When extremists win

Iterated learning with heterogenous agents

Dan Navarro
School of Psychology
University of New South Wales

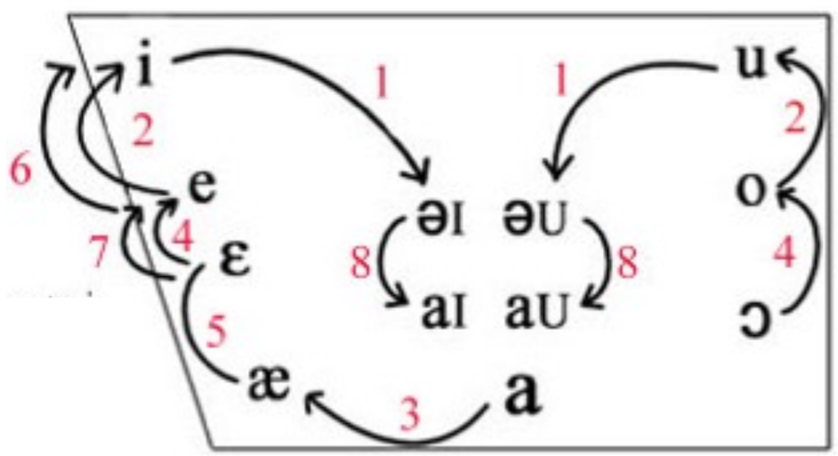
Amy Perfors
School of Psychology
University of Adelaide

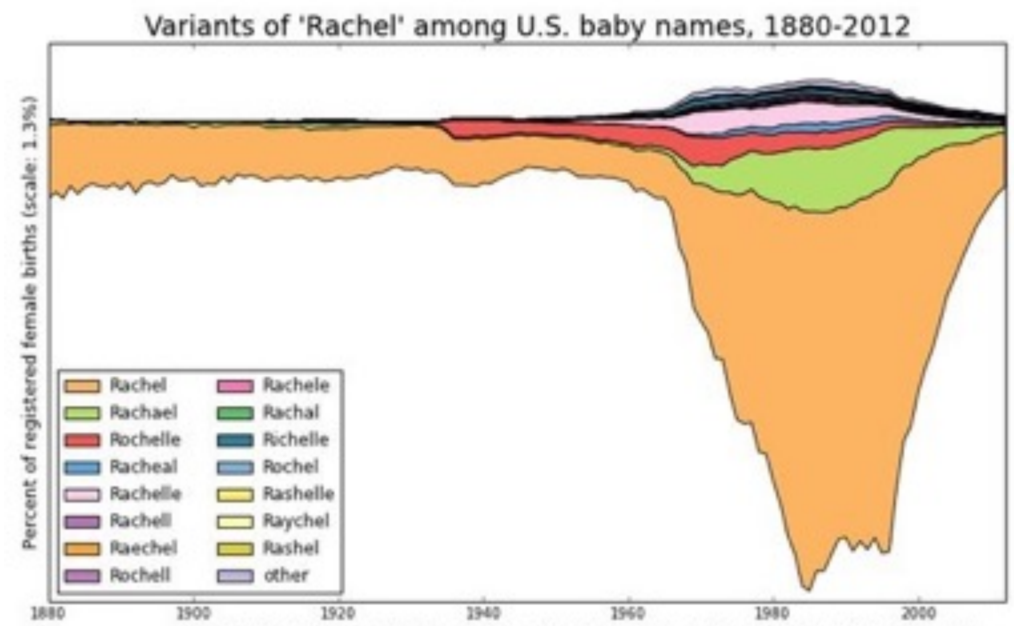
Arthur Kary
School of Psychology
University of New South Wales

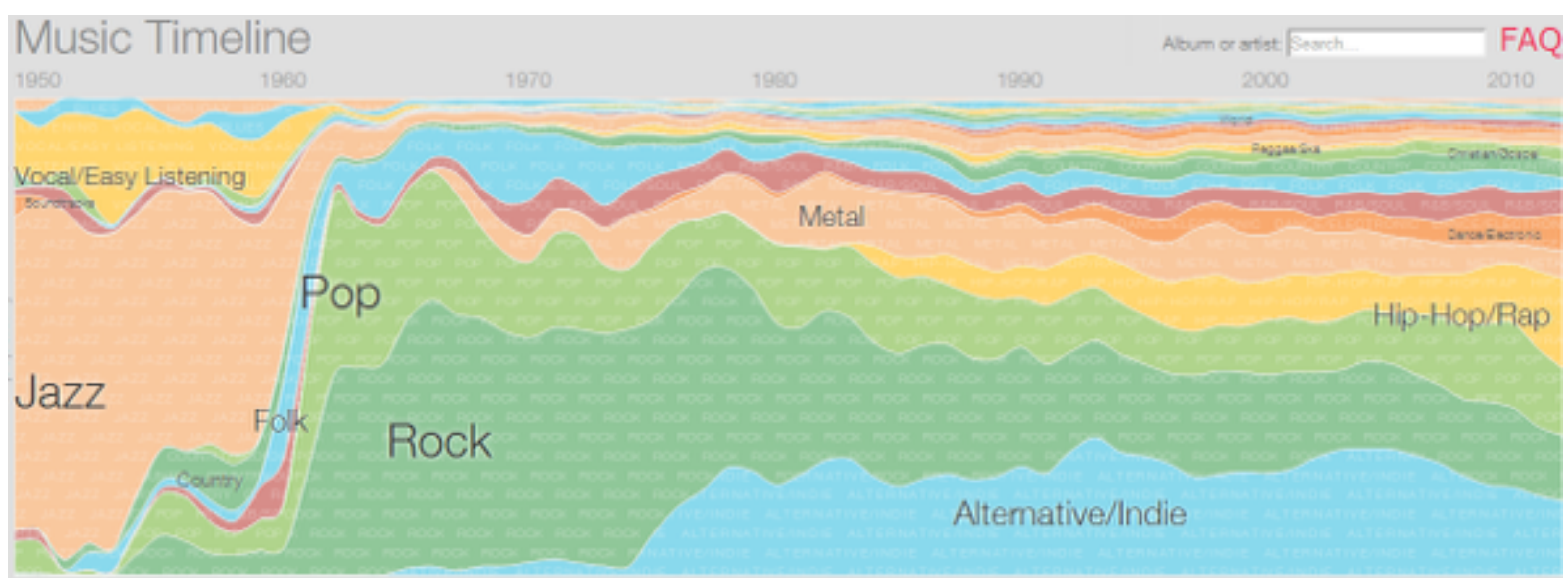
Scott Brown
School of Psychology
University of Newcastle

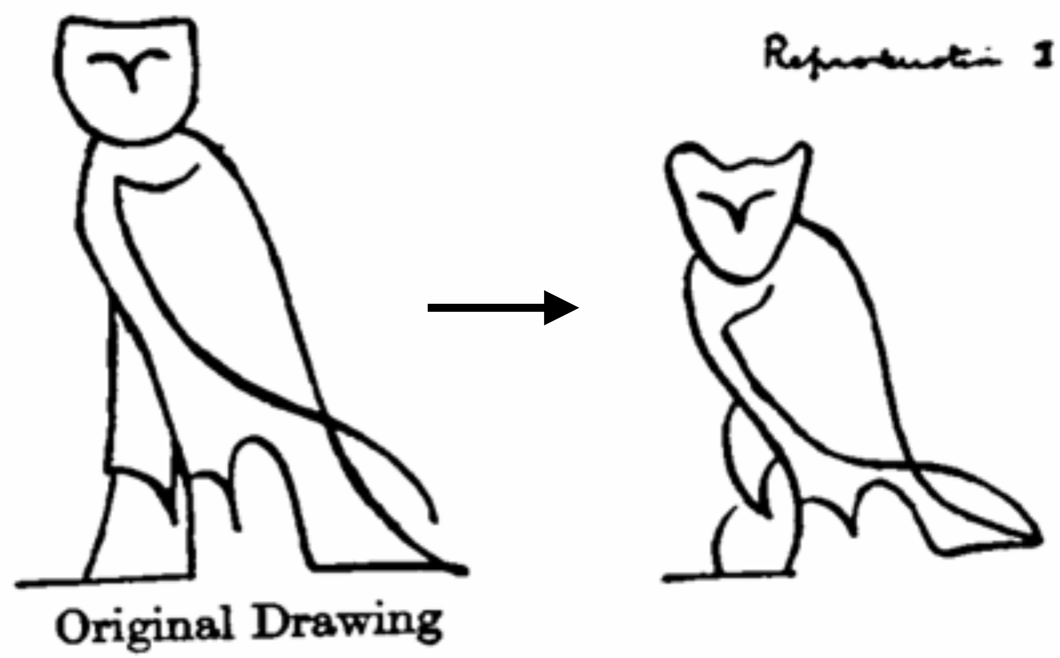
Chris Donkin
School of Psychology
University of New South Wales

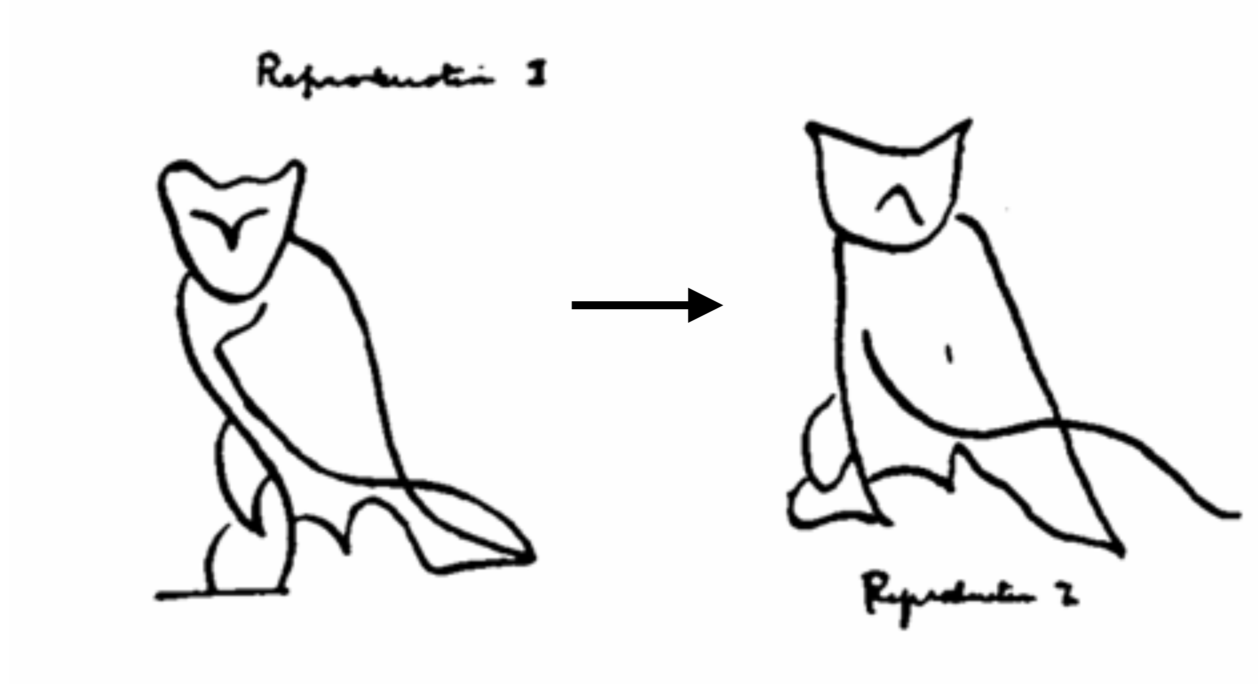
What dynamics underpin cultural and linguistic change? What do they say about the mind?





