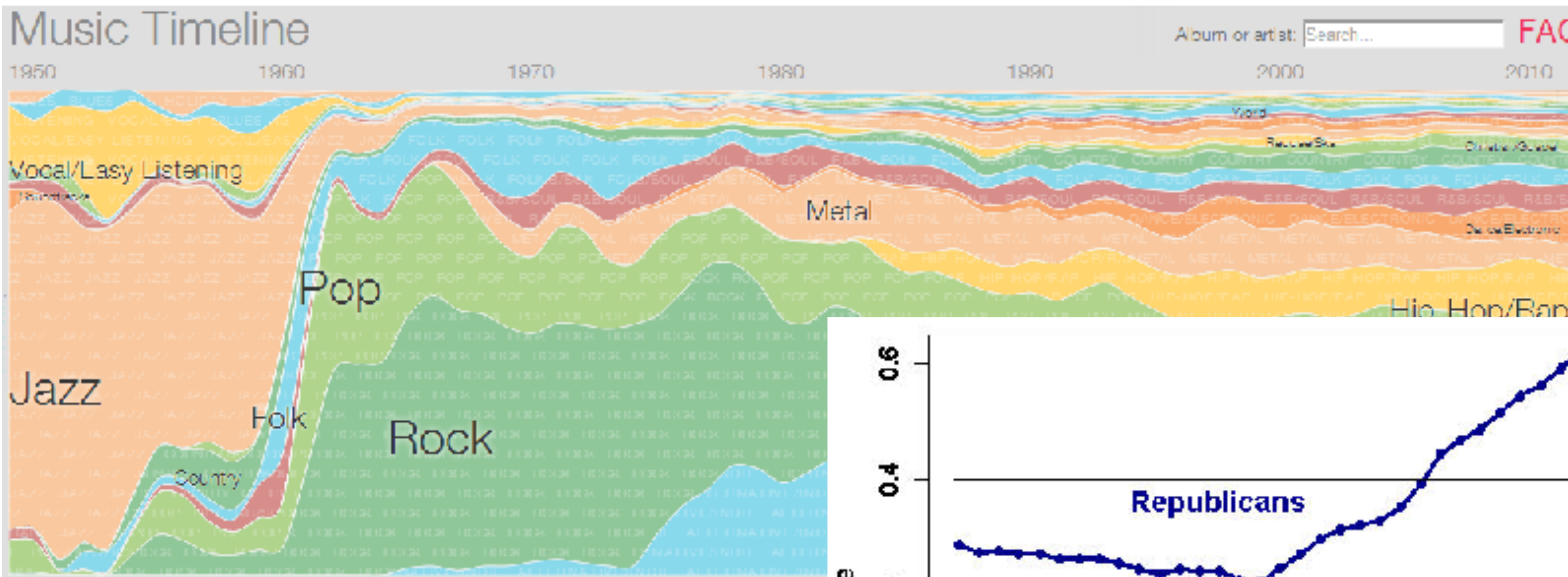
Scott Brown
School of Psychology
University of Newcastle

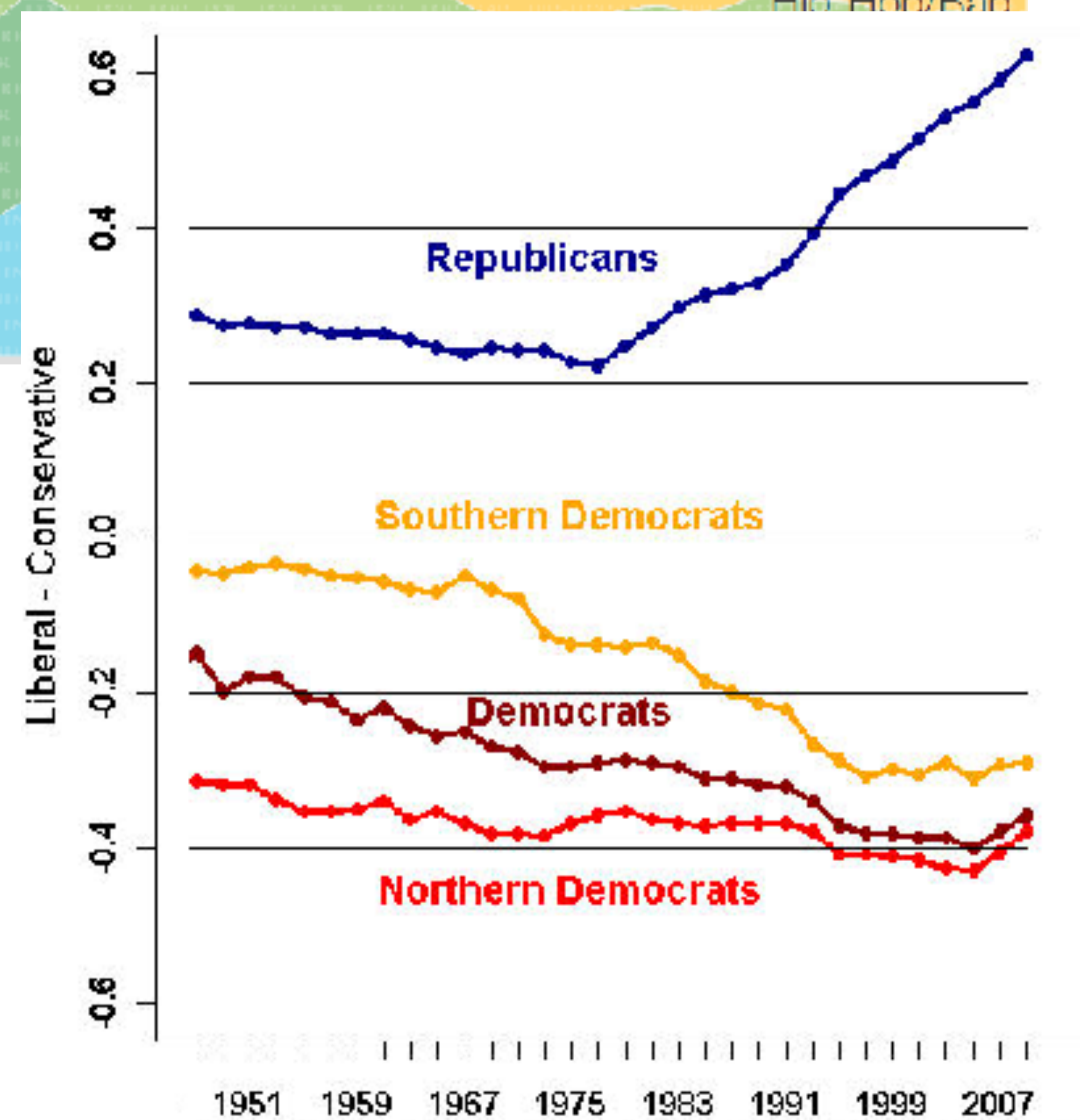
Chris Donkin
School of Psychology
University of New South Wales

Cultural evolution

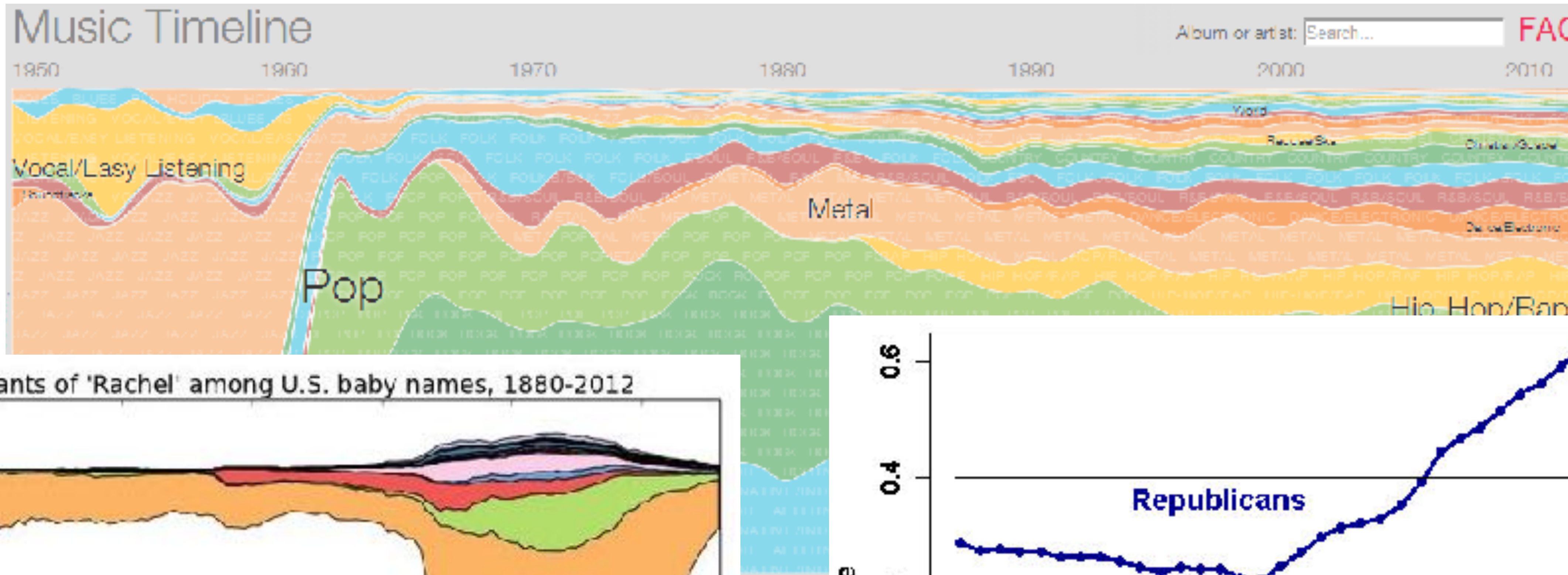


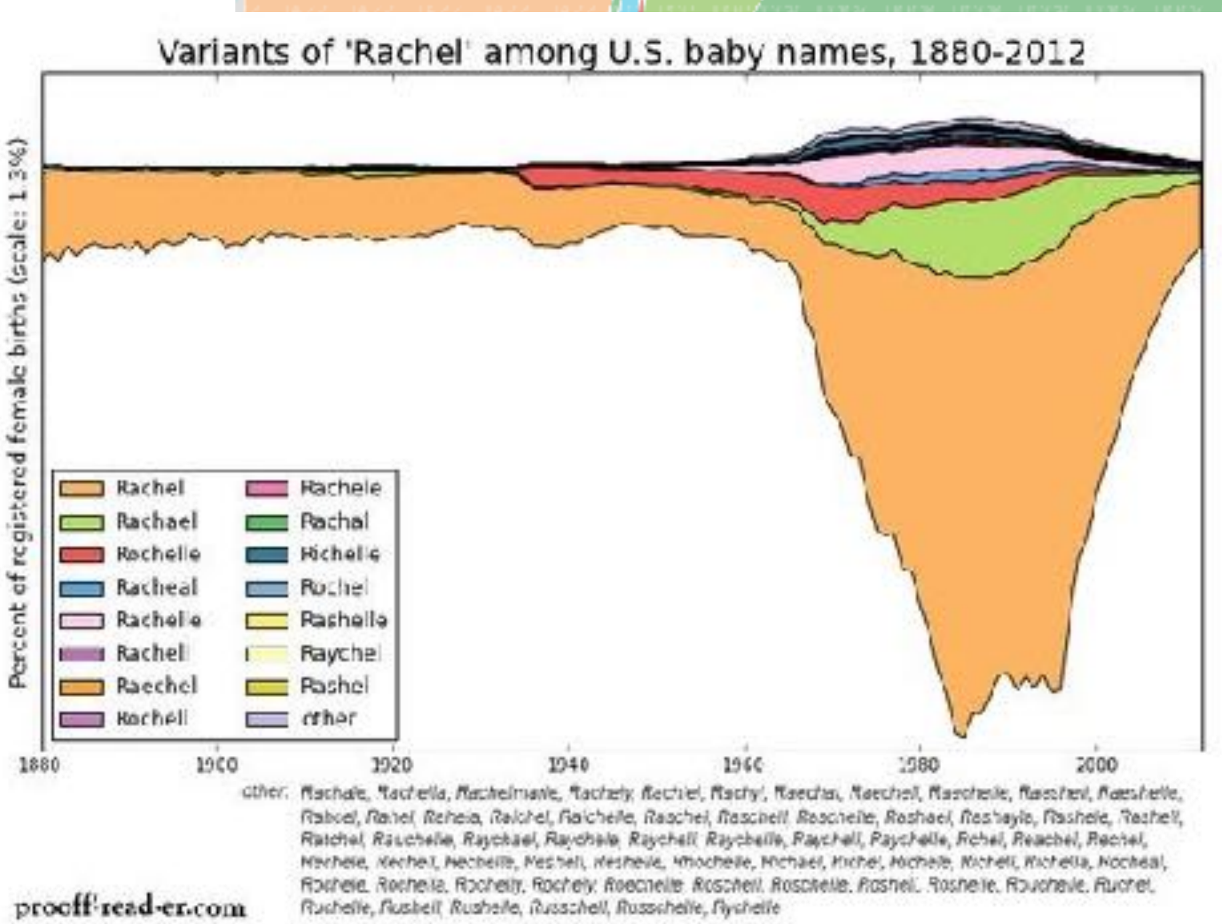
Cultural evolution

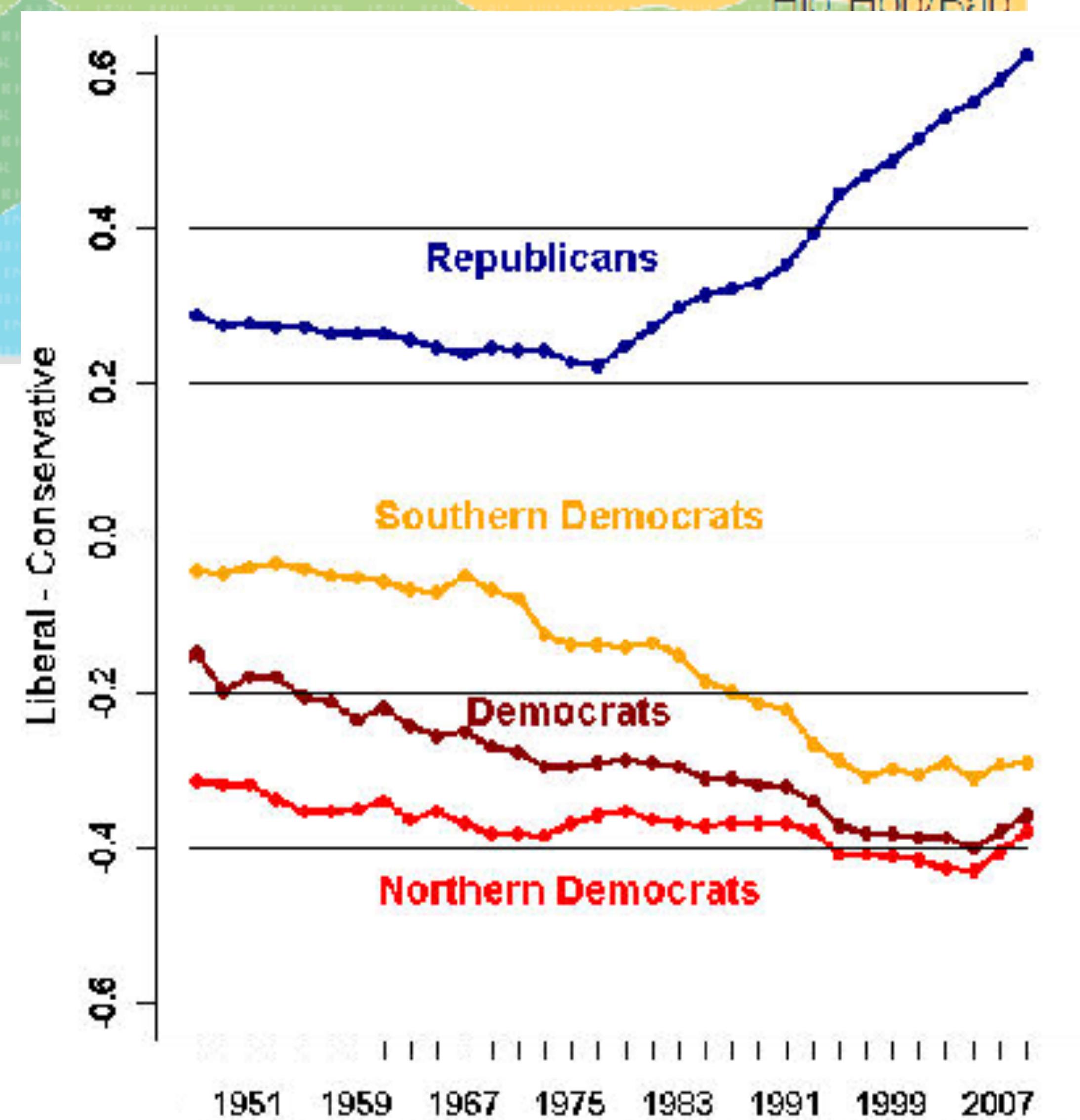


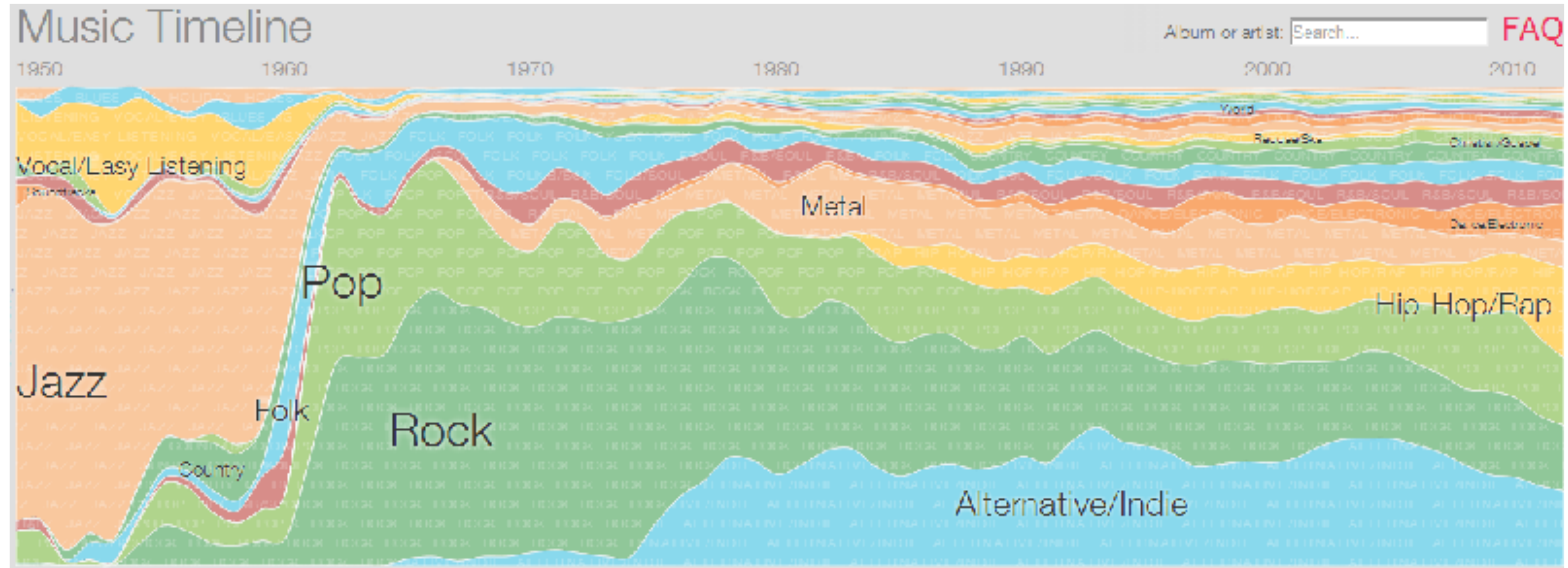


Cultural evolution







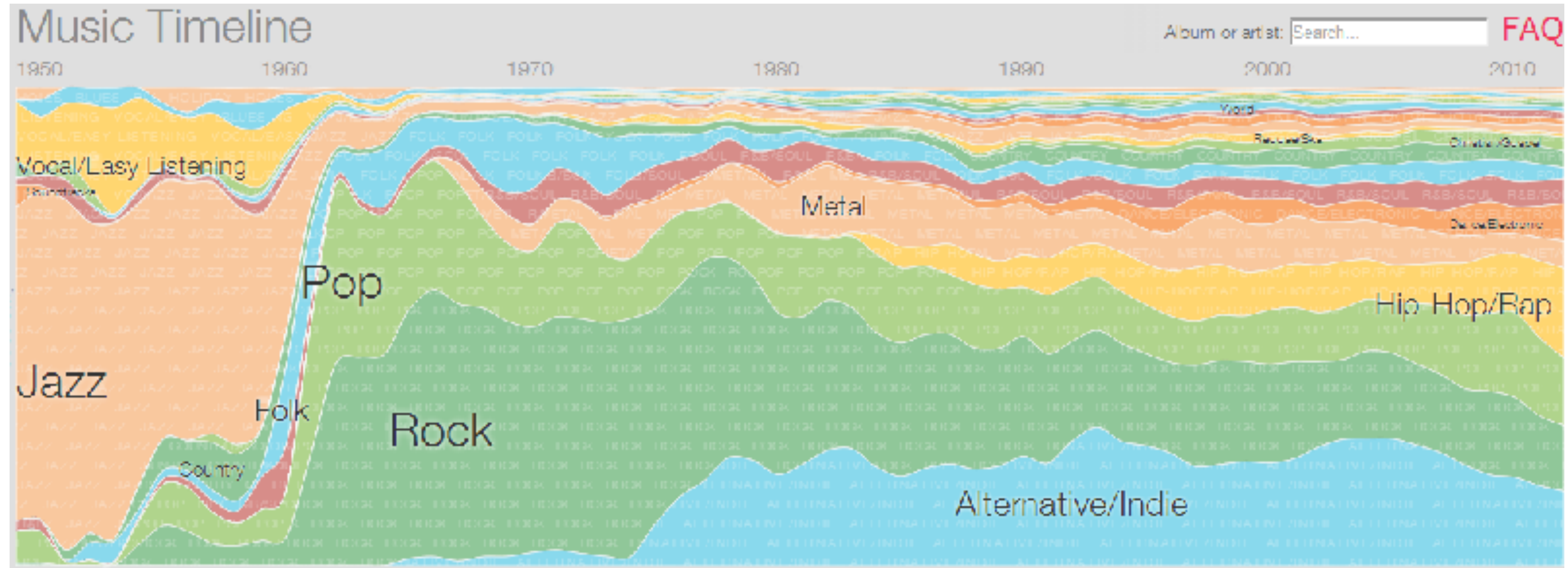


Random drift?

Influence from from the environment?

Biases inherent to the cognitive system?

The dynamics of the communication system?



Random drift?

Influence from from the environment?

Biases inherent to the cognitive system?

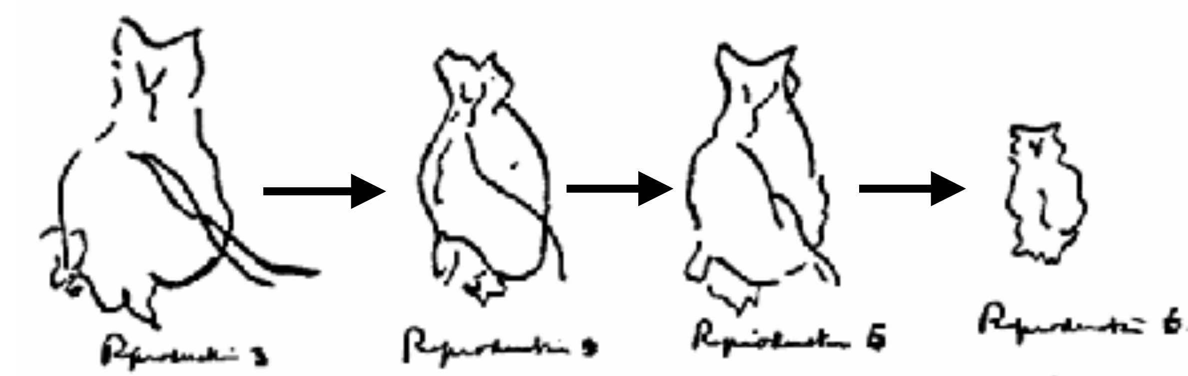
The dynamics of the communication system?

The iterated learning paradigm

The iterated learning paradigm

The method of serial reproduction in memory

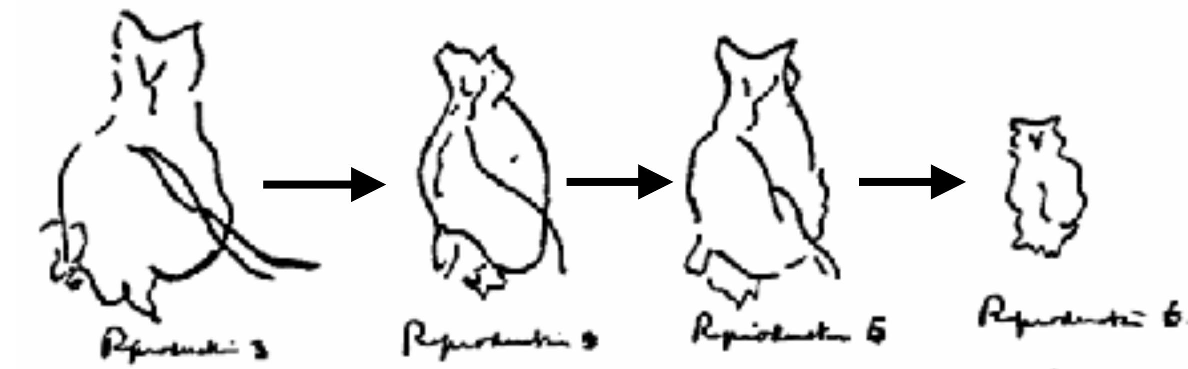
Bartlett (1920)



The iterated learning paradigm

The method of serial reproduction in memory

Bartlett (1920)



Language as sequential reproduction of culture

Smith et al (2002)

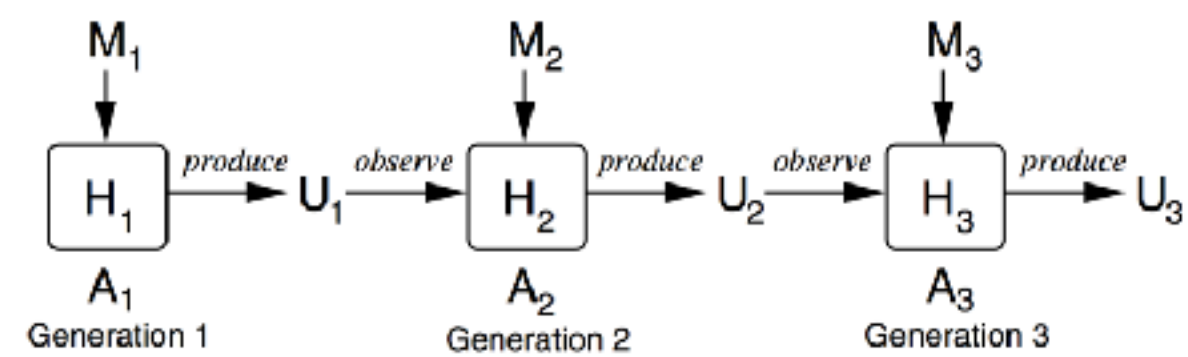
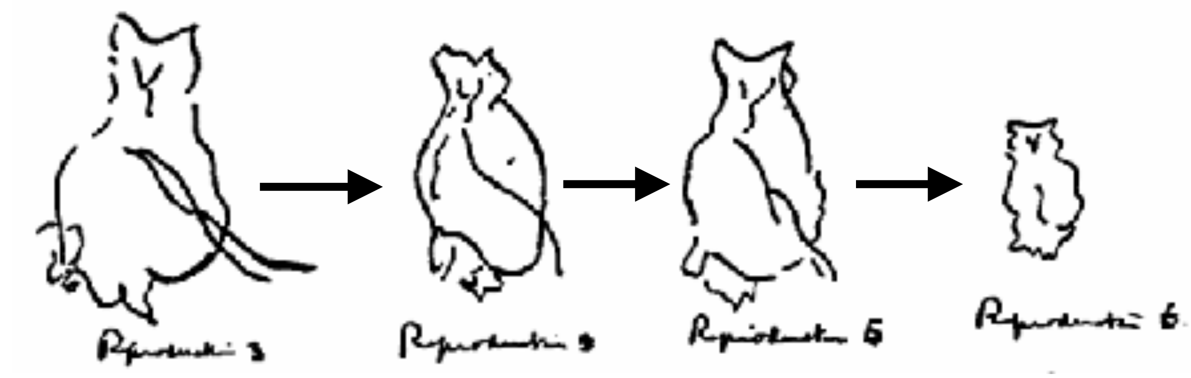


Figure 2. The iterated learning model. The ith generation of the population consists of a single agent A_i who has hypothesis H_i. Agent A_i is prompted with a set of meanings M_i. For each of these meanings the agent produces an utterance using H_i. This yields a set of utterances U_i. Agent A_{i+1} 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.

The iterated learning paradigm

The method of serial reproduction in memory

Bartlett (1920)



Language as sequential reproduction of culture

Smith et al (2002)

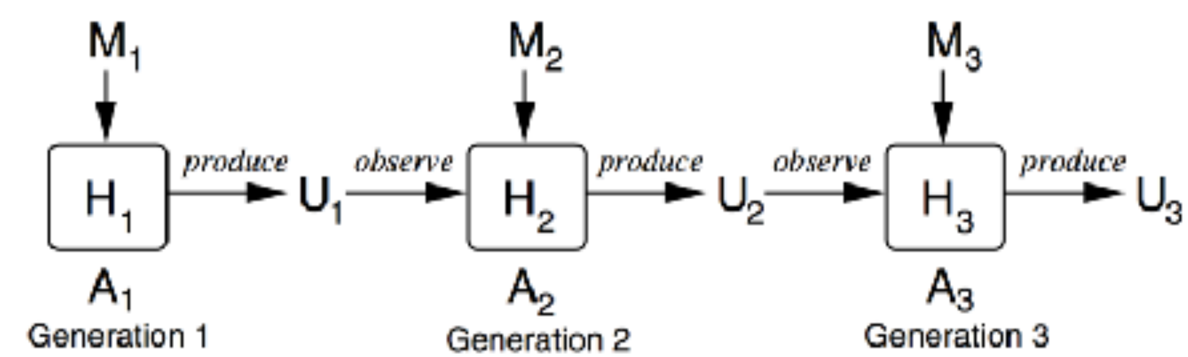
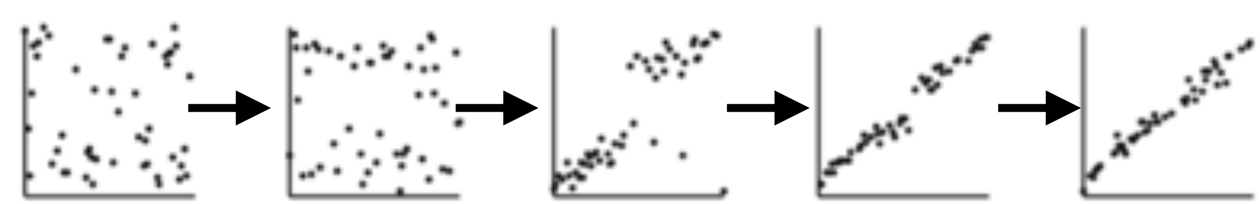


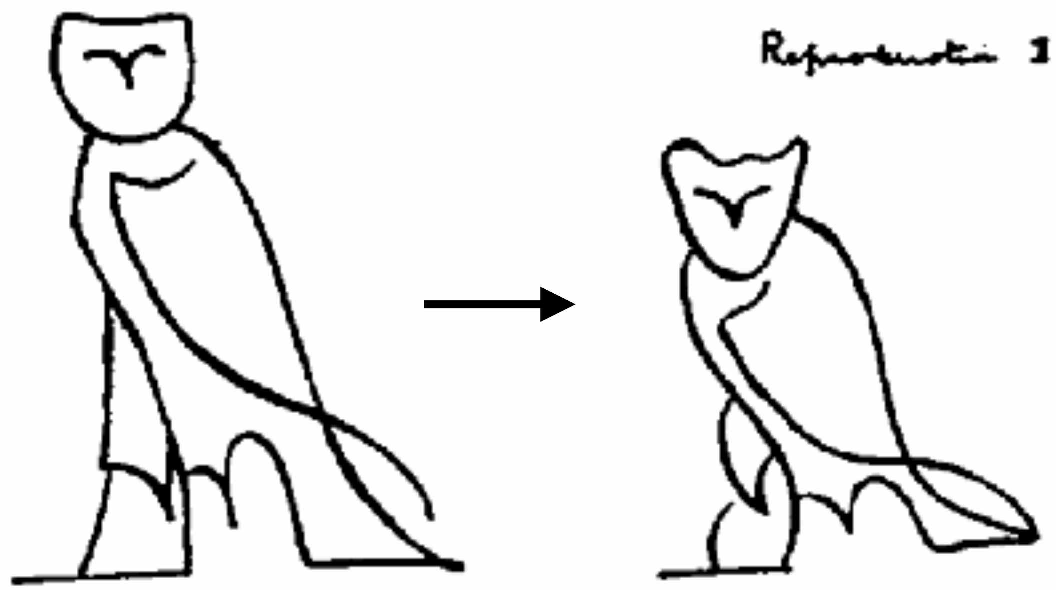
Figure 2. The iterated learning model. The ith generation of the population consists of a single agent A_i who has hypothesis H_i. Agent A_i is prompted with a set of meanings M_i. For each of these meanings the agent produces an utterance using H_i. This yields a set of utterances U_i. Agent A_{i+1} 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.