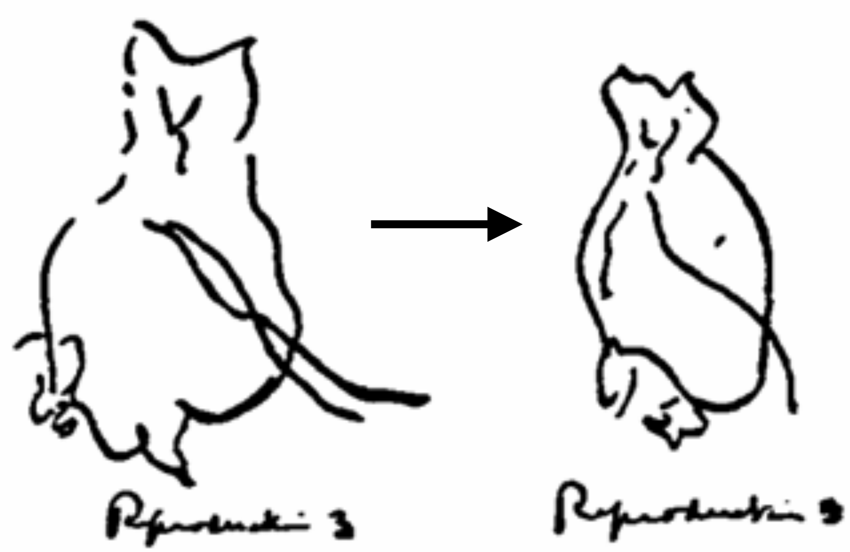


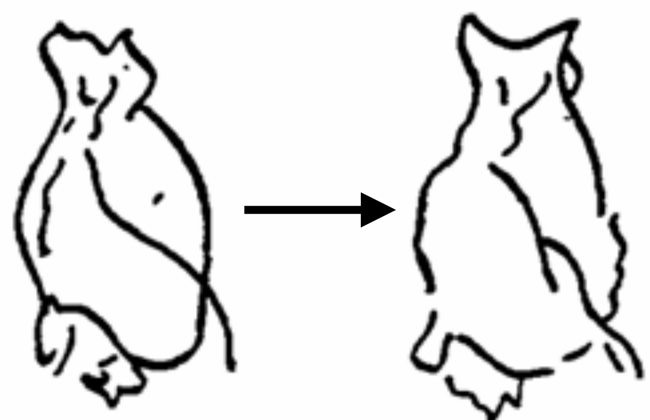

Original Drawing
Reproduction 1


Reproduction 1
Reproduction 2

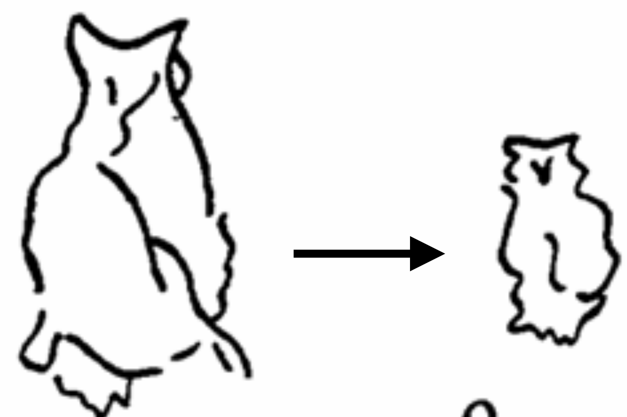
Reproduction 2

Reproduction 3


Reproduction 3
Reproduction 9



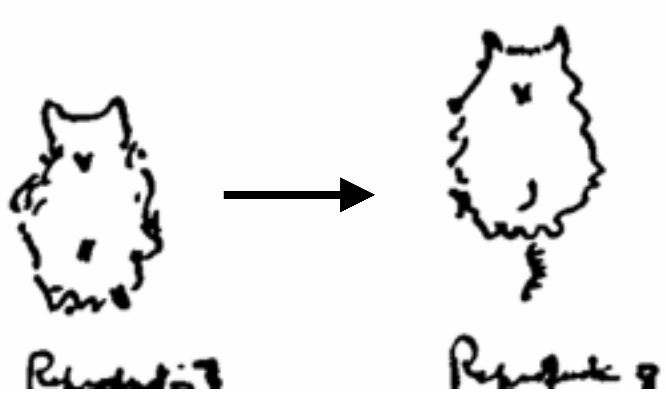
Reproduction 3 Reproduction 5

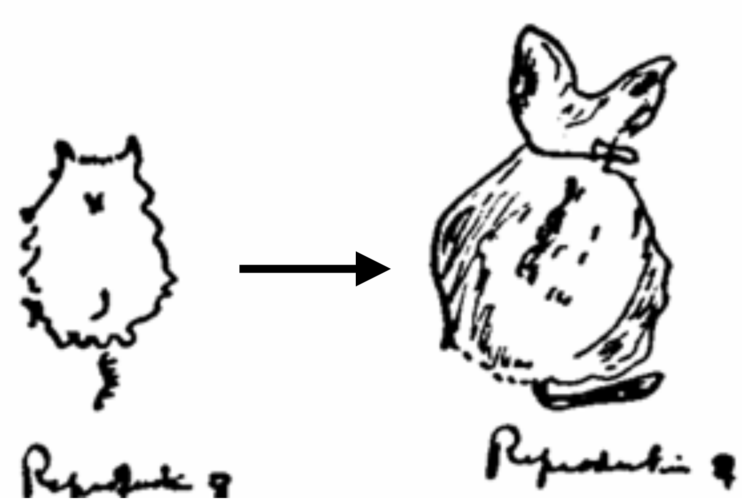


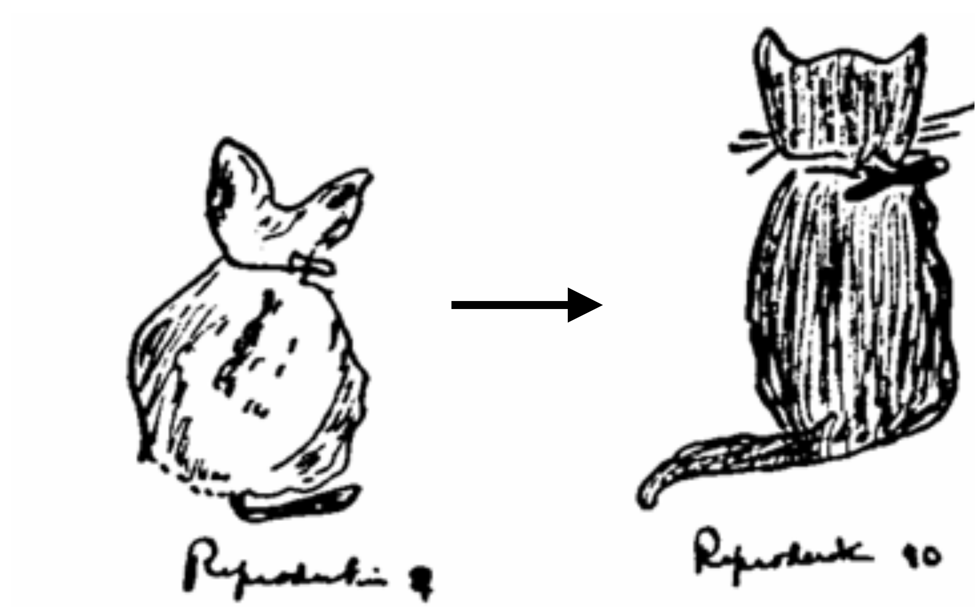
Reproduction 5 Reproduction 6.

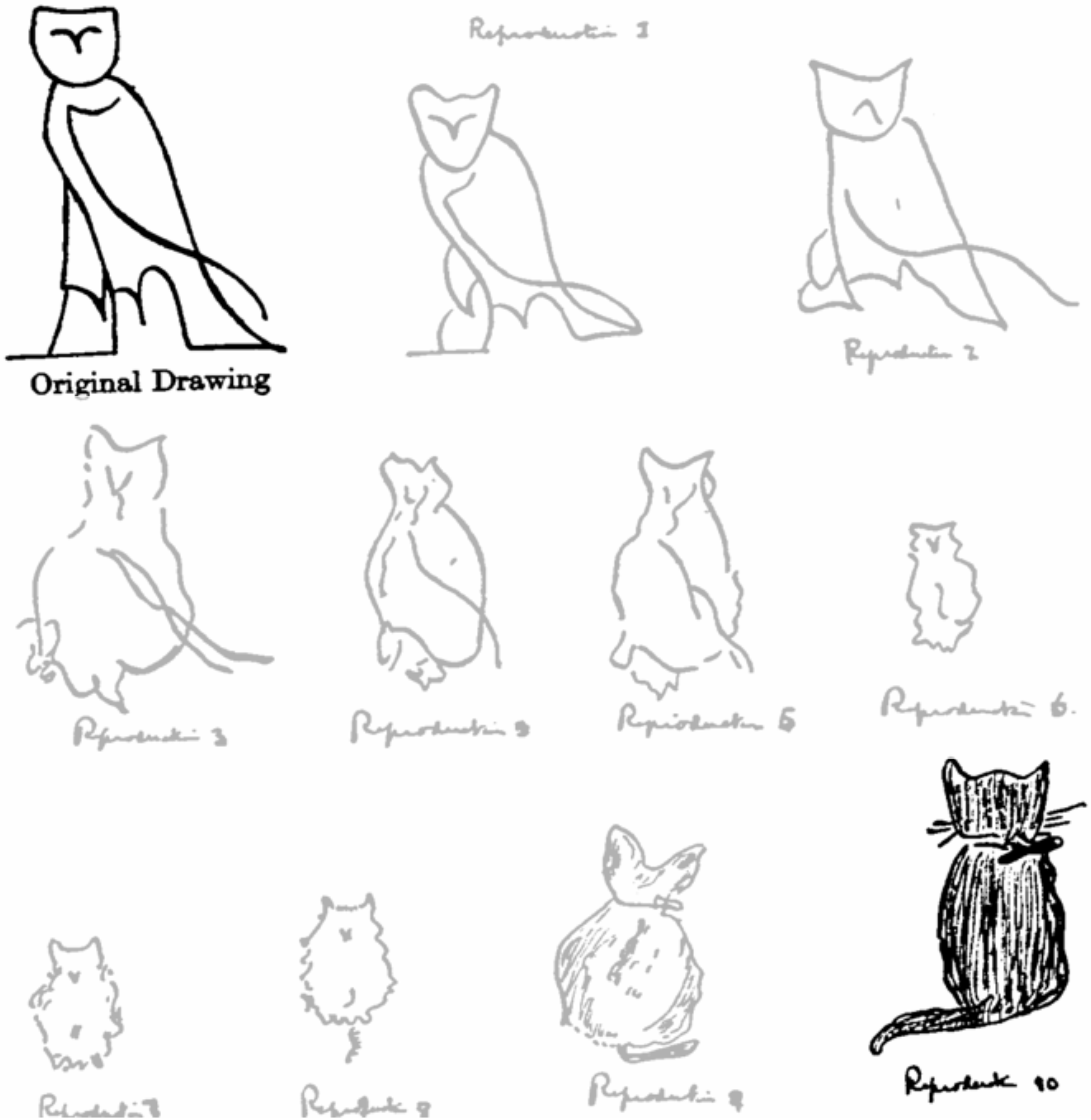
Reproduction 6.

Reproduction 7



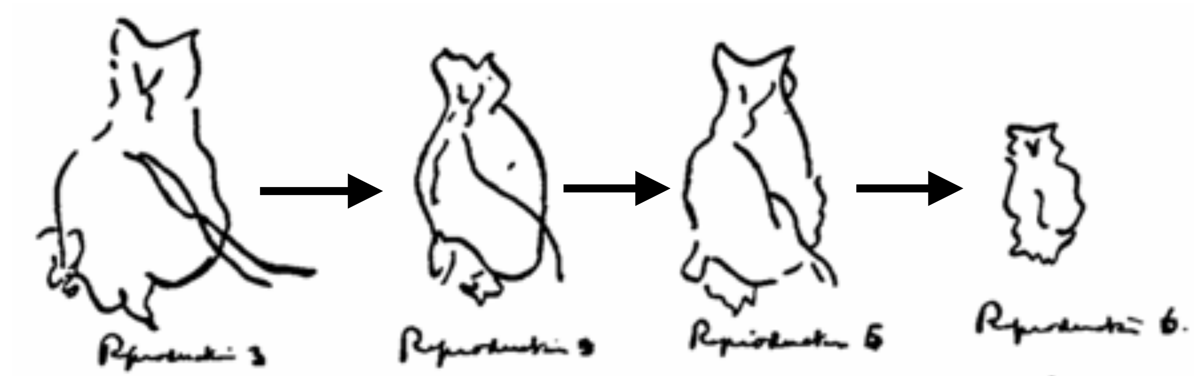



Reproduction 9
Reproduction 10


Original Drawing
Reproduction 1
Reproduction 2
Reproduction 3
Reproduction 5
Reproduction 6.
Reproduction 10

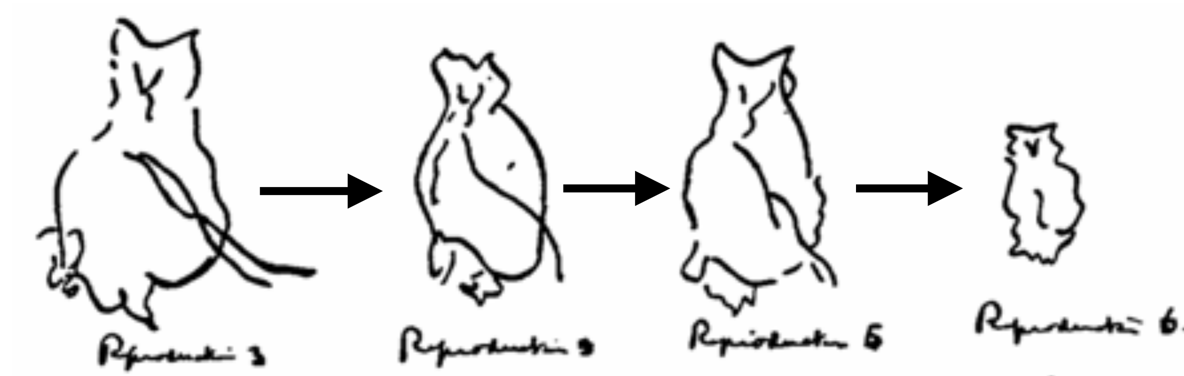
The method of serial reproduction in memory

Bartlett (1920)



The method of serial reproduction in memory

Bartlett (1920)



Language as sequential reproduction of culture

Smith et al (2002)

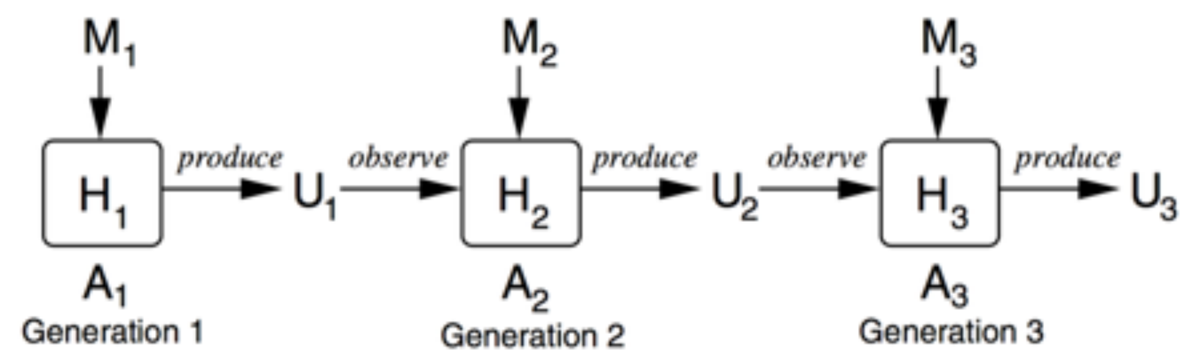
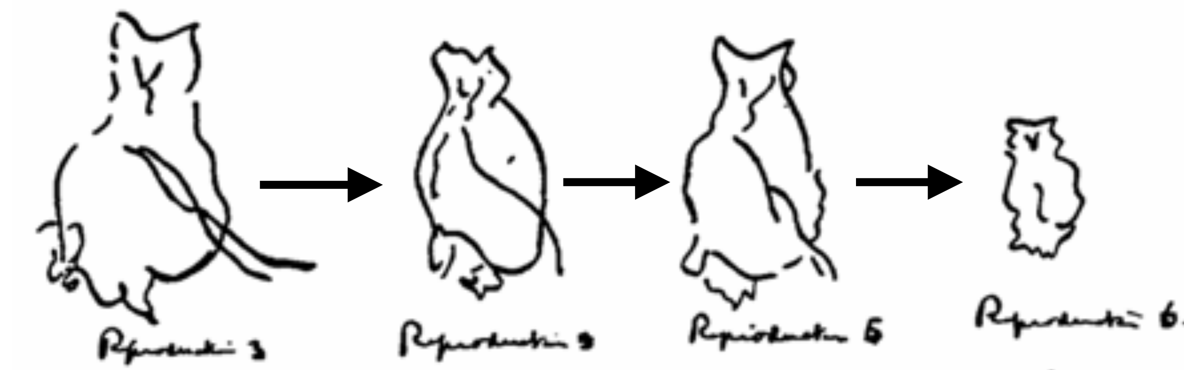


Figure 2. The iterated learning model. The *i*th generation of the population consists of a single agent A_i who has hypothesis H_i. Agent A_i is prompted with a set of meanings M_i. For each of these meanings the agent produces an utterance using H_i. This yields a set of utterances U_i. Agent A_{i+1}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

The method of serial reproduction in memory

Bartlett (1920)



Language as sequential reproduction of culture

Smith et al (2002)

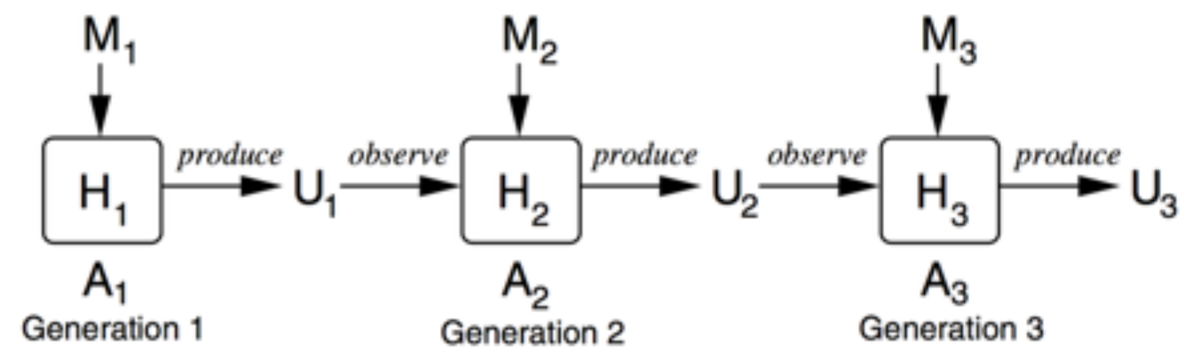
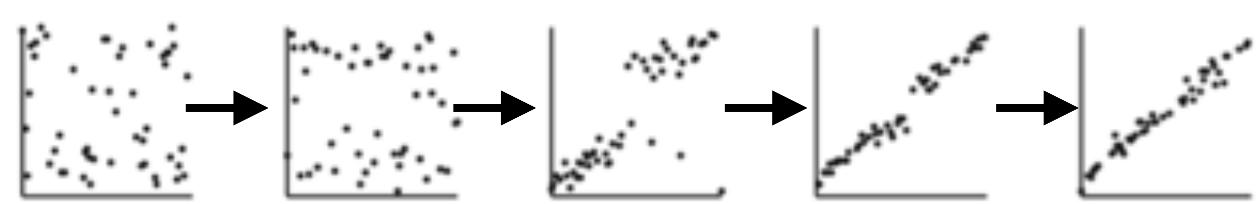


Figure 2. The iterated learning model. The *i*th generation of the population consists of a single agent A_i who has hypothesis H_i. Agent A_i is prompted with a set of meanings M_i. For each of these meanings the agent produces an utterance using H_i. This yields a set of utterances U_i. Agent A_{i+1}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

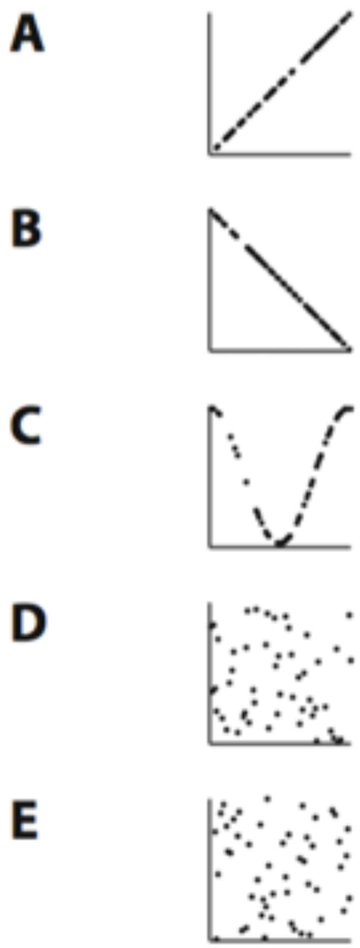
The method of iterated learning reveals inductive bias

Kalish et al (2007)



Example: function learning

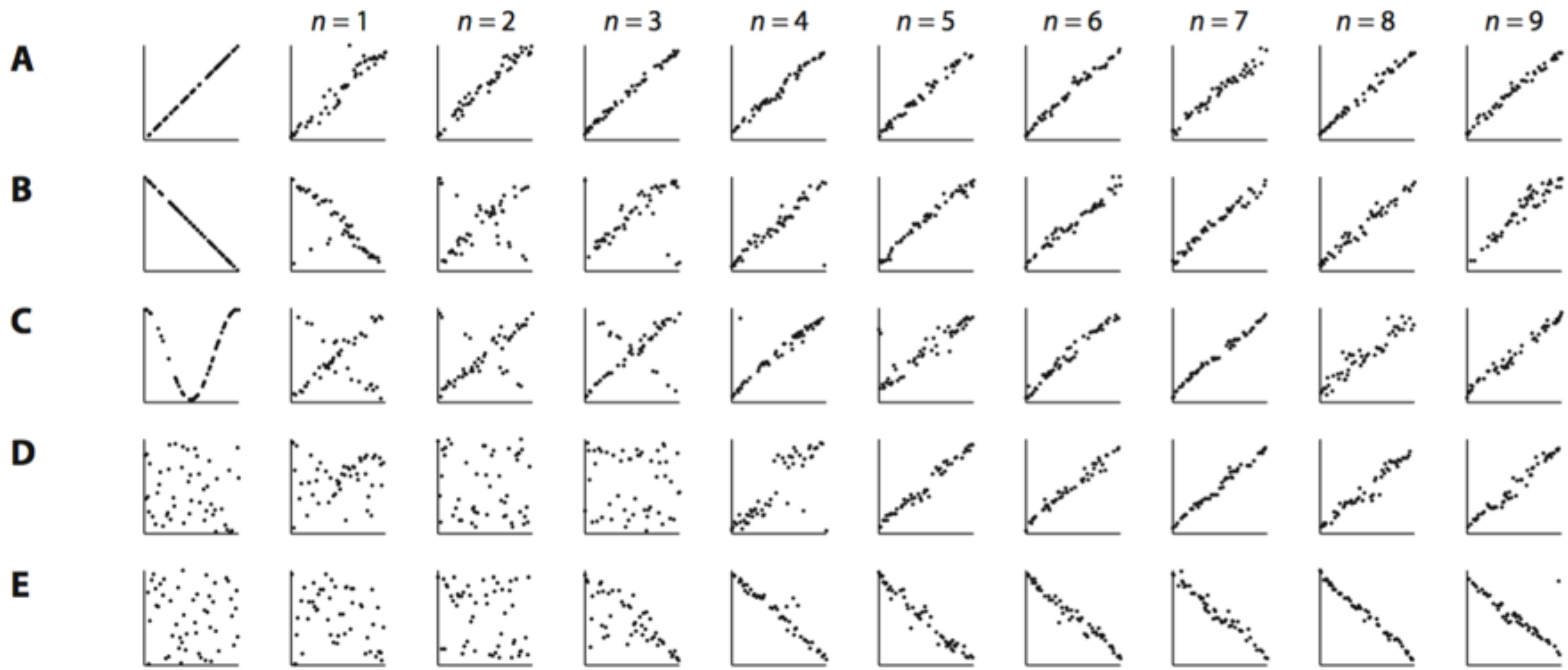
(Kalish et al 2007)



original

Example: function learning

(Kalish et al 2007)

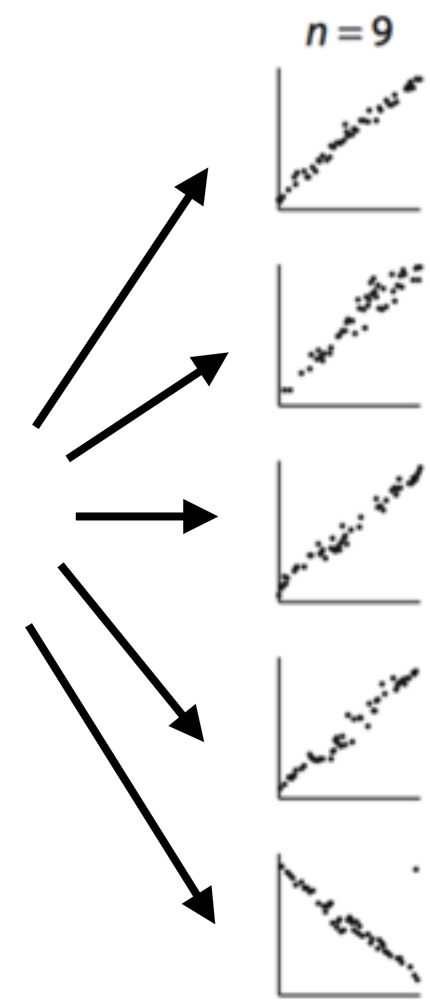


original → final

Example: function learning

(Kalish et al 2007)

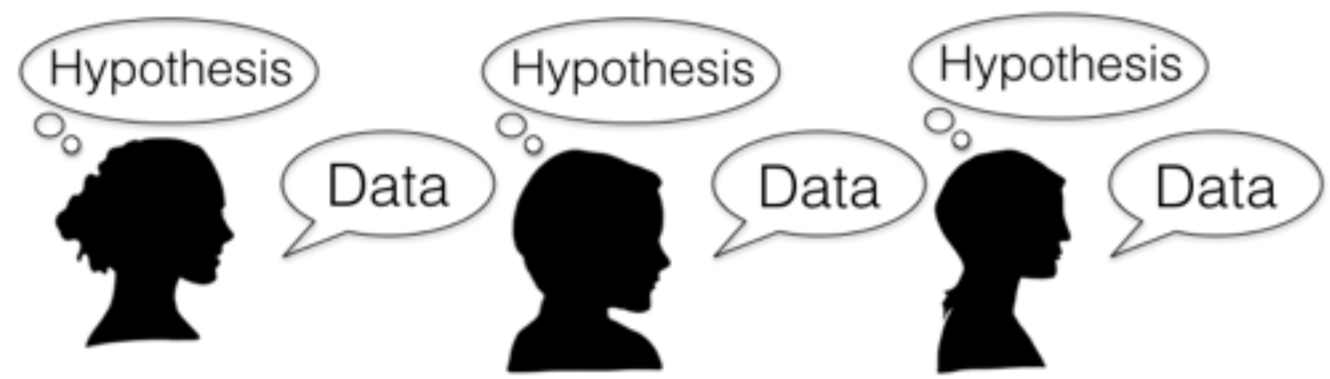
Conclusion: we have an inductive bias for linear functions



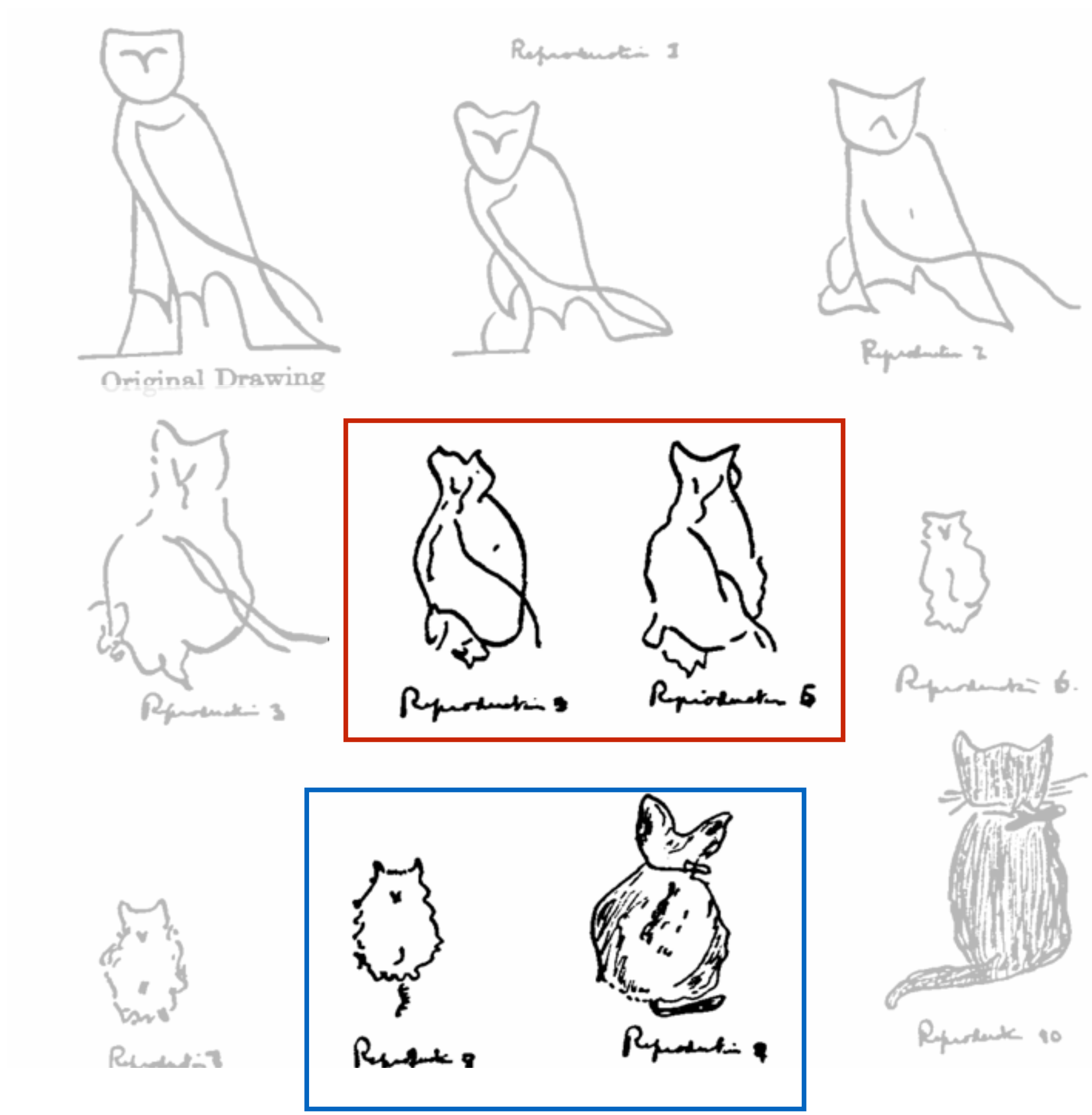
Proof that iterated learning with Bayesian agents reveals the prior

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

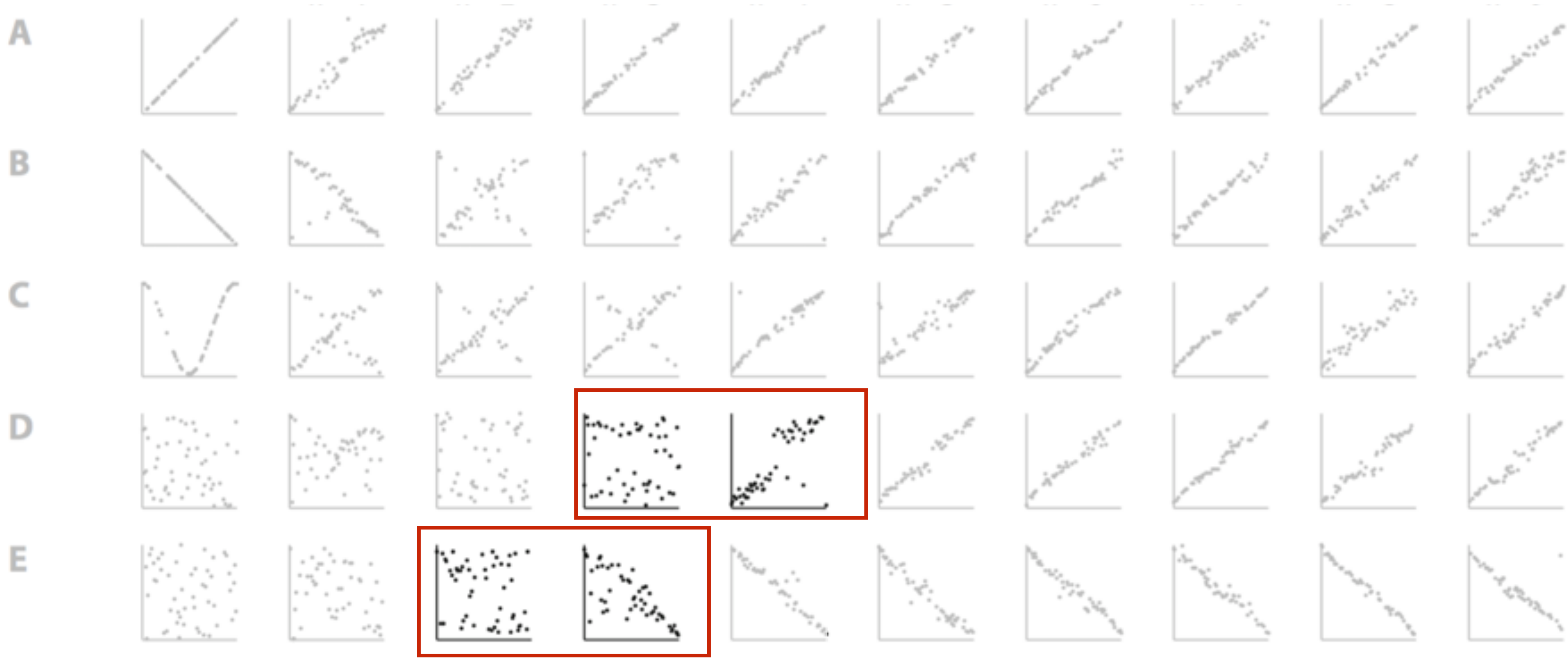
(Griffiths & Kalish 2007)



… as long as everyone has the same prior

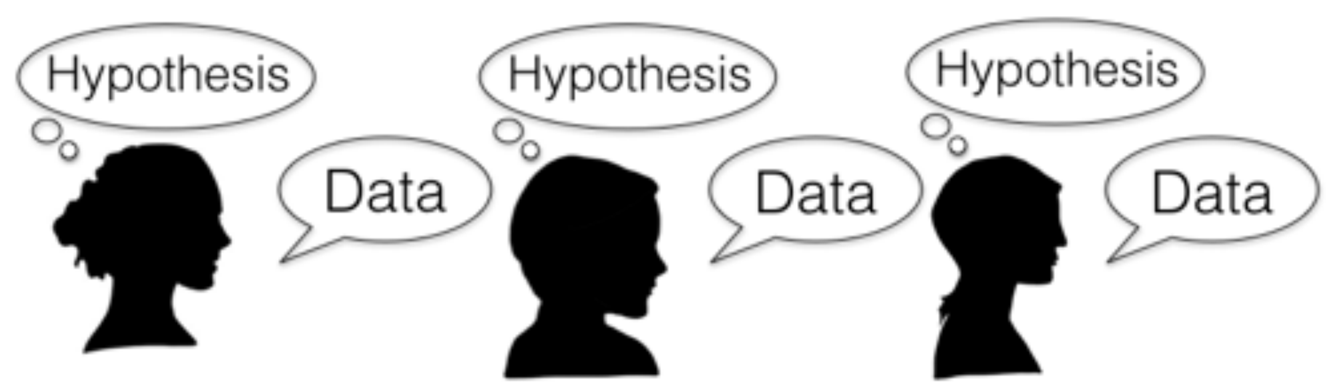


Hm.