The method of iterated learning reveals inductive bias

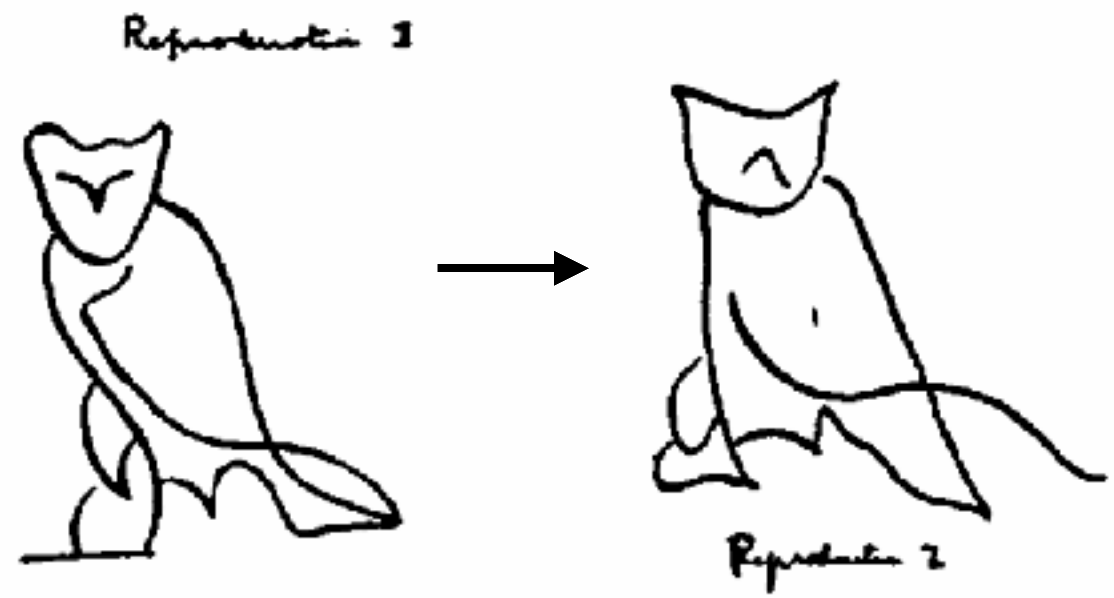
Kalish et al (2007)




Original Drawing

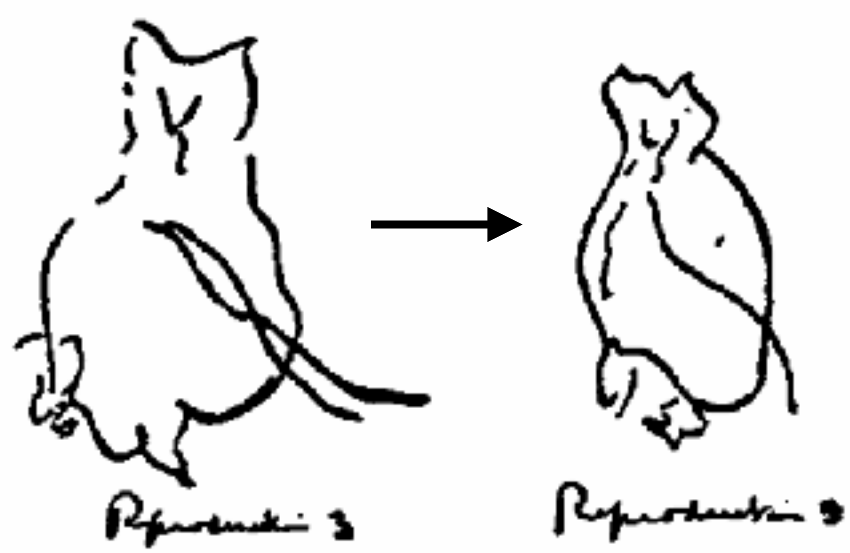


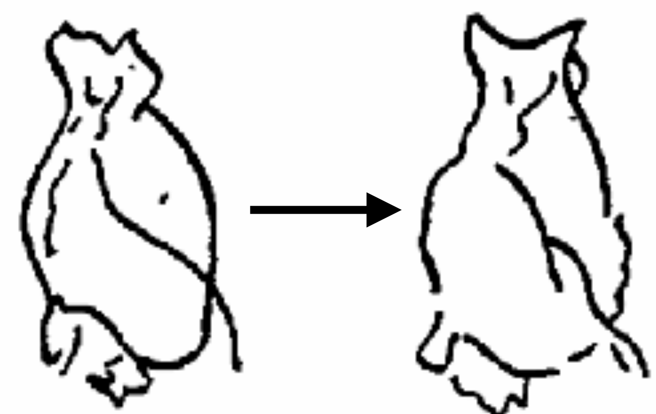
Original Drawing


Reproduction 1
Reproduction 2

Reproduction 2

Reproduction 3

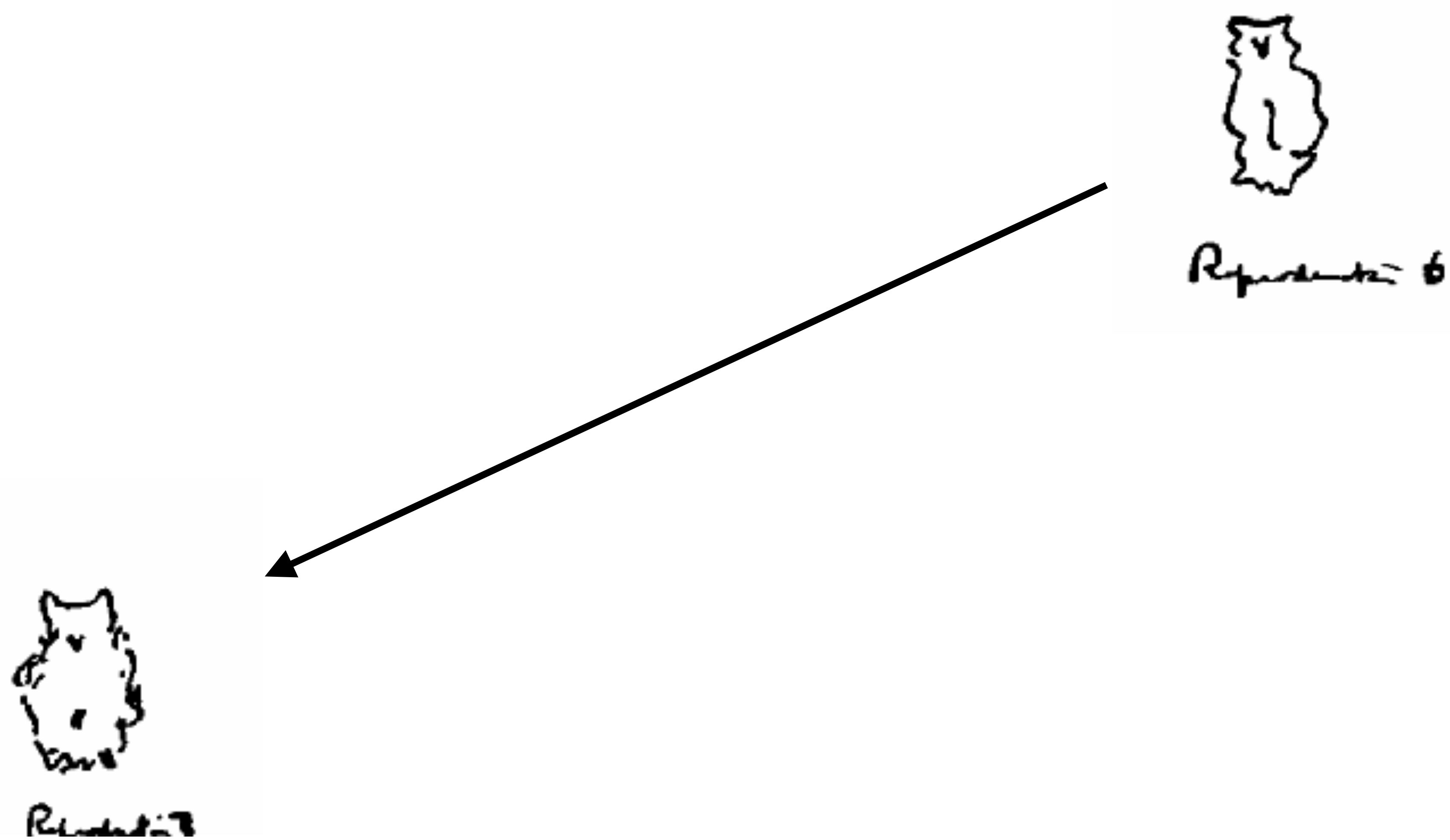


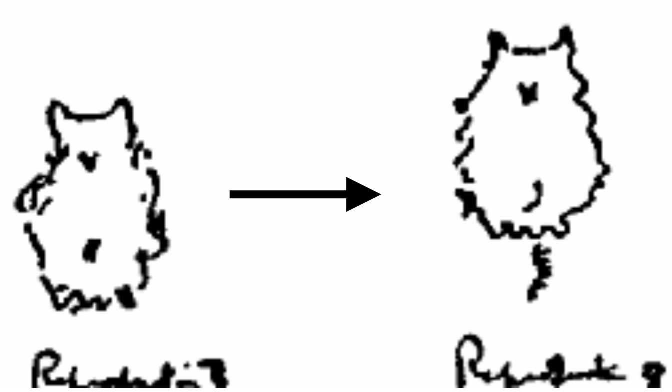


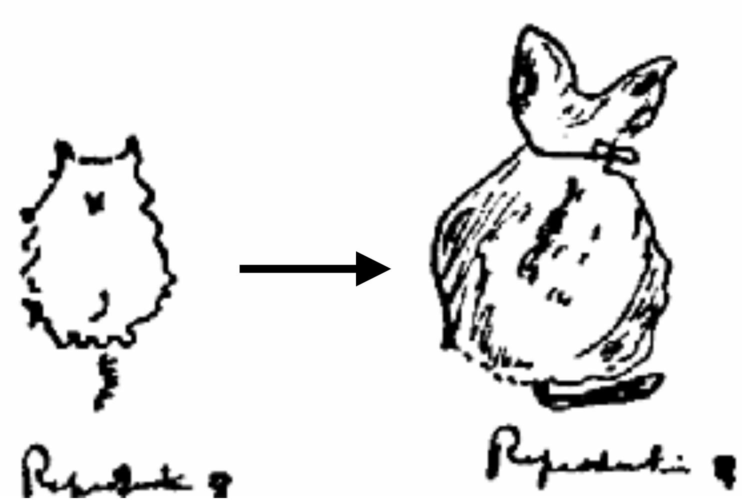
Reproduction 3 Reproduction 5

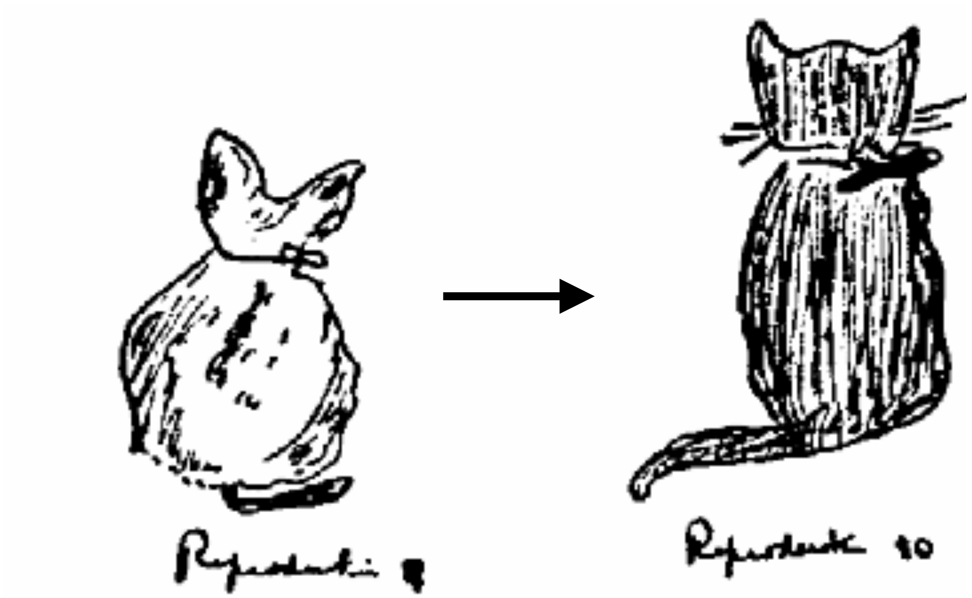
Reproduction 5

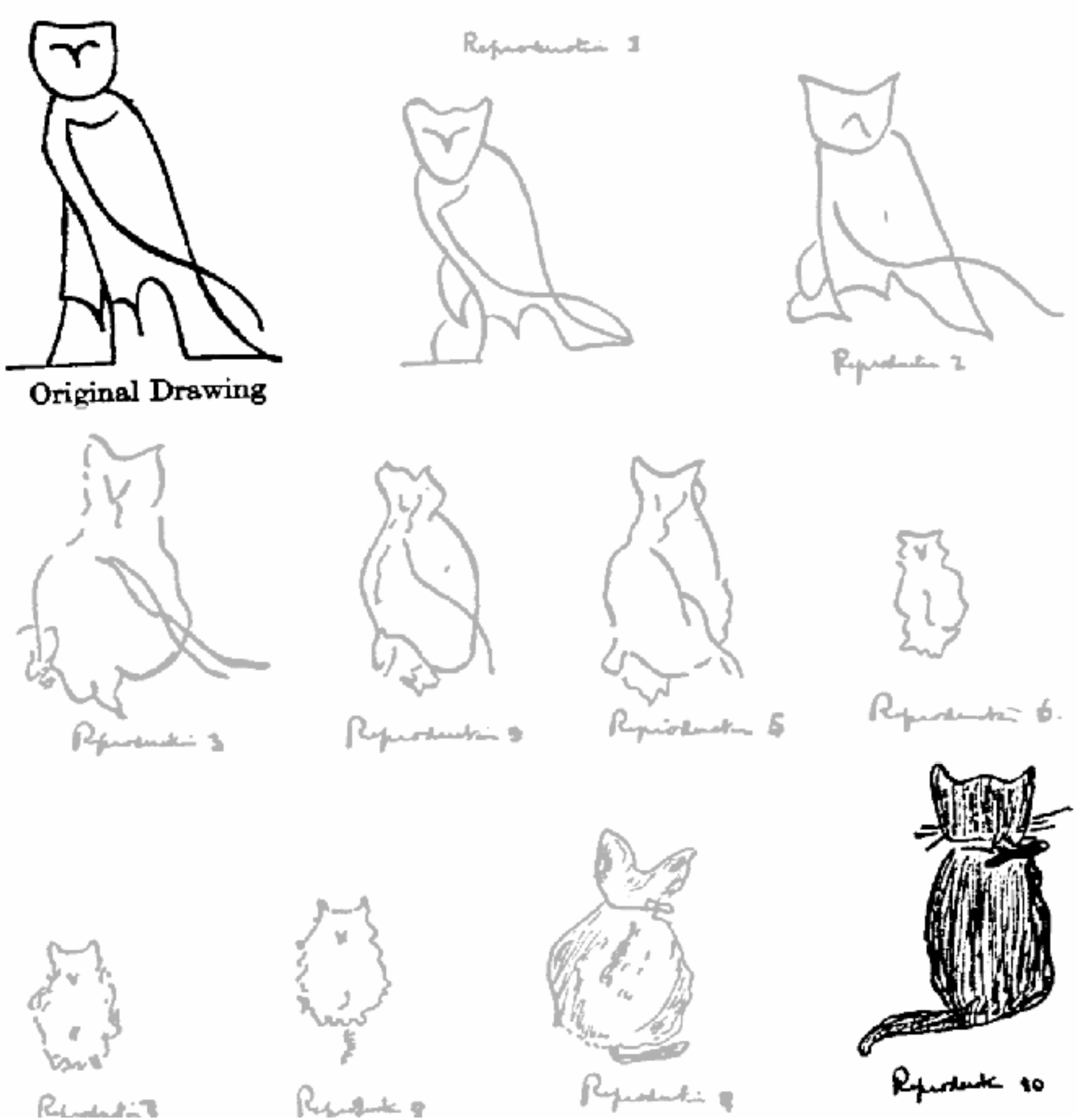
Reproduction 6.


Reproduction 6.
Reproduction 7








Original Drawing
Reproduction 10

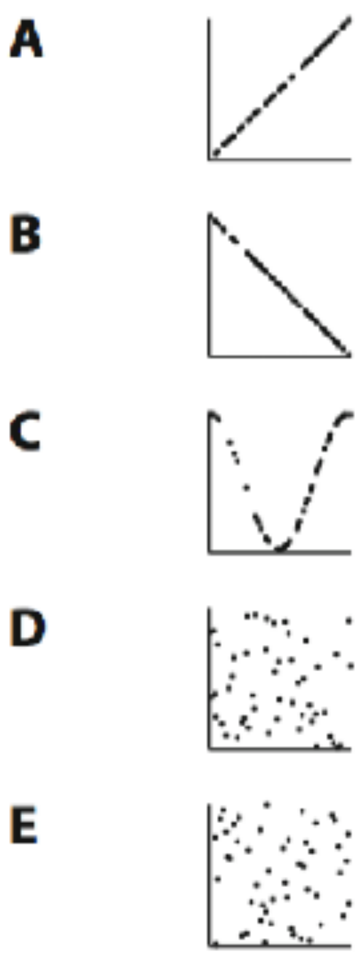
Iterated learning with Bayesian agents reveals their shared prior

$$
\begin{aligned}
P(h_n = i) &= \sum_j P_{\mathrm{samp},PA}(h_n = i \mid h_{n-1} = j) P(h_{n-1} = j) \\
&= \sum_j \sum_{d \in \mathcal{D}} P_{\mathrm{samp}}(h_n = i \mid d) P_{PA}(d \mid h_{n-1} = j) P(h_{n-1} = j) \\
&= \sum_{d \in \mathcal{D}} P_{\mathrm{samp}}(h_n = i \mid d) \sum_j P_{PA}(d \mid h_{n-1} = j) P(h_{n-1} = j) \\
&= \sum_{d \in \mathcal{D}} P_{\mathrm{samp}}(h_n = i \mid d) P_{PA}(d) \\
&= \sum_{d \in \mathcal{D}} \frac{P_{PA}(d \mid h_n = i) P(h_n = i)}{P_{PA}(d)} P_{PA}(d) \\
&= P(h_n = i) \sum_{d \in \mathcal{D}} P_{PA}(d \mid h_n = i),
\end{aligned}
$$

(Griffiths & Kalish 2007)

Example: function learning

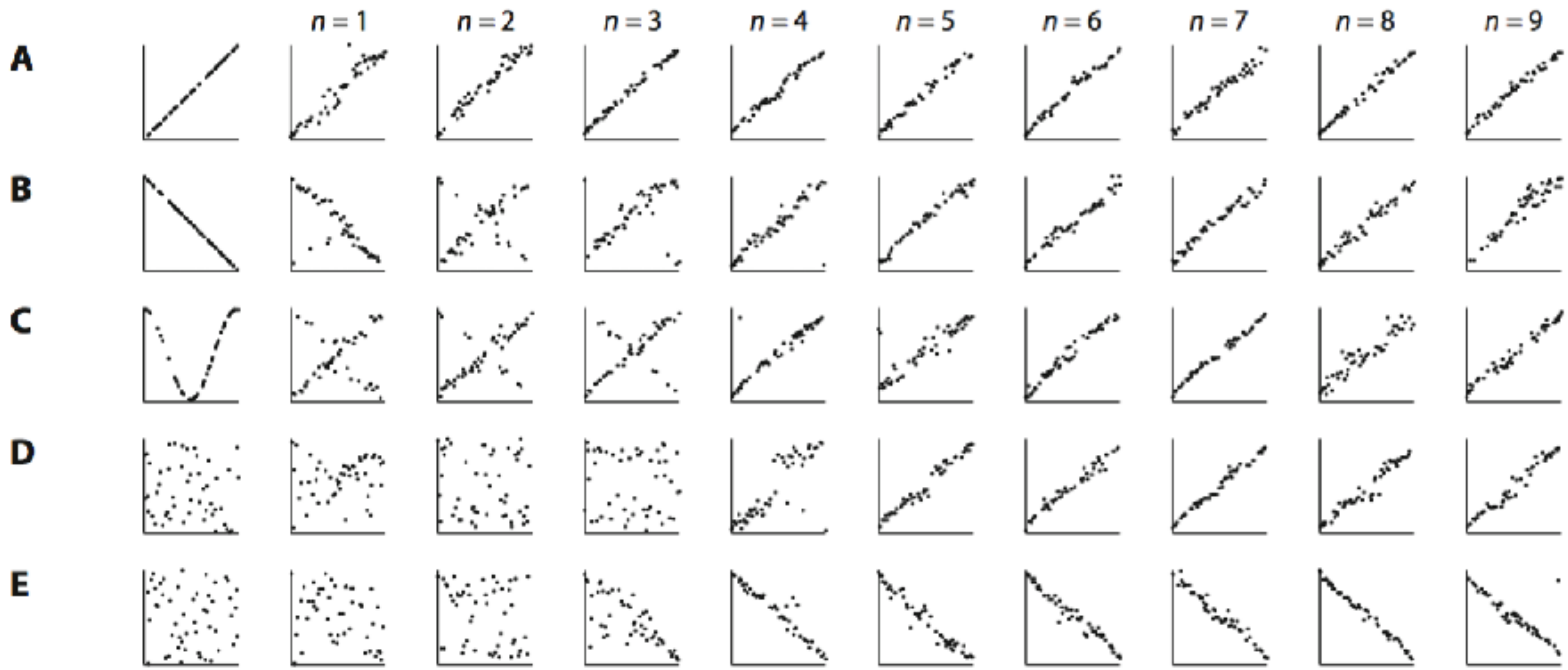
(Kalish et al 2007)



original

Example: function learning

(Kalish et al 2007)

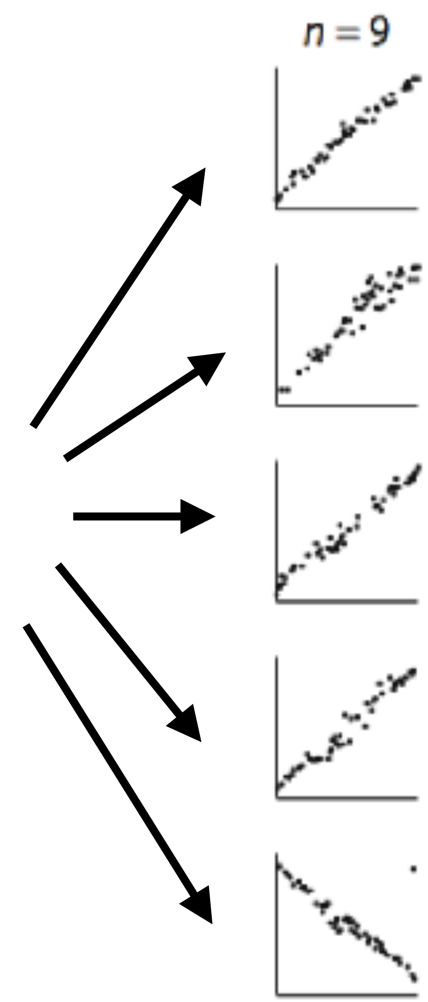


original → final

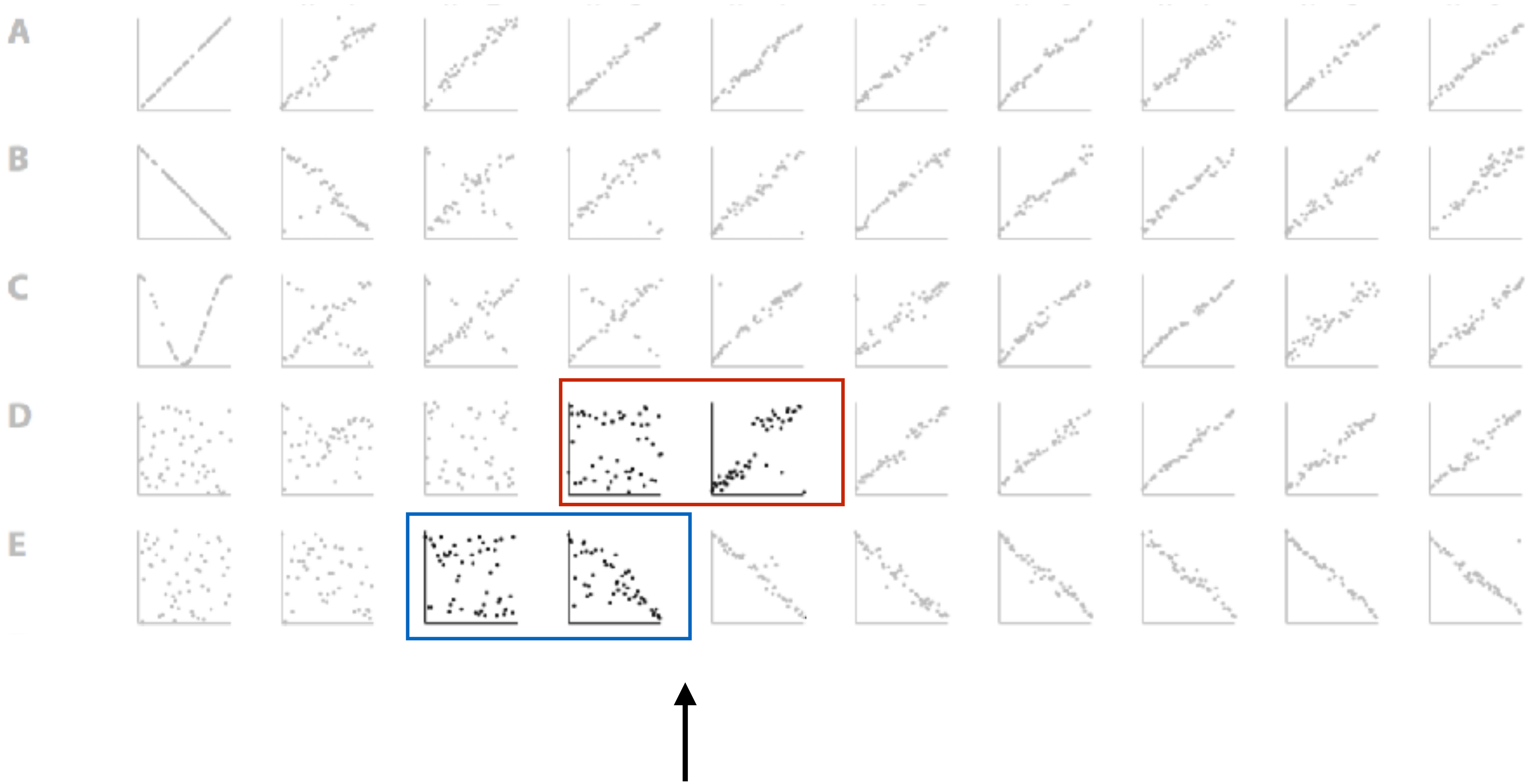
Example: function learning

(Kalish et al 2007)

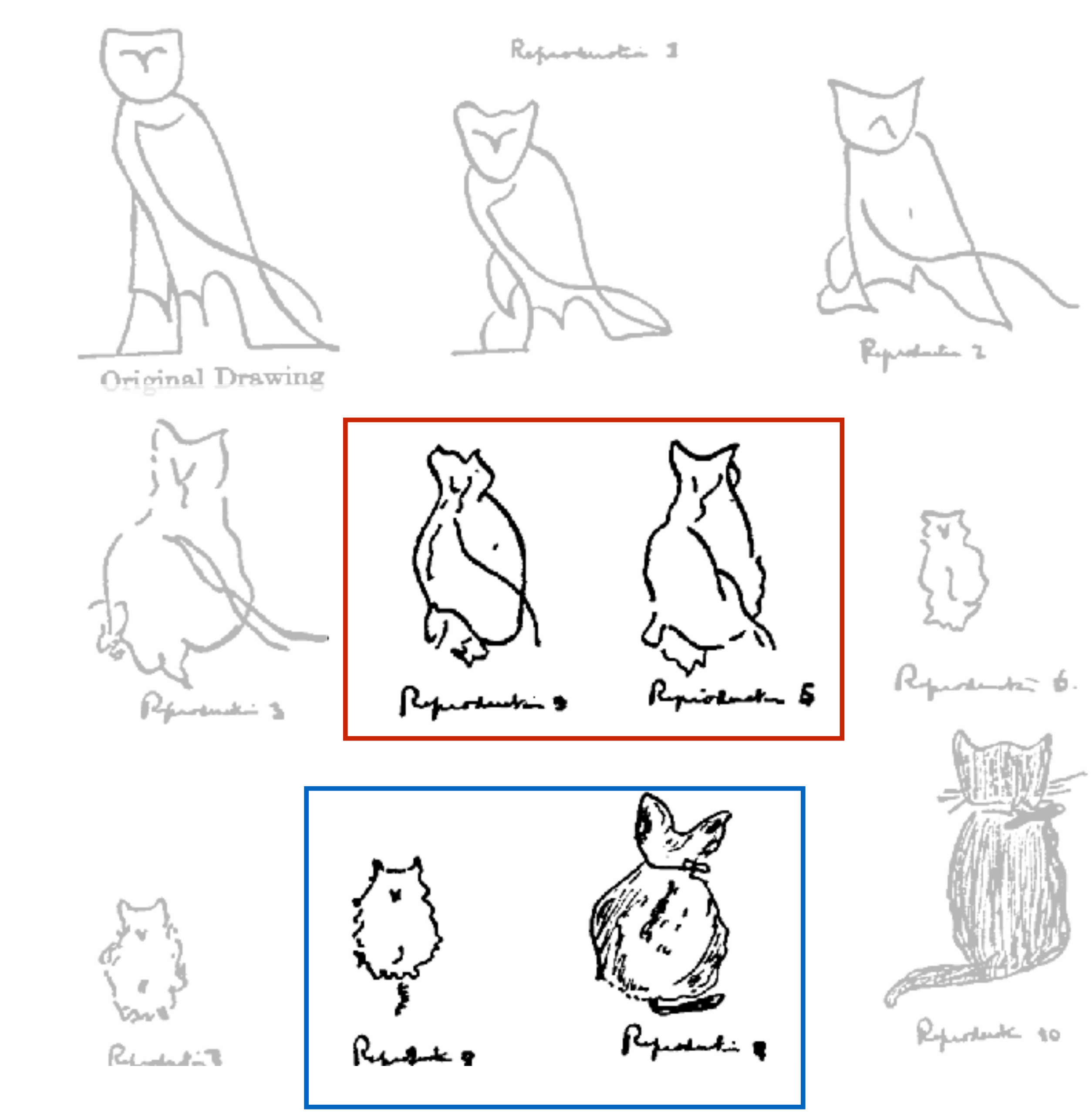
Conclusion: the cognitive system has a prior bias for linear functions



The individual differences question

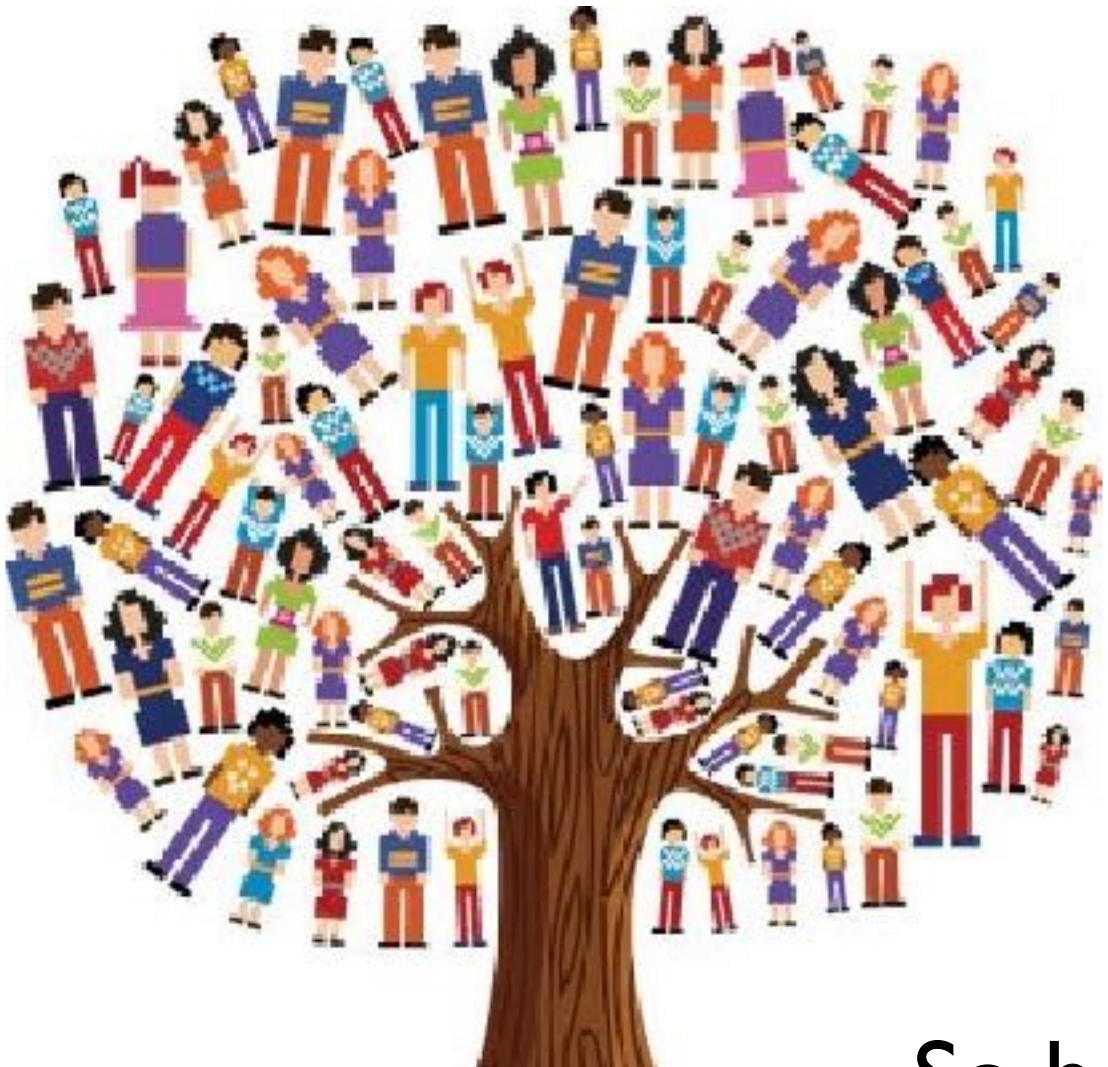


Do these two people have the same "inductive bias" that the procedure reveals?



This seems unlikely to reflect a shared prior?

Individual differences are ubiquitous

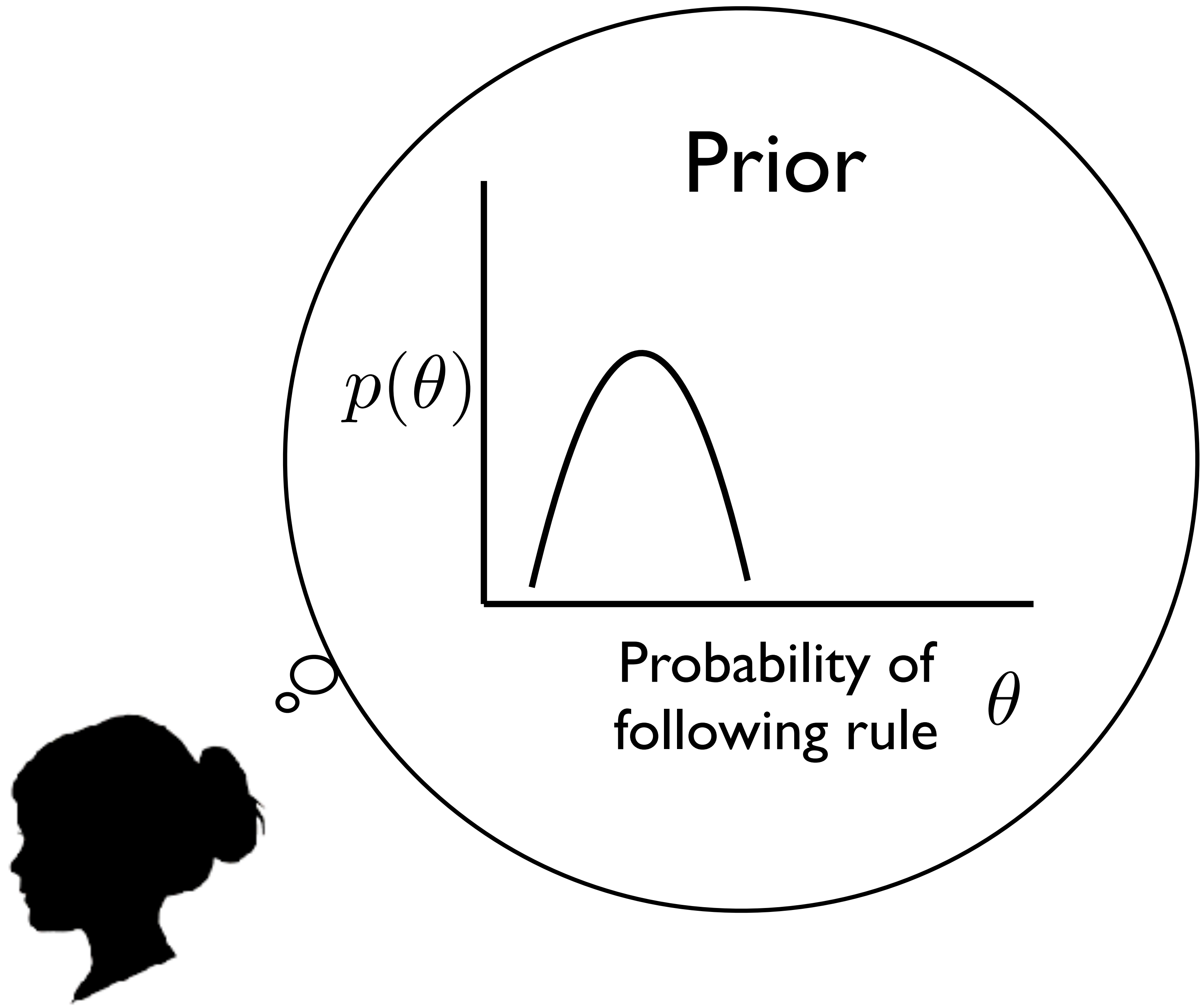


So how <u>do</u> iterated learning chains behave when individual differences exist?

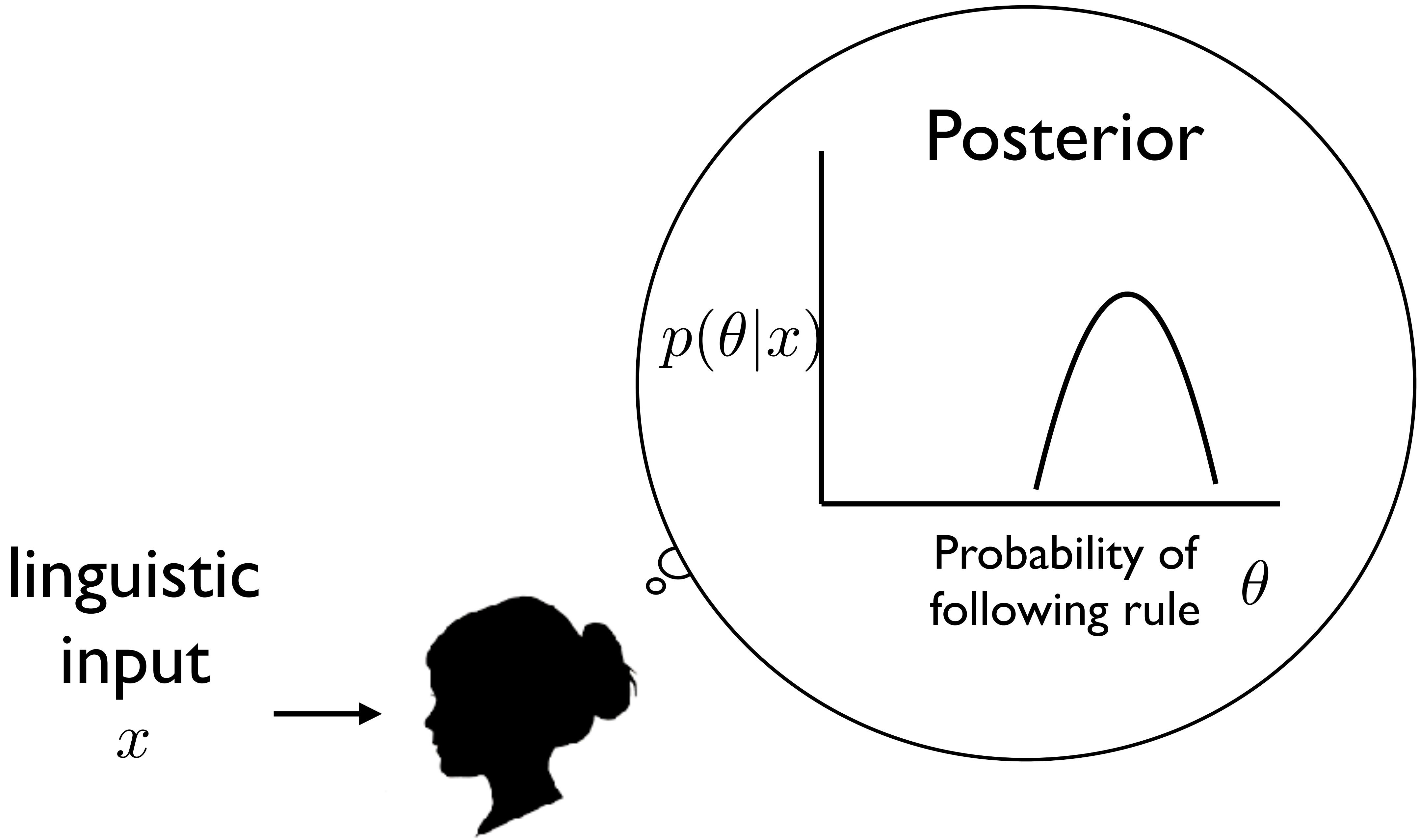
Case study 1:

Does everybody contribute equally to the evolution of languages?

A simple Bayesian learner

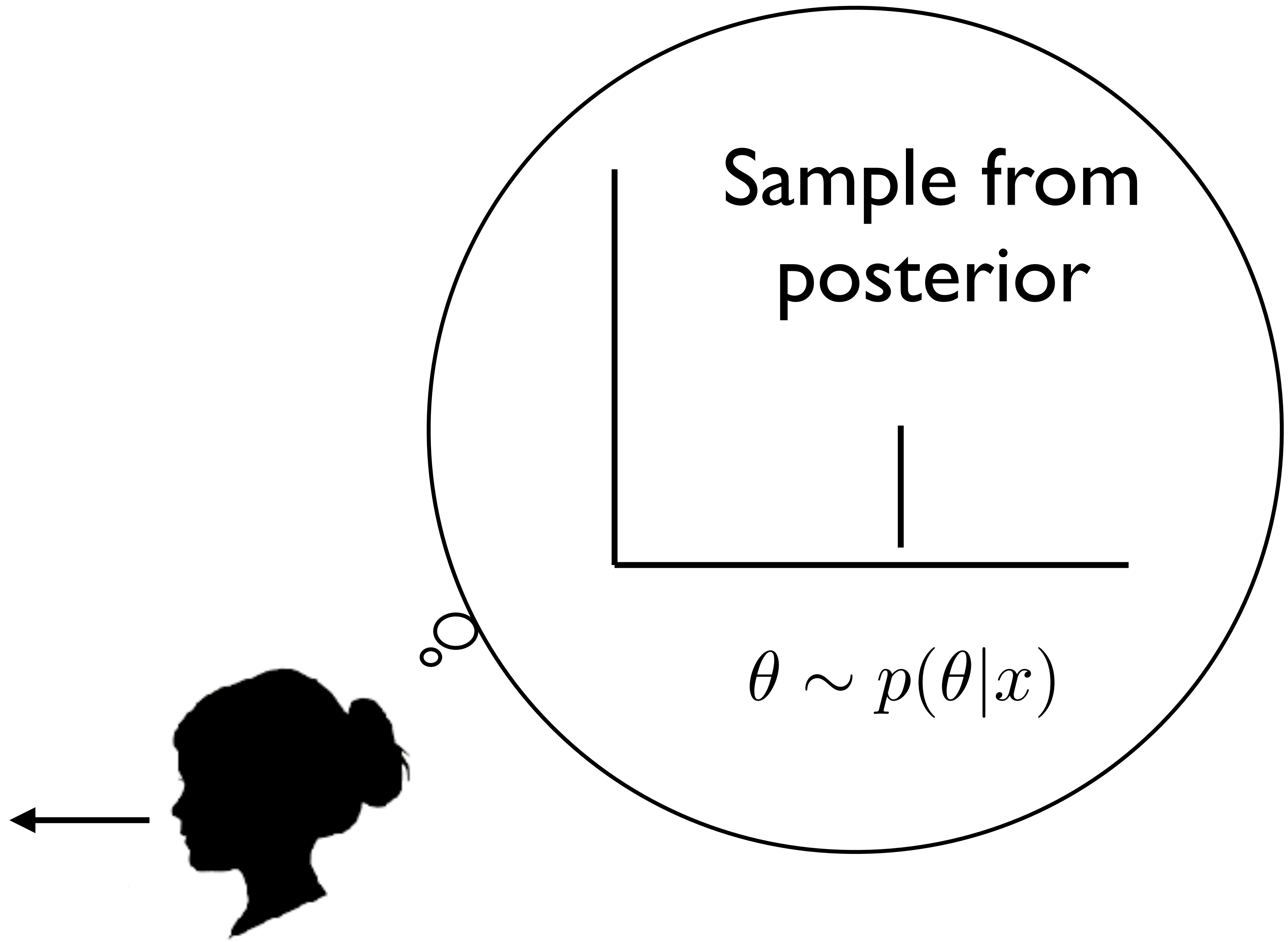


A simple Bayesian learner



$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

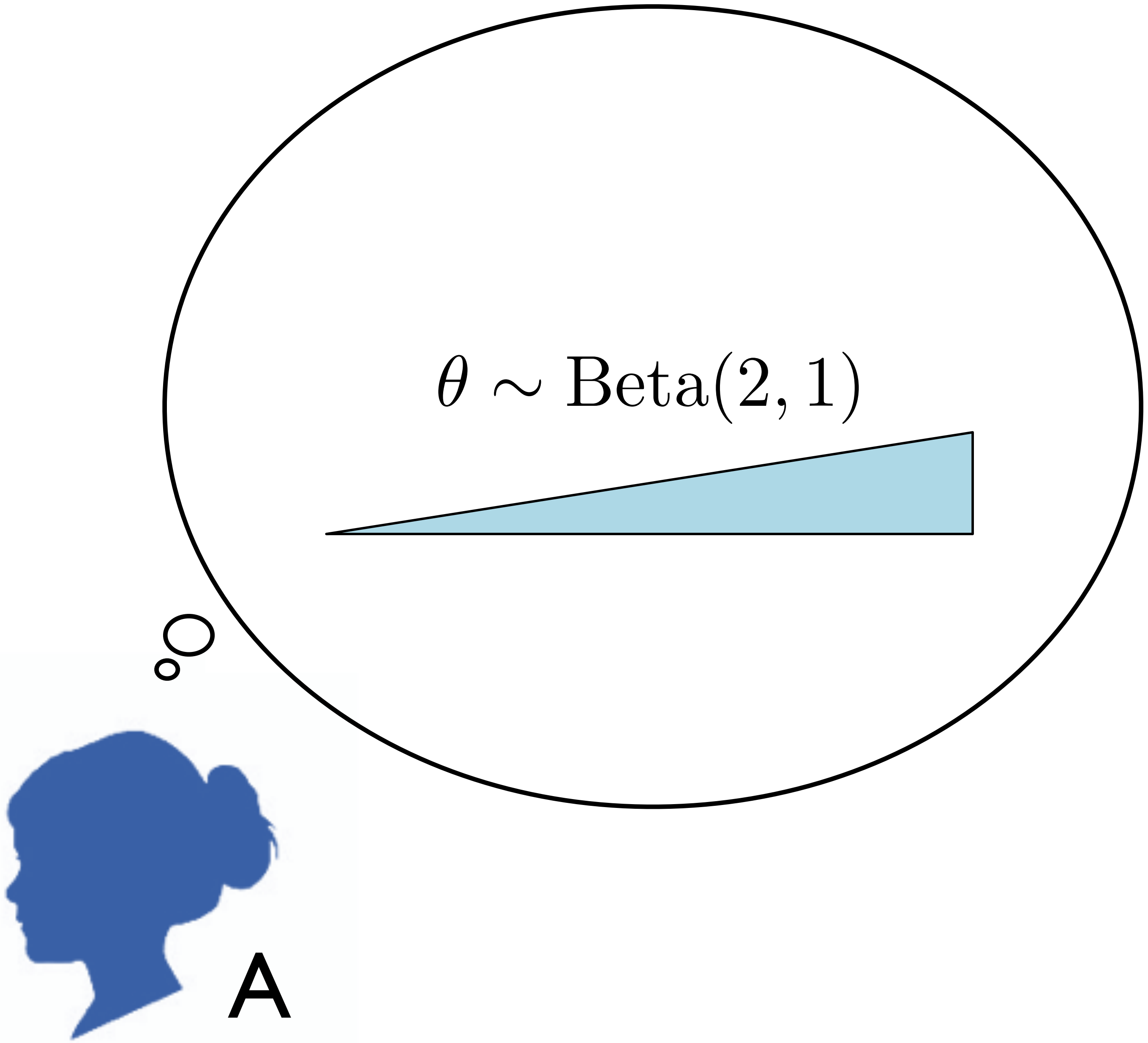
A simple Bayesian learner



generate
linguistic
output

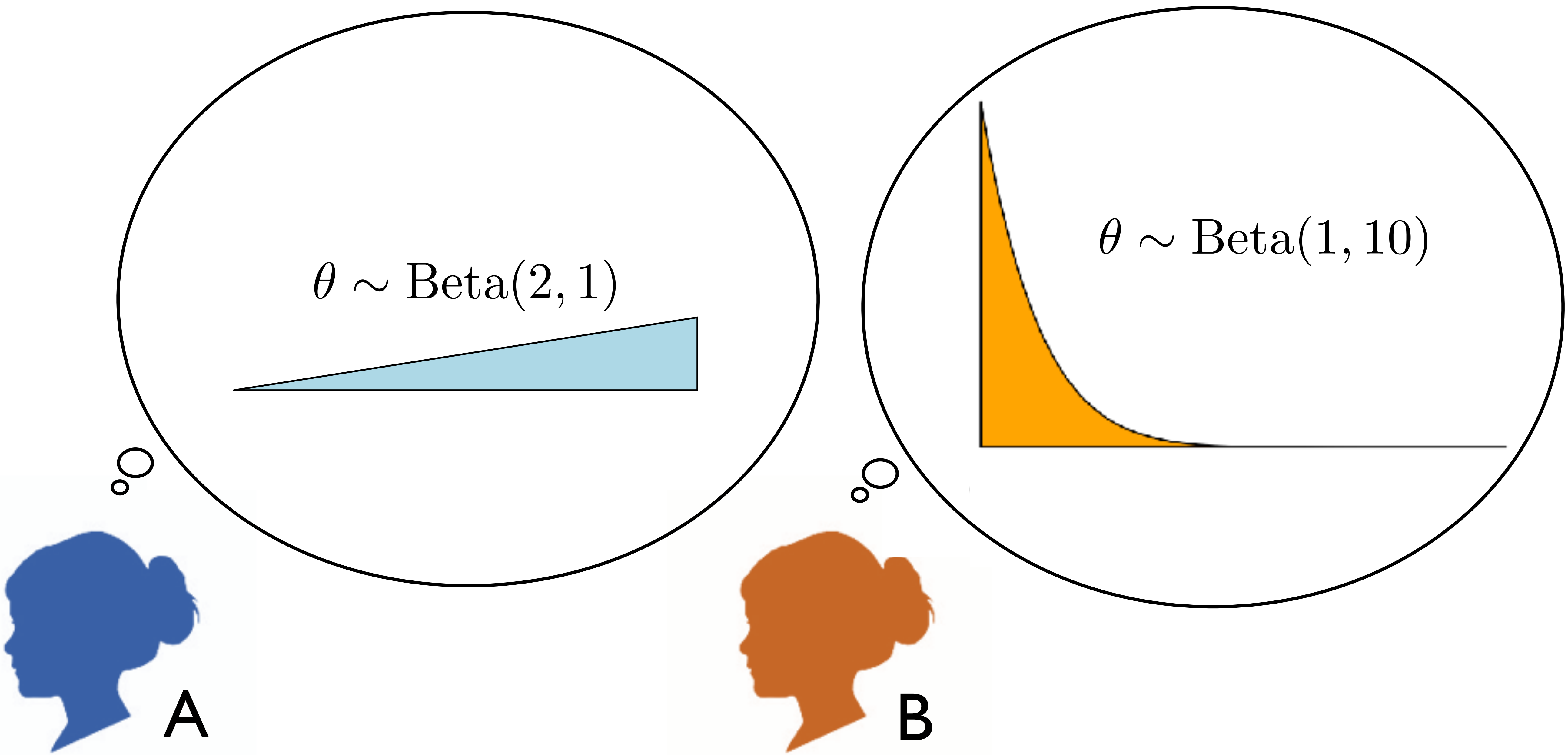
$y \sim p(y|\theta)$

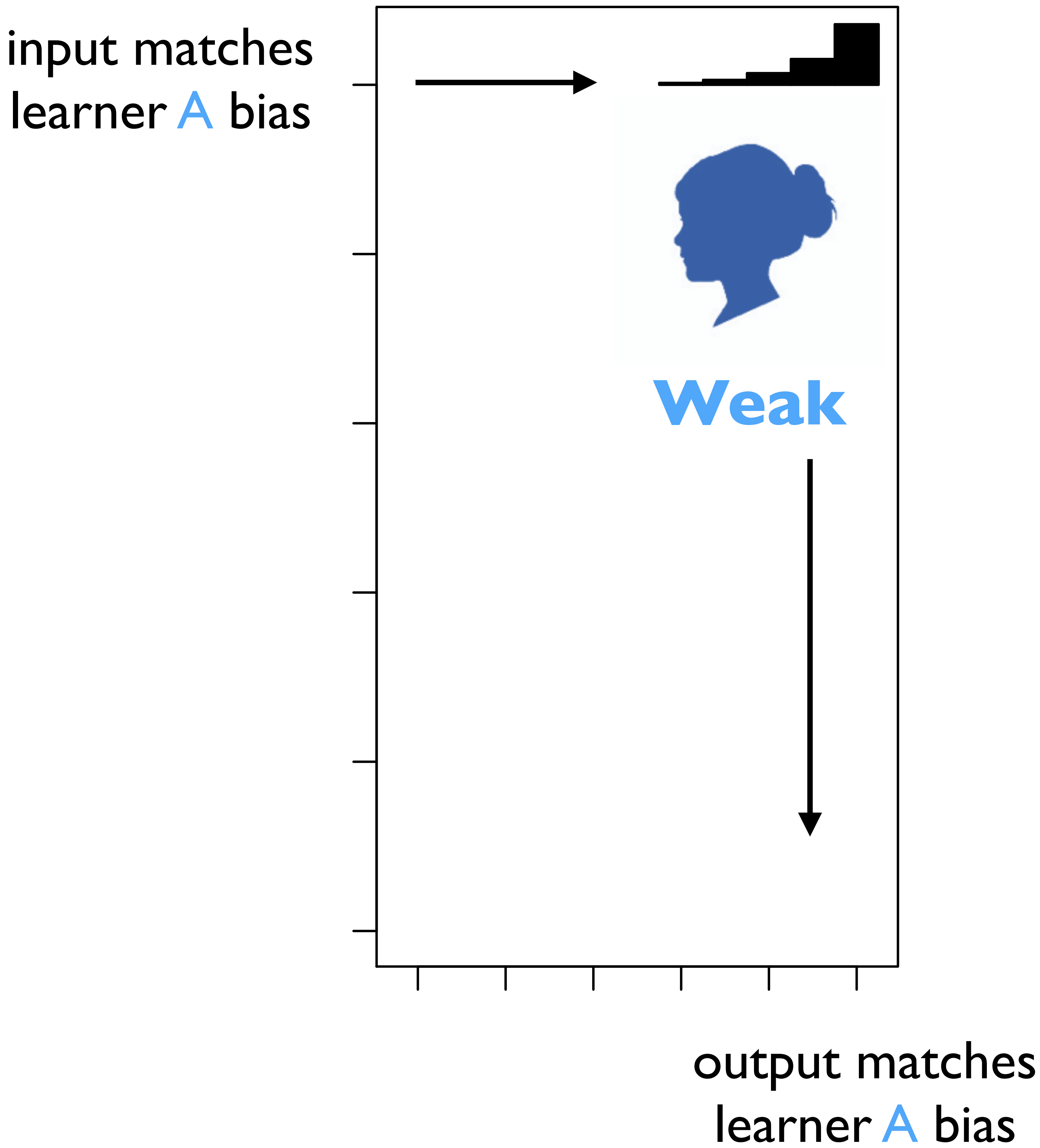
Some learners use a prior that imposes a weak bias

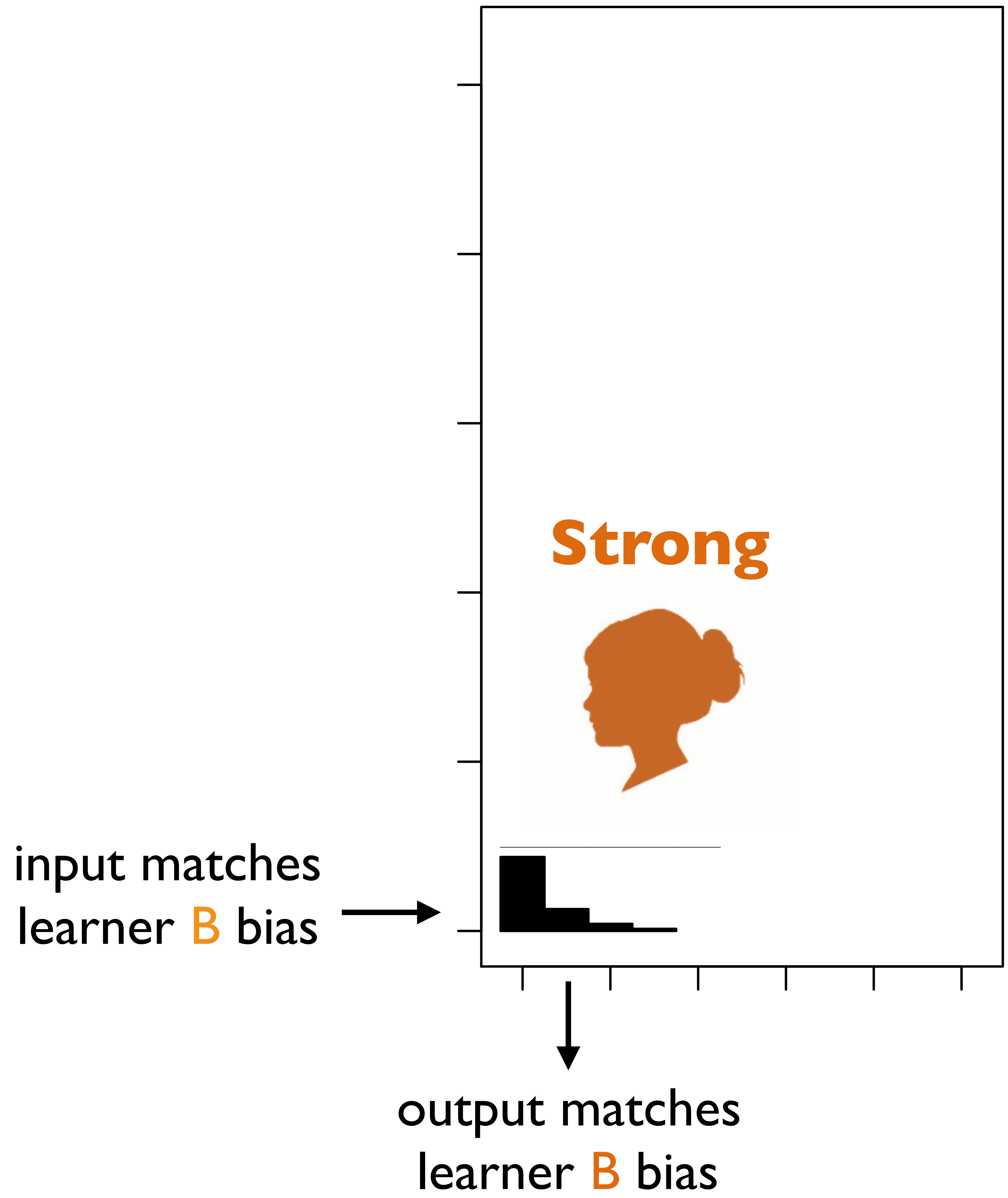


Some learners use a prior that imposes a weak bias

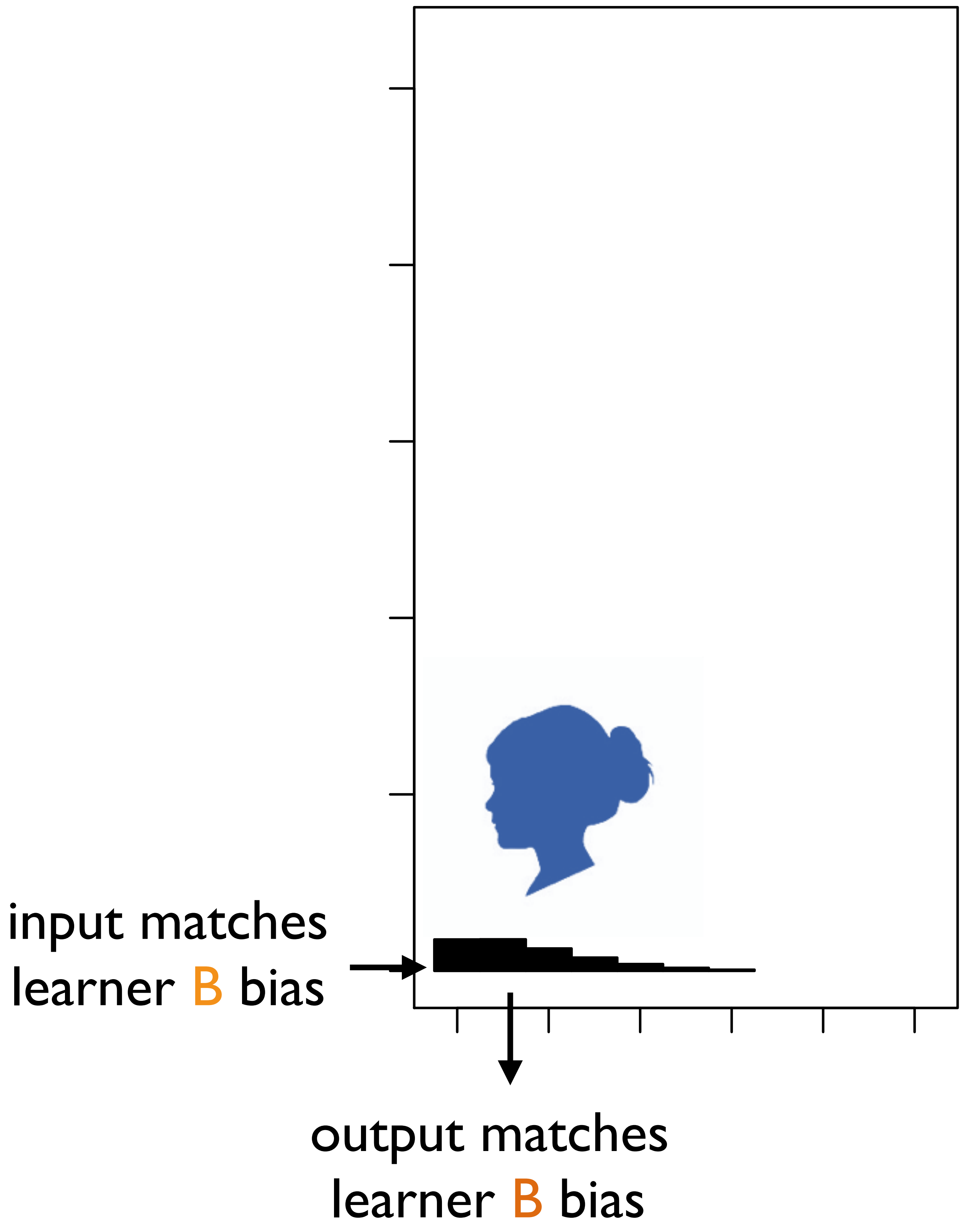
Some learners use a prior that imposes a strong bias




input matches
learner A bias
Weak
output matches
learner A bias

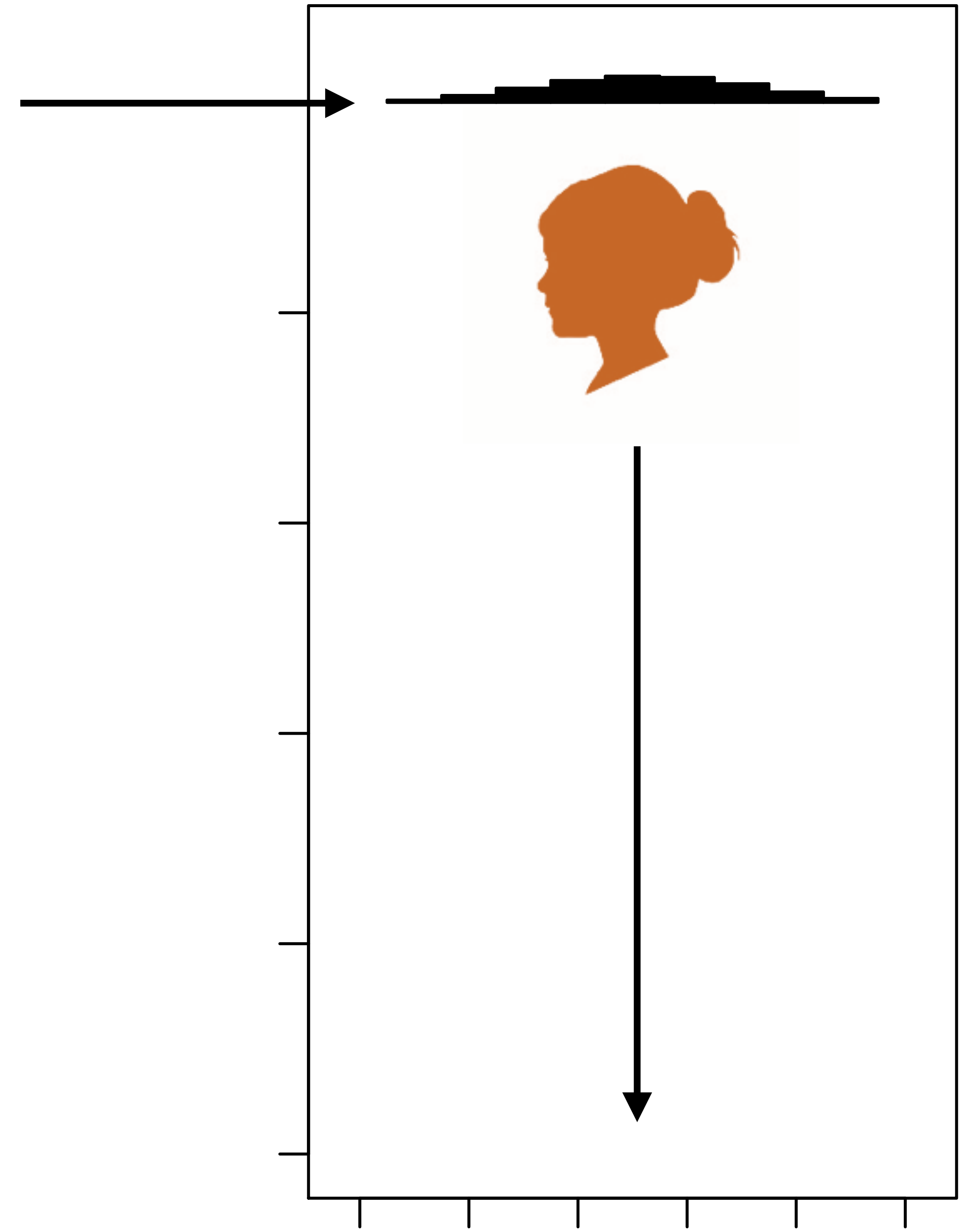


Strong
input matches
learner B bias
output matches
learner B bias



Learners with weak biases tend to mirror input even when it disagrees with the learner bias

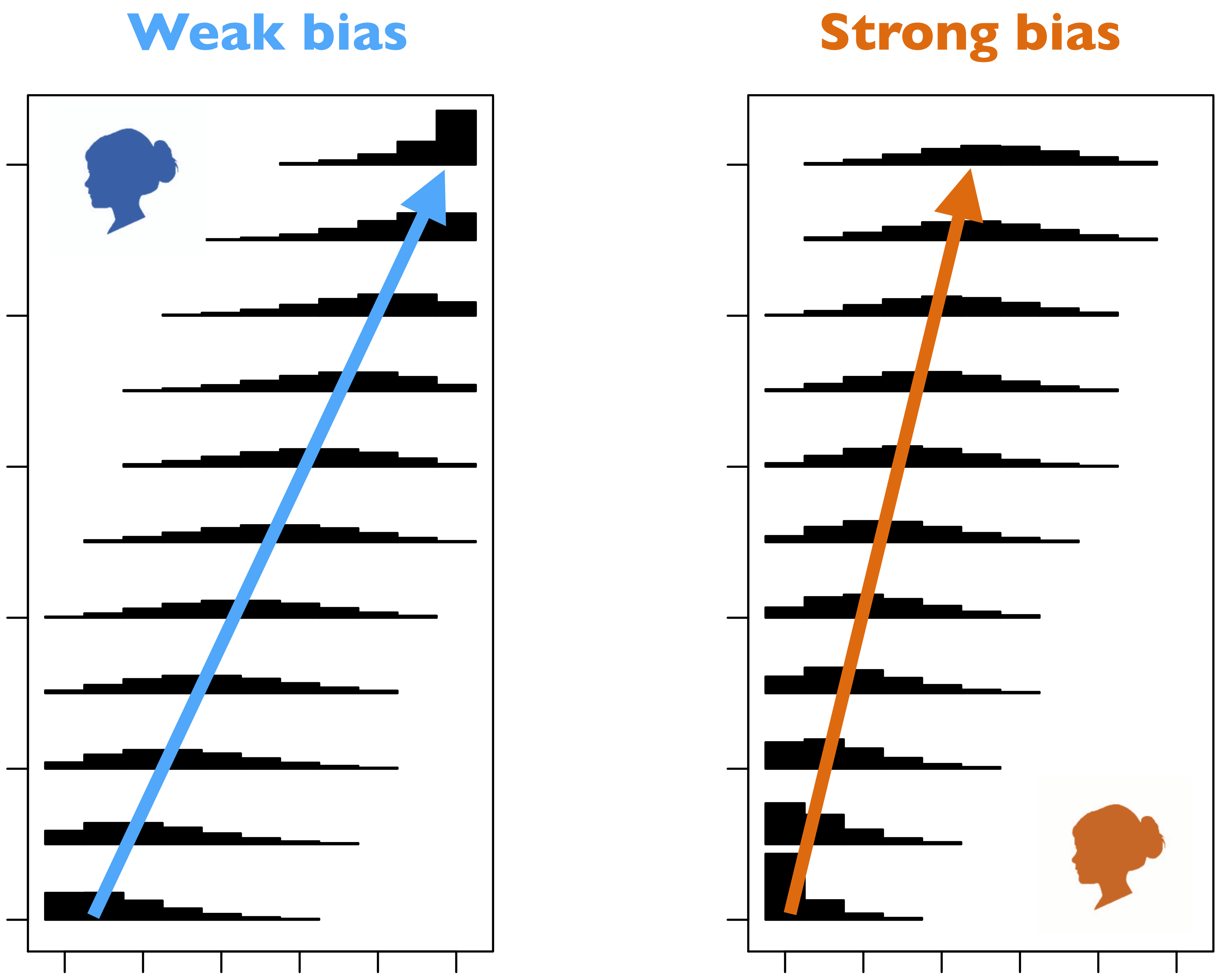
input matches
learner A bias



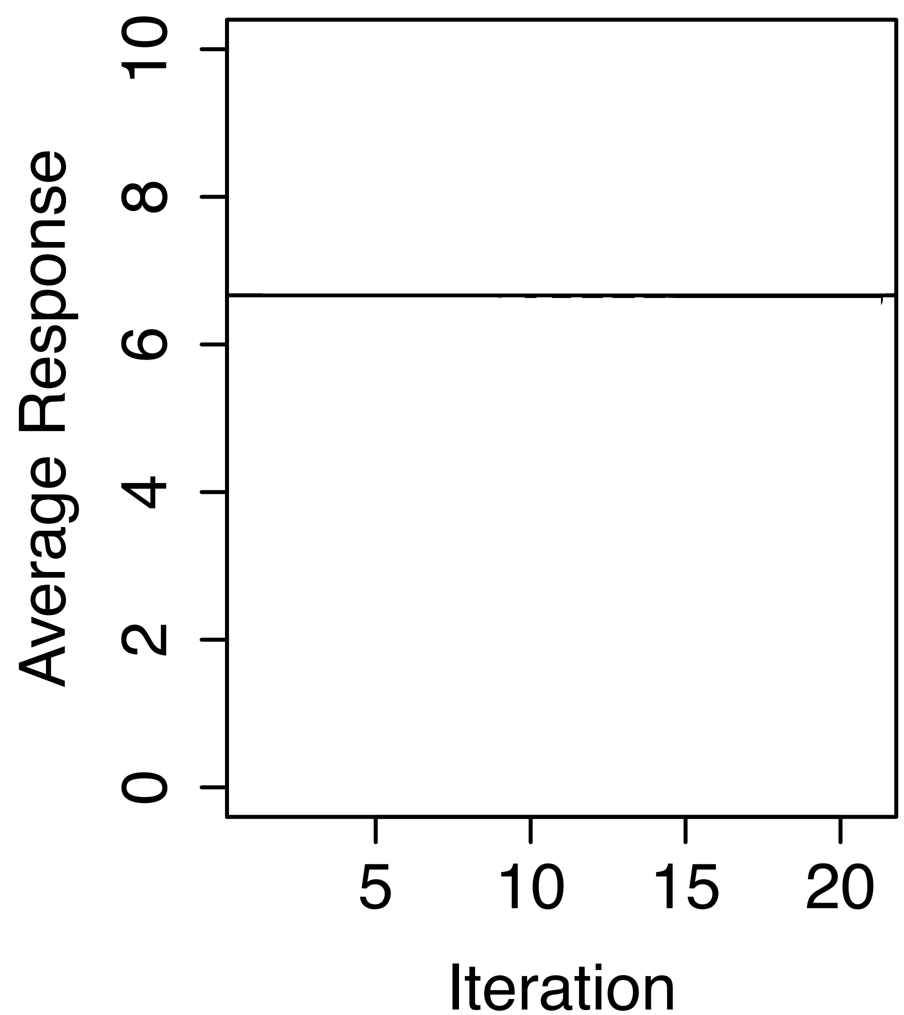
Learners with strong biases do not:

They (partially) impose their own biases

output is a compromise between learner B bias and the input

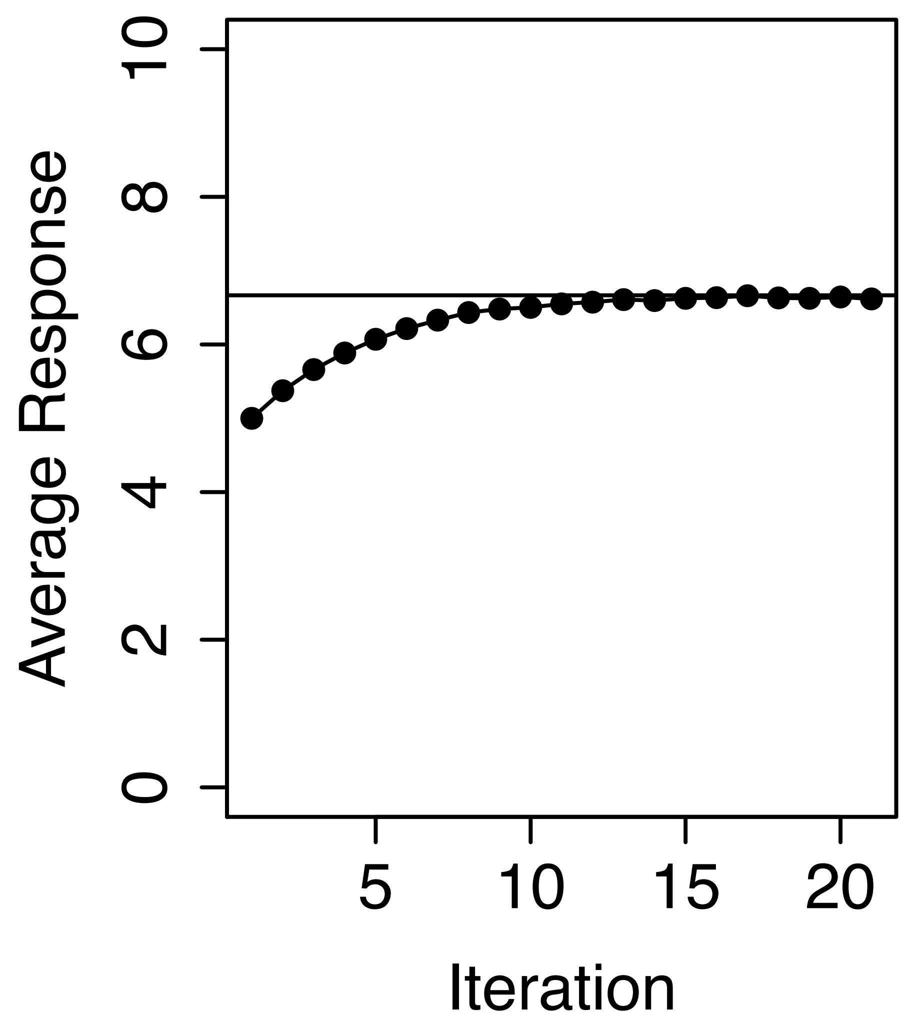

Weak bias
Strong bias

Weak bias

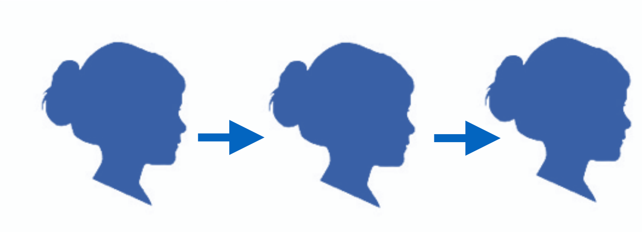


Homogenous population with weak bias

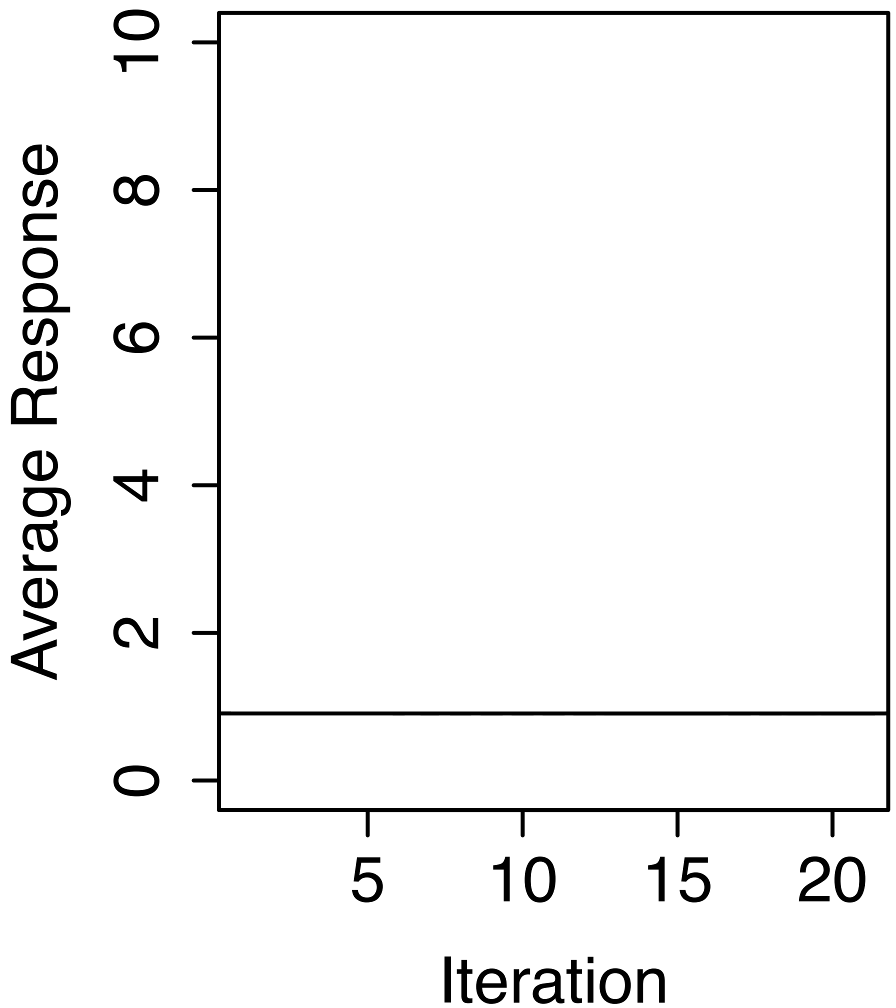
Weak bias



Iterated learning chain converges to the prior

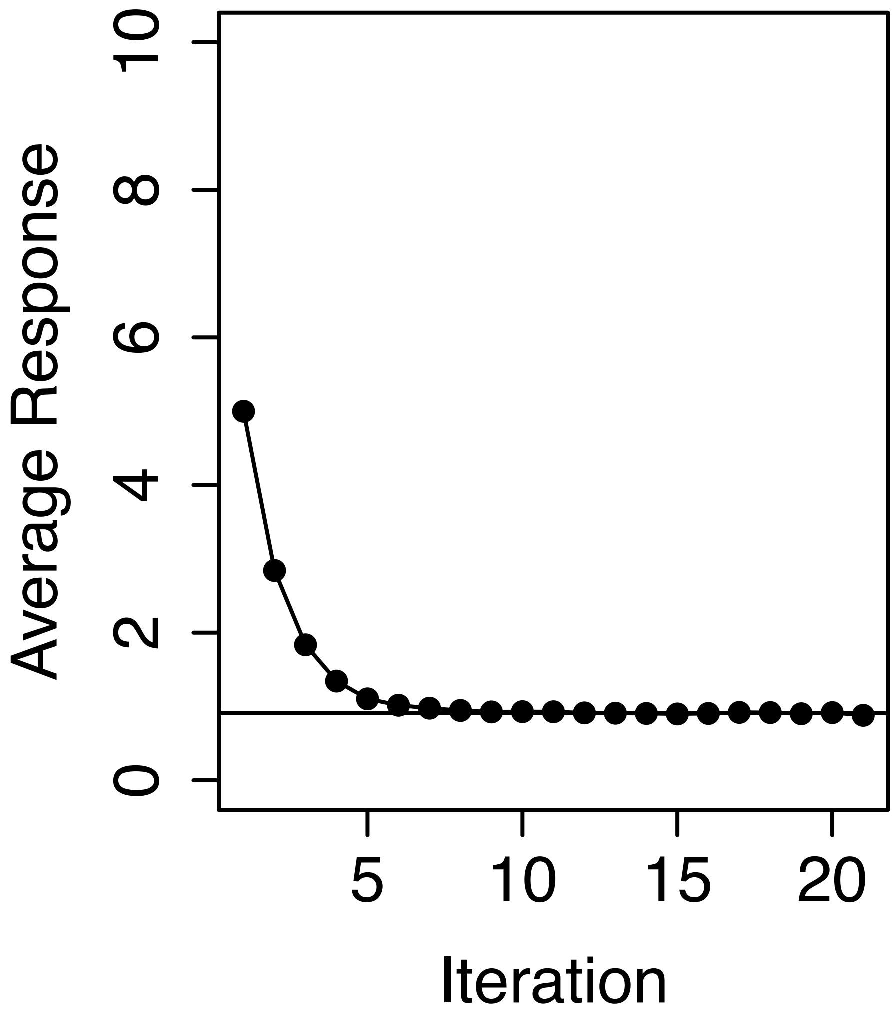


Strong bias

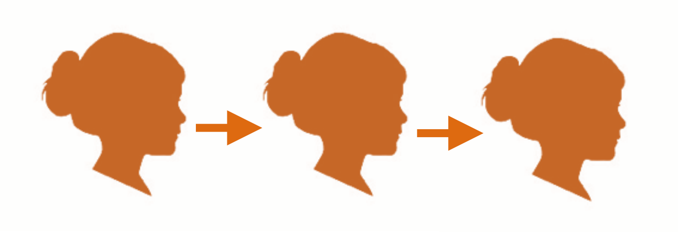


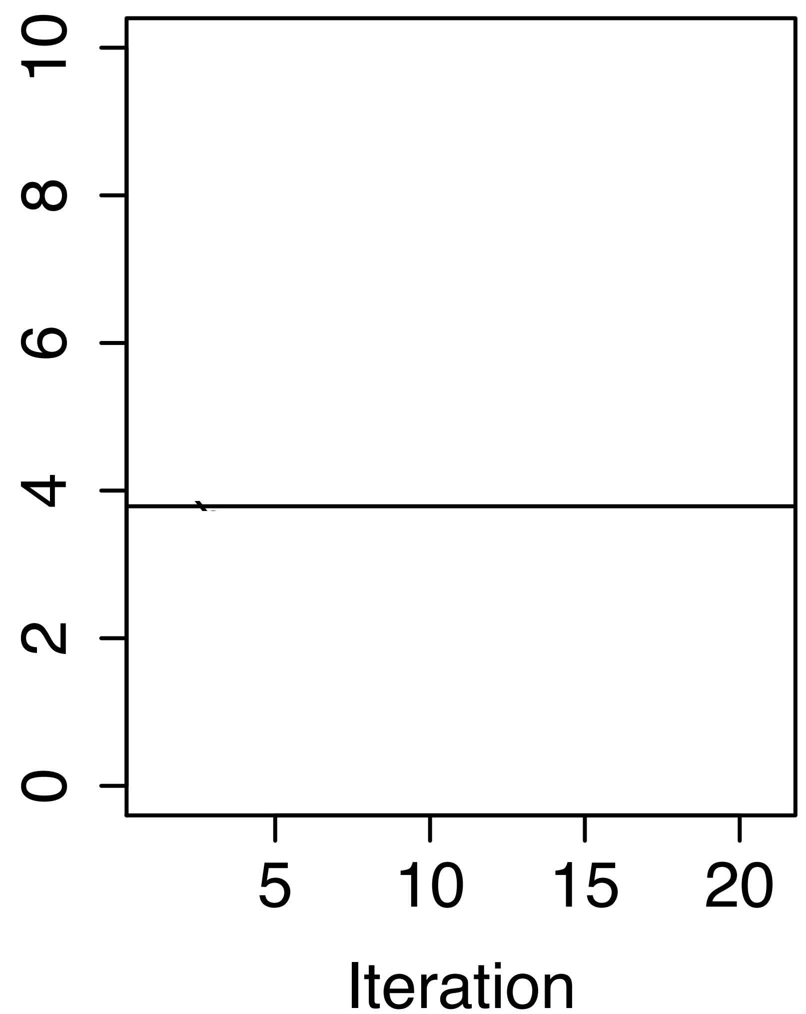
Homogenous population with strong bias

Strong bias

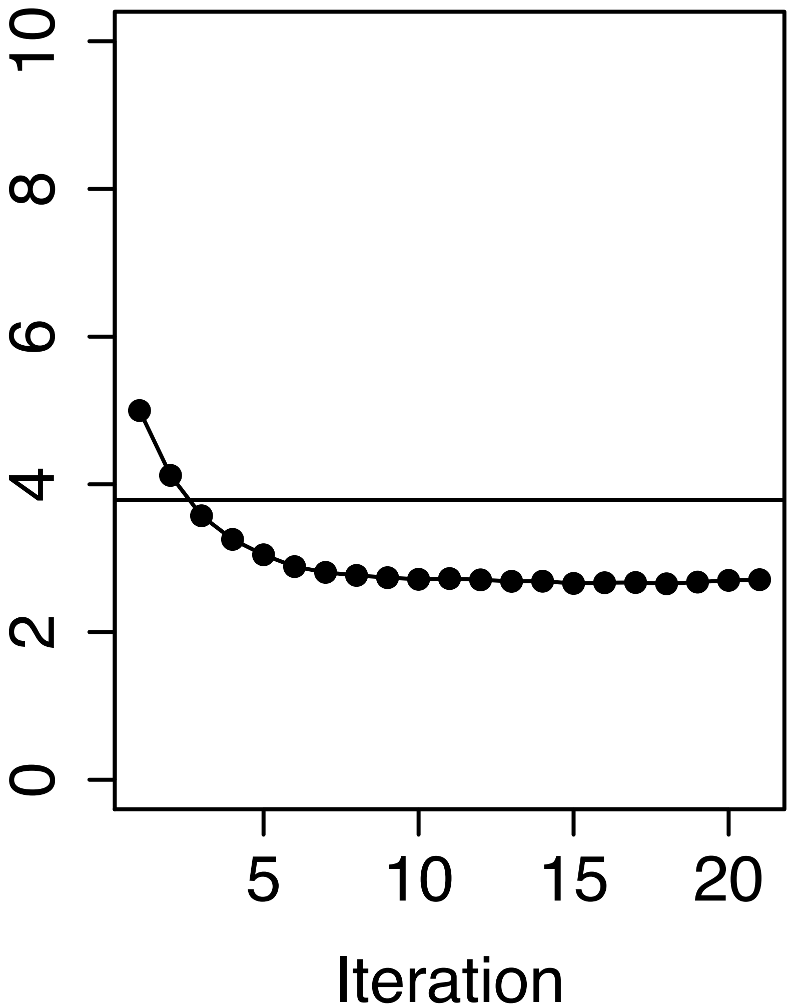


Iterated learning chain converges to the prior

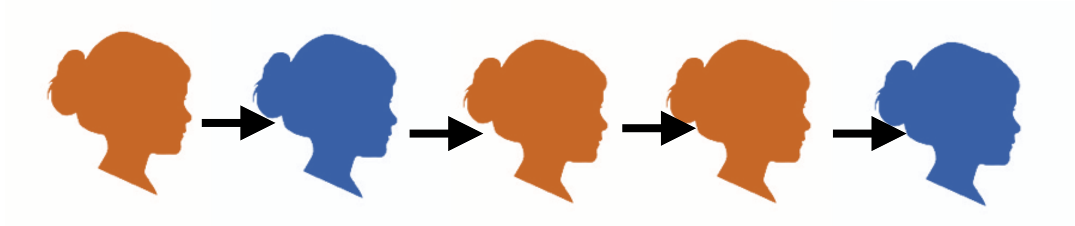


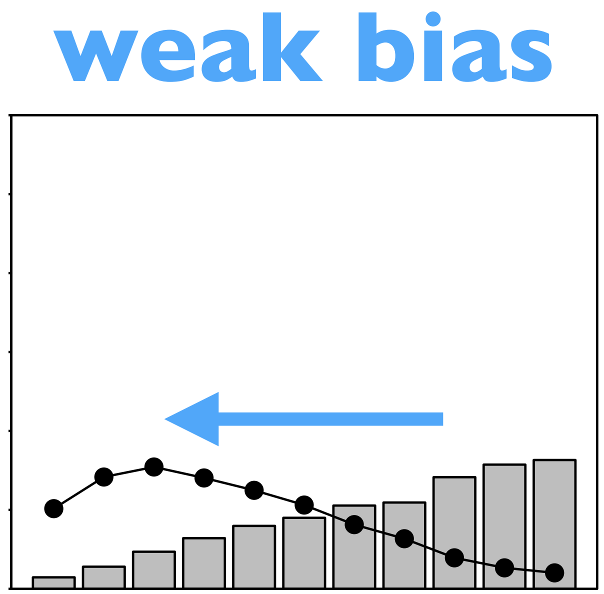


Heterogenous population with equal proportions of both learner types

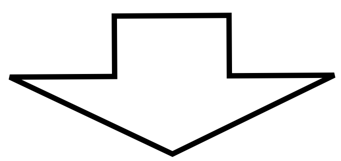


Mixed chain does not converge to the prior



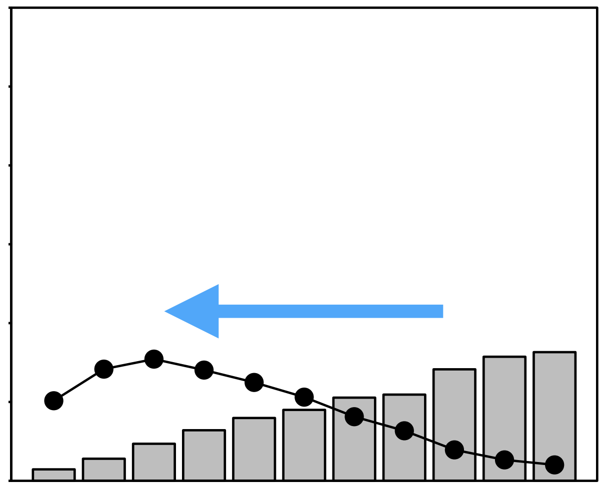


weak bias

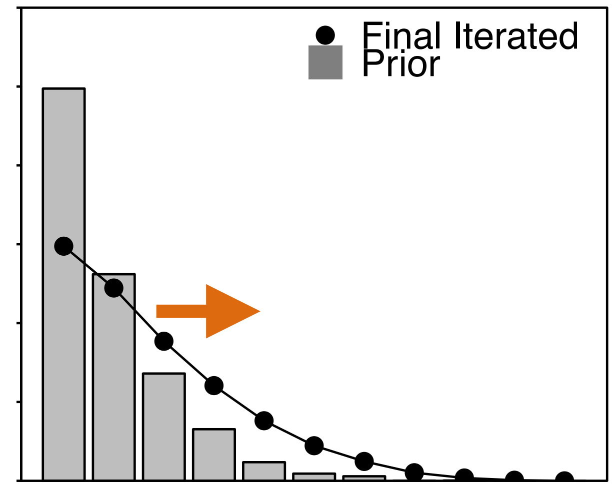


very responsive to input

weak bias



strong bias

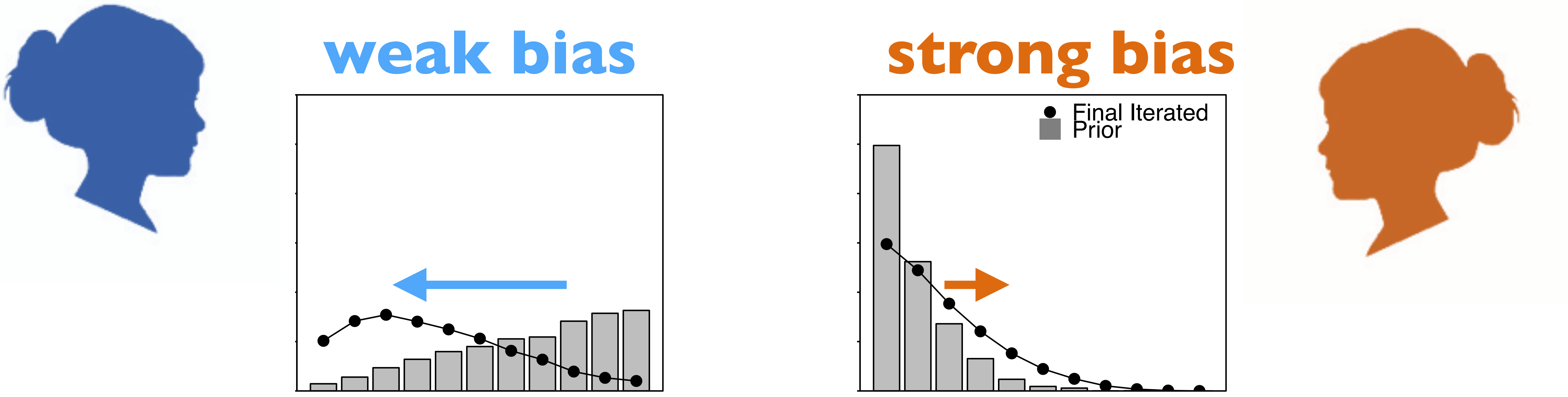


weak bias

very responsive to input

strong bias

insensitivity to input



weak bias

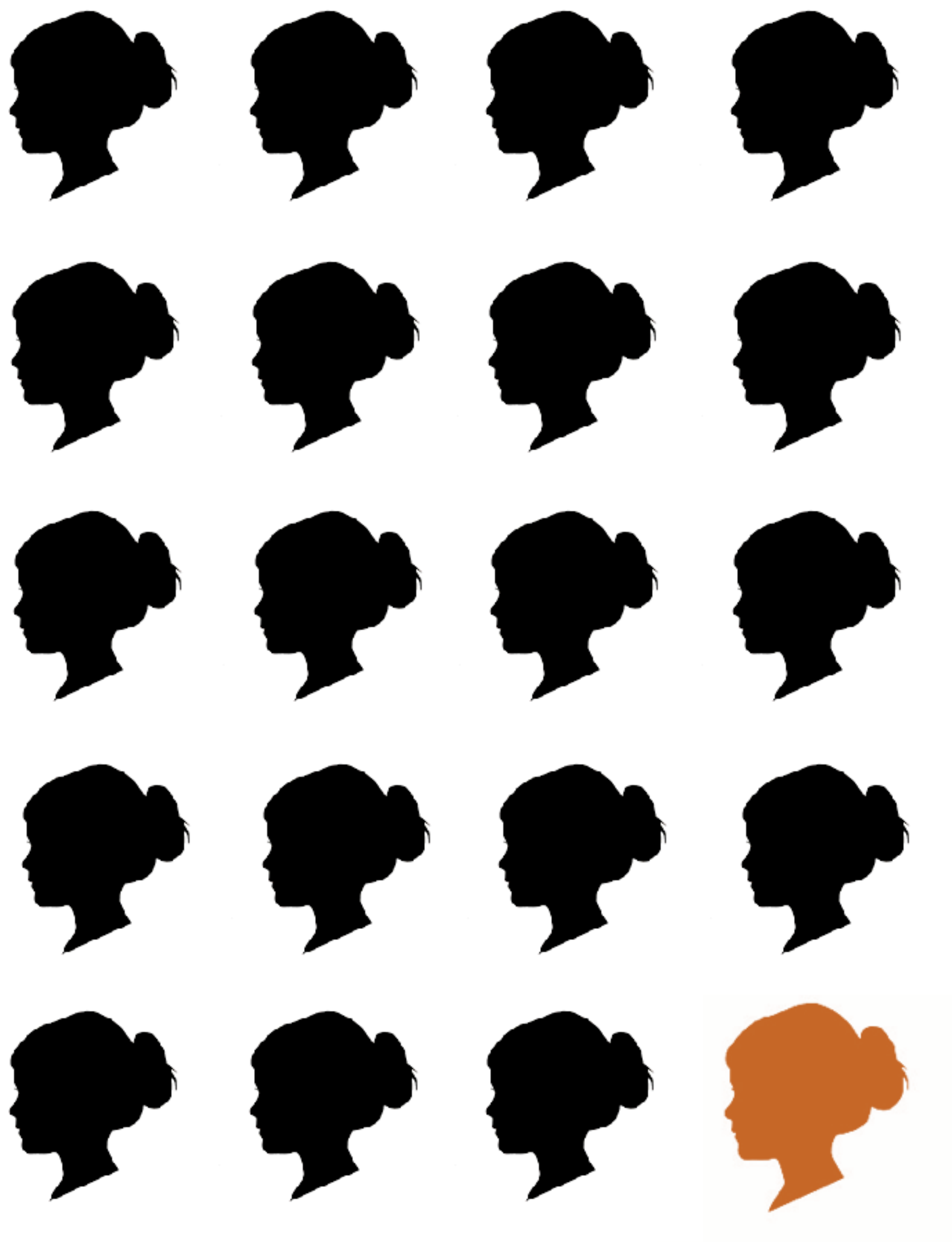
very responsive to input

small influence on
the chain

strong bias

insensitivity to input

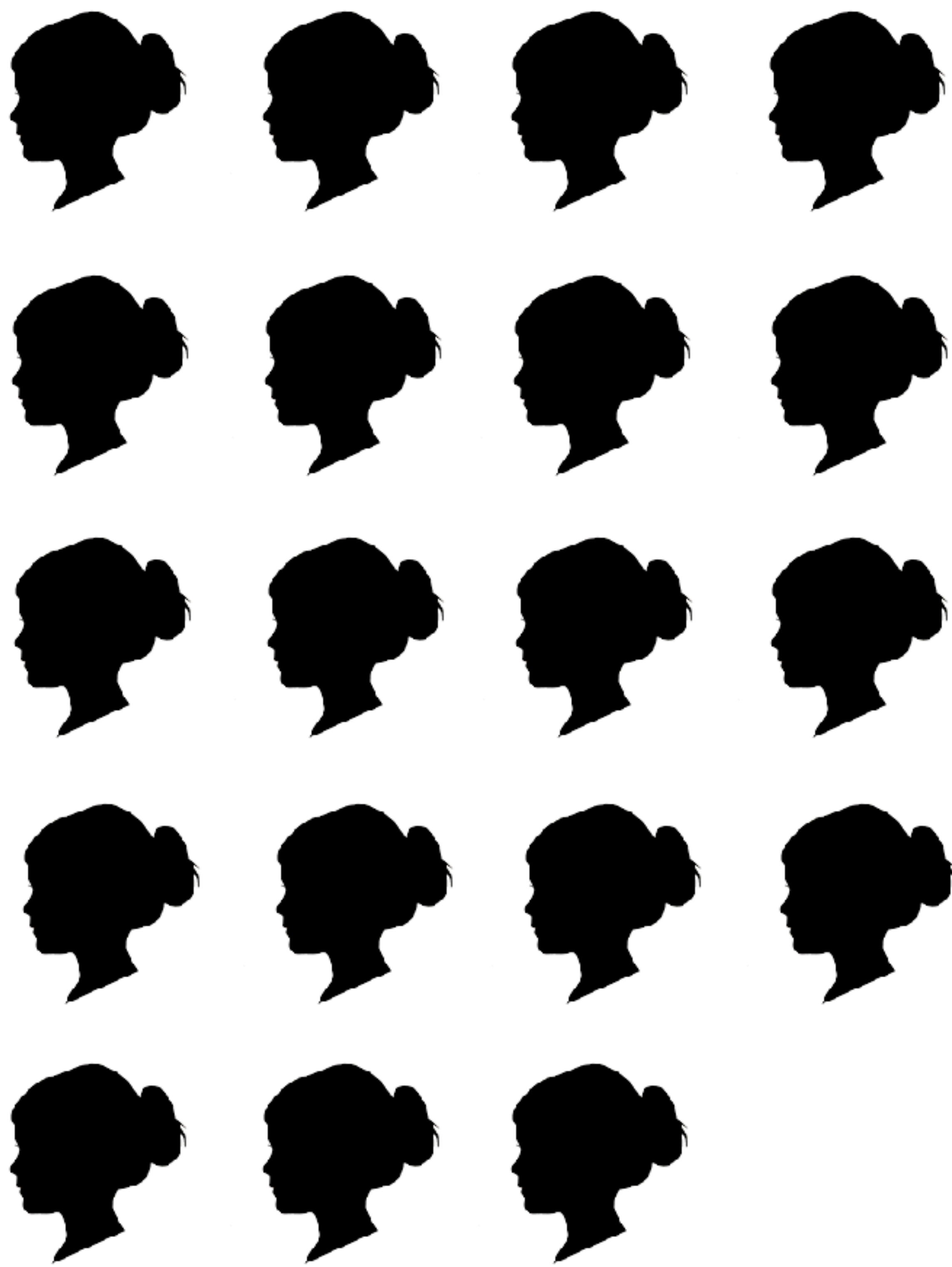
greater influence on
the chain

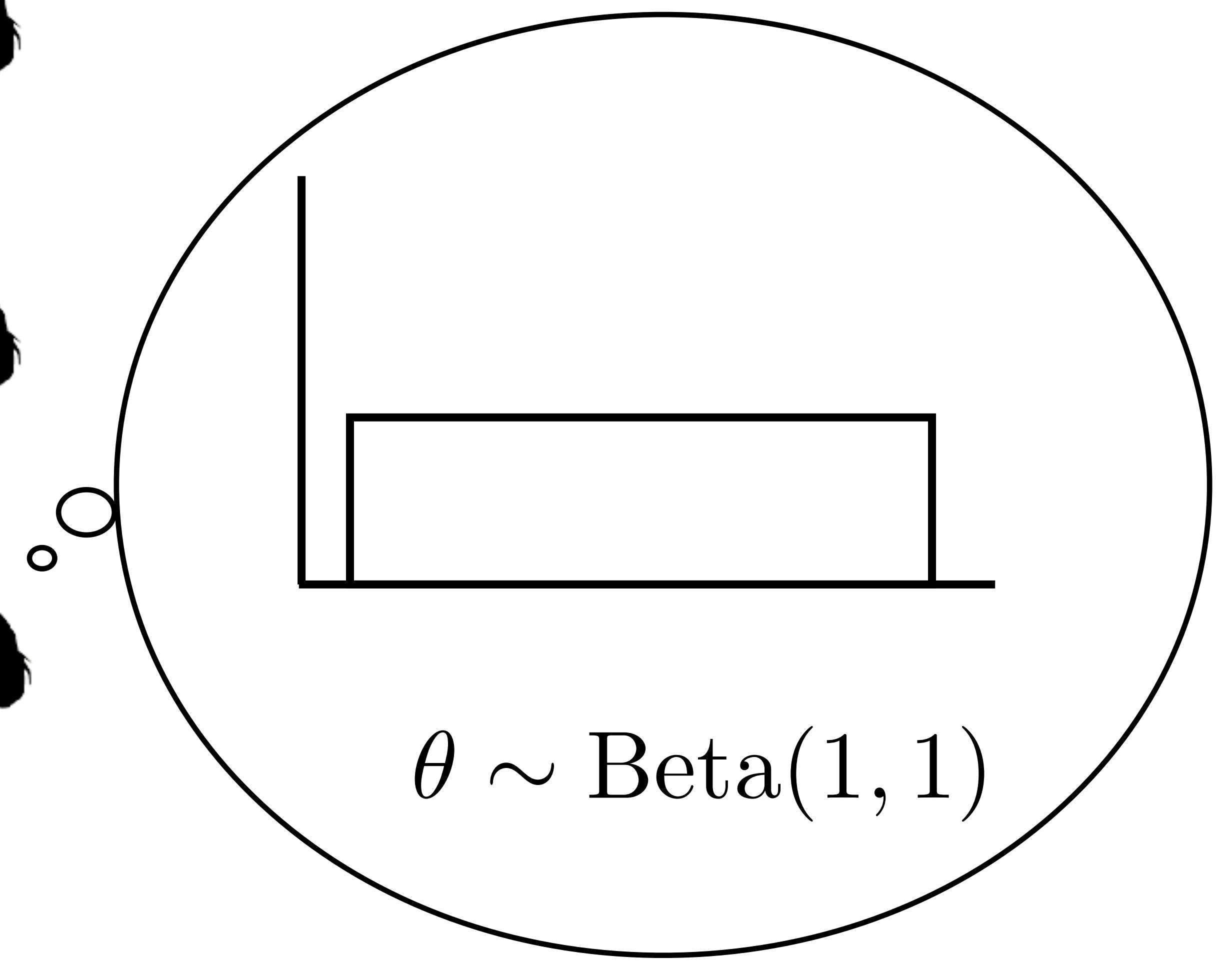


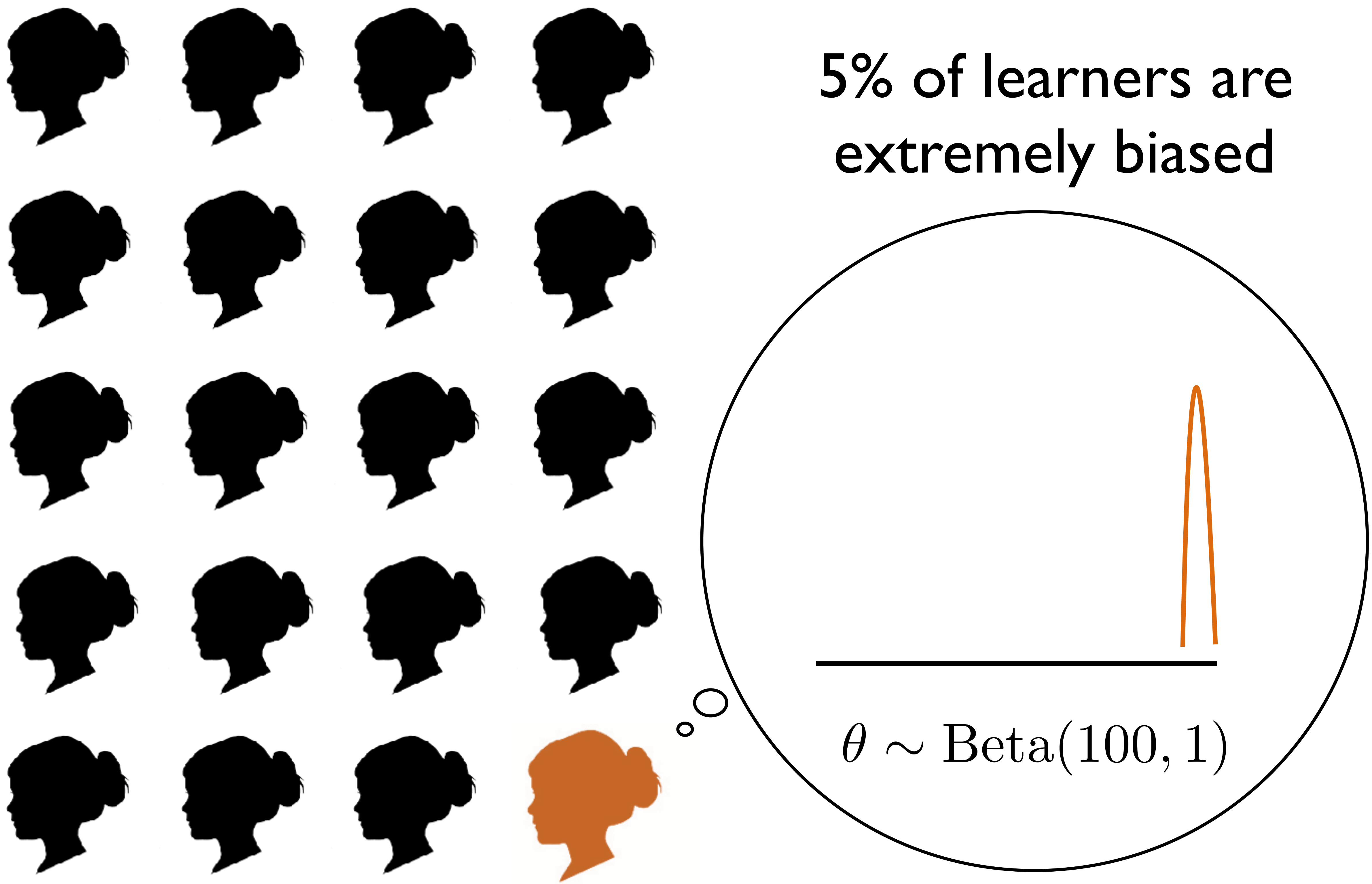
How much influence can a strong bias confer?