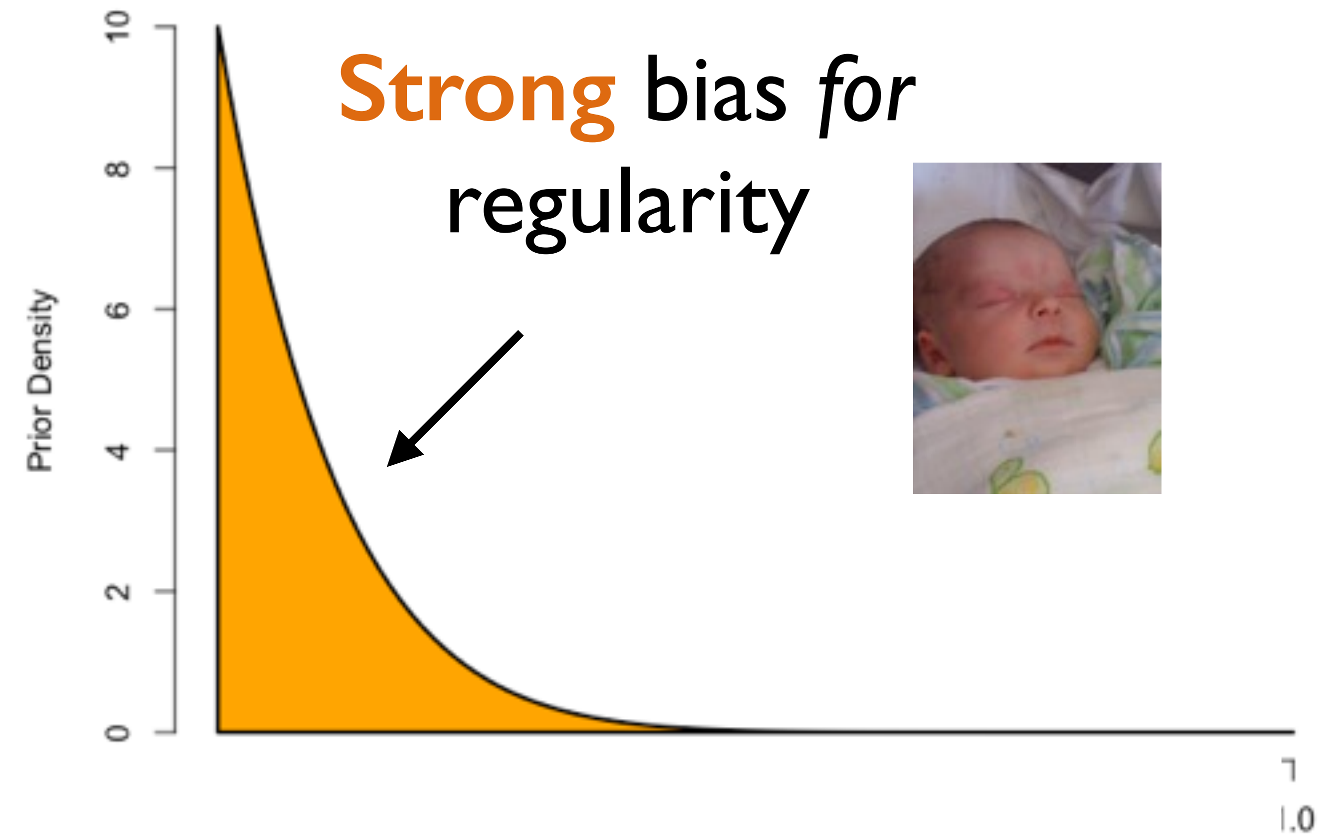
Hm.

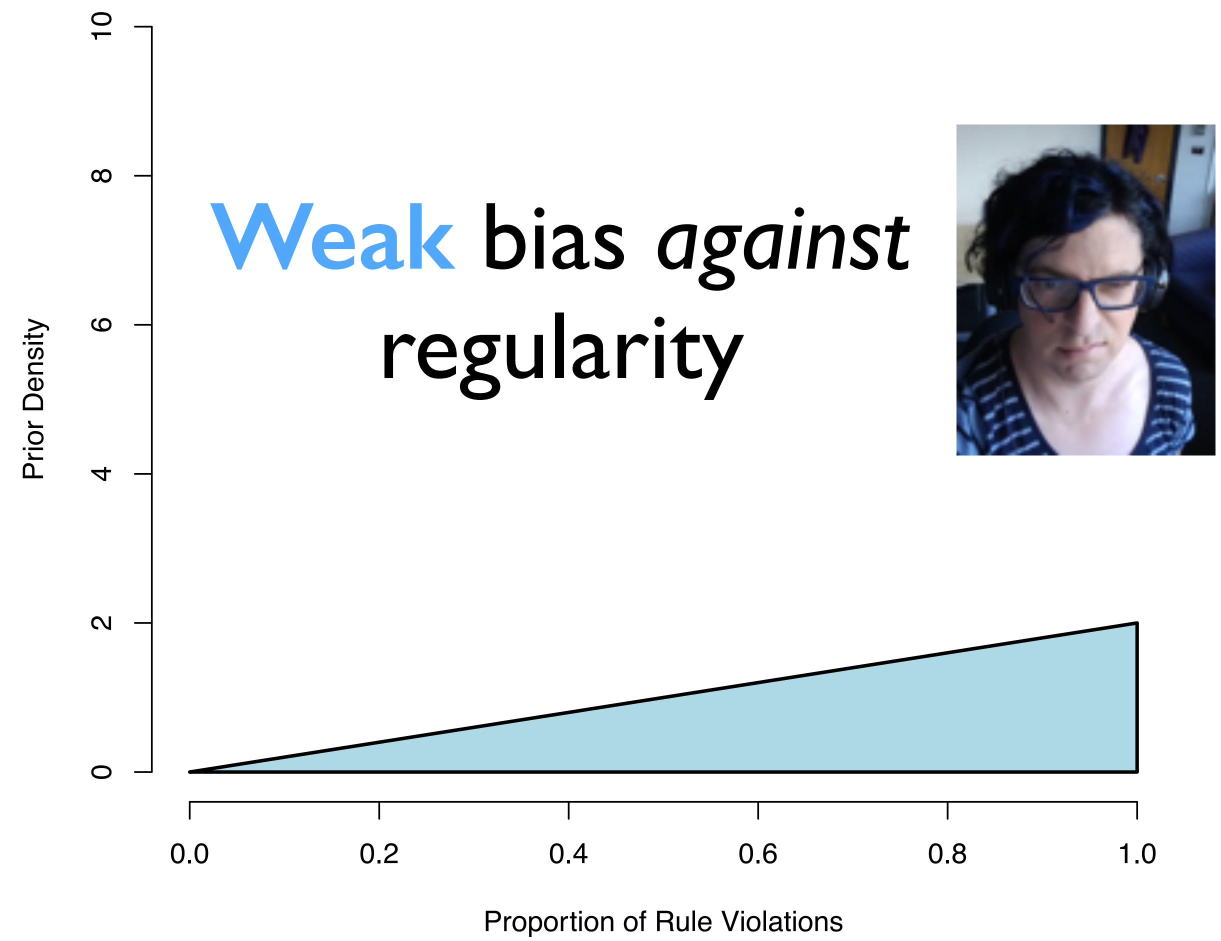
So how do iterated learning chains behave when individual differences exist?

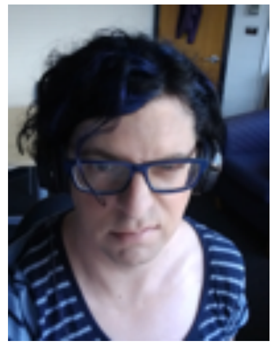


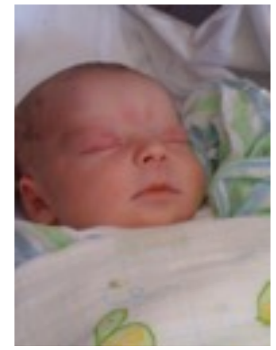
Case study 1:
Does everybody contribute equally to the evolution of languages?

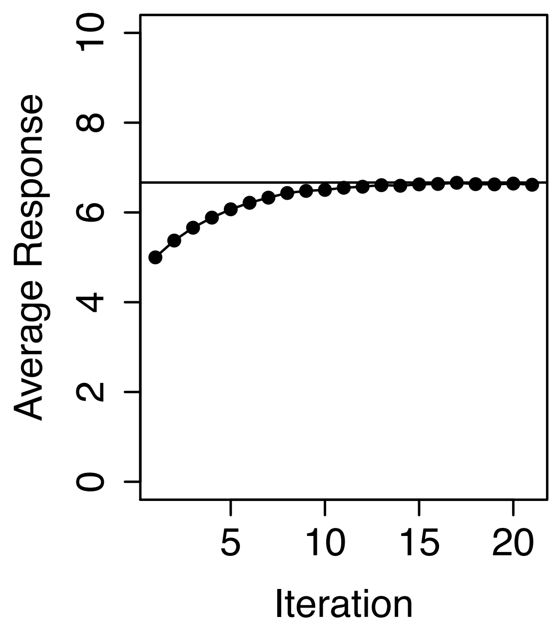
Bayesian models for language regularisation with two different kinds of bias