An extreme example

95% of learners are unbiased

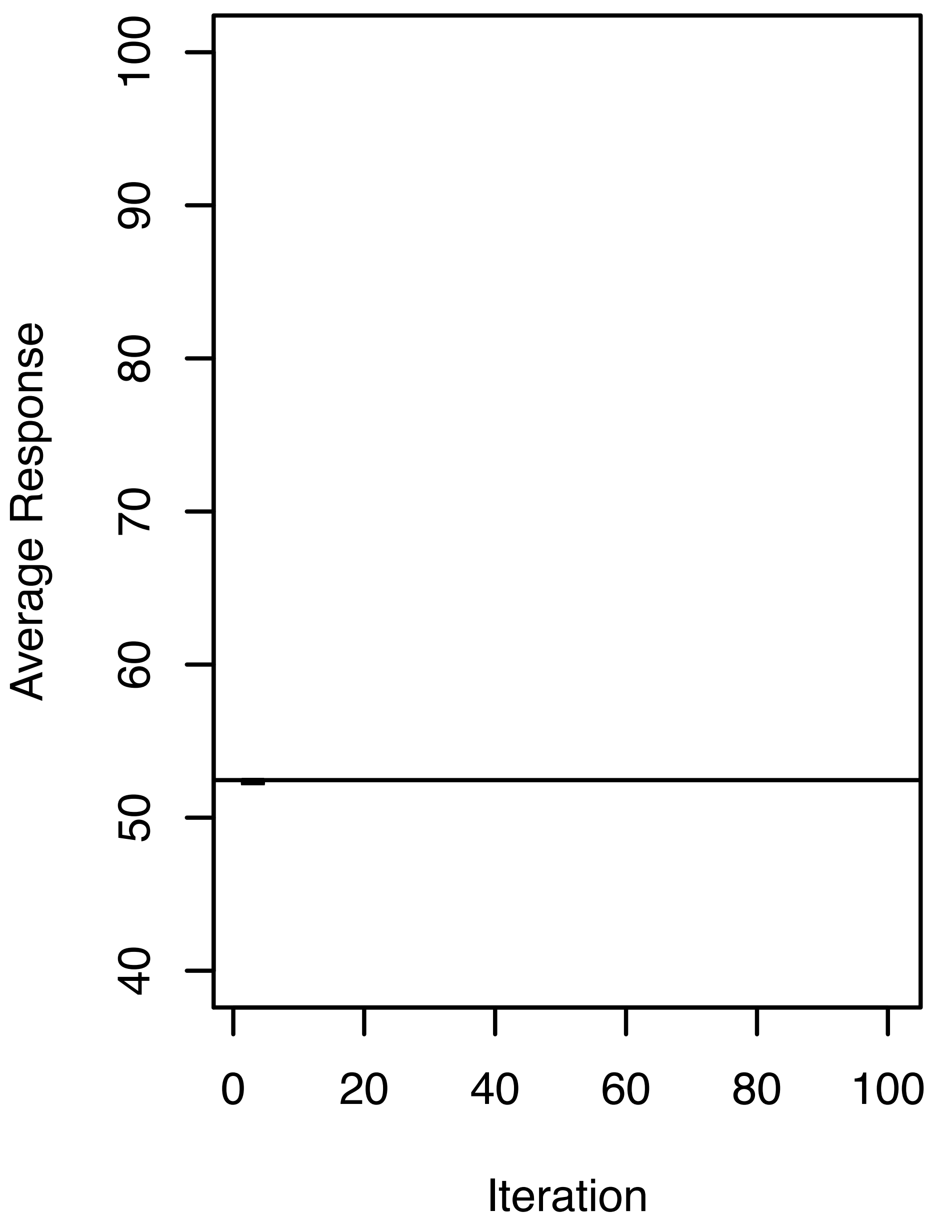


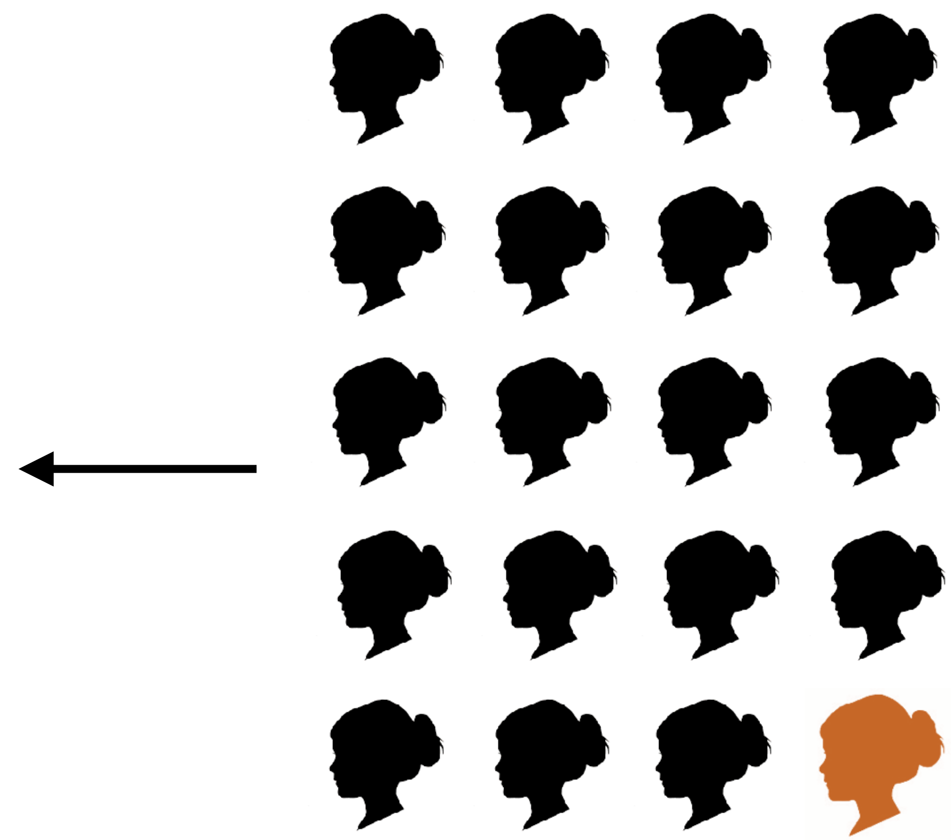




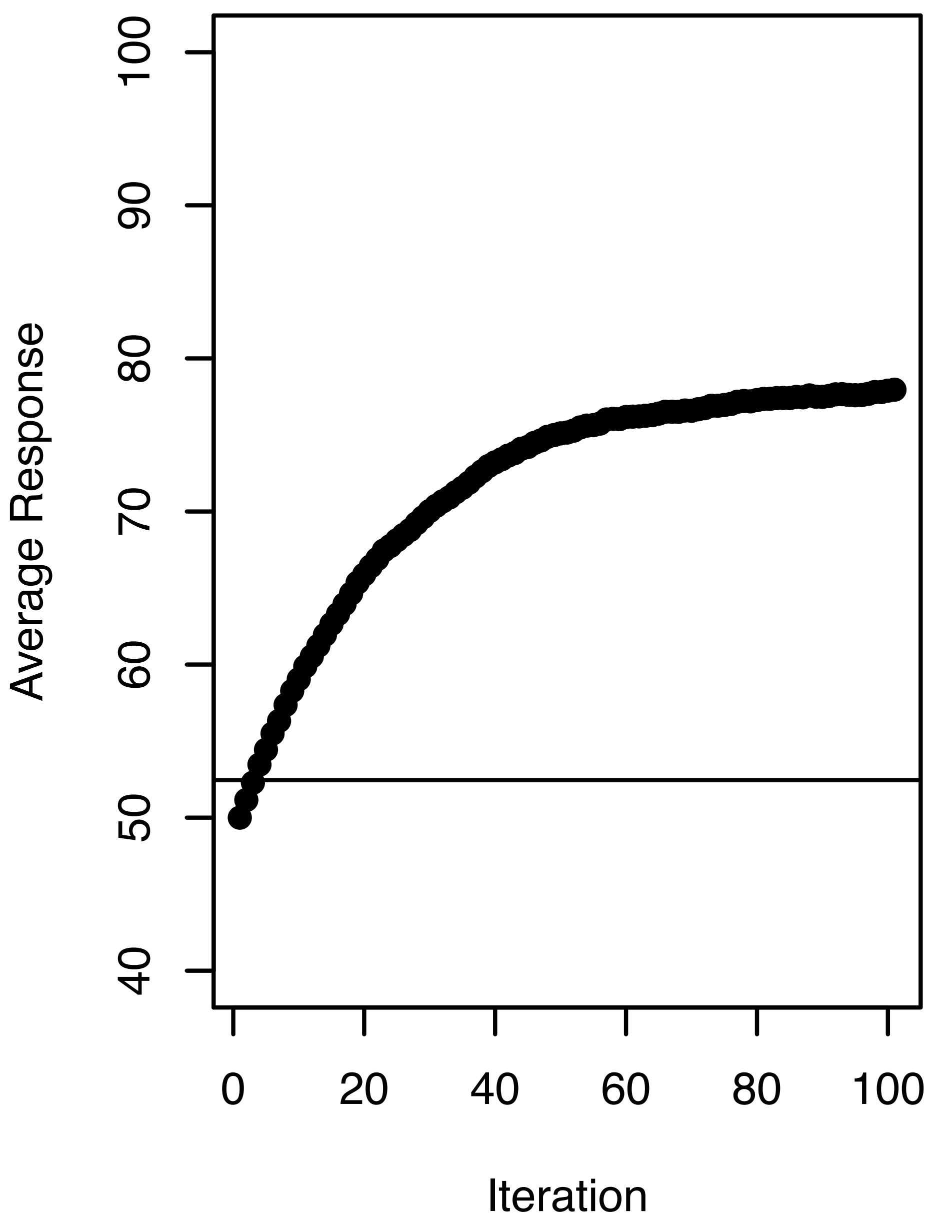
5% of learners are extremely biased
$\theta \sim \mathrm{Beta}(100, 1)$

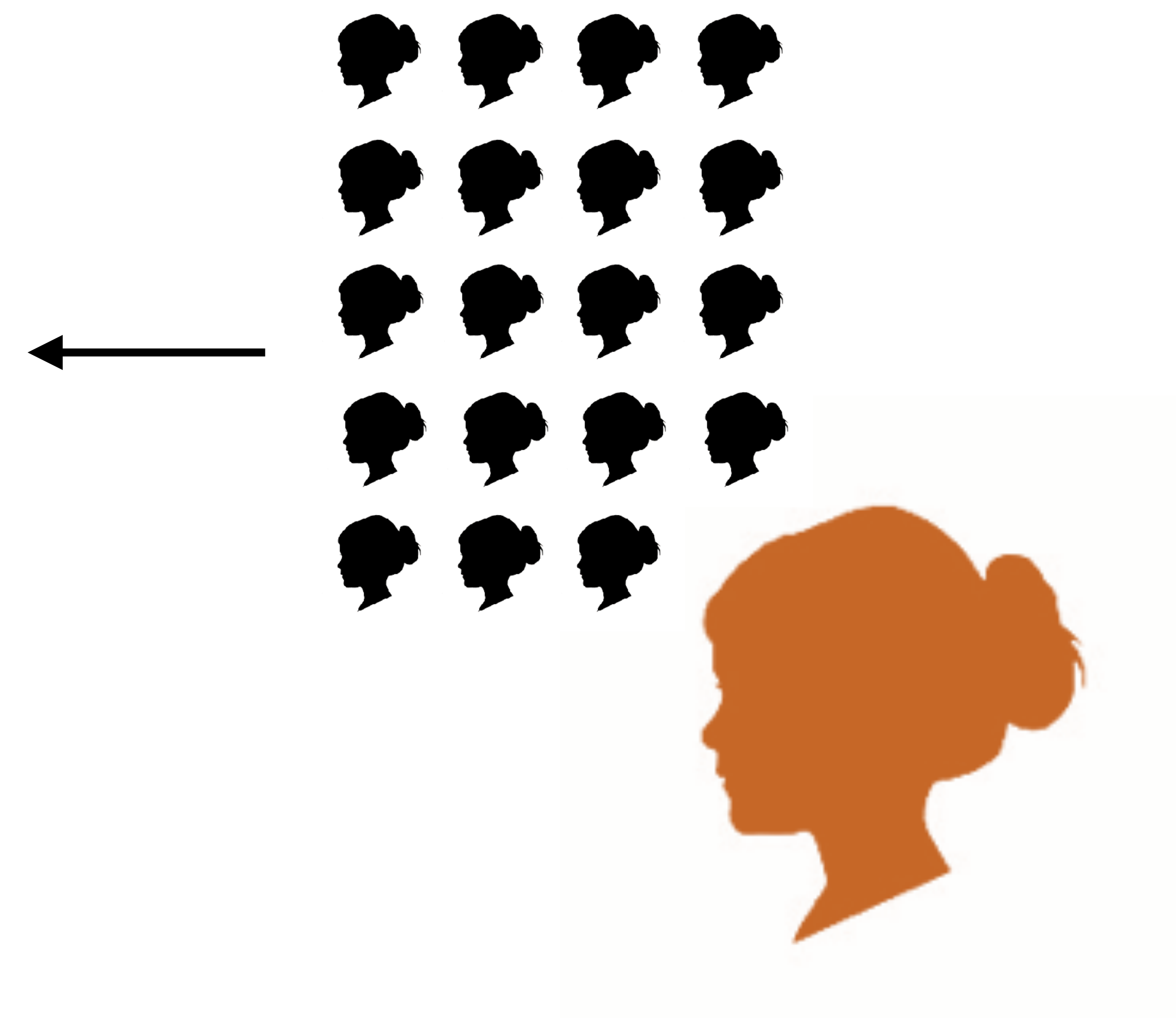
The average response if everyone samples from their prior





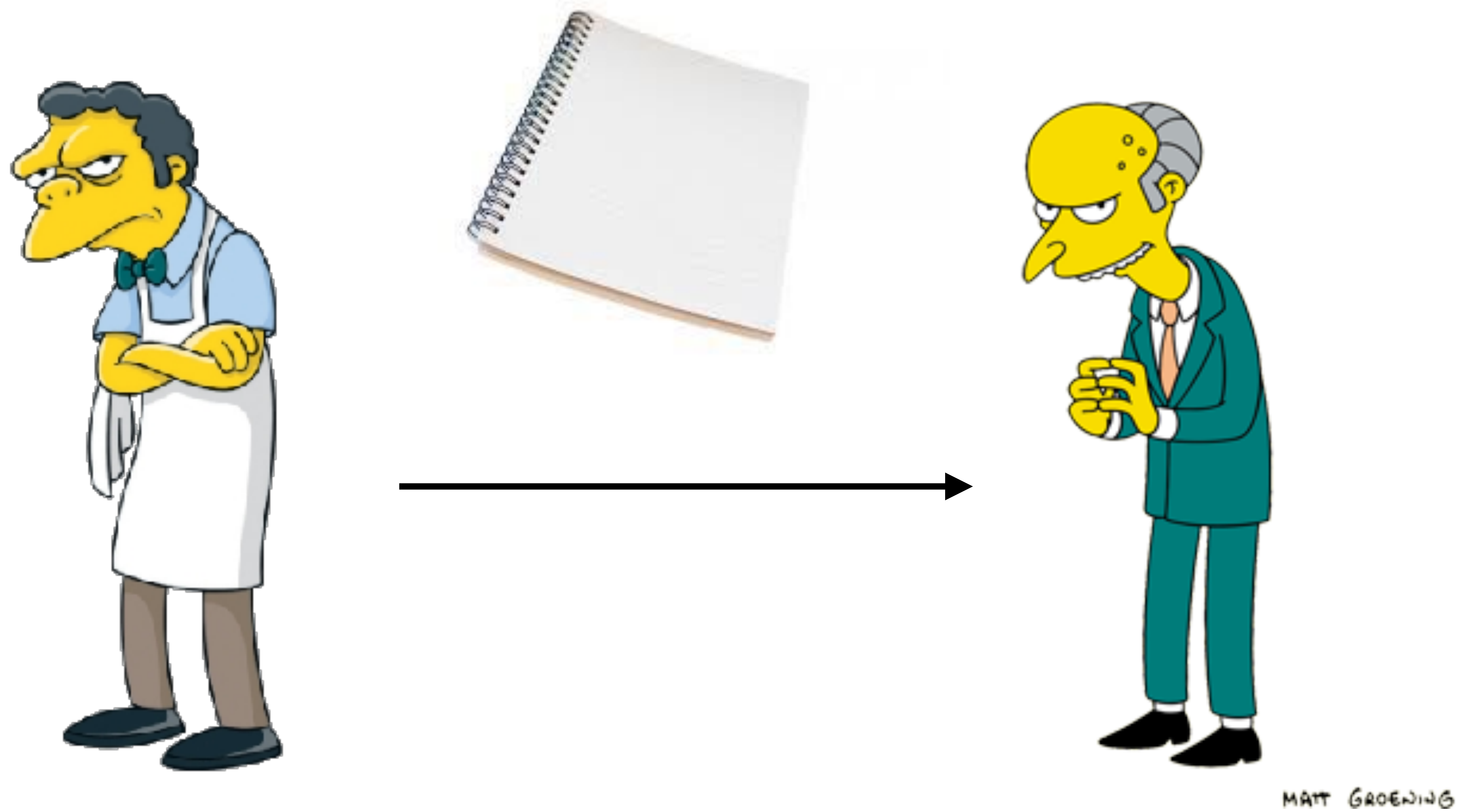
Iterated learning chain is dominated by the extreme bias learners





Case study 2: How to induce Bayesian groupthink

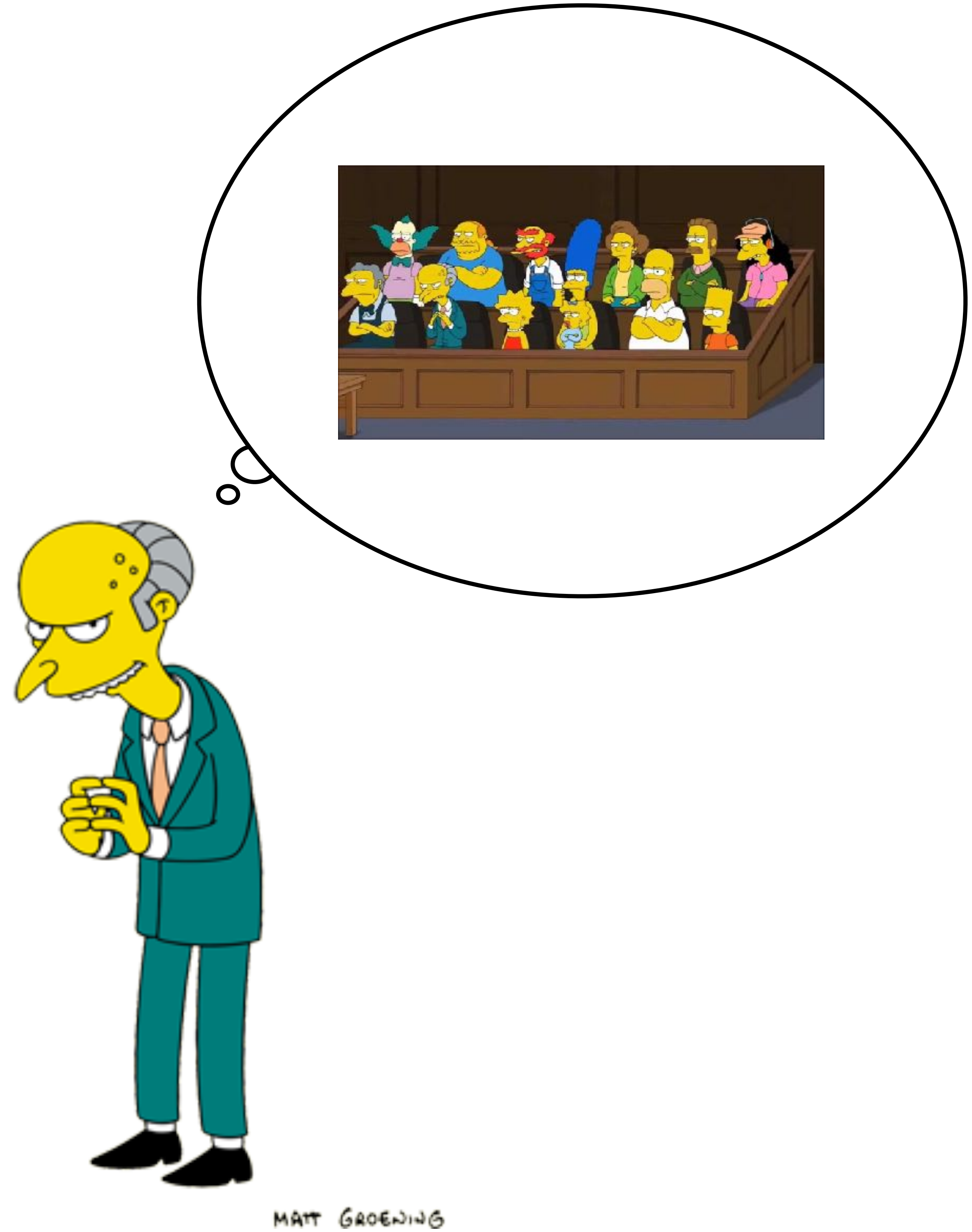
Juror i records vote, removes sheet, passes notebook



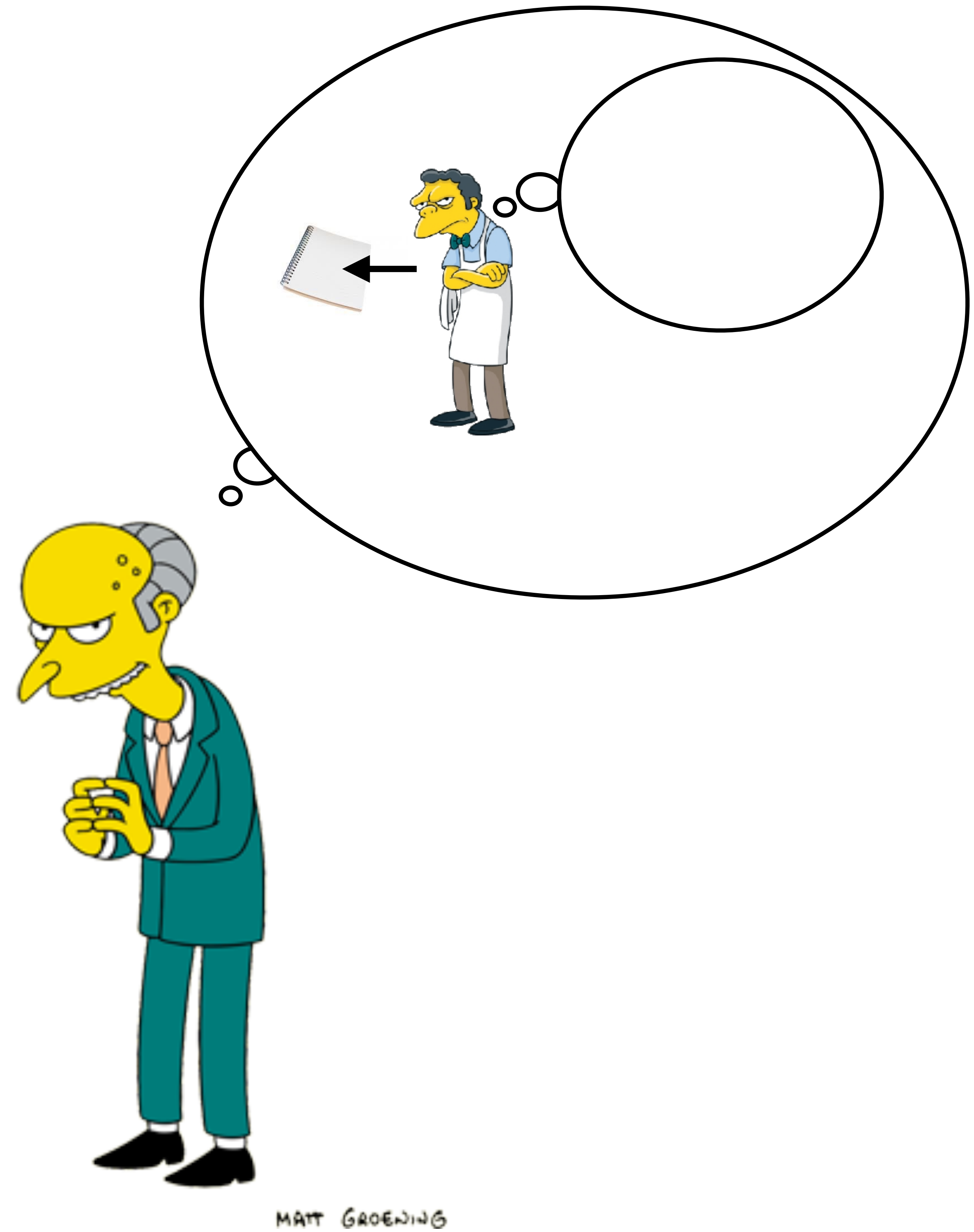
Juror i records vote, removes sheet, passes notebook

Juror $i+1$ can see the previous vote via indentations…

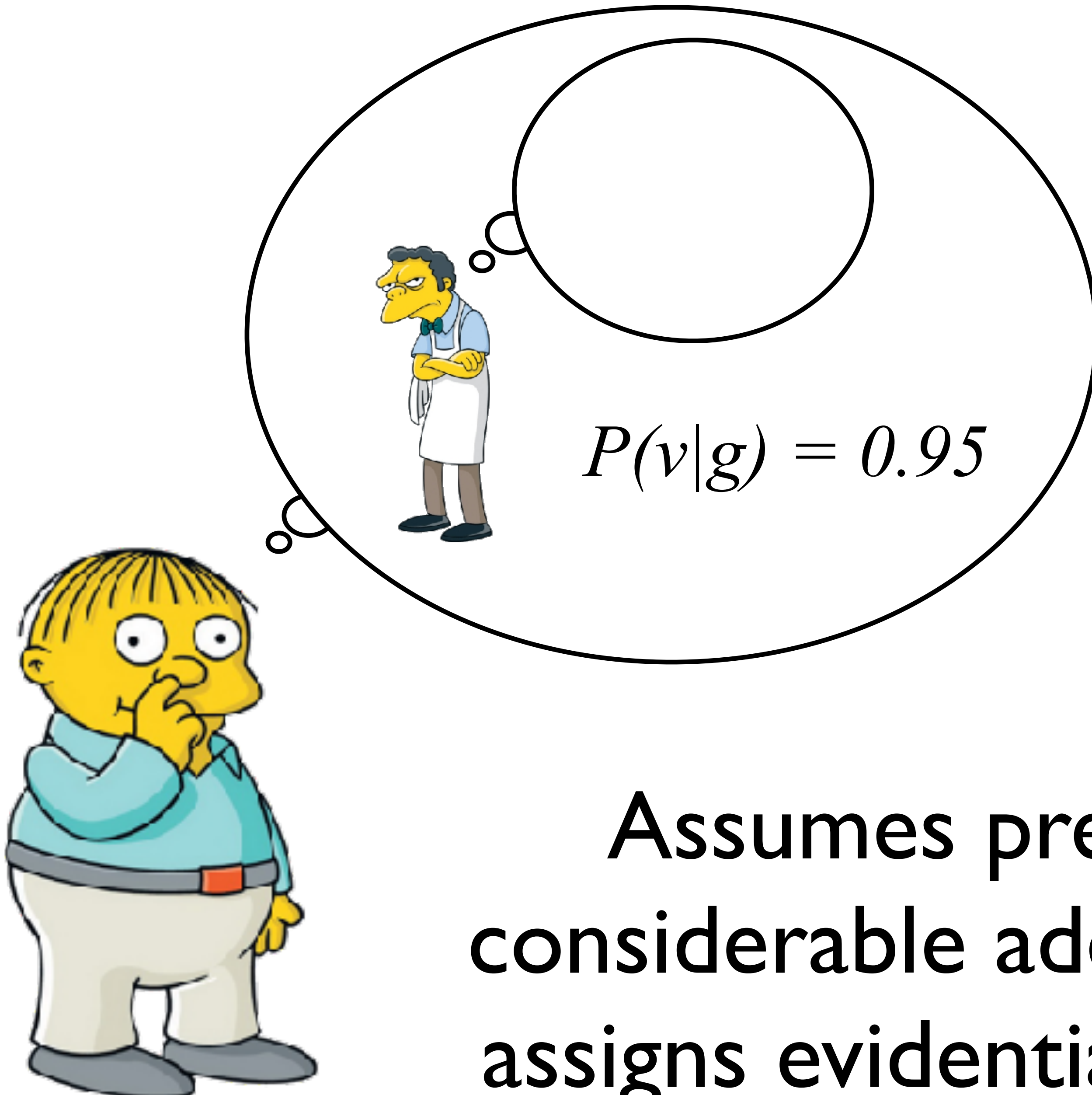
Prior belief about guilt
$P(g)$ is set by the trial



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?

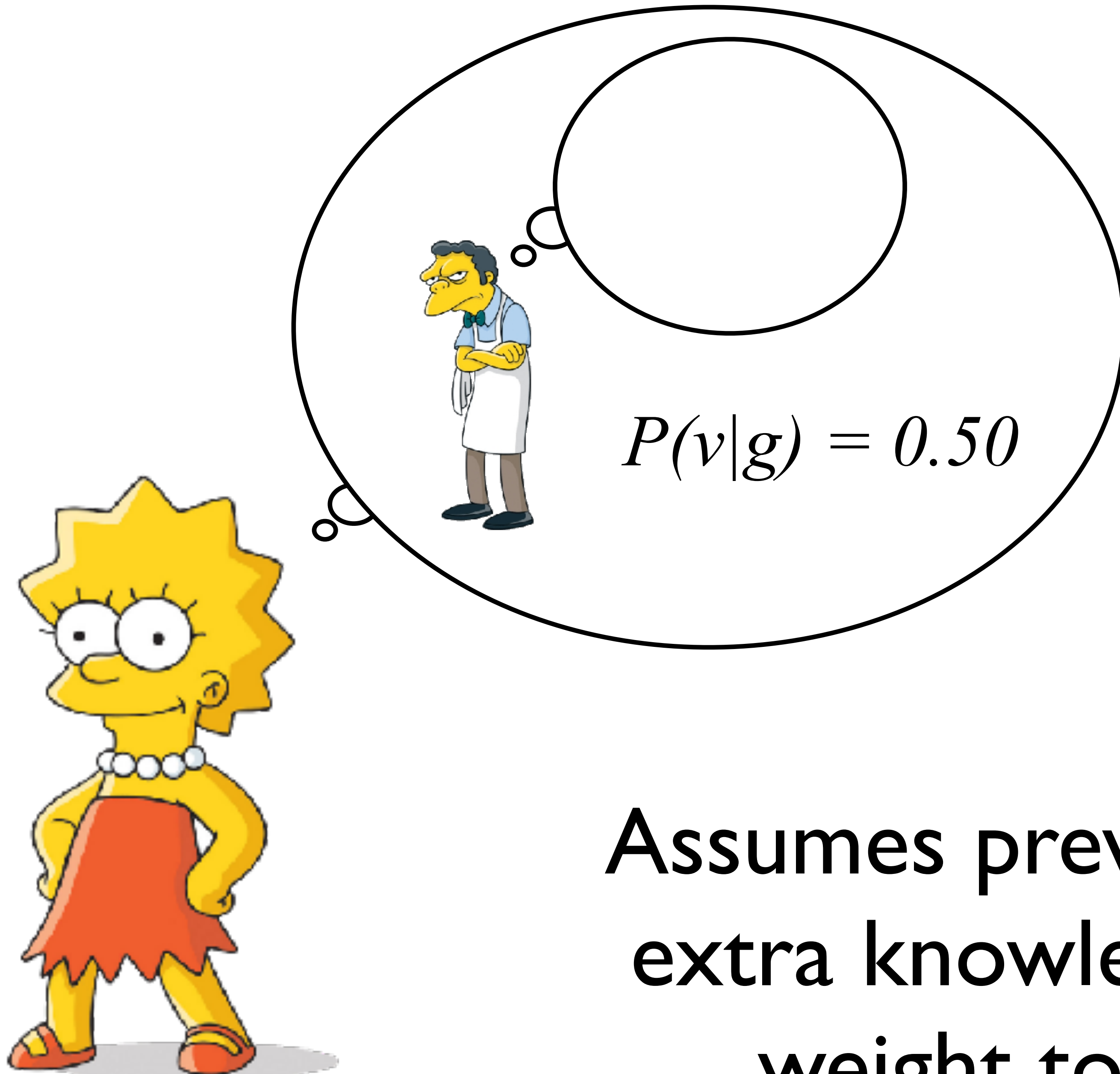


Bayesian “sheep”

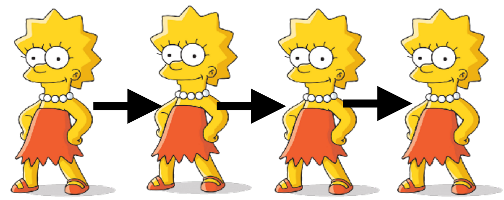


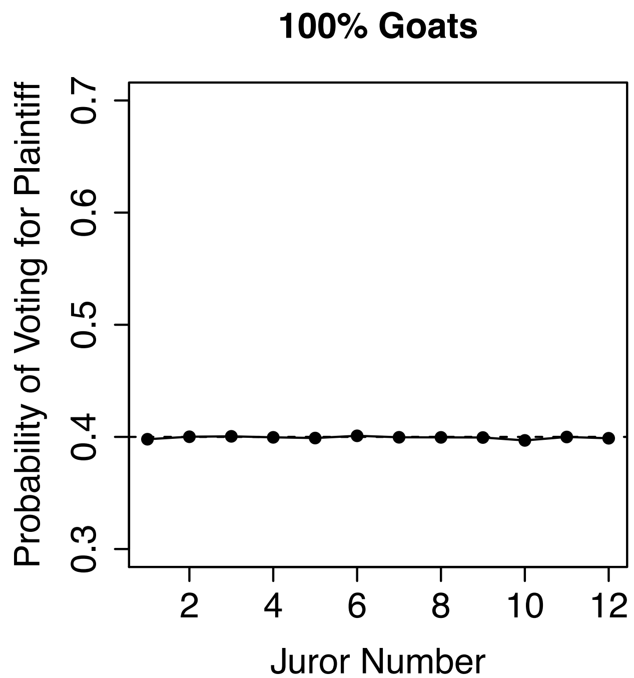
Assumes previous juror has considerable additional knowledge, assigns evidentiary weight to their opinion

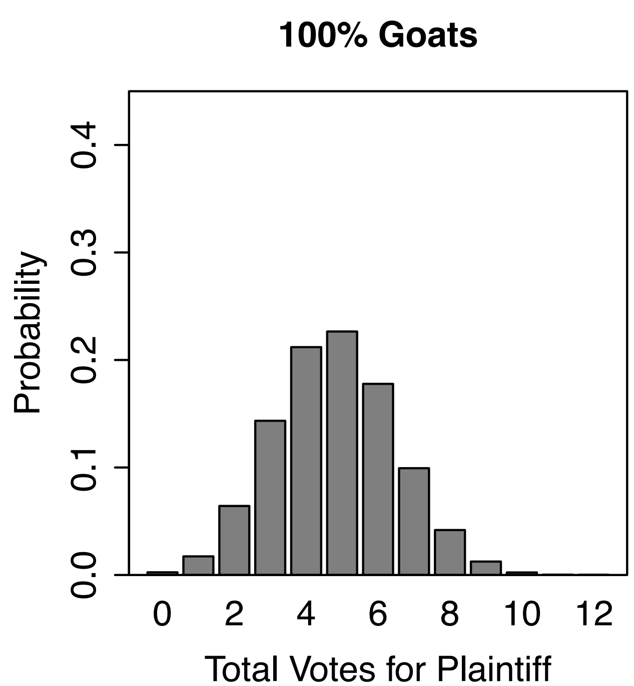
Bayesian “goat”



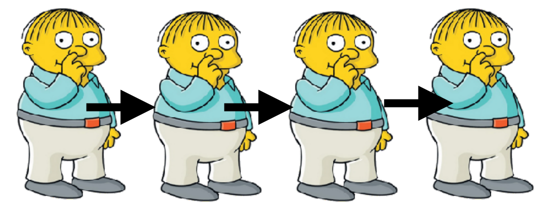
Assumes previous juror has no extra knowledge, assigns zero weight to their opinion

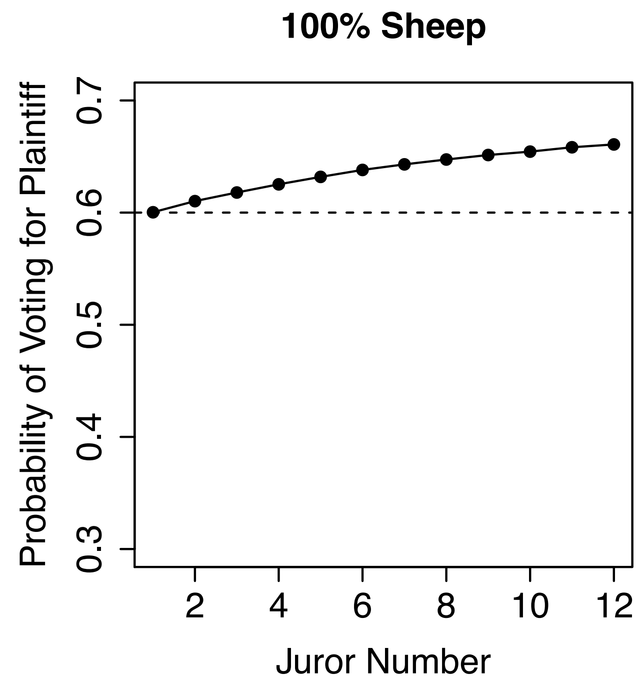


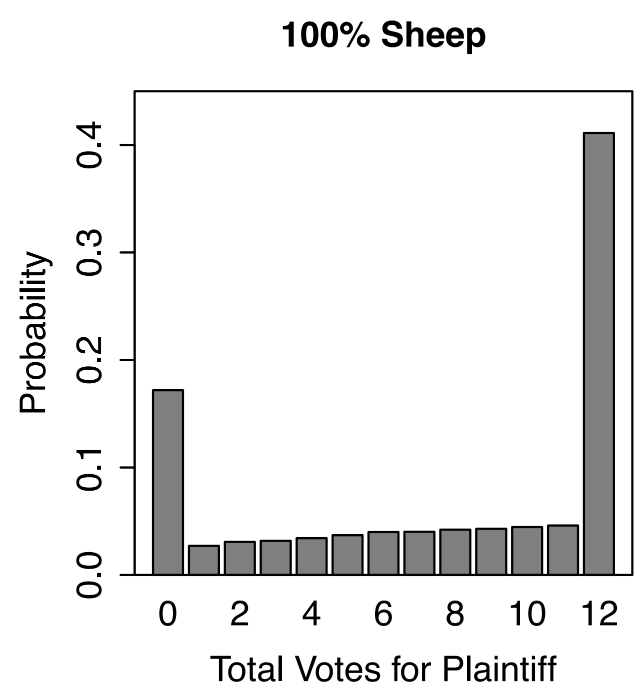




A jury of goats ignores one another and the “chain” converges just fine



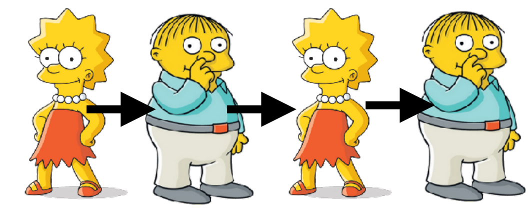


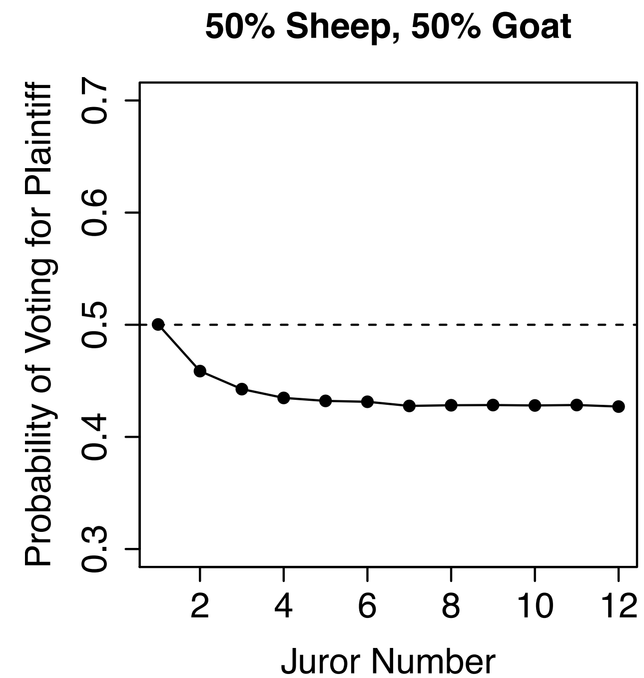


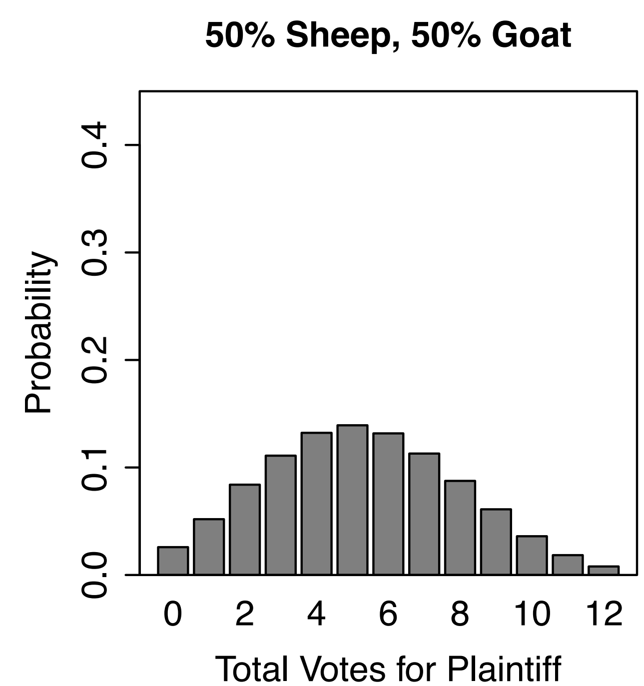
A jury of sheep displays **groupthink**

$$\begin{aligned}\boldsymbol{\pi T} &\propto [d,p]\begin{bmatrix}1-p & p\\ d & 1-d\end{bmatrix}\\ &= [d(1-p)+pd, dp+p(1-d)]\\ &= [d,p] \propto \boldsymbol{\pi}\end{aligned}$$

A mixed jury is dominated by goats



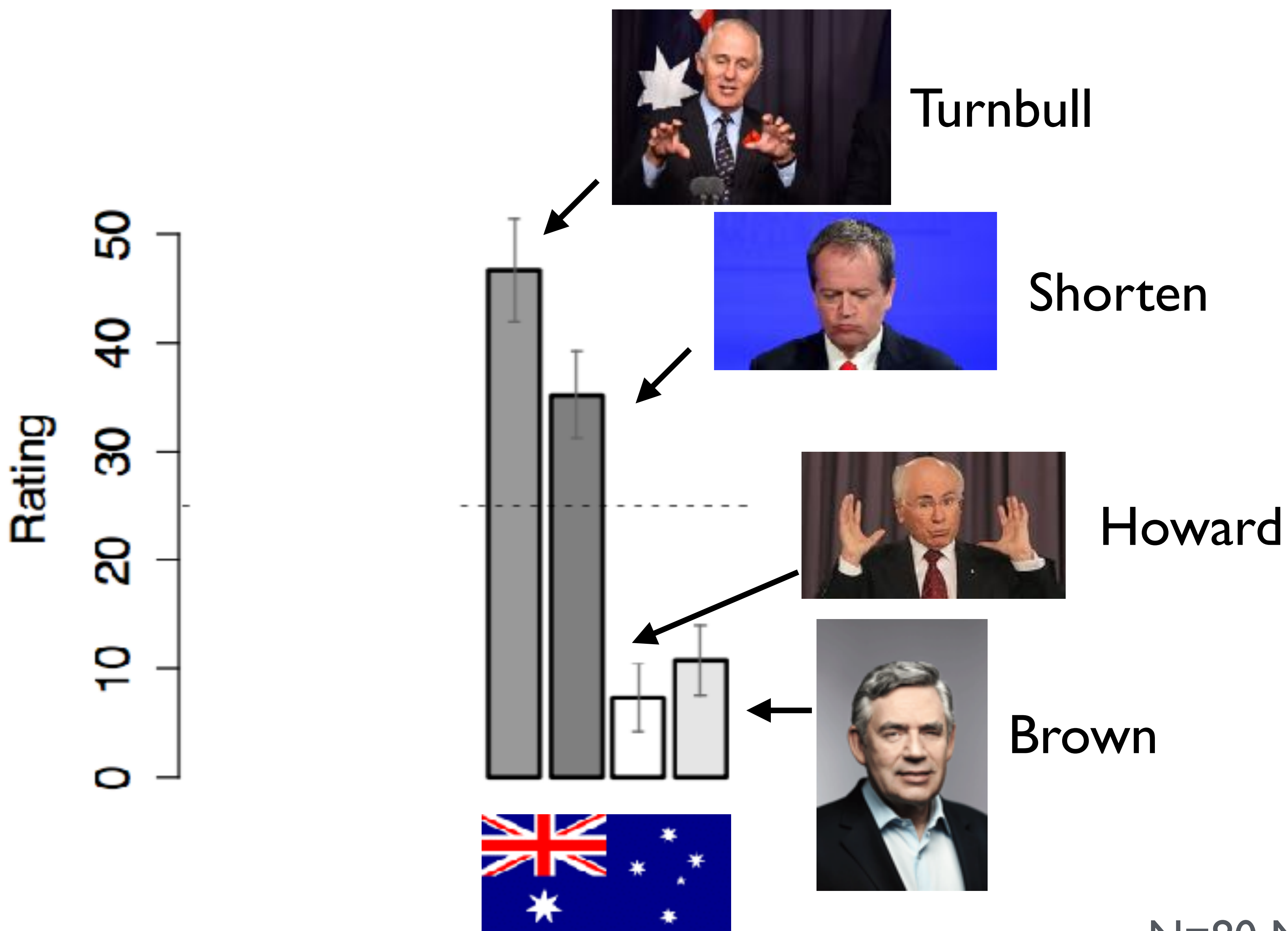




Case study 3:

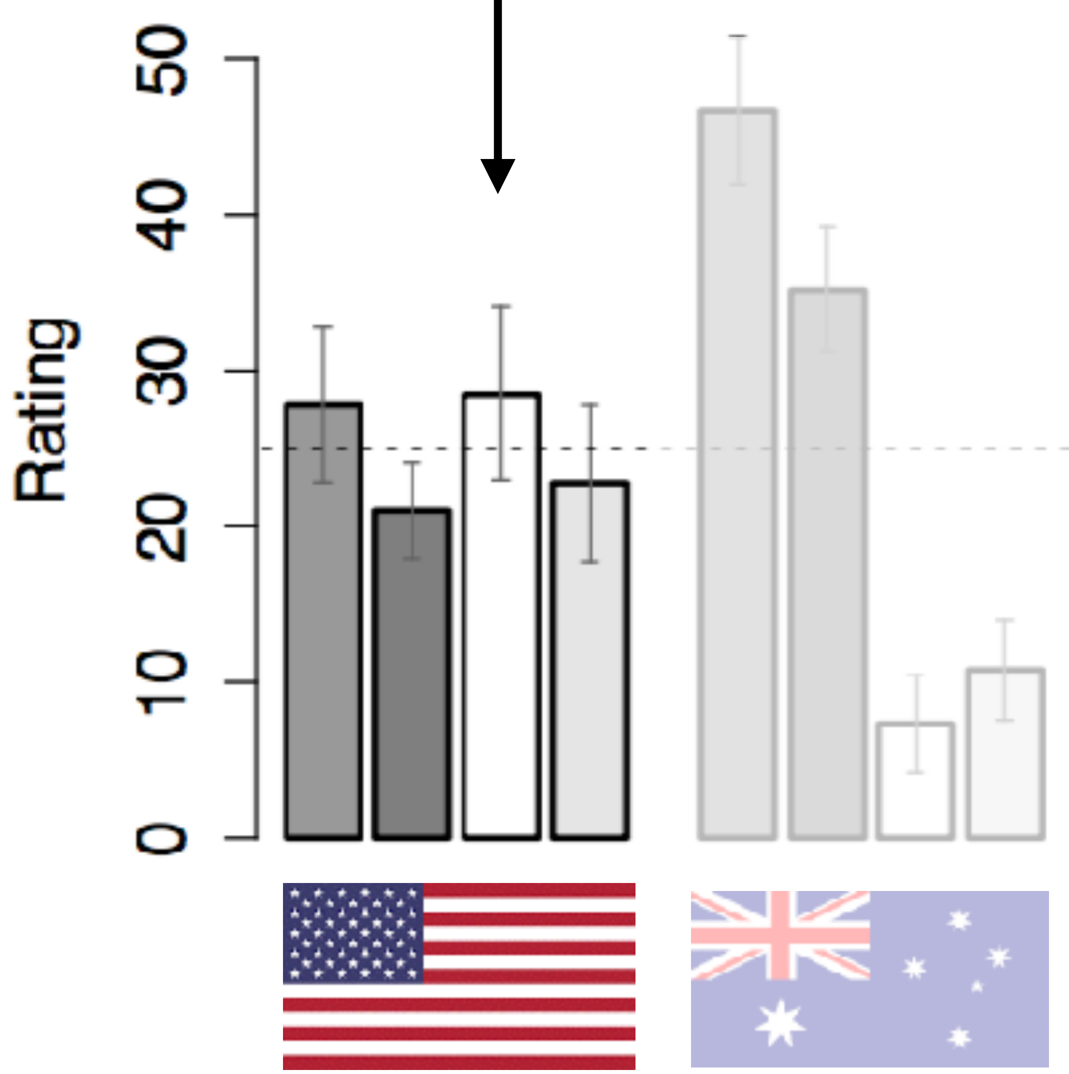
Using differential expertise to create a sheep/goat split in an empirical context

“Who will win the 2016 Australian election?”



N=80 MTurk workers
and UNSW students

Andy?



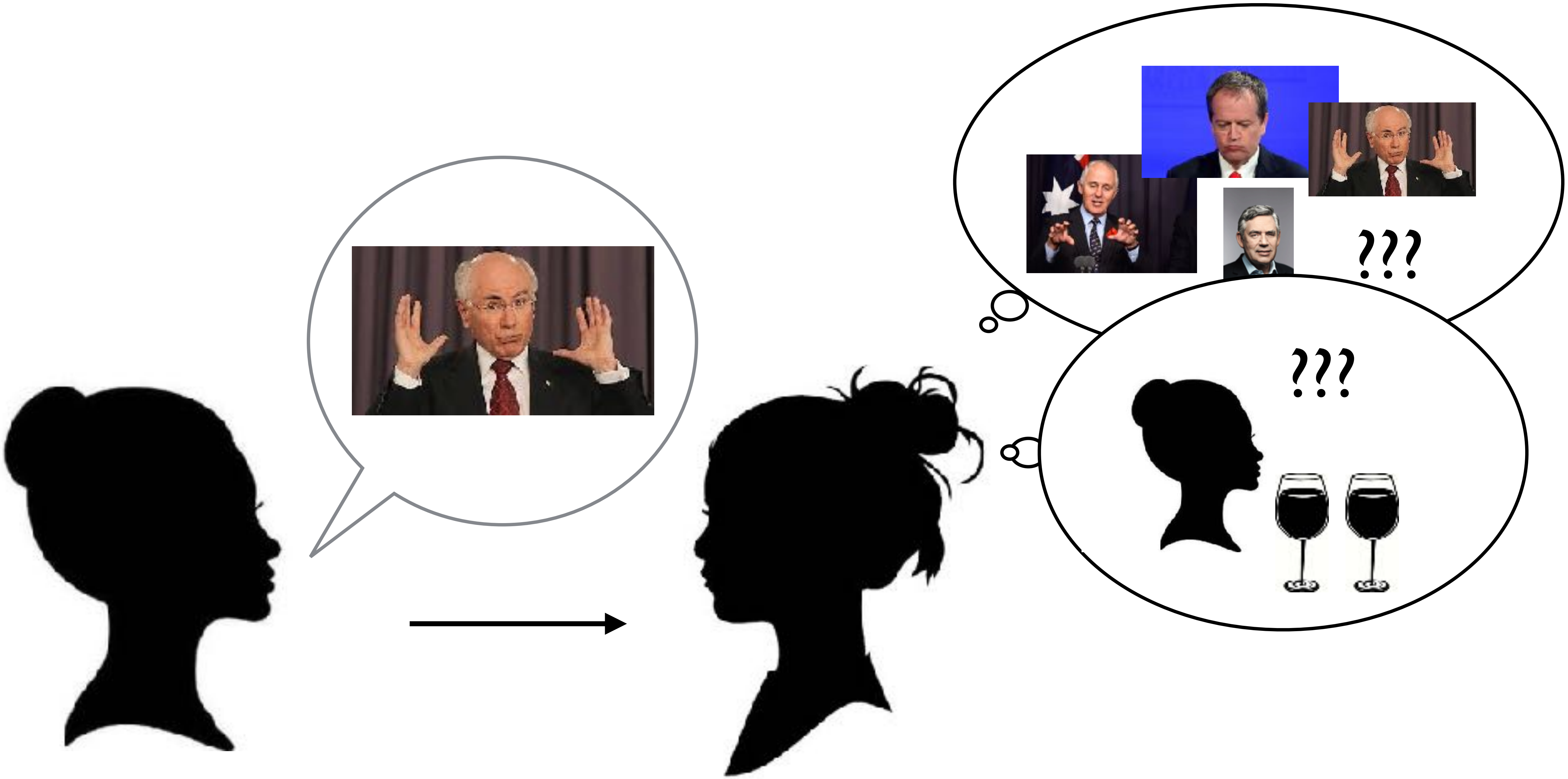
N=80 MTurk workers
and UNSW students

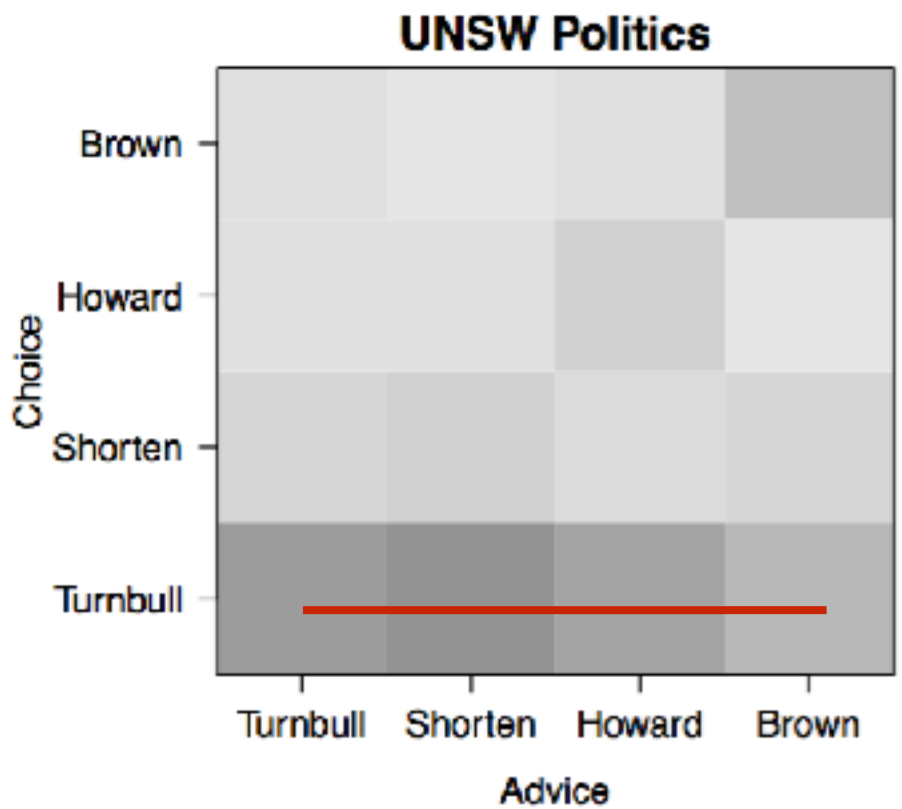
The advisor task

"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?"

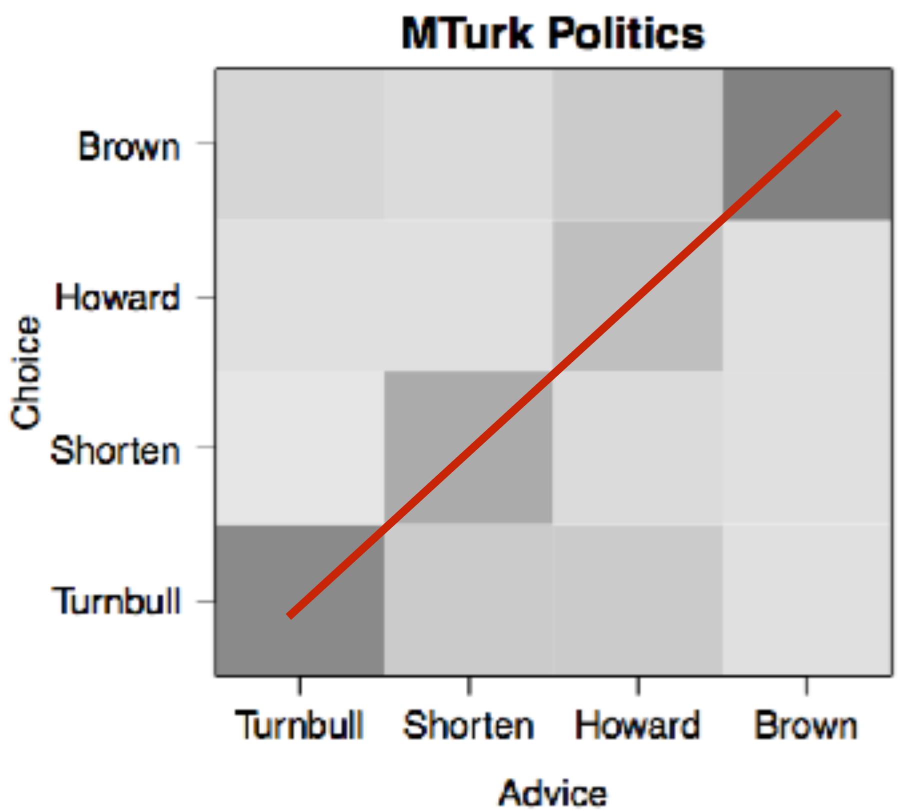
The advisor task



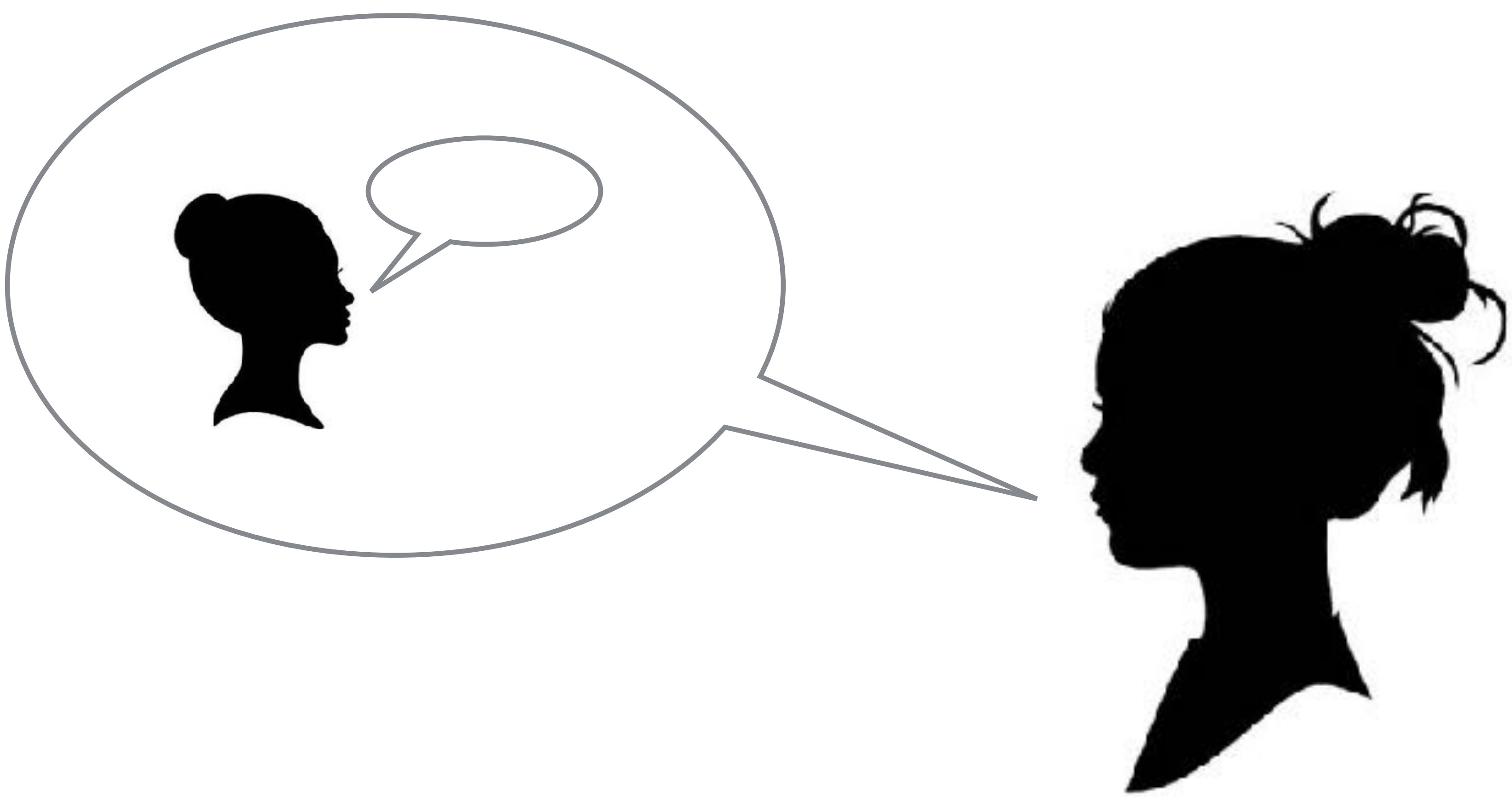


Australians ignored the advisor and predicted a Turnbull victory

N=124 UNSW students

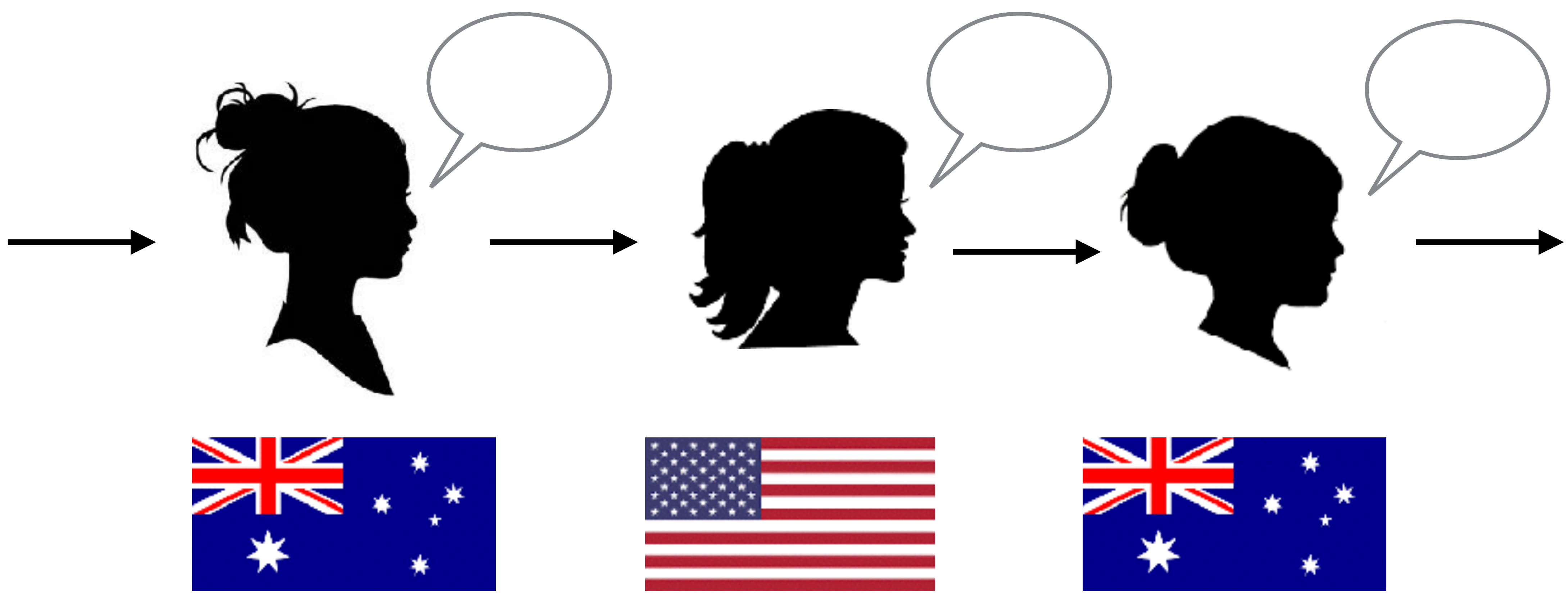


Americans followed the advisor regardless

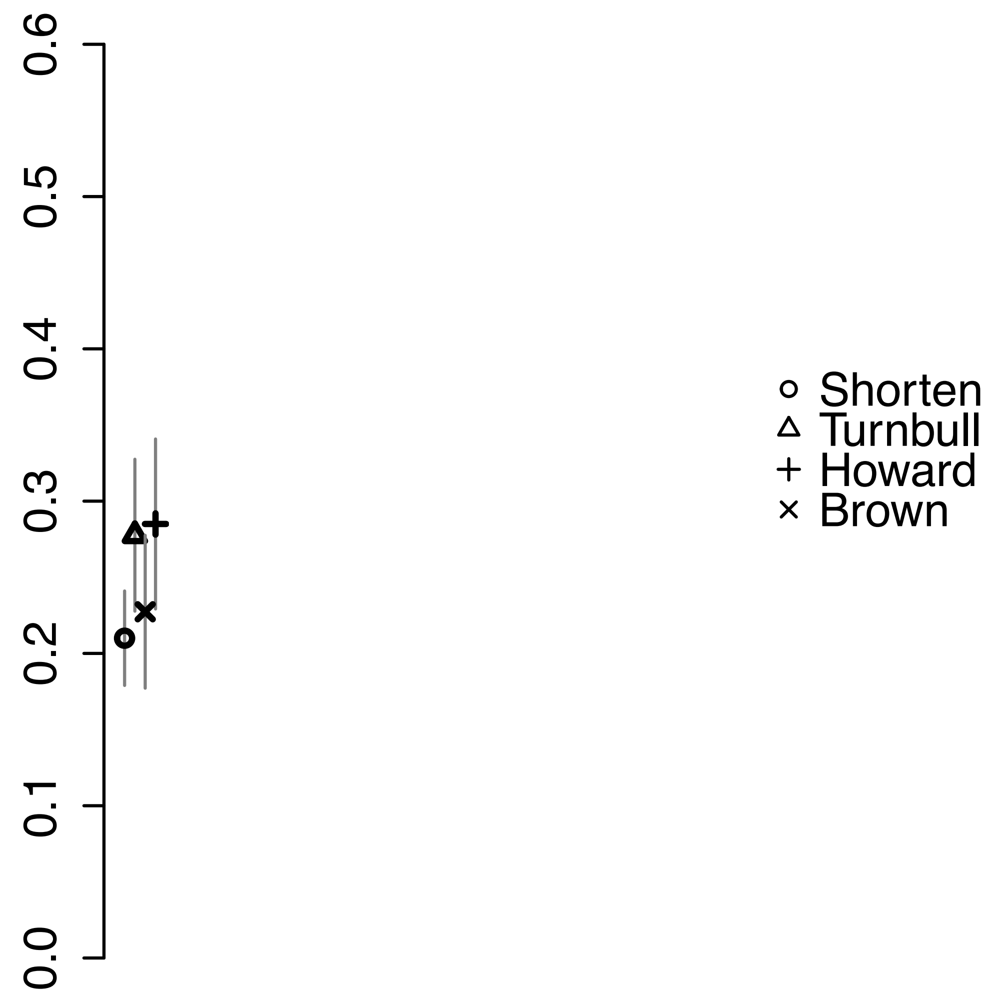


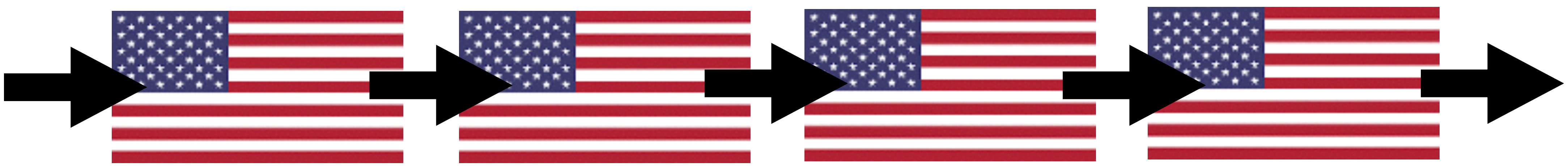
N=196 MTurk workers

Using these empirical transition matrices we can construct iterated learning chains with any mixture of nationalities

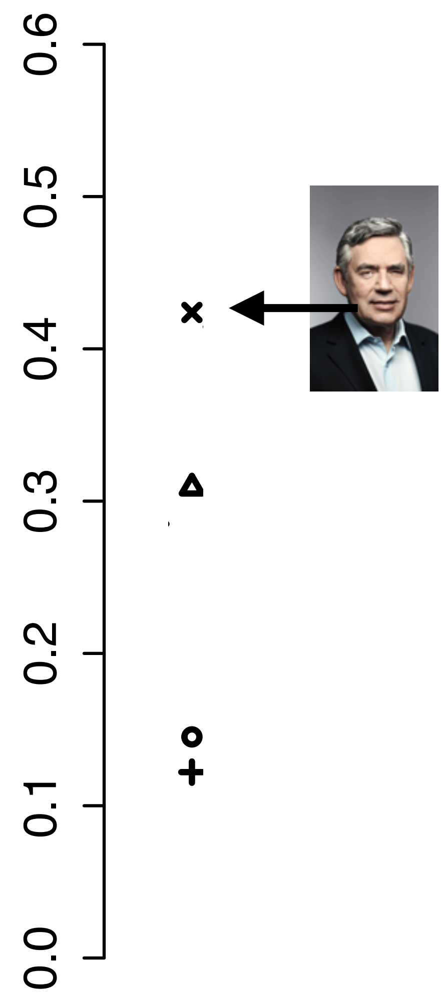


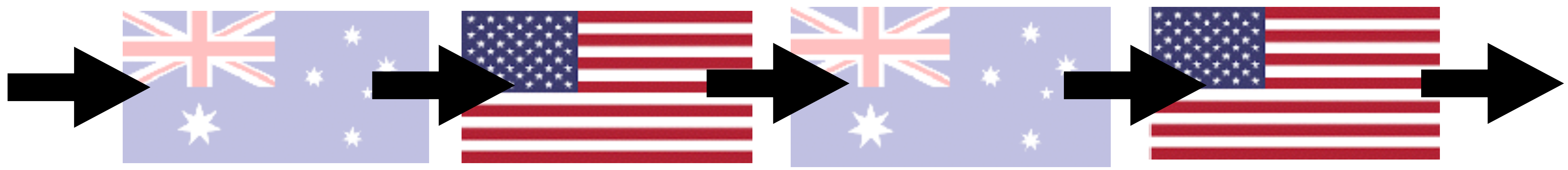
Americans _claim_ to be totally ignorant about Australian politics…