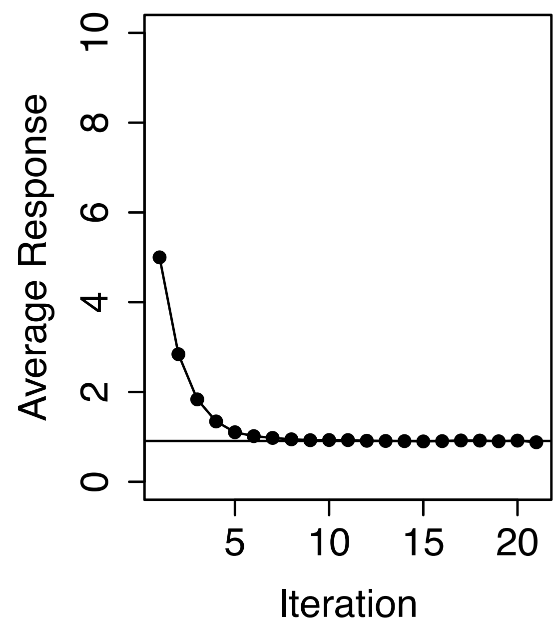




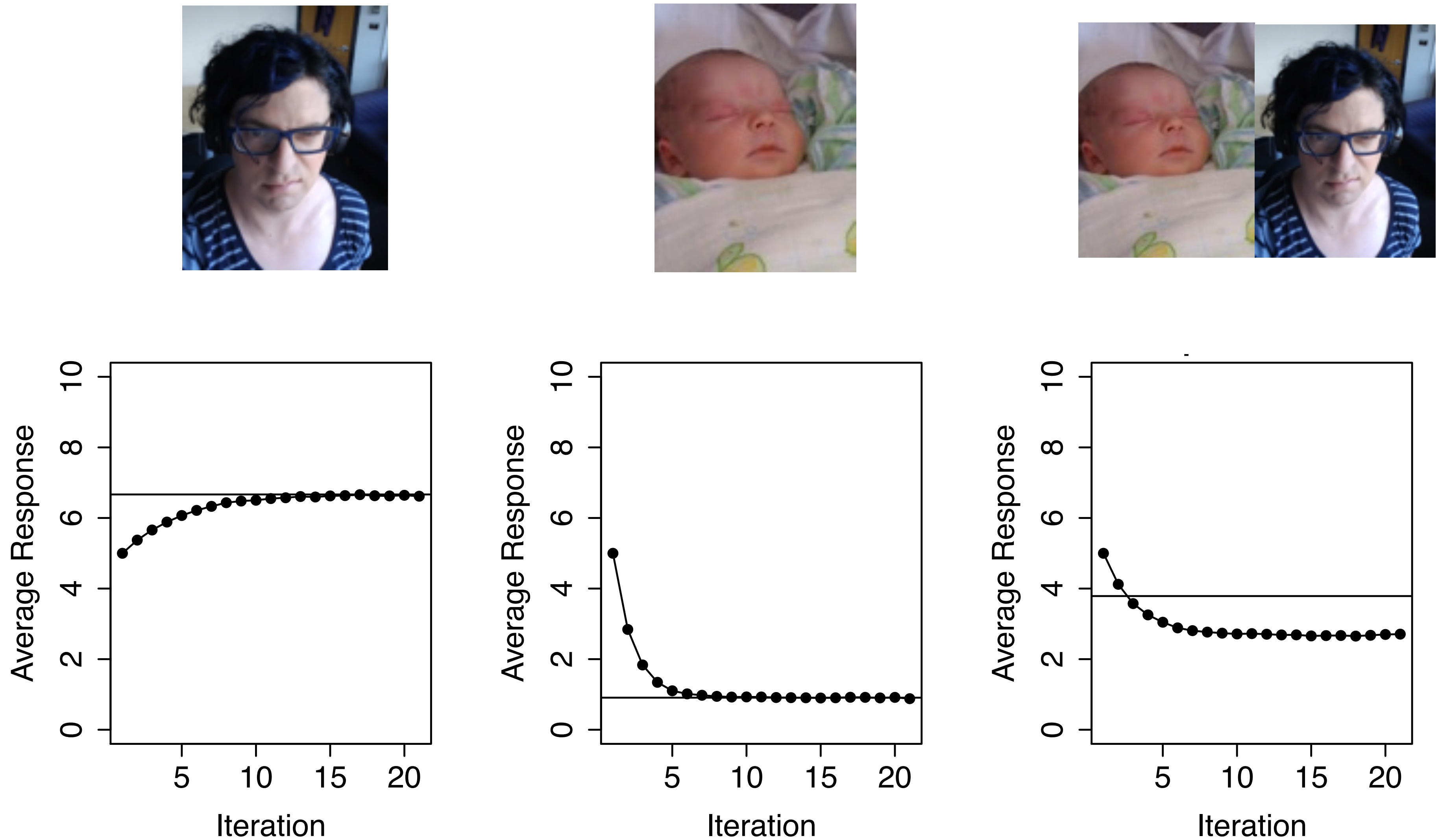




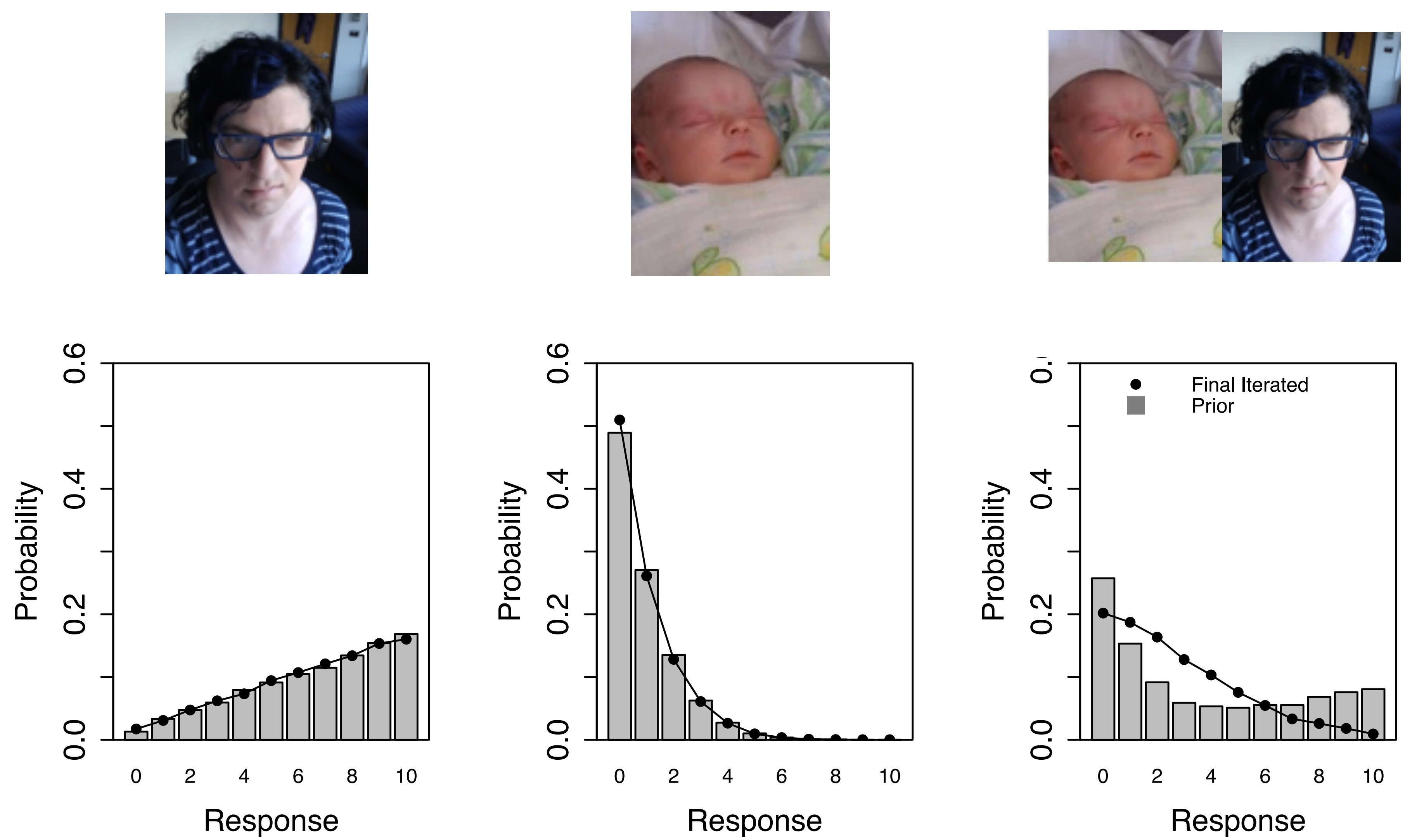




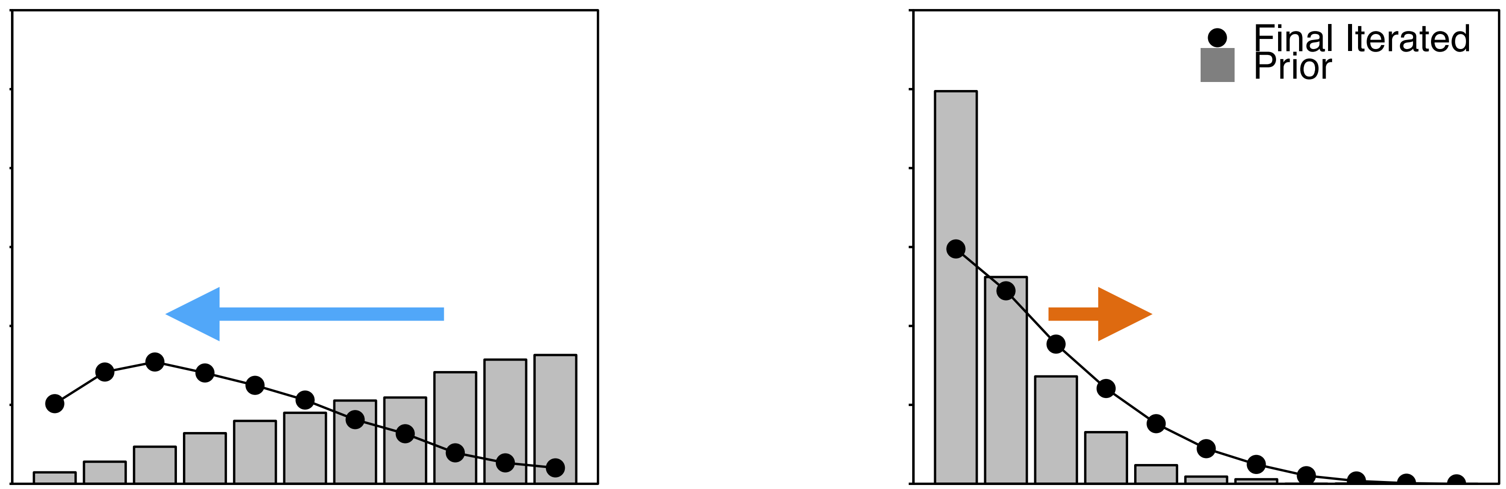
Homogenous iterated learning chains converge to the prior



A heterogenous chain does not converge to the average of the prior biases



…and the distribution of responses is severely distorted weak bias strong bias

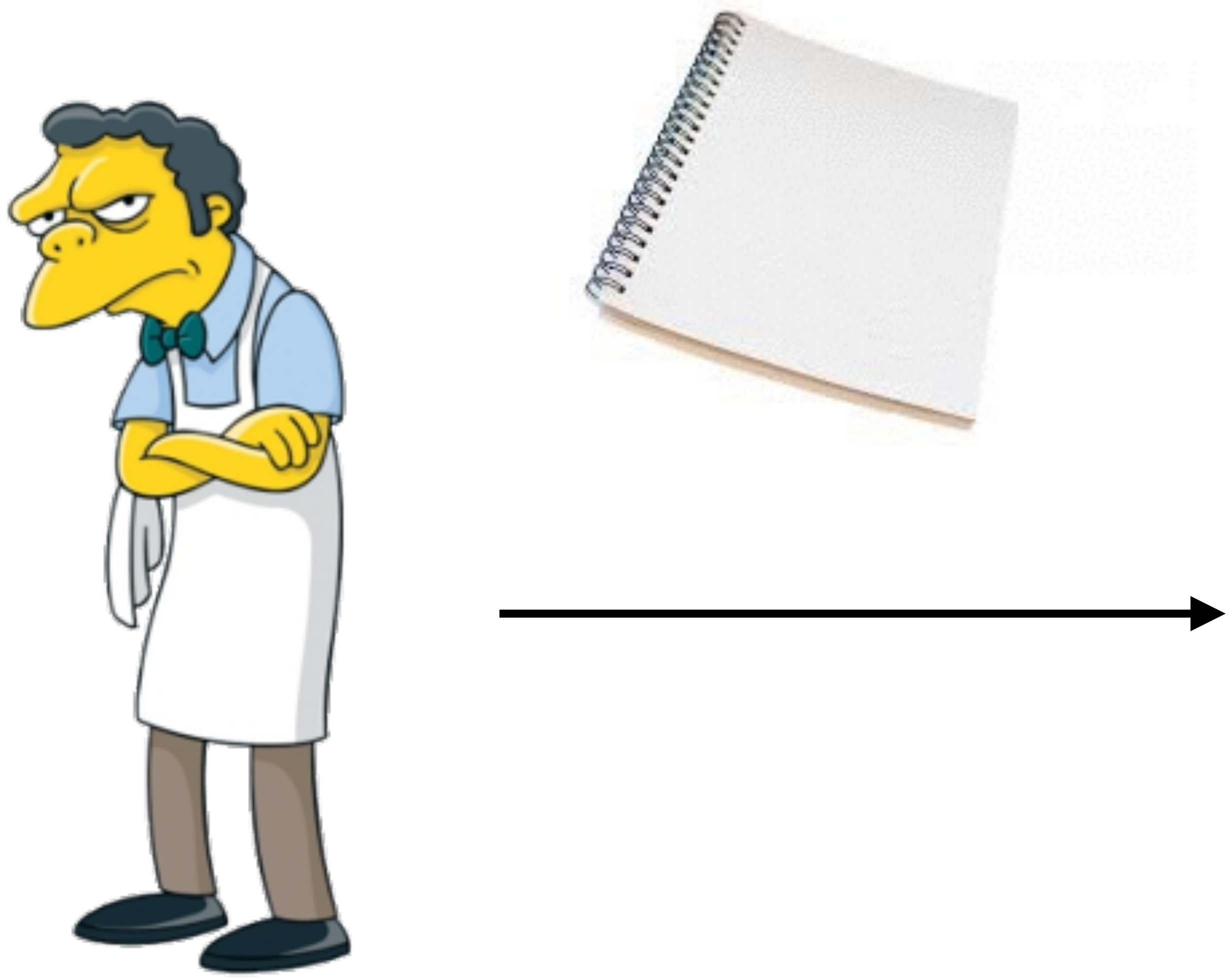


strong bias

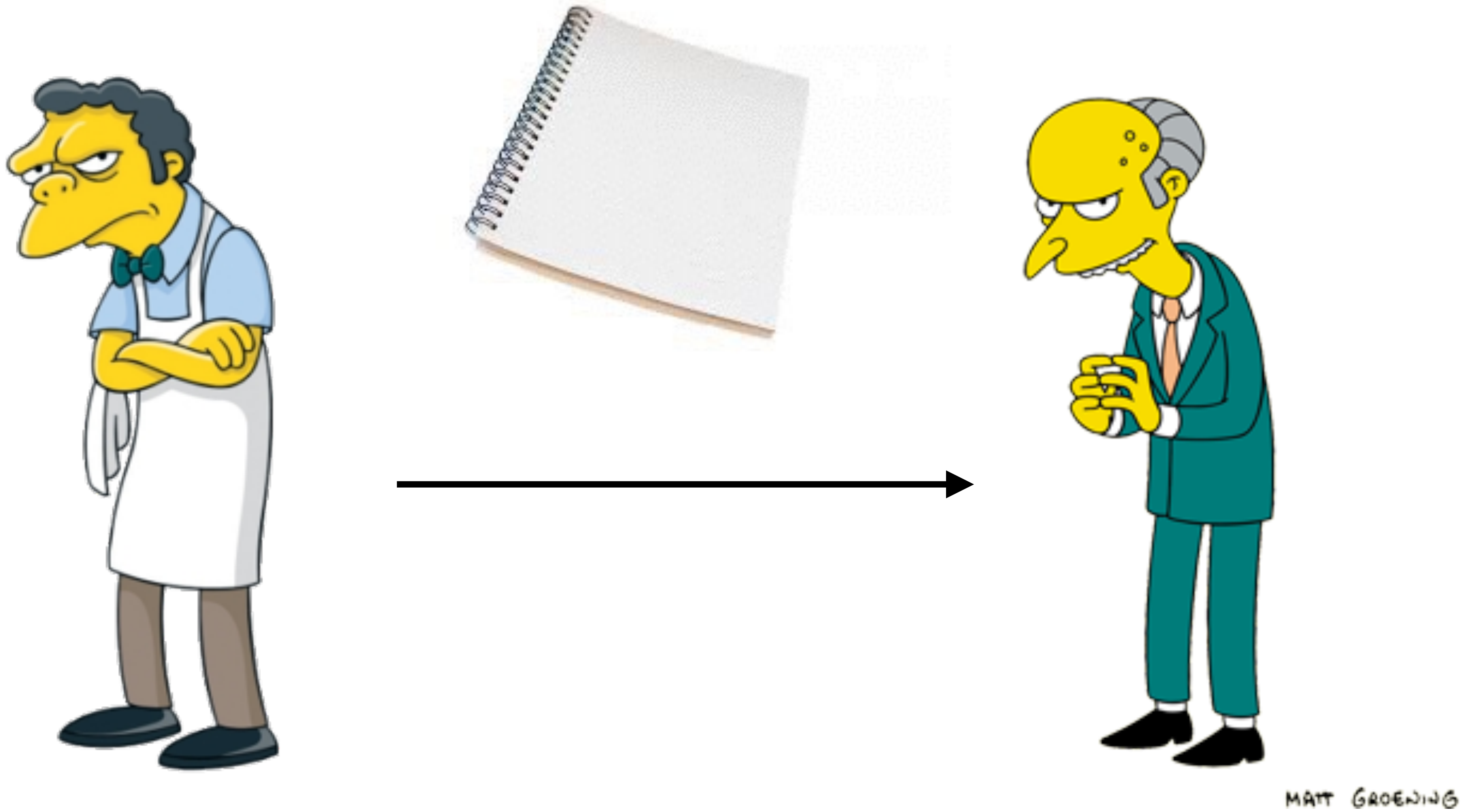
insensitivity to input

greater influence on the chain

<u>Case study 2</u>: Bayesian groupthink



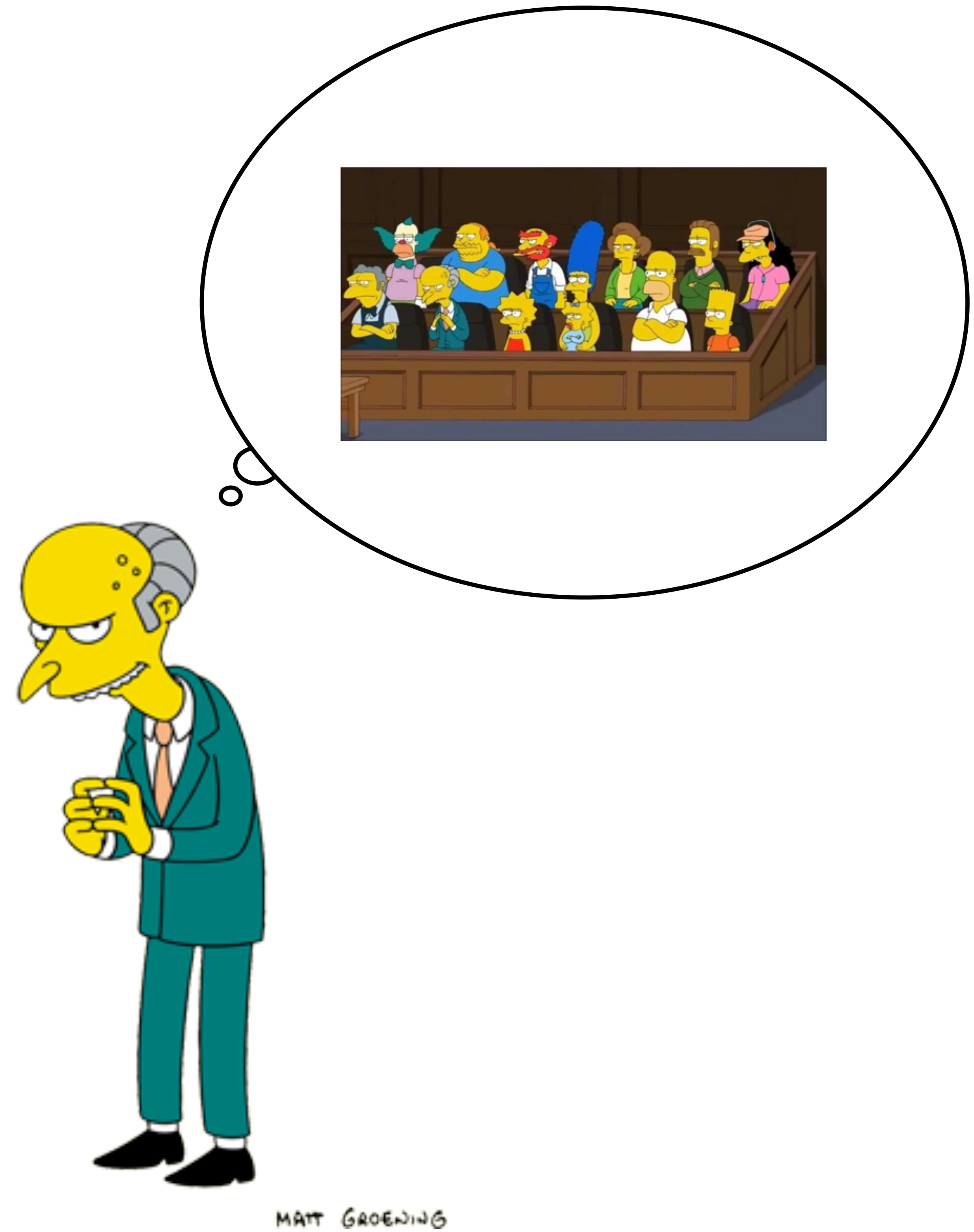
Juror i records vote,
removes sheet, passes
notebook



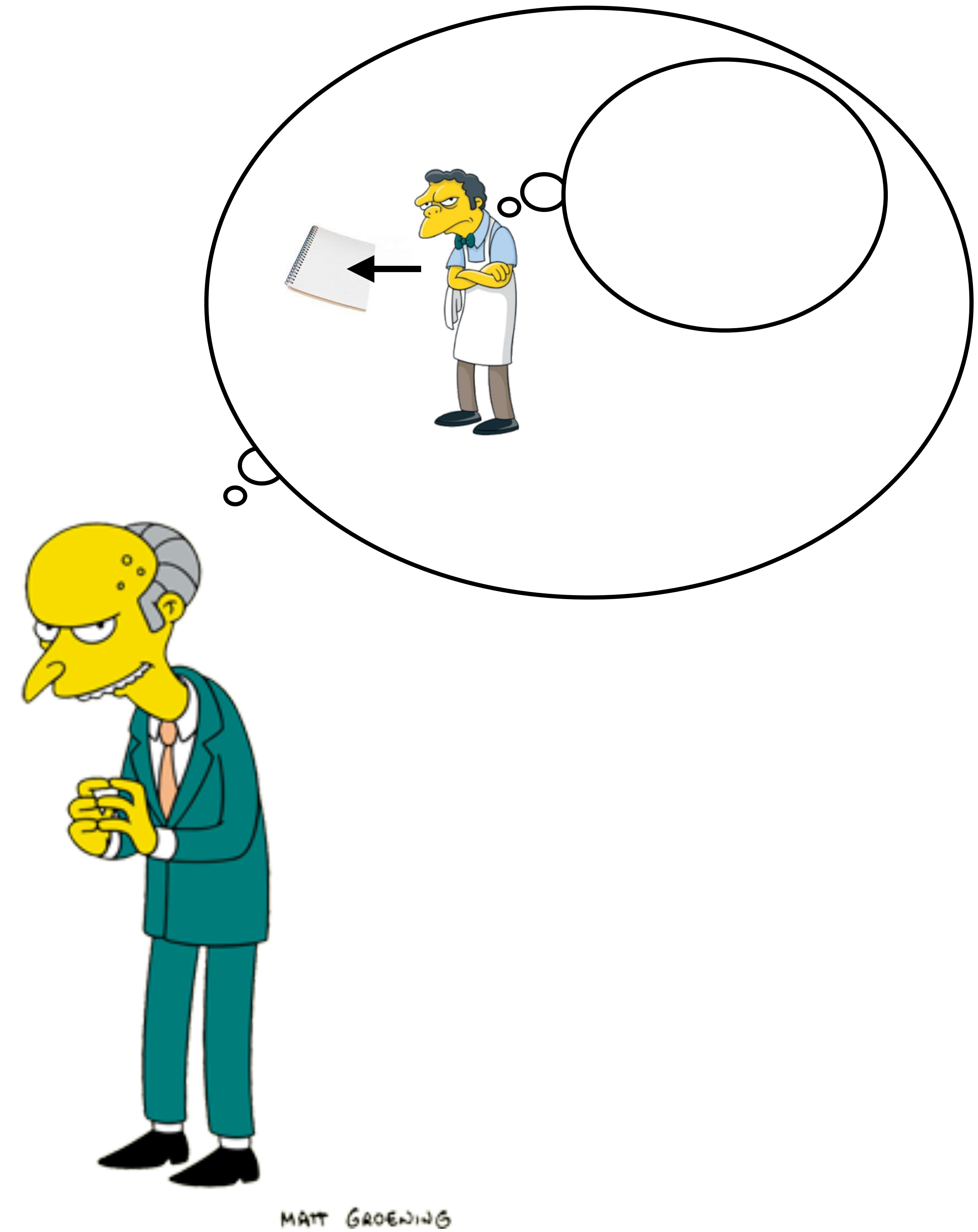
Juror i records vote, removes sheet, passes notebook