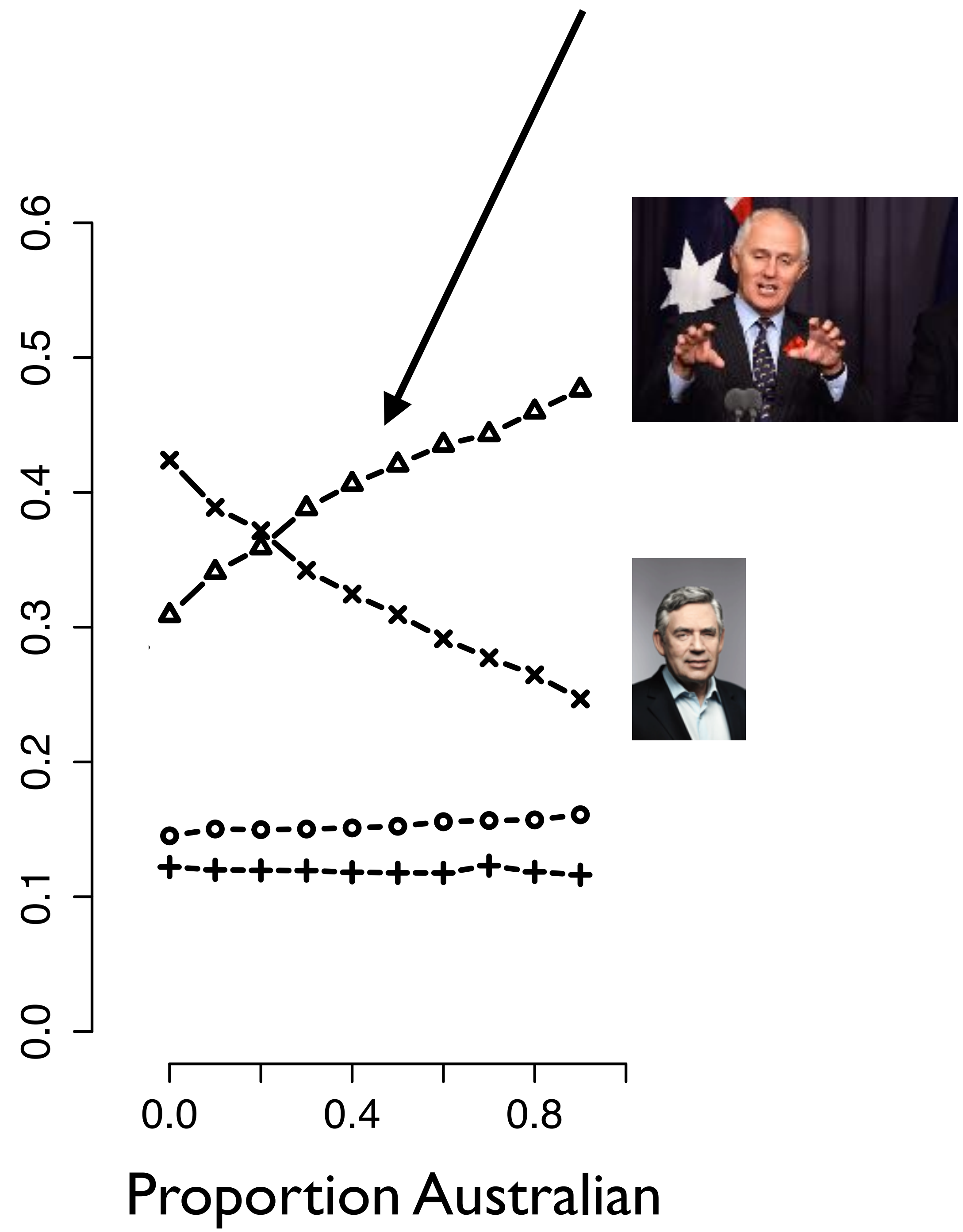


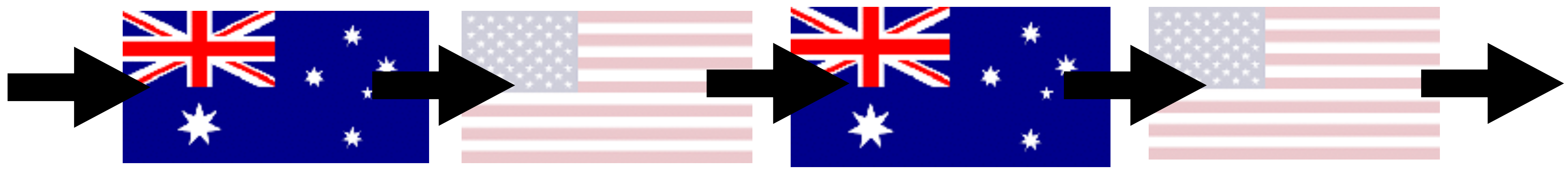
… and an all American iterated learning chain "reveals" a "preference" for <u>*Gordon Brown*</u> …

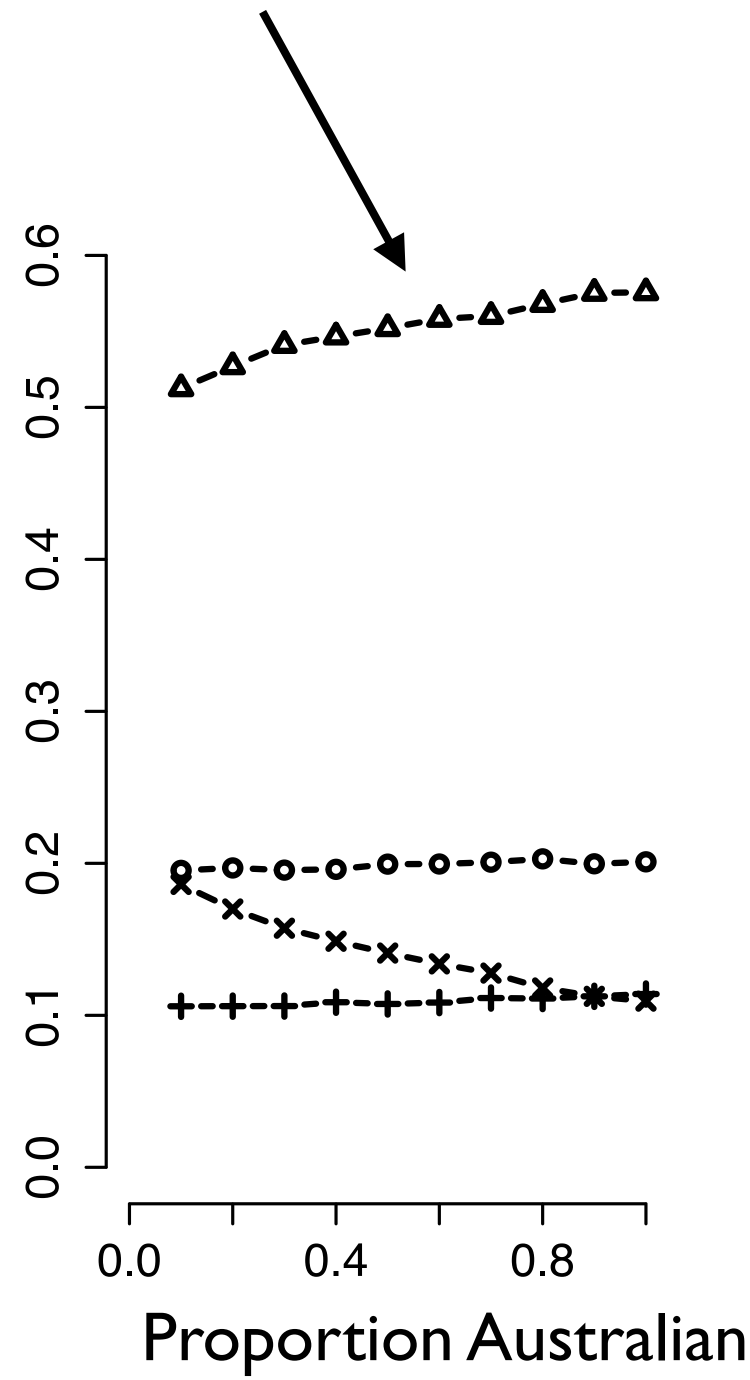




… but if we mix some Australians into the chain the Americans endorse *<u>Malcolm Turnbull</u>*







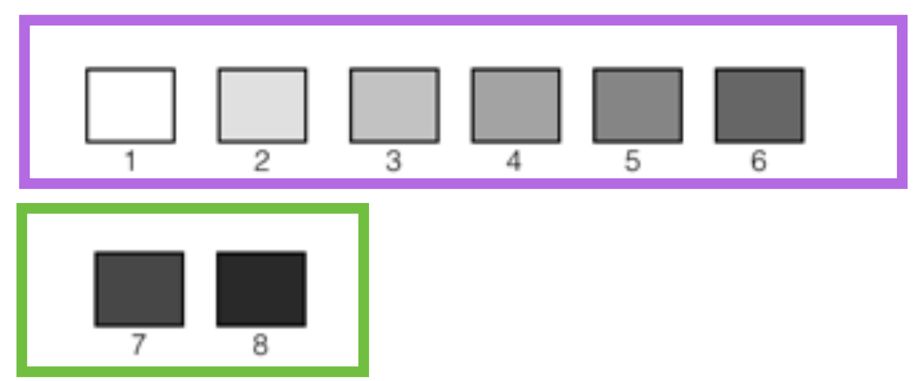
Australians choose Turnbull no matter how many Americans are included

Case study 4:

It’s not always obvious which inductive biases are distorted by heterogeneity

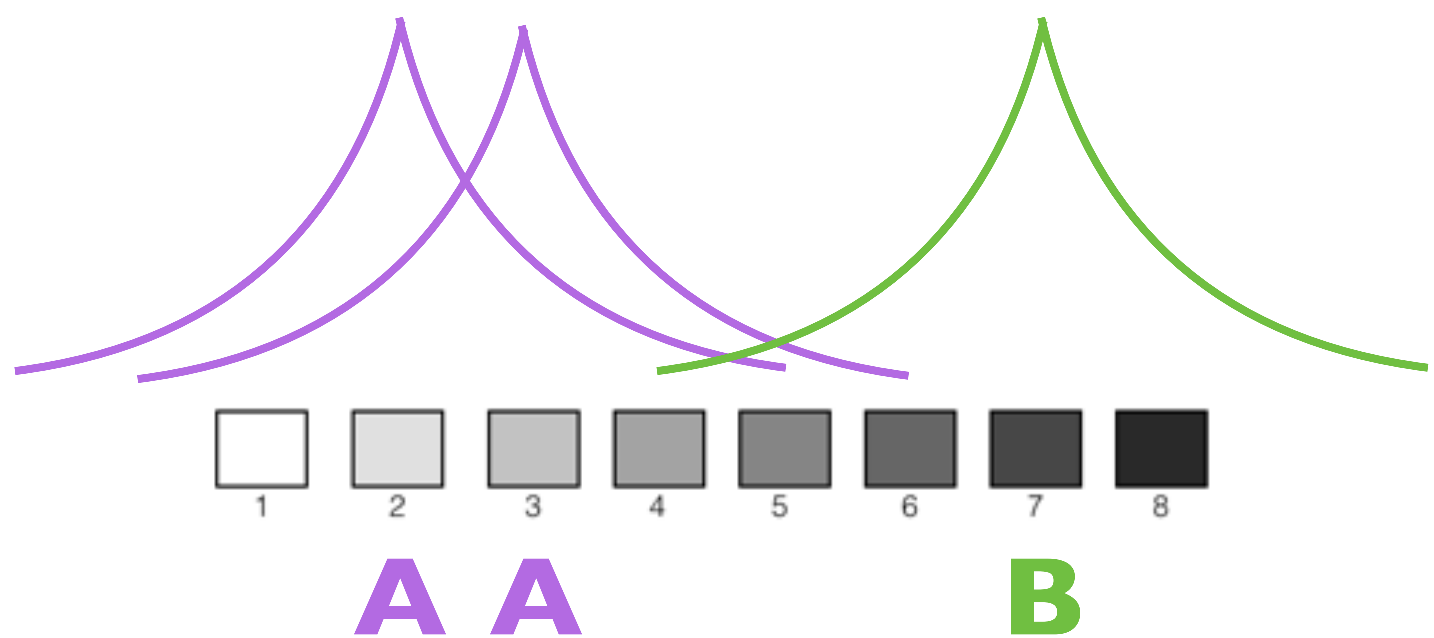
Iterated learning can be used to study the biases people bring to categorisation problems

(e.g., Austerweil 2014)



Exemplar model of categorisation

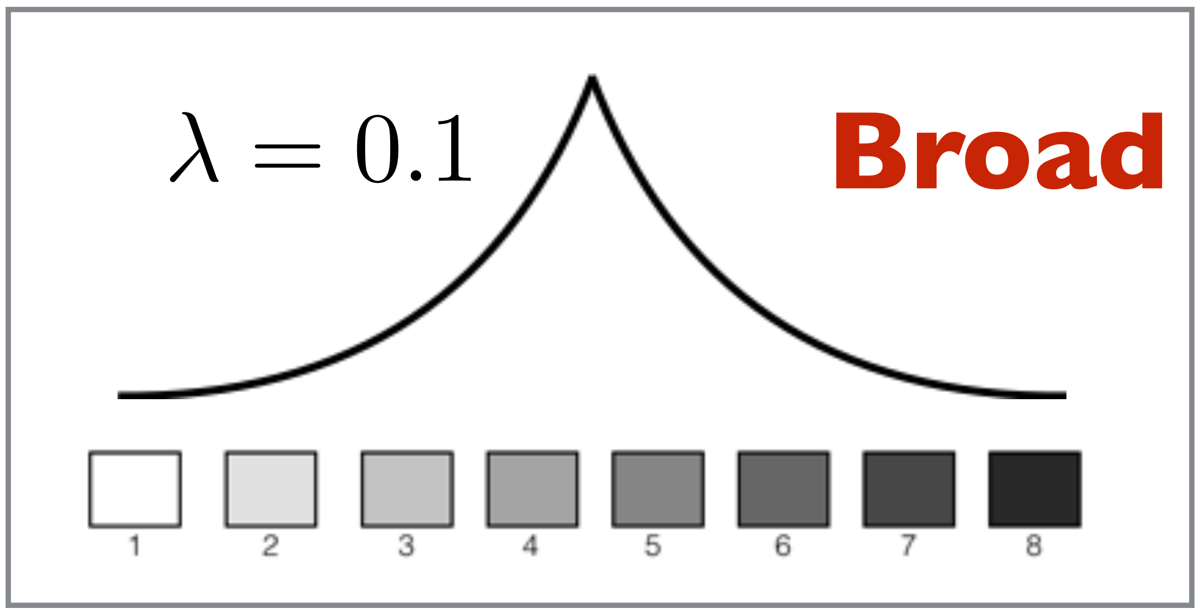
(Nosofsky 1986; Pothos & Bailey 2009)



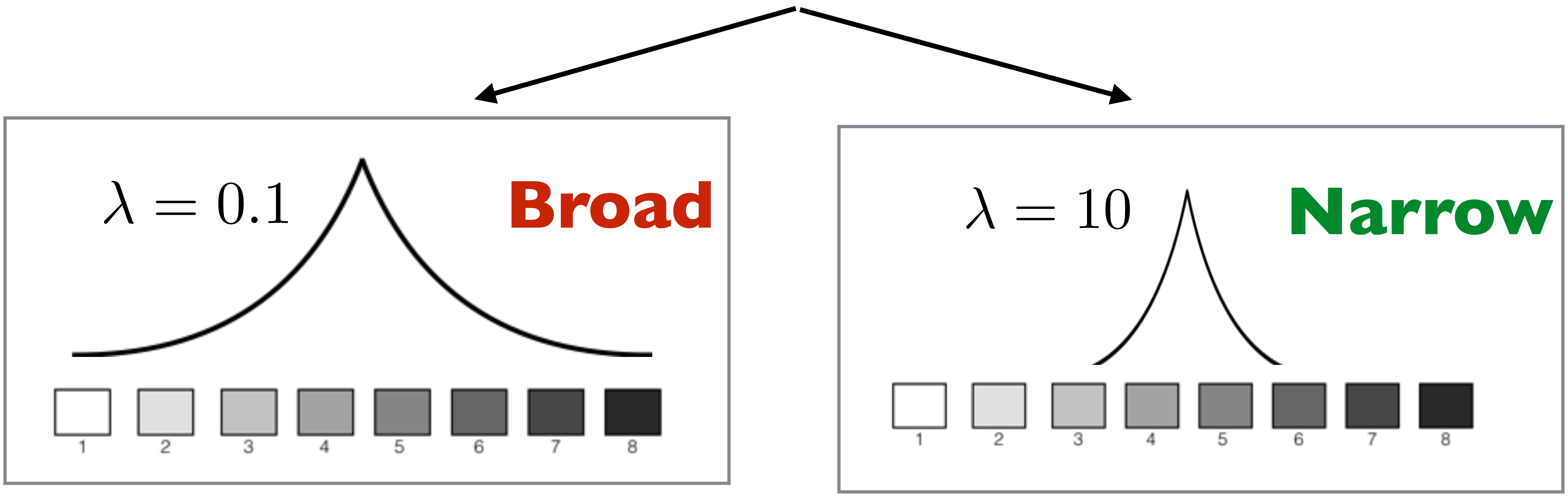
GCM: categorisation probability is proportional to sum similarity

$$P(y \in A) = \frac{\sum_{a \in A} s(a, y)}{\sum_{X} \sum_{x \in X} s(x, y)}$$

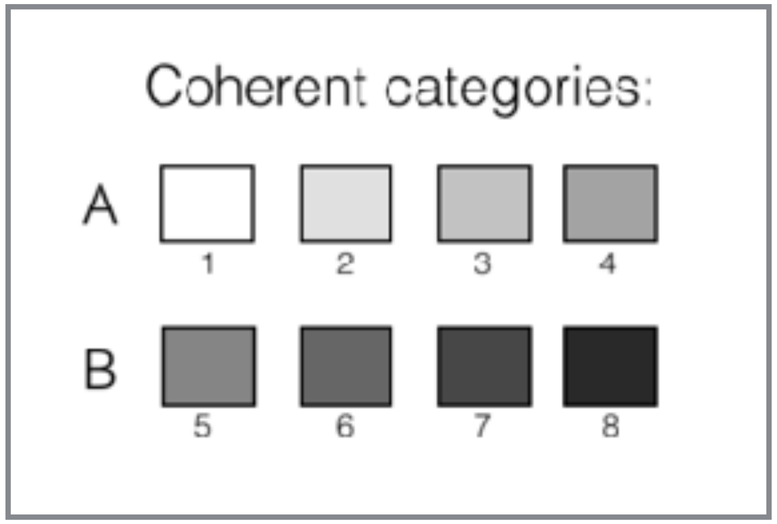
GCM allows learners to vary in how broadly they generalise from a stimulus



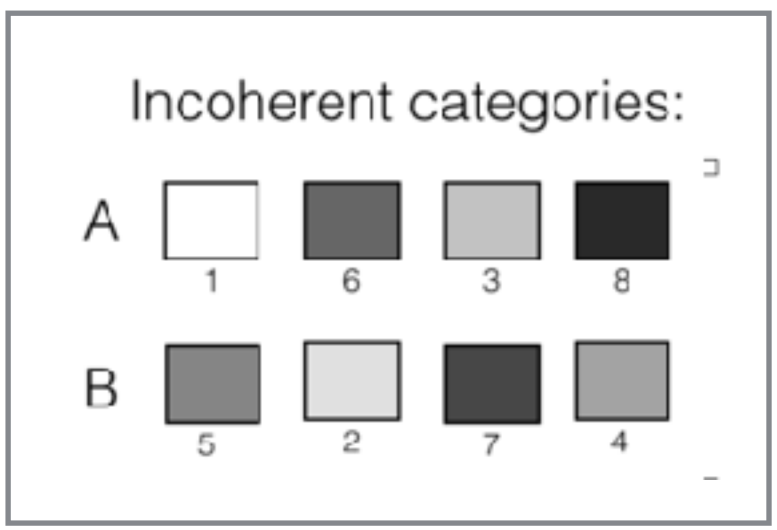
GCM allows learners to vary in how broadly they generalise from a stimulus



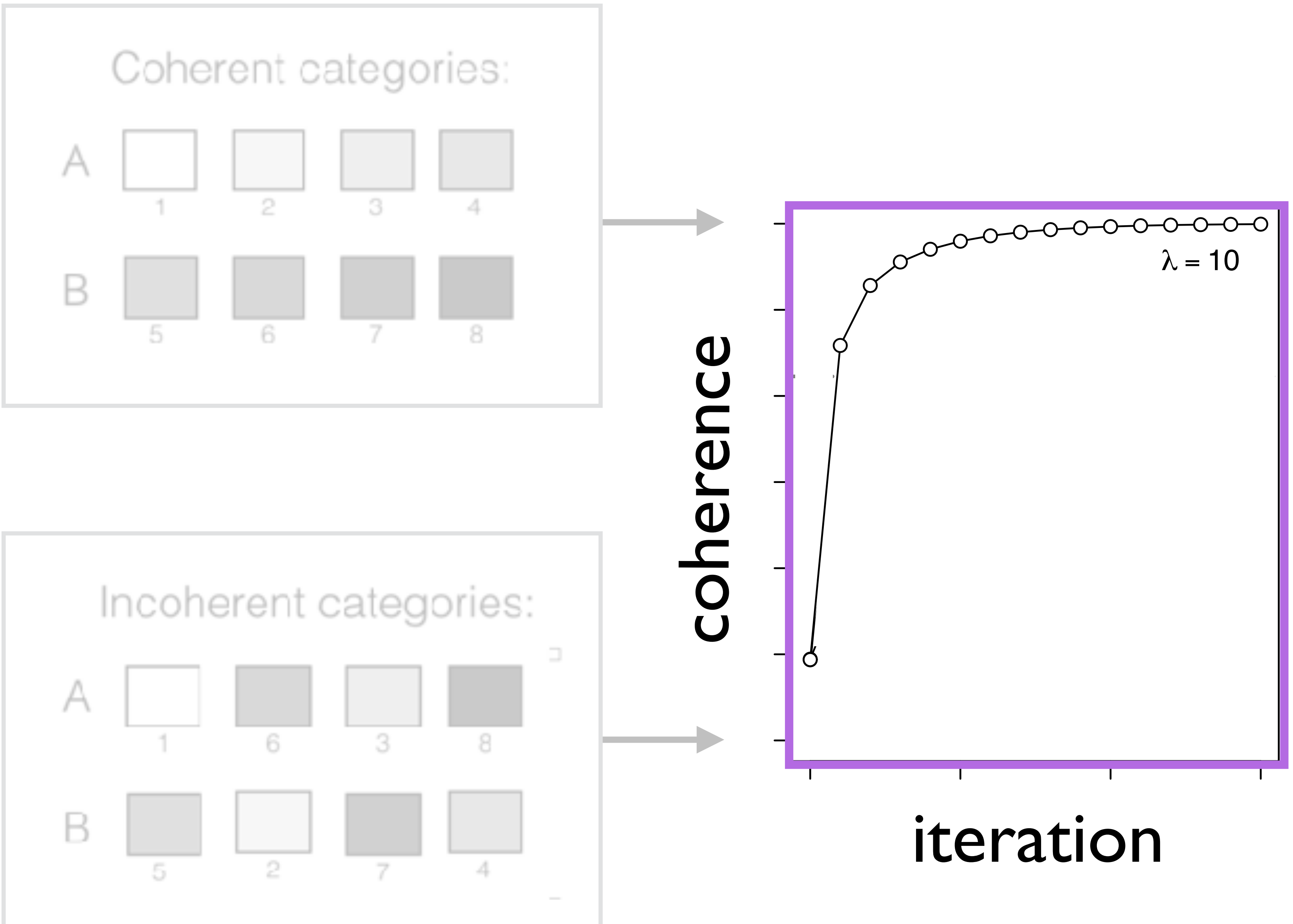
Categorisation bias #1



Coherent systems assign similar items to the same category

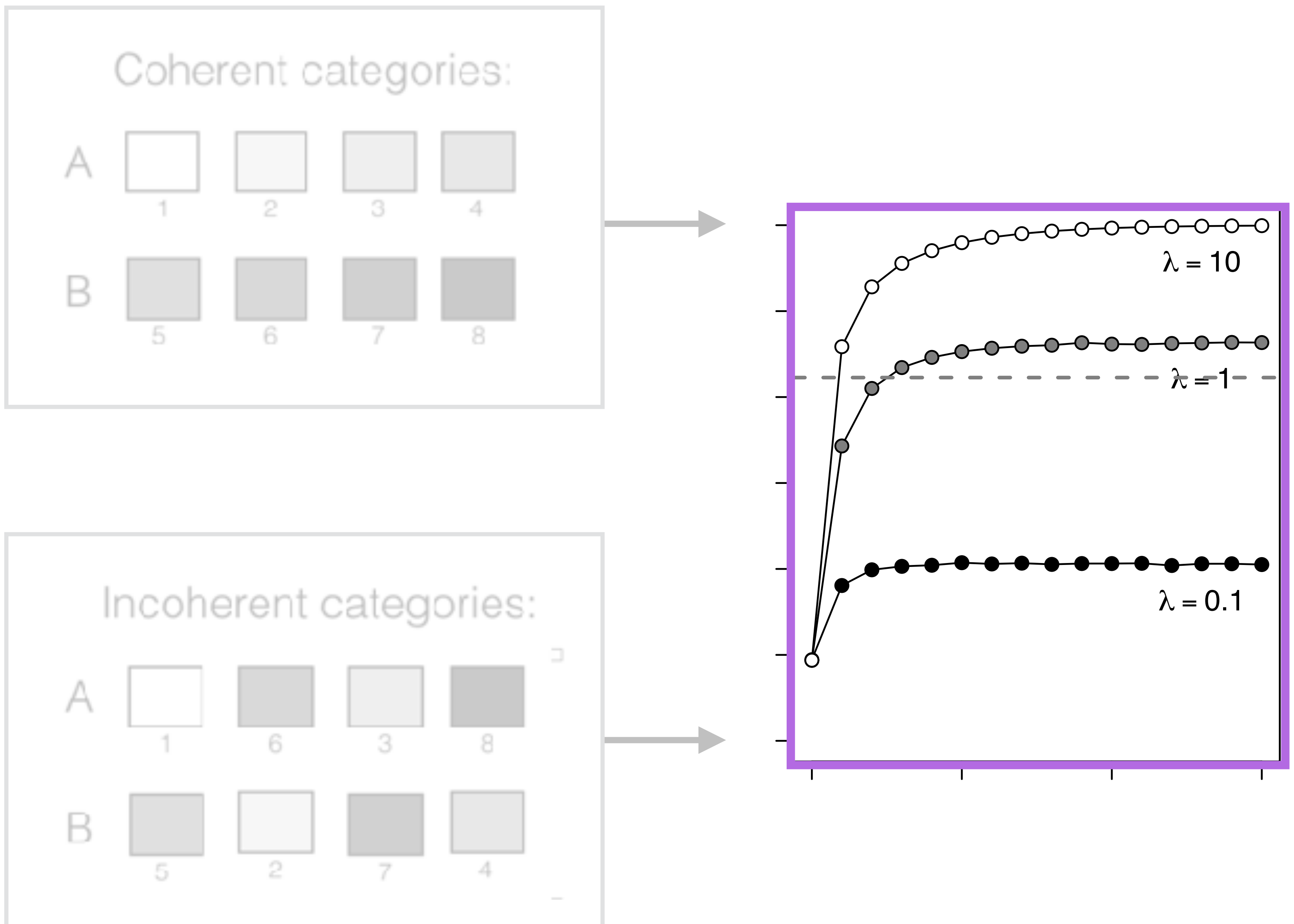


Homogenous population



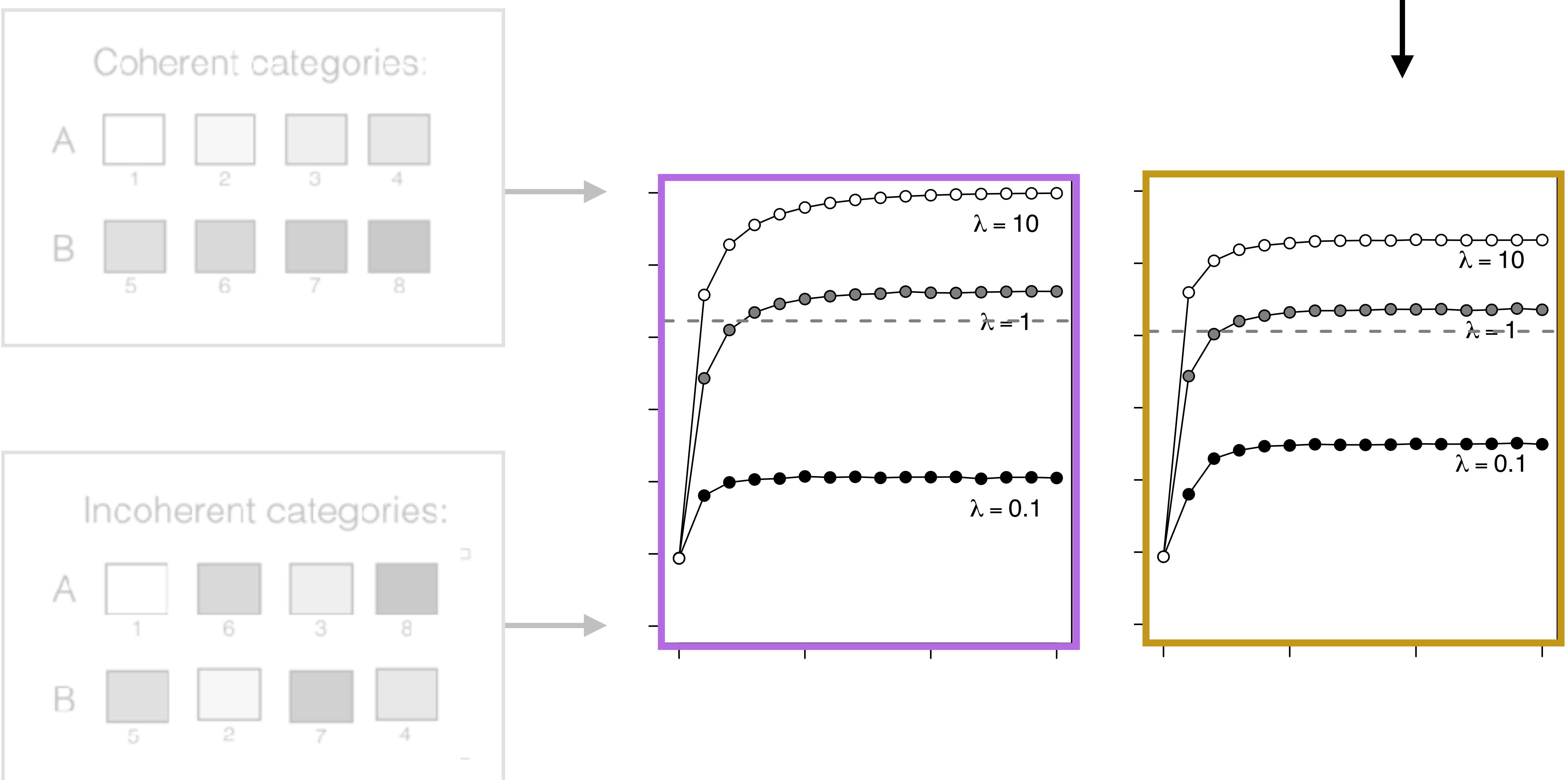
Narrow generalisation produces a strong coherence bias in GCM

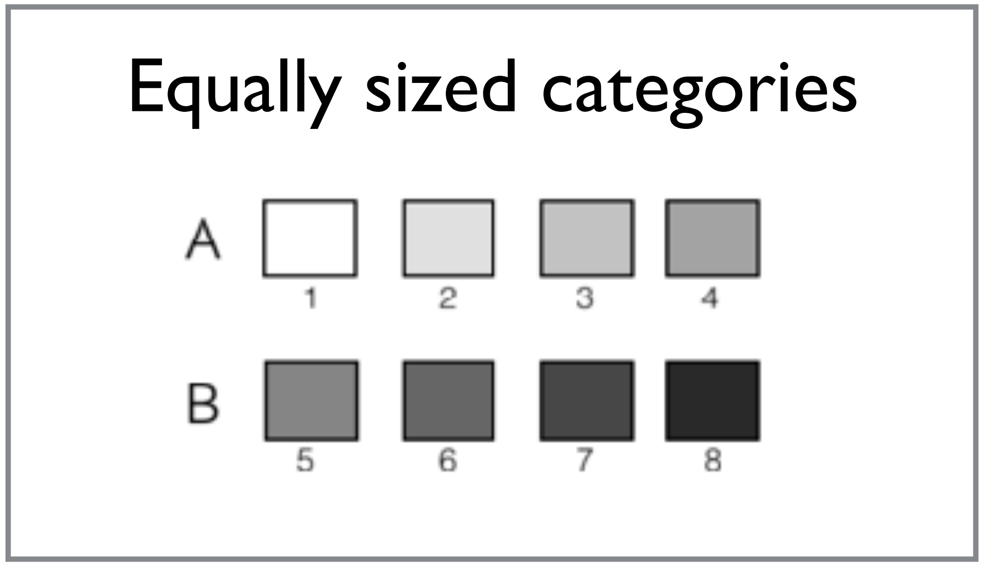
Homogenous population

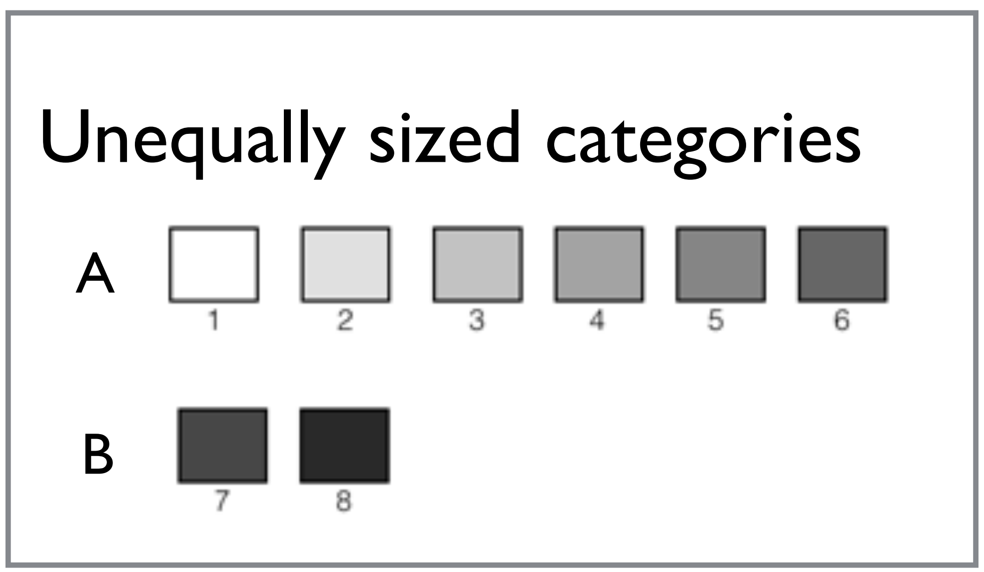


Broad generalisation produces a weak coherence bias in GCM

Heterogeneity isn't much of a problem here

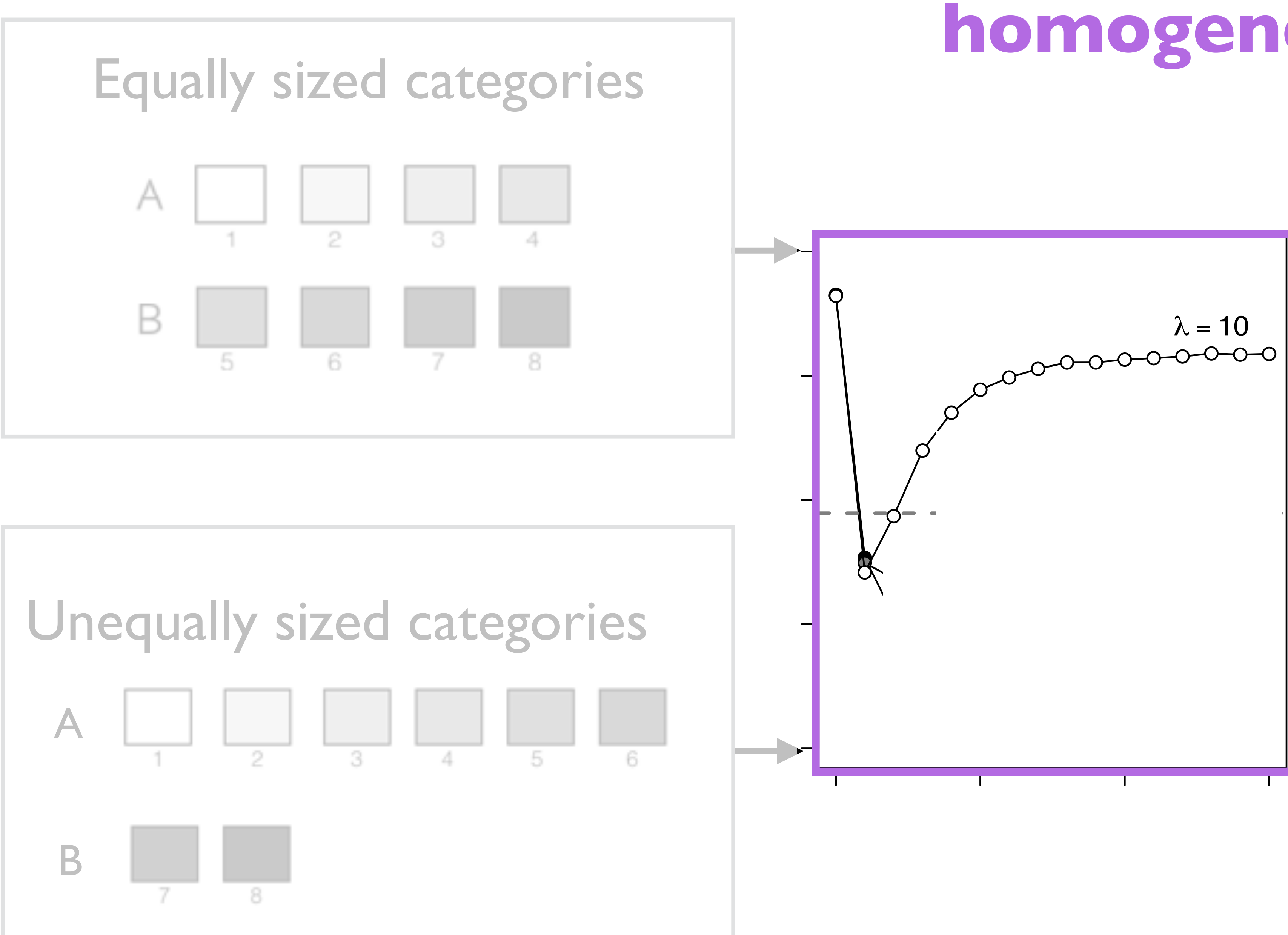






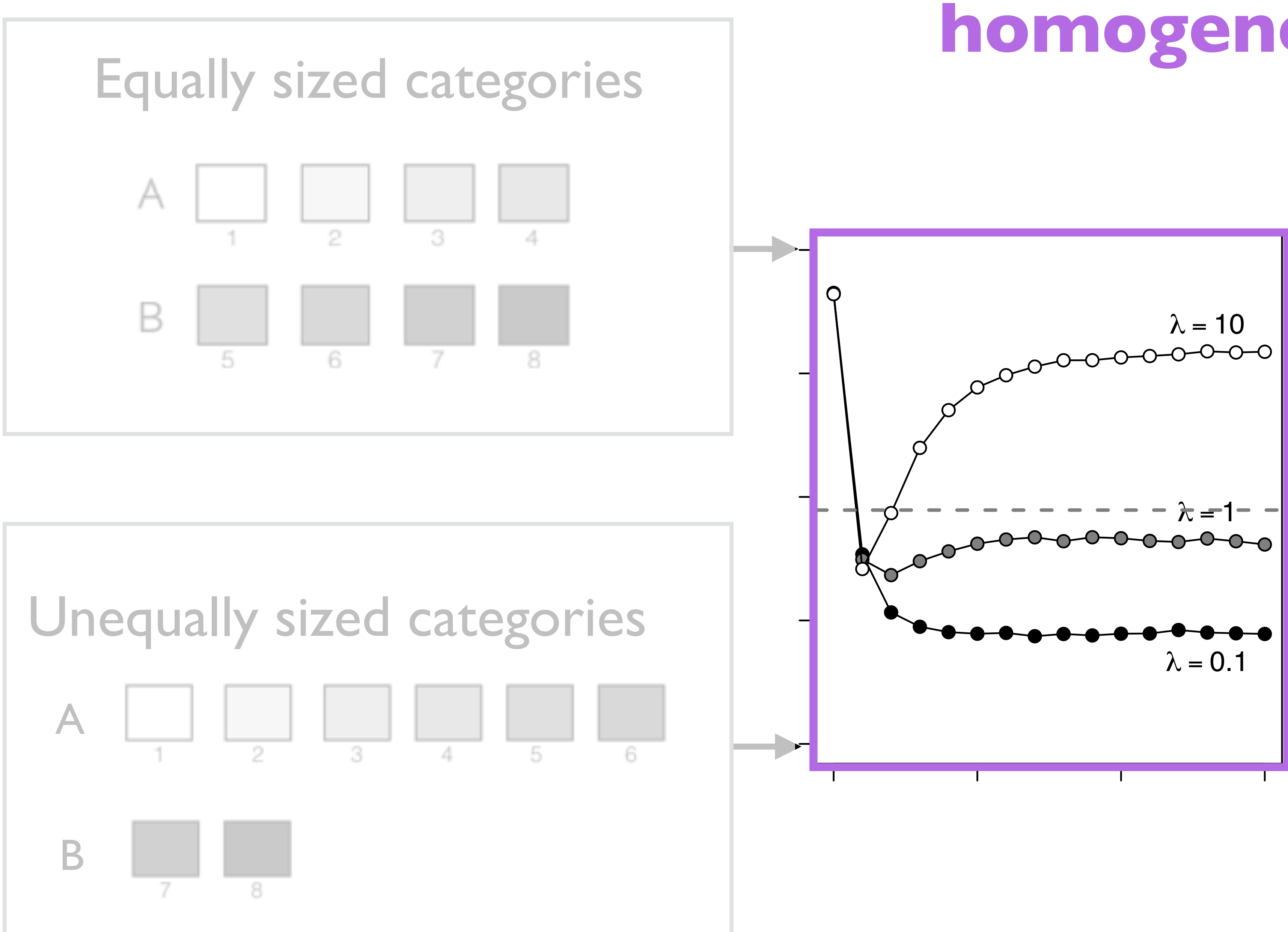
Categorisation bias #2

Iterated learning chains with **homogenous** populations



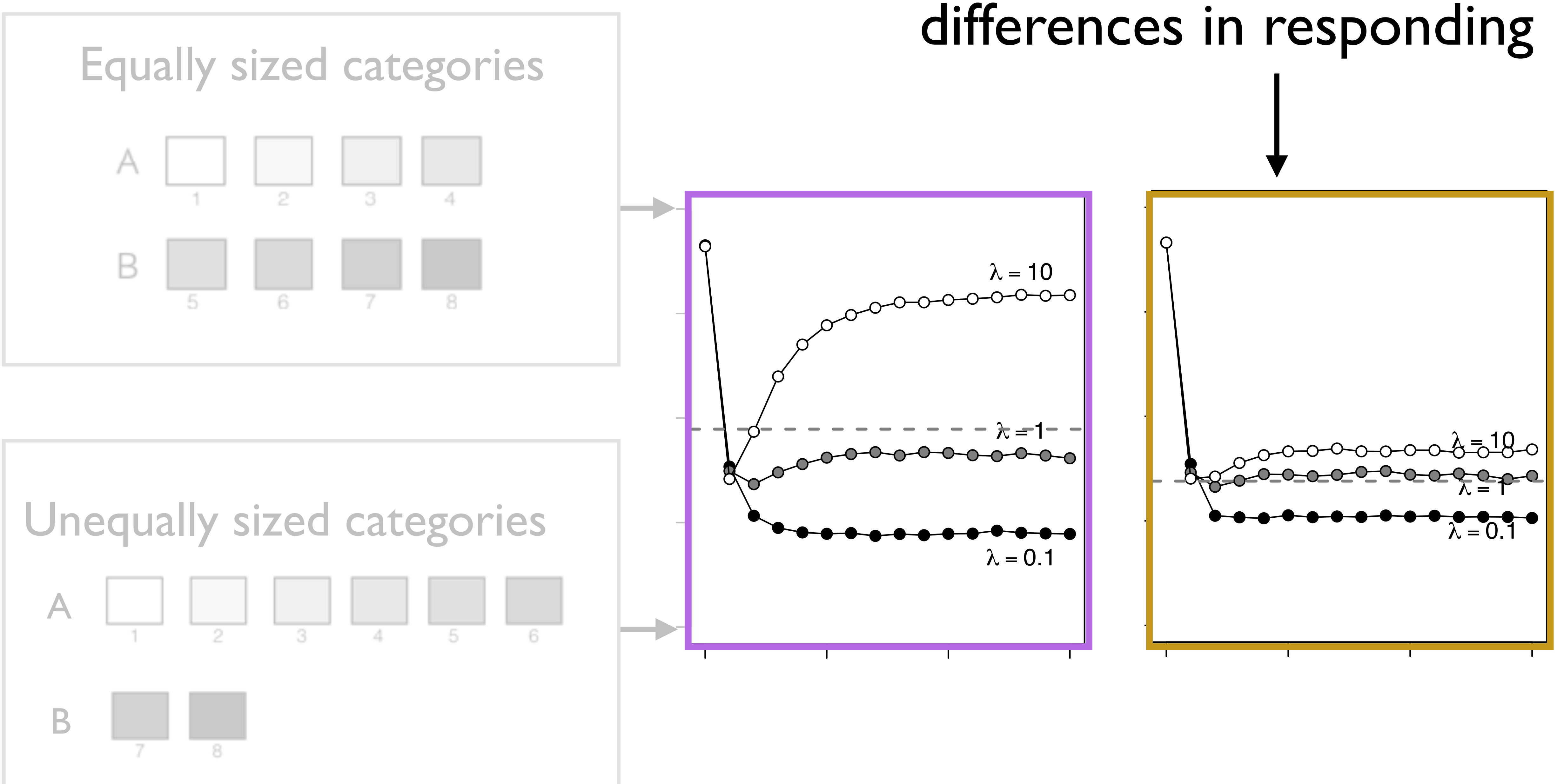
Narrow
generalisation in GCM
produces bias for
equally sized categories

Iterated learning chains with **homogenous** populations

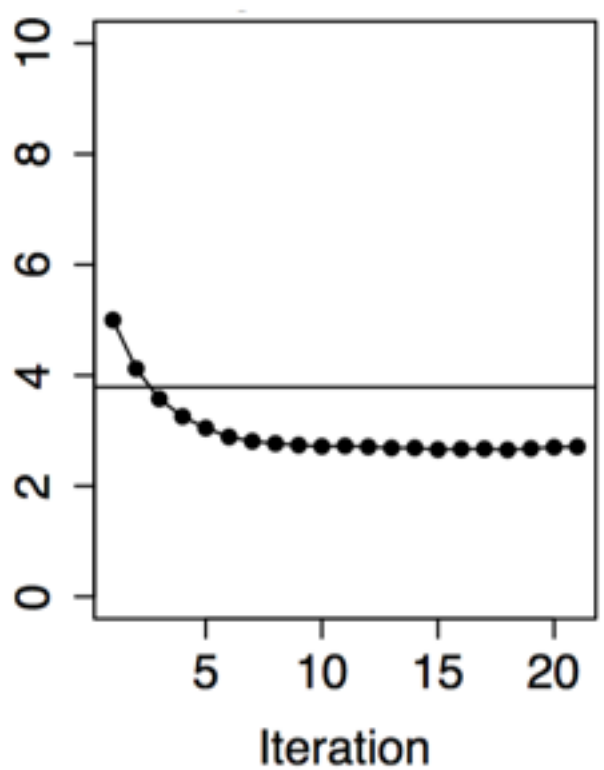


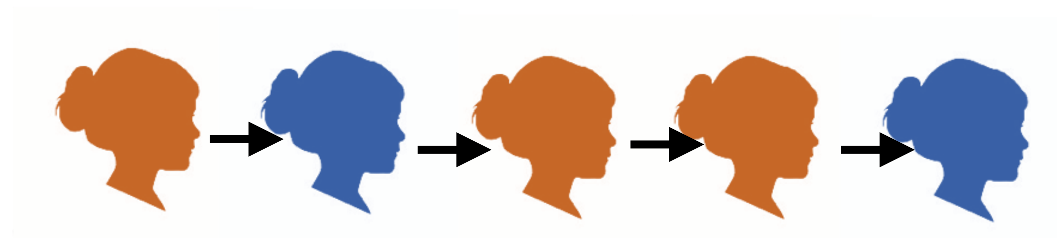
Broad generalisation produces bias for unequal size

Heterogeneity in the population erases the individual differences in responding

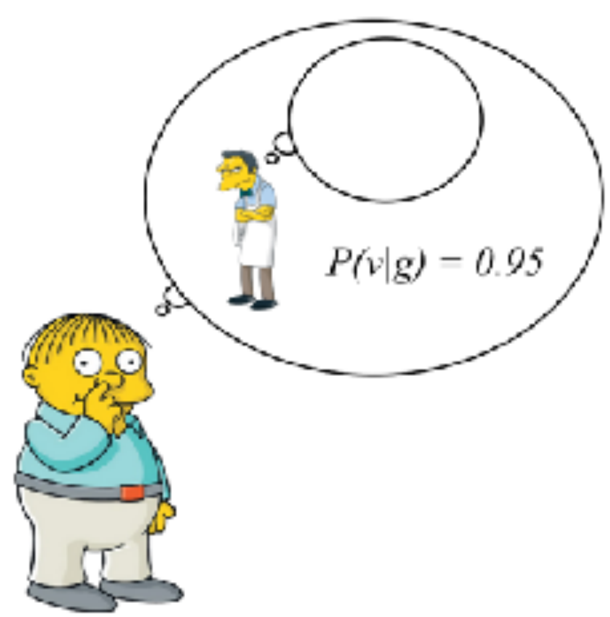


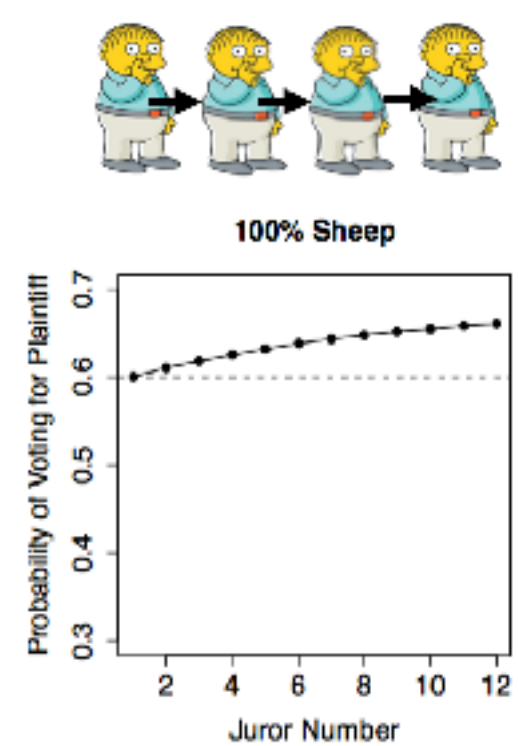
- <u>Summary</u>:
  - Iterated learning distorts inductive bias when individual differences are present





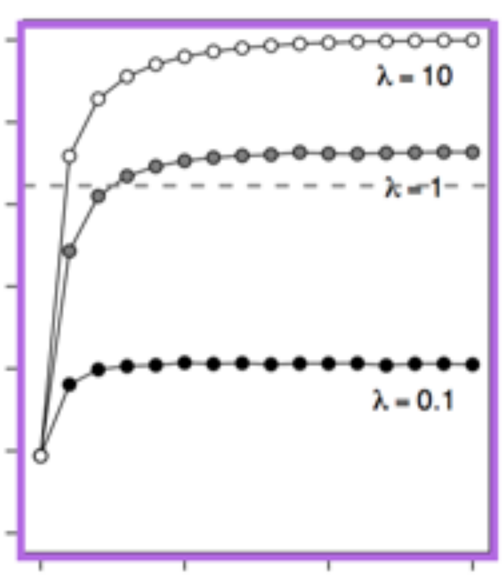
- Summary:
  - Iterated learning distorts inductive bias when individual differences are present
  - Miscalibrated agents can distort their own inductive biases even in homogenous chains

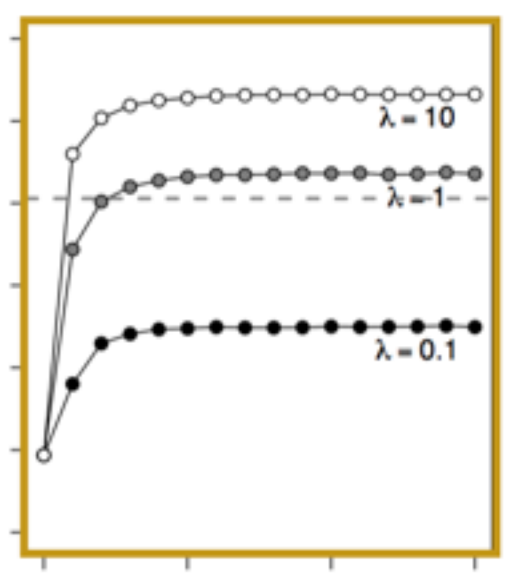


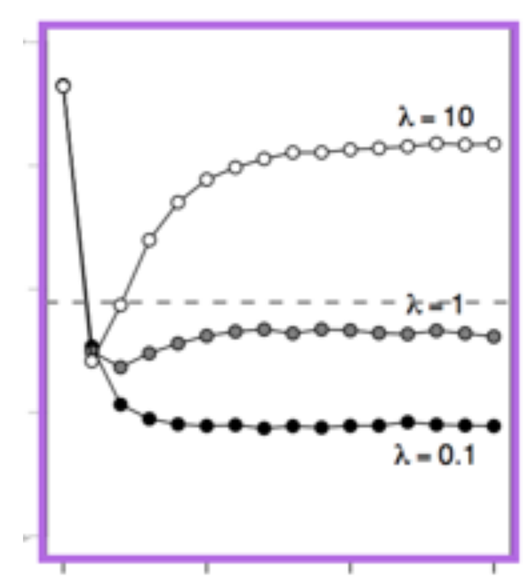


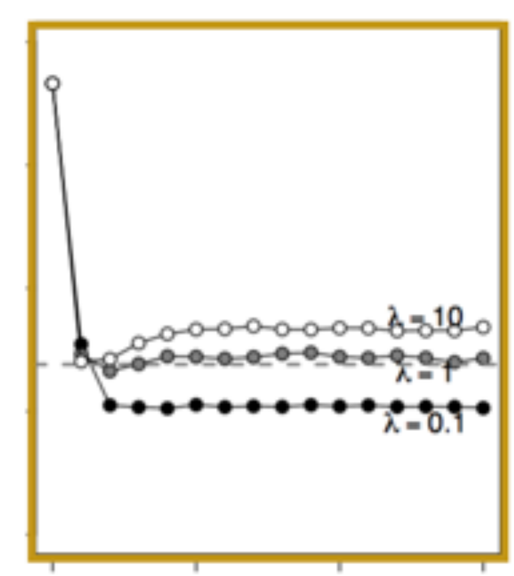
- Summary:
  - Iterated learning distorts inductive bias when individual differences are present
  - Miscalibrated agents can distort their own inductive biases even in homogenous chains
  - IL chains favour learners with extreme biases

- <u>Summary</u>:
  - 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





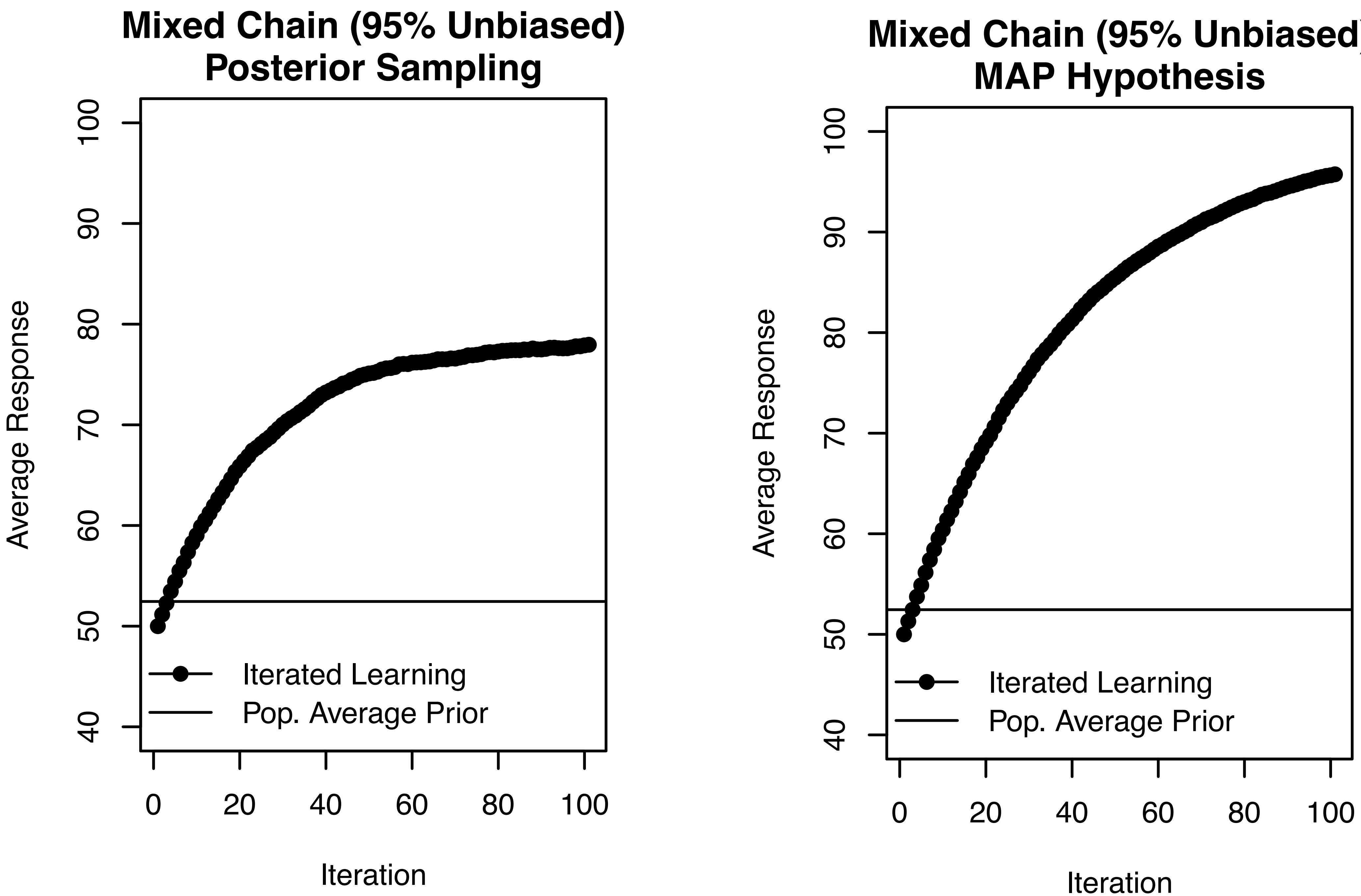


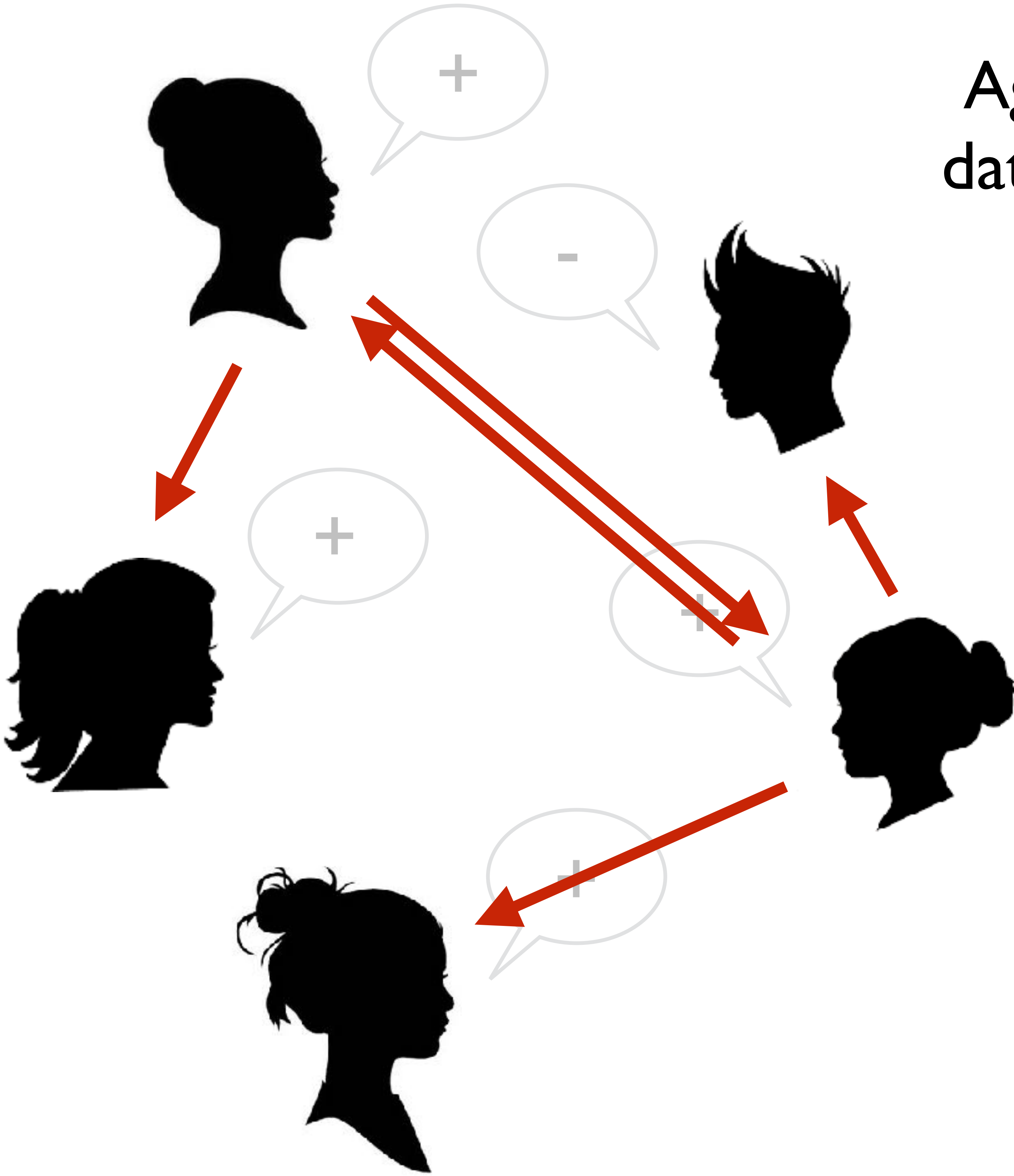


- Summary:
    - Iterated learning distorts inductive bias when individual differences are present
    - Miscalibrated agents can distort their own inductive biases even in homogenous chains
    - IL chains favour learners with extreme biases
    - The magnitude of the distortion is variable
- Implications:
    - IL has limits as a tool for “revealing priors”
    - IL is useful for studying “distortions” in cultural and linguistic evolution

Thanks!

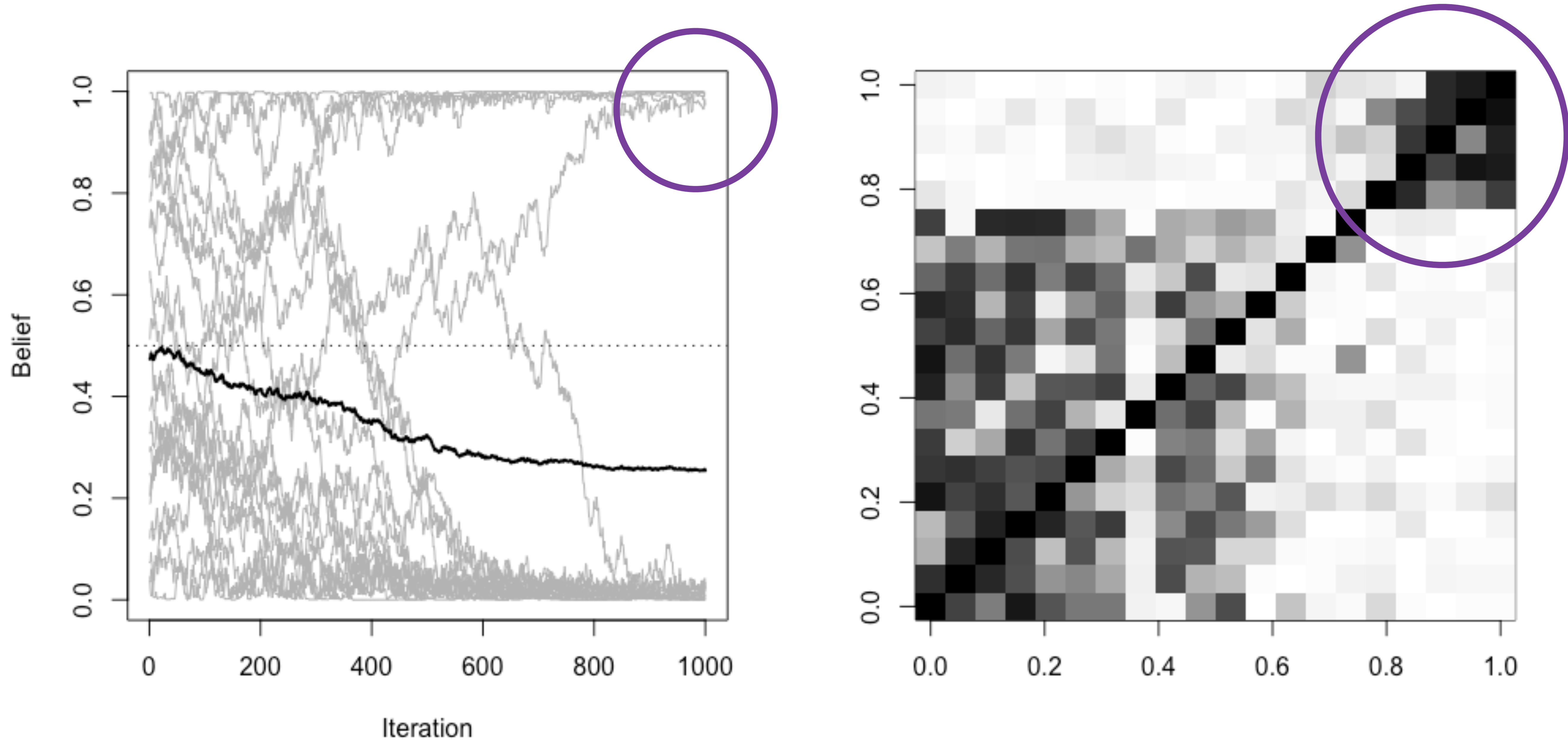
The effect is exaggerated if learners maximise rather than sample

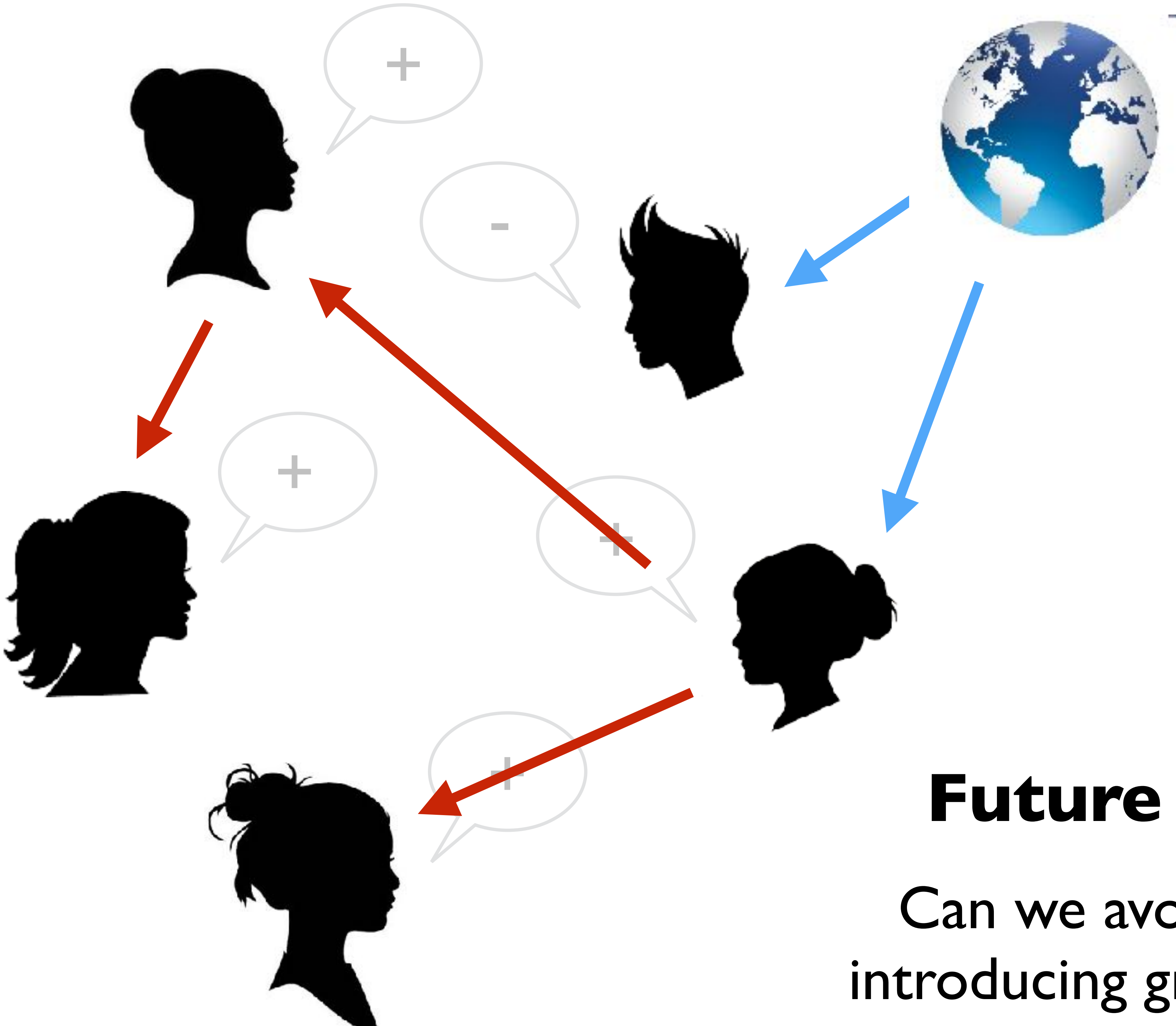




Agents prefer to receive data from trusted sources

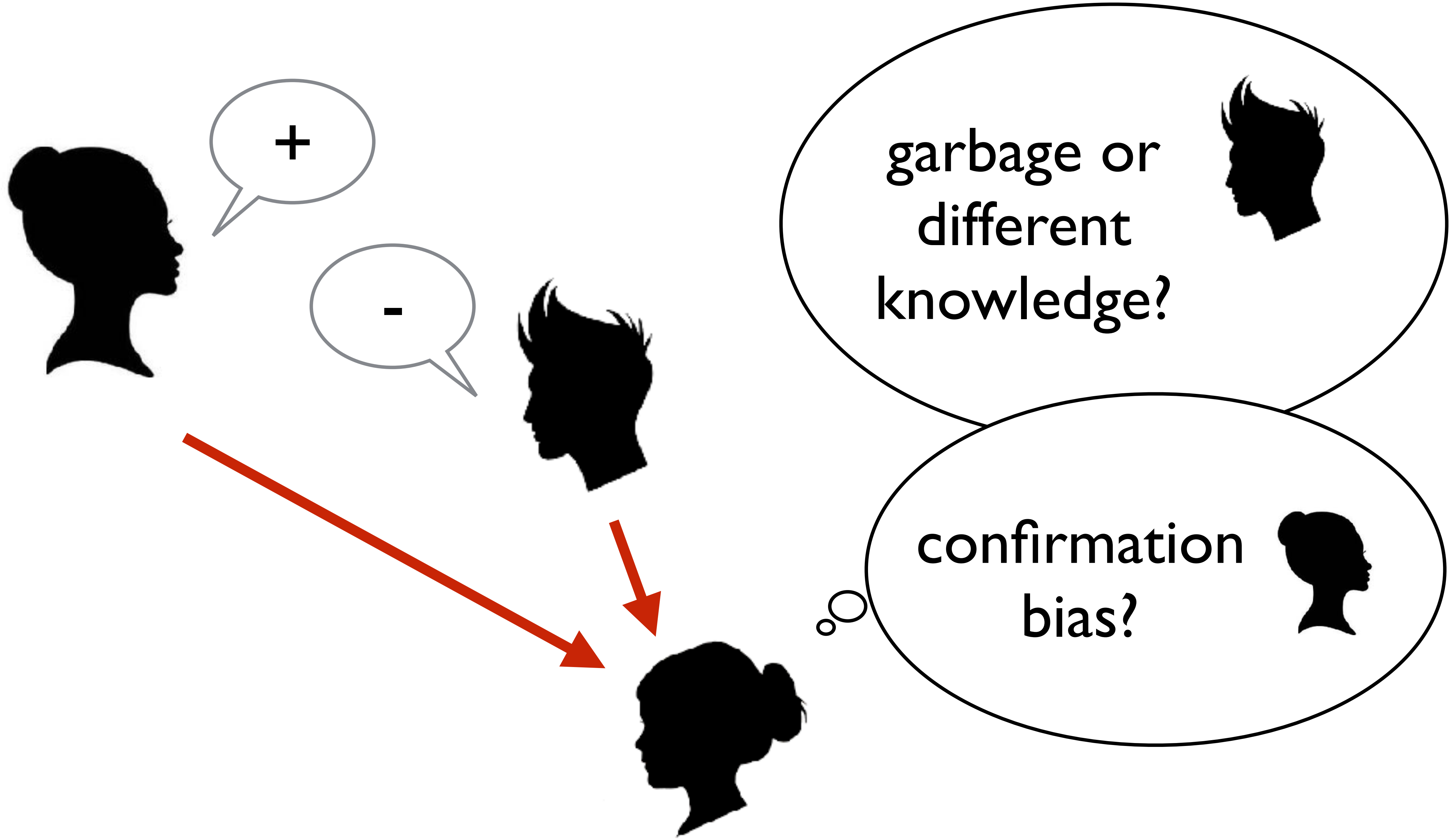
Simple ToM to update trustworthiness


Belief
Iteration
0
200
400
600
800
1000
0.0
0.2
0.4
0.6
0.8
1.0



Future work:

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?