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

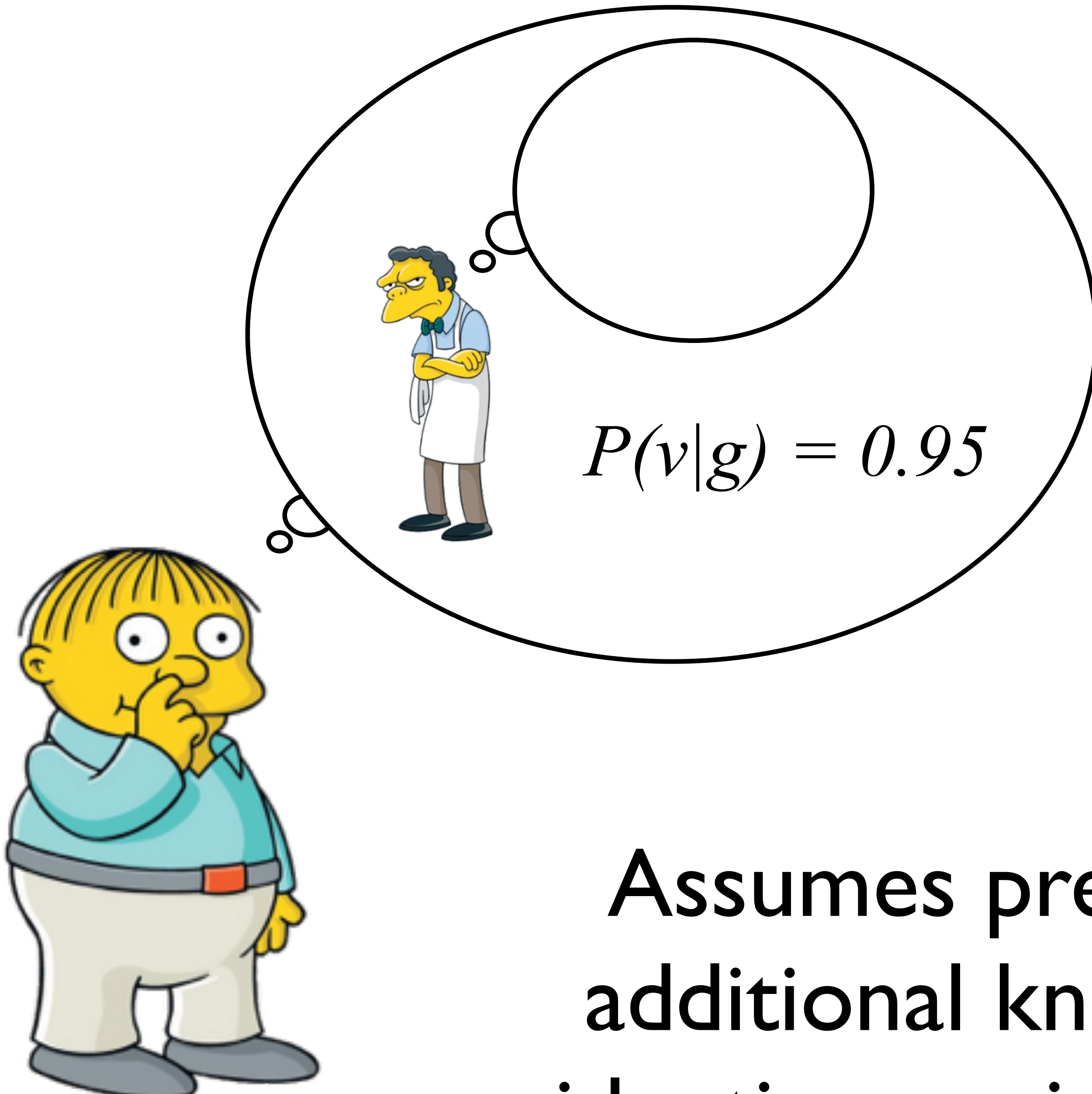
Prior belief about guilt
$P(g)$ is set by the trial



Likelihood of previous juror's vote $P(v|g)$ requires ***theory of mind***… what do they know that I don't know?

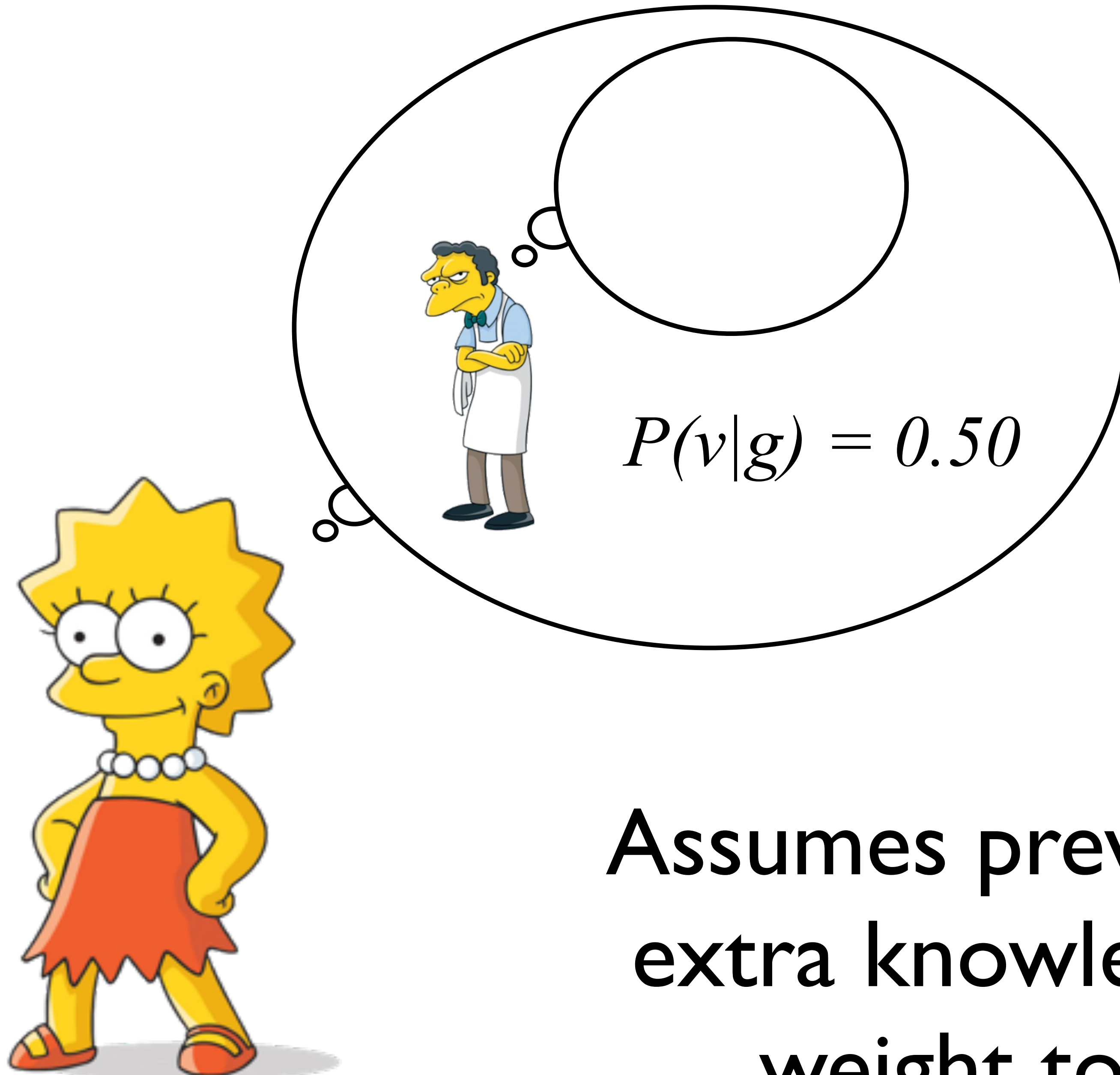


Bayesian “sheep”

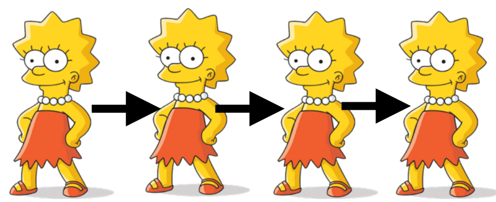


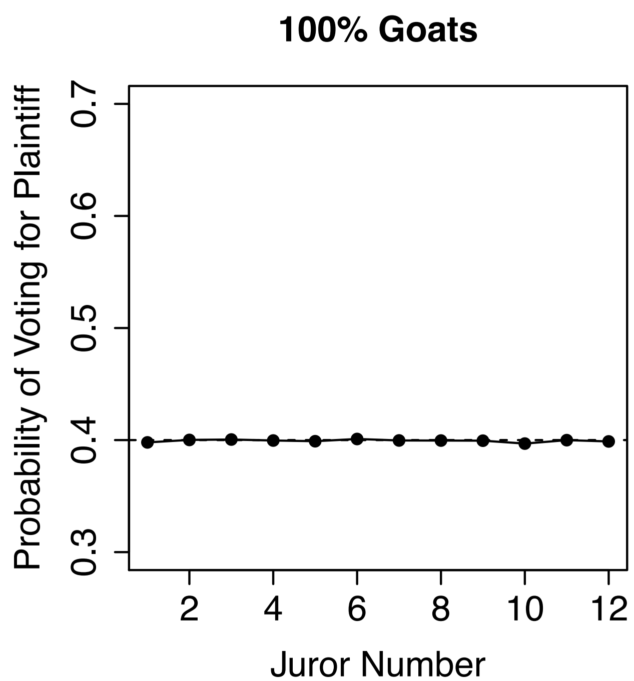
Assumes previous juror has additional knowledge, assigns evidentiary weight to their opinion

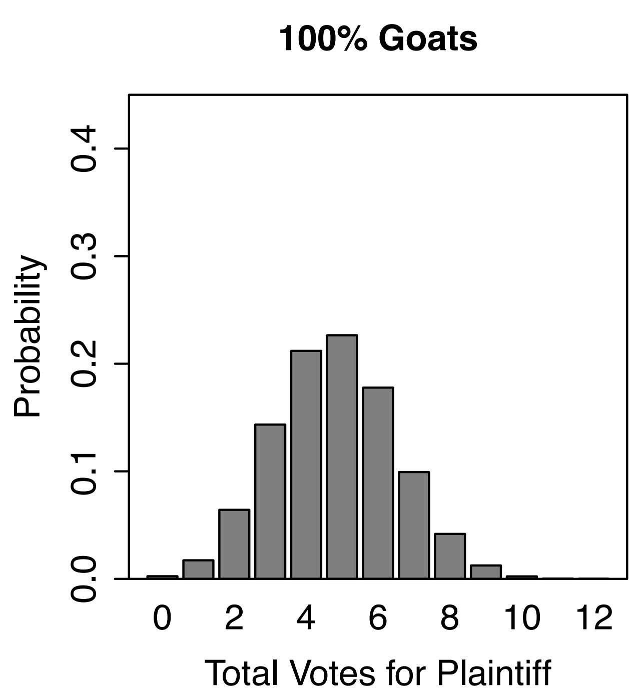
Bayesian "goat"



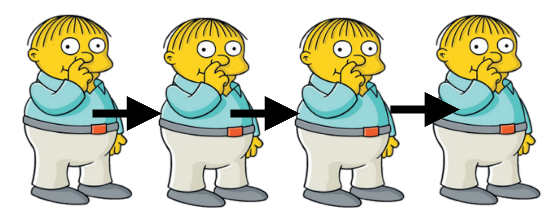
Assumes previous juror has no extra knowledge, assigns zero weight to their opinion

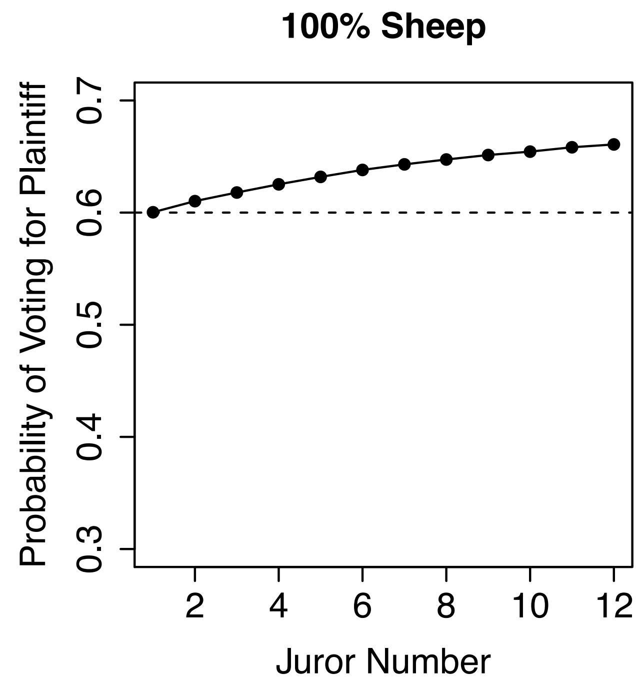


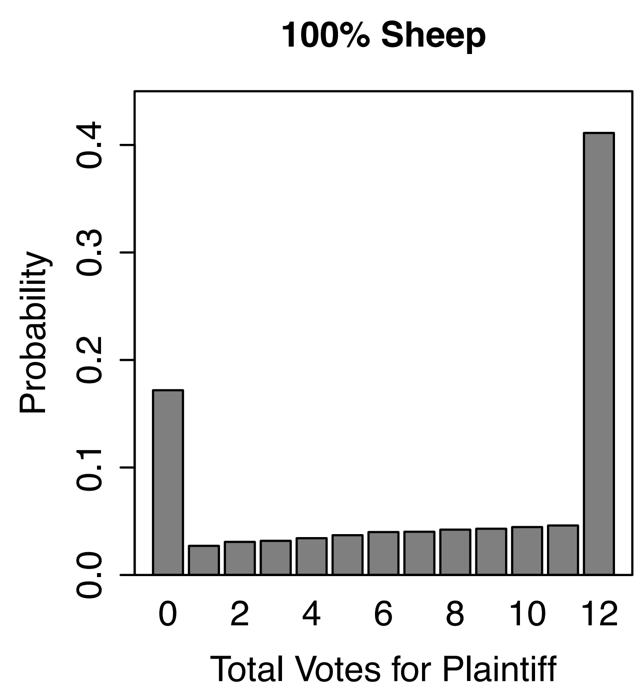




A jury of goats ignores one another and the "chain" converges just fine



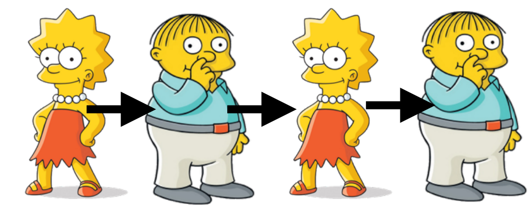


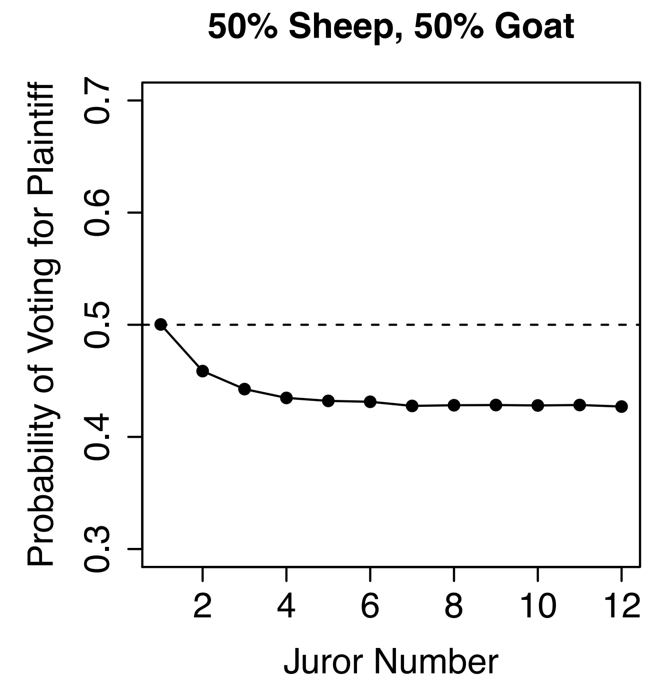


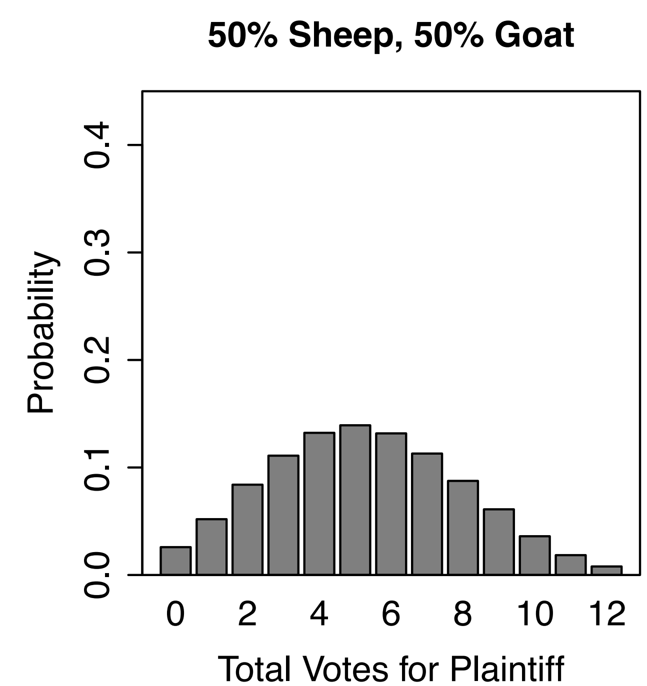
A jury of sheep displays **groupthink**

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

A mixed jury is dominated by goats

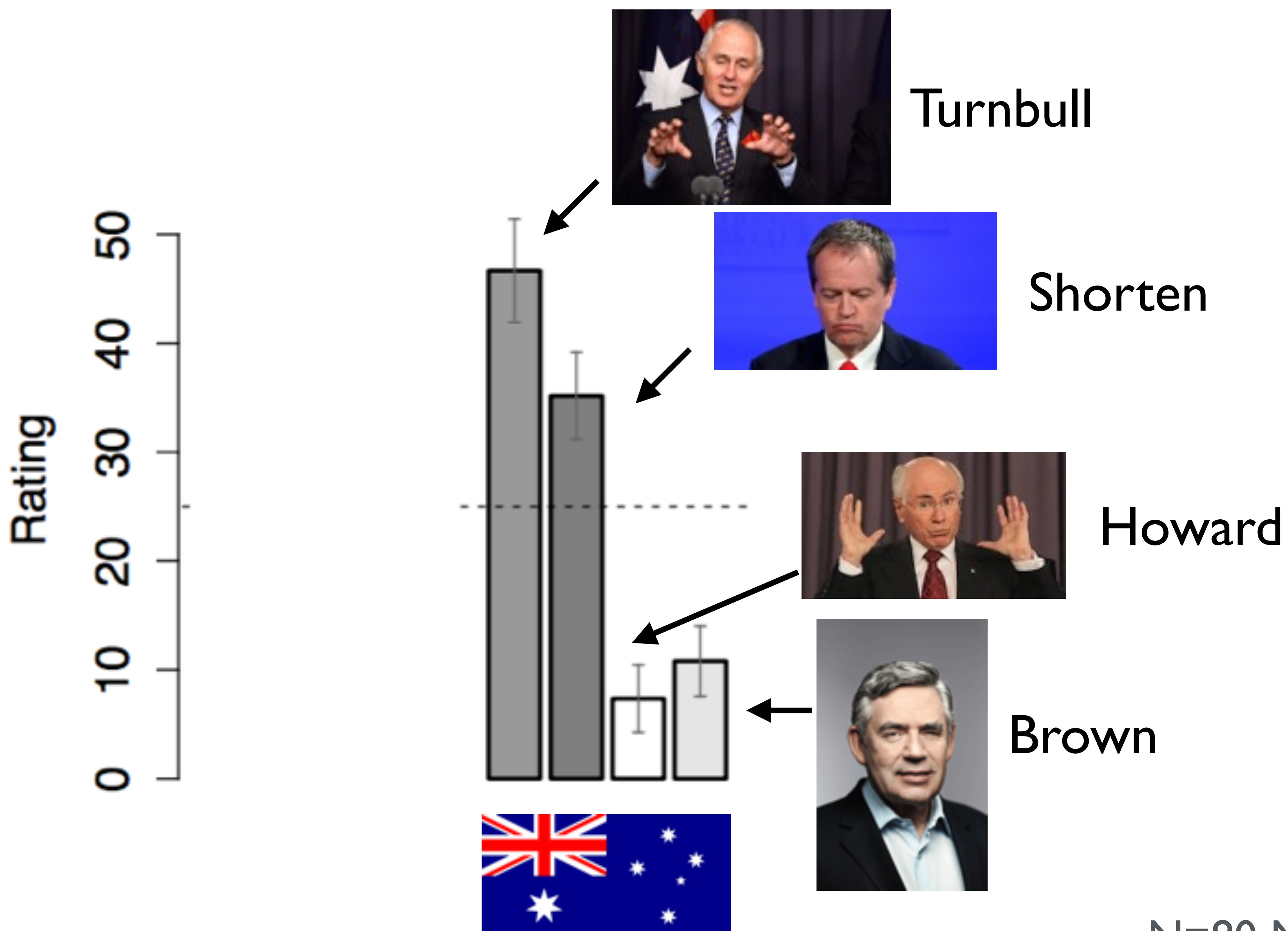






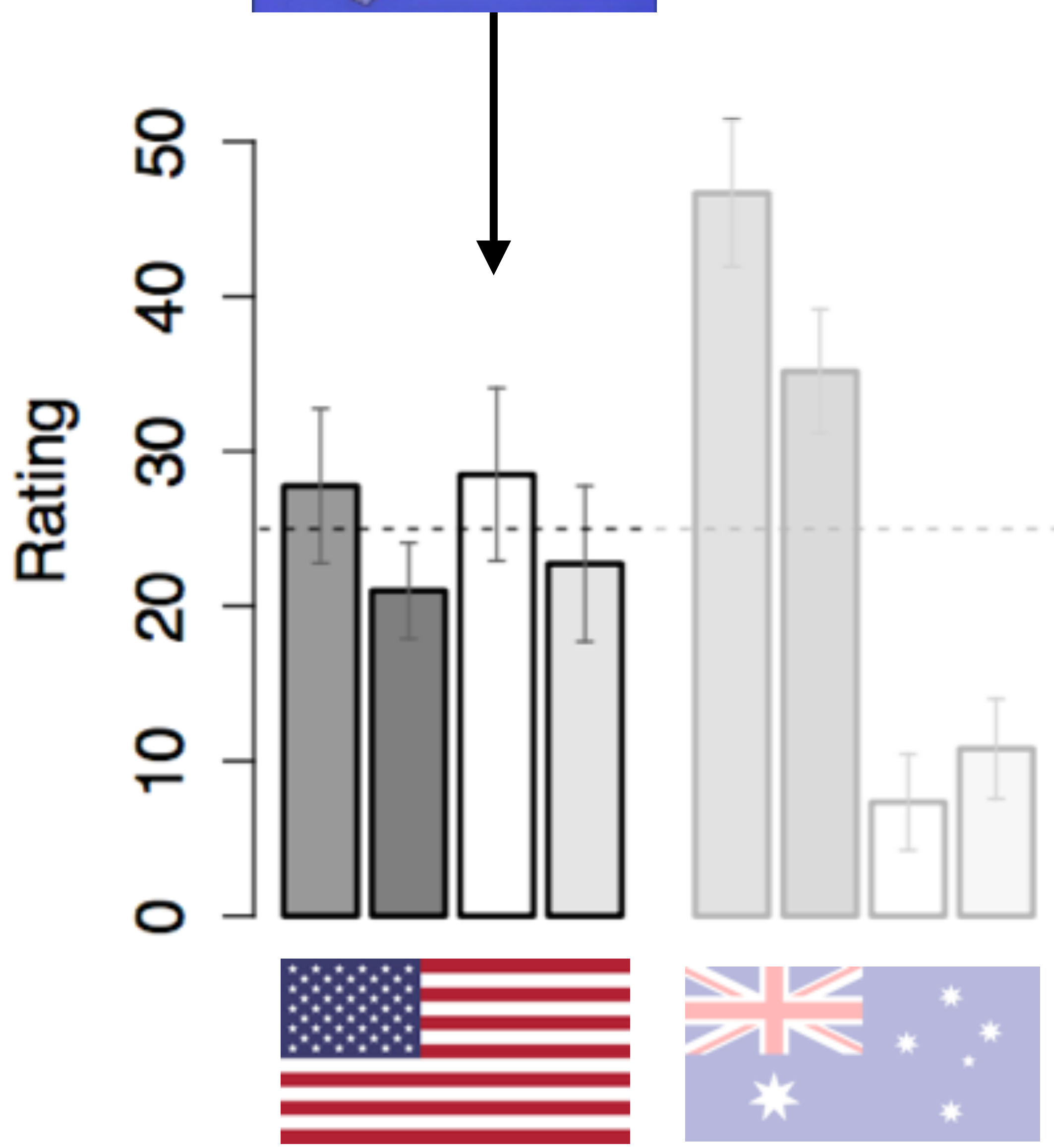
<u>Case study 3</u>:
An empirical illustration

“Who will win the 2016 Australian election?”



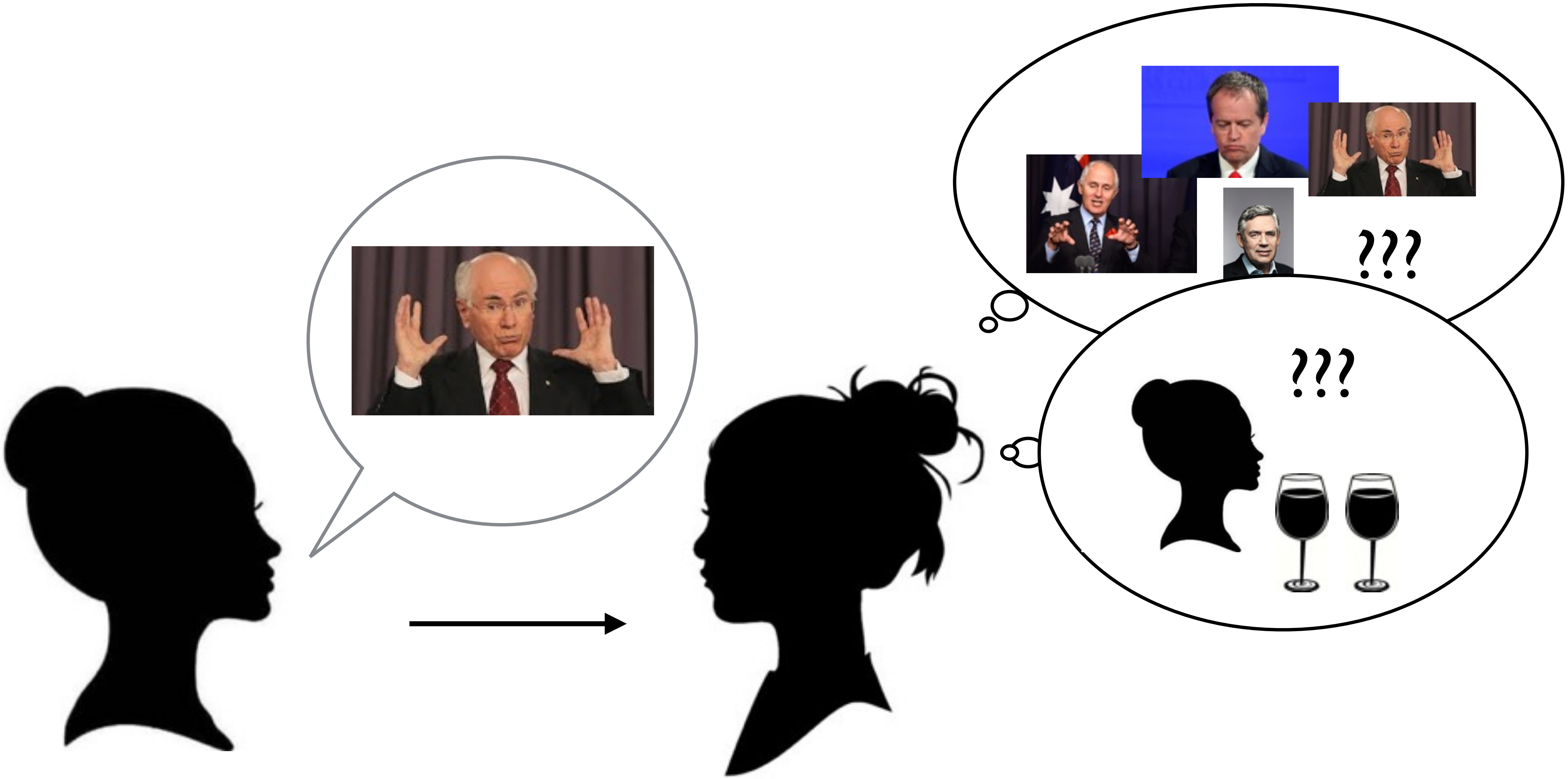
N=80 MTurk workers
and UNSW students

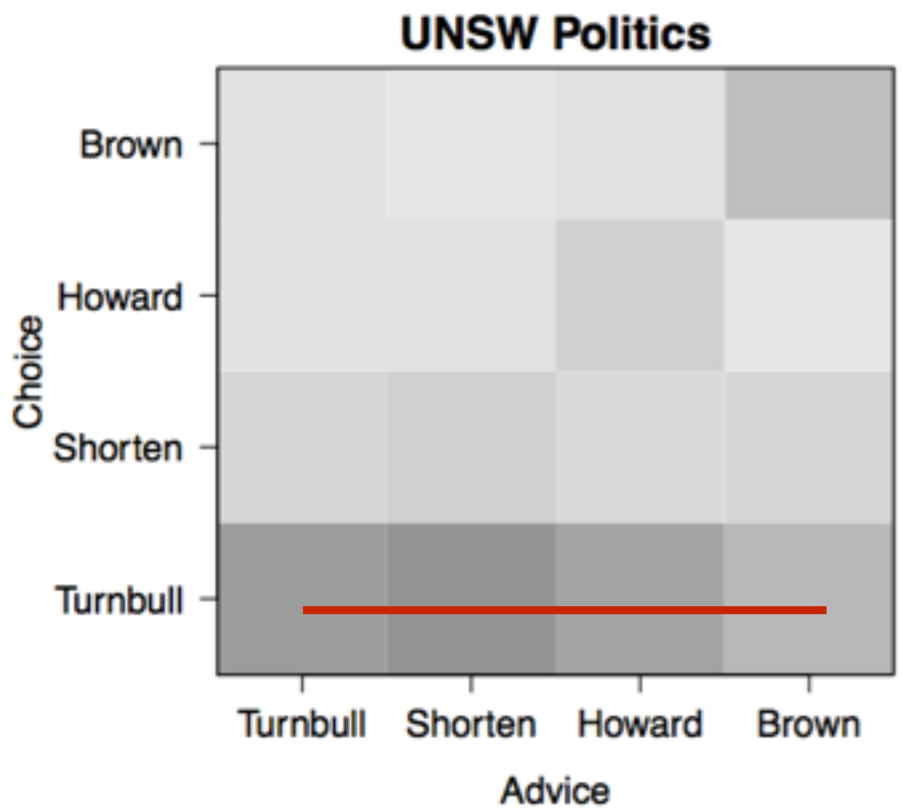
Andy?



N=80 MTurk workers
and UNSW students

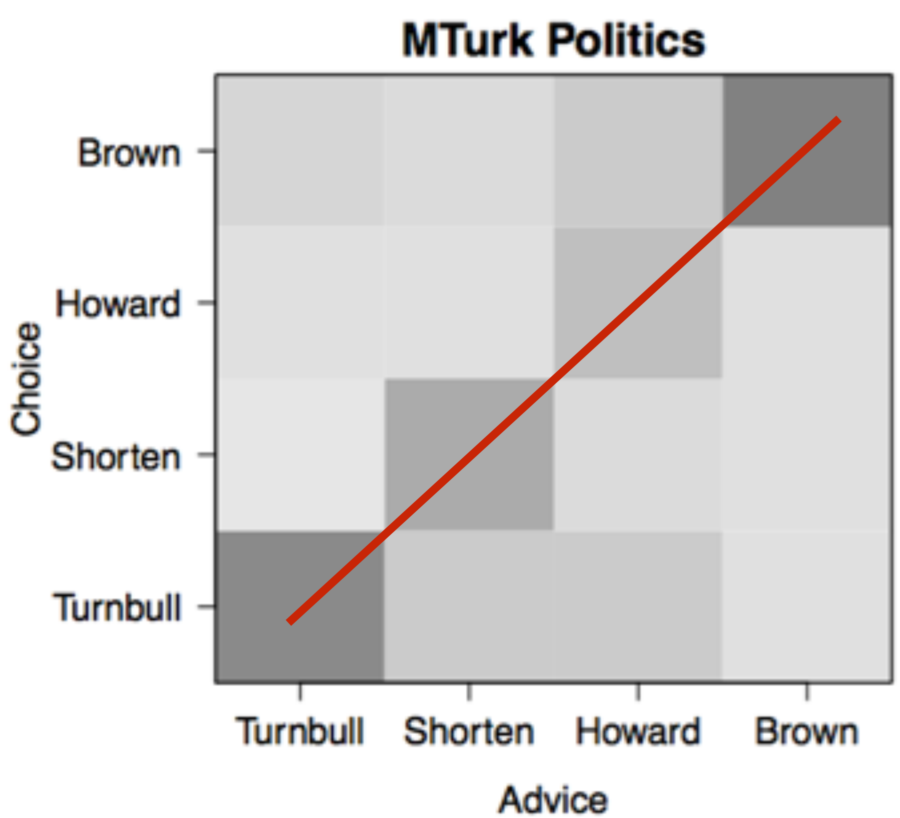
The advisor task



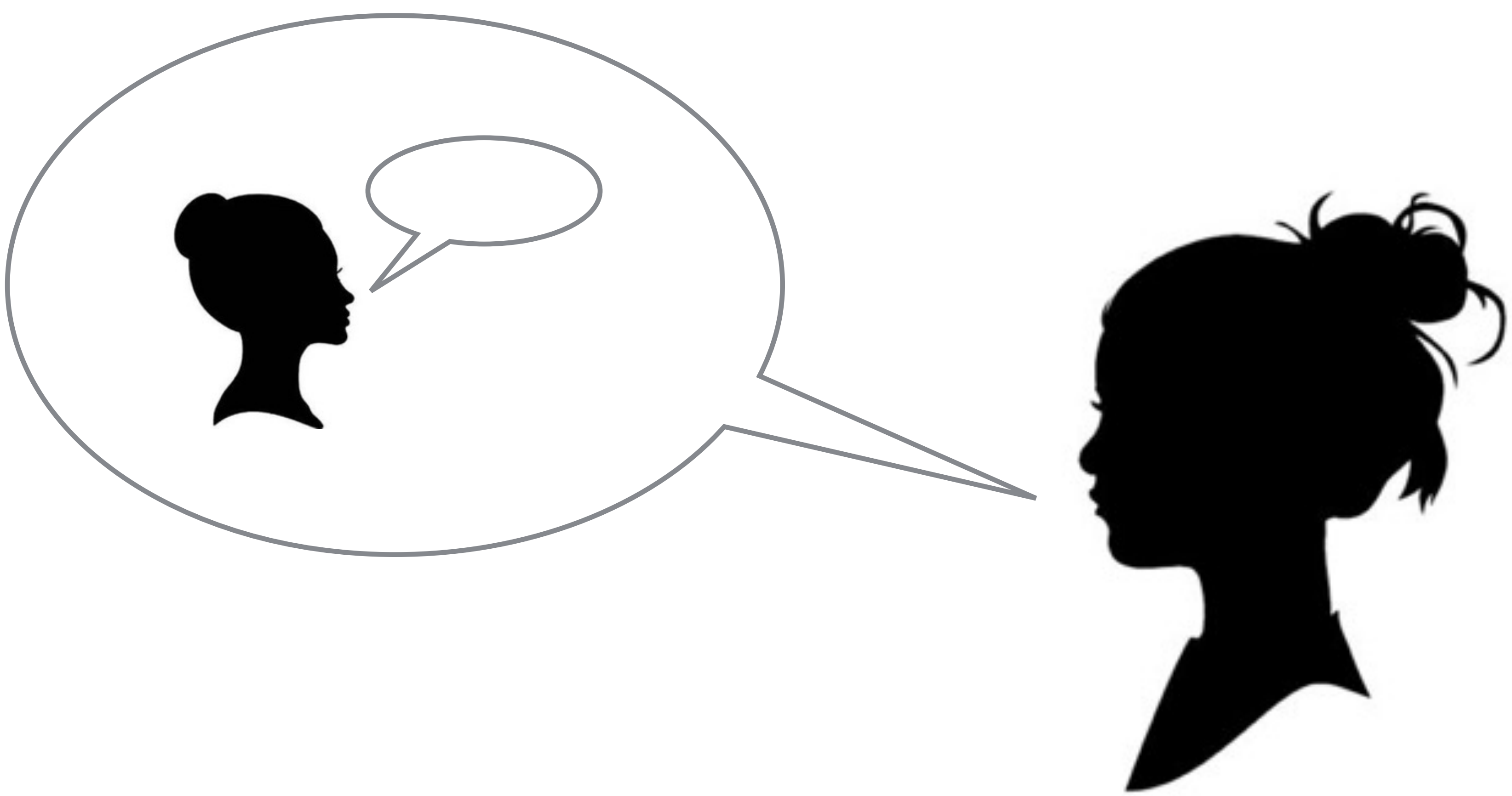


Australians ignored the advisor and predicted a Turnbull victory

N=124 UNSW students

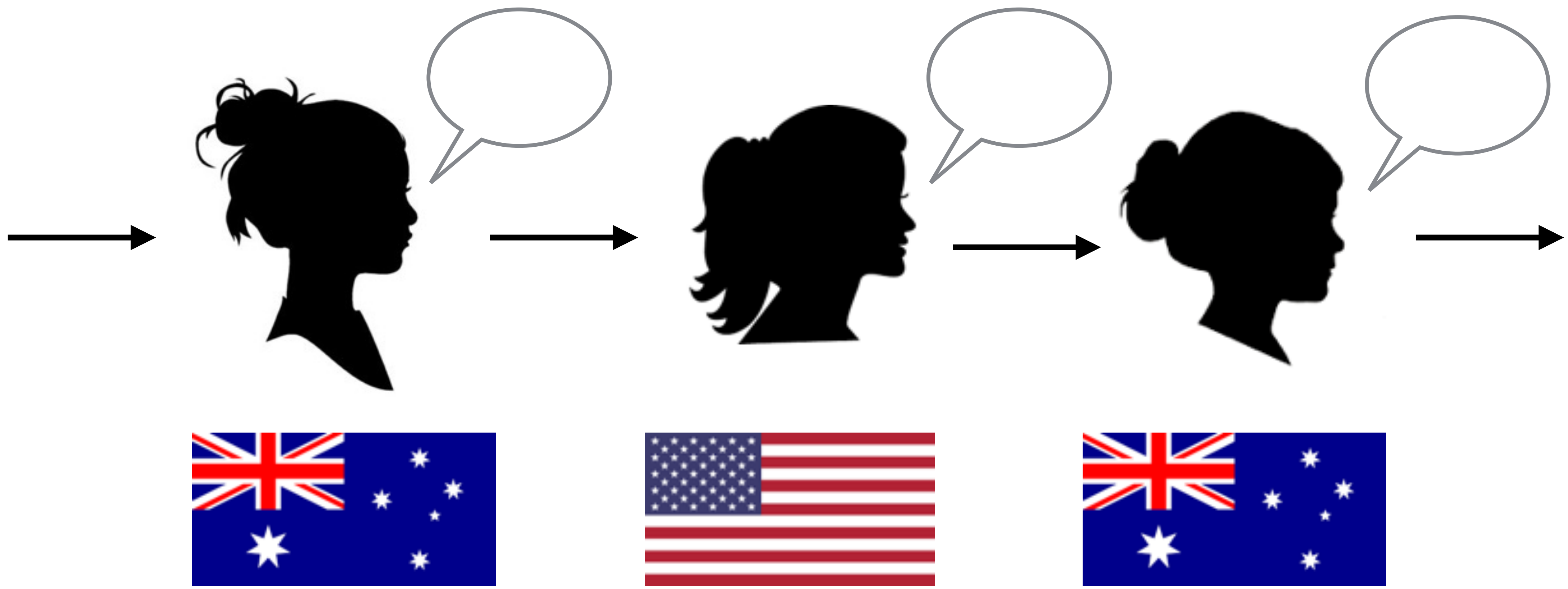


Americans followed the advisor regardless

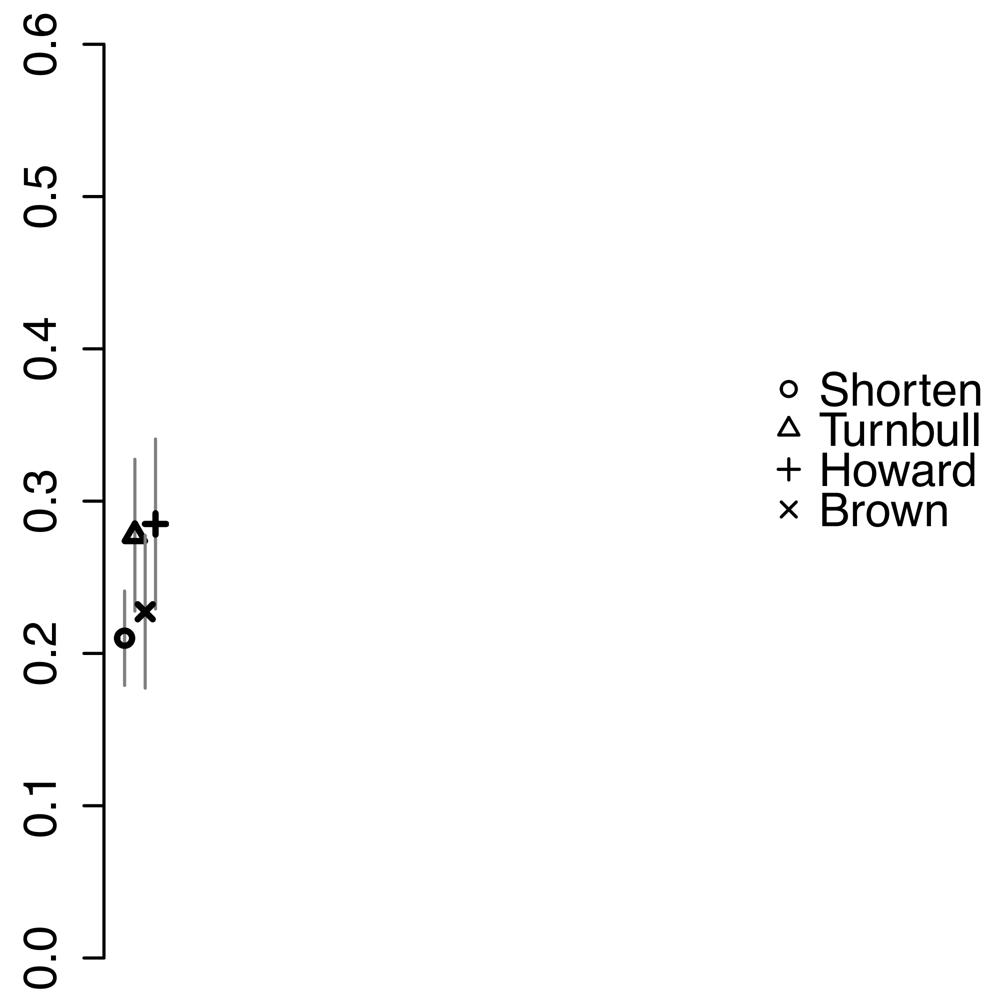


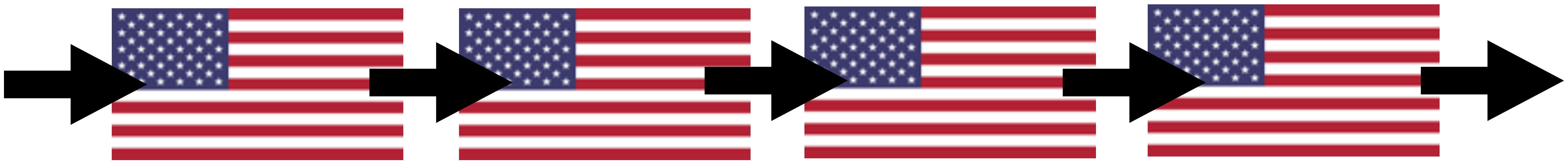
N=196 MTurk workers

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

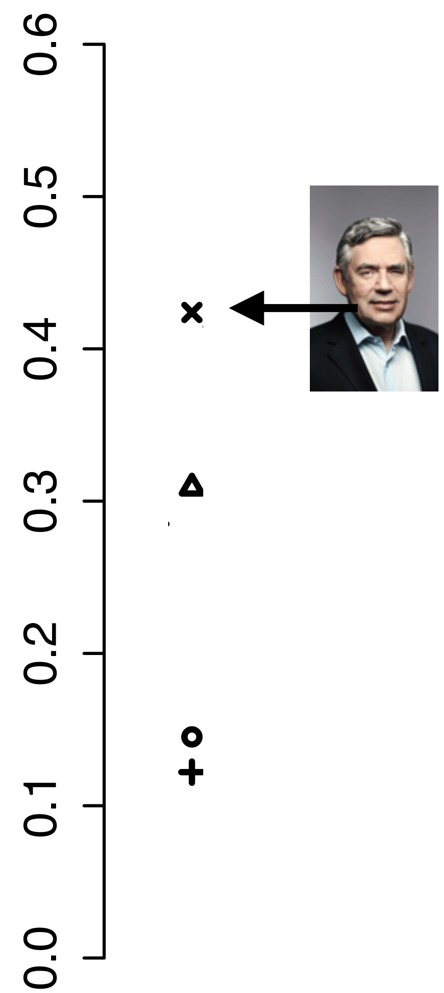


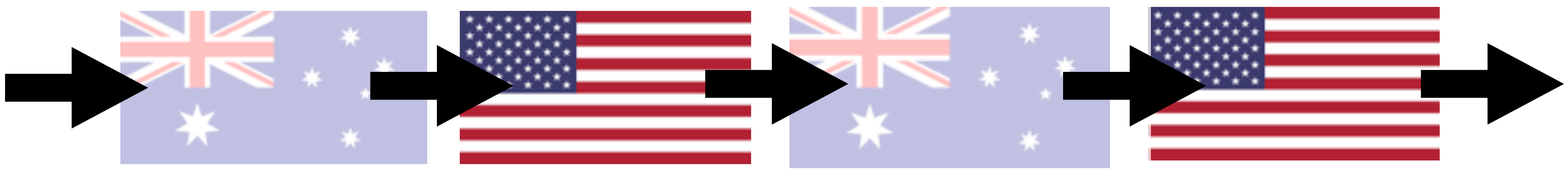
Americans *claim* to be totally ignorant about Australian politics…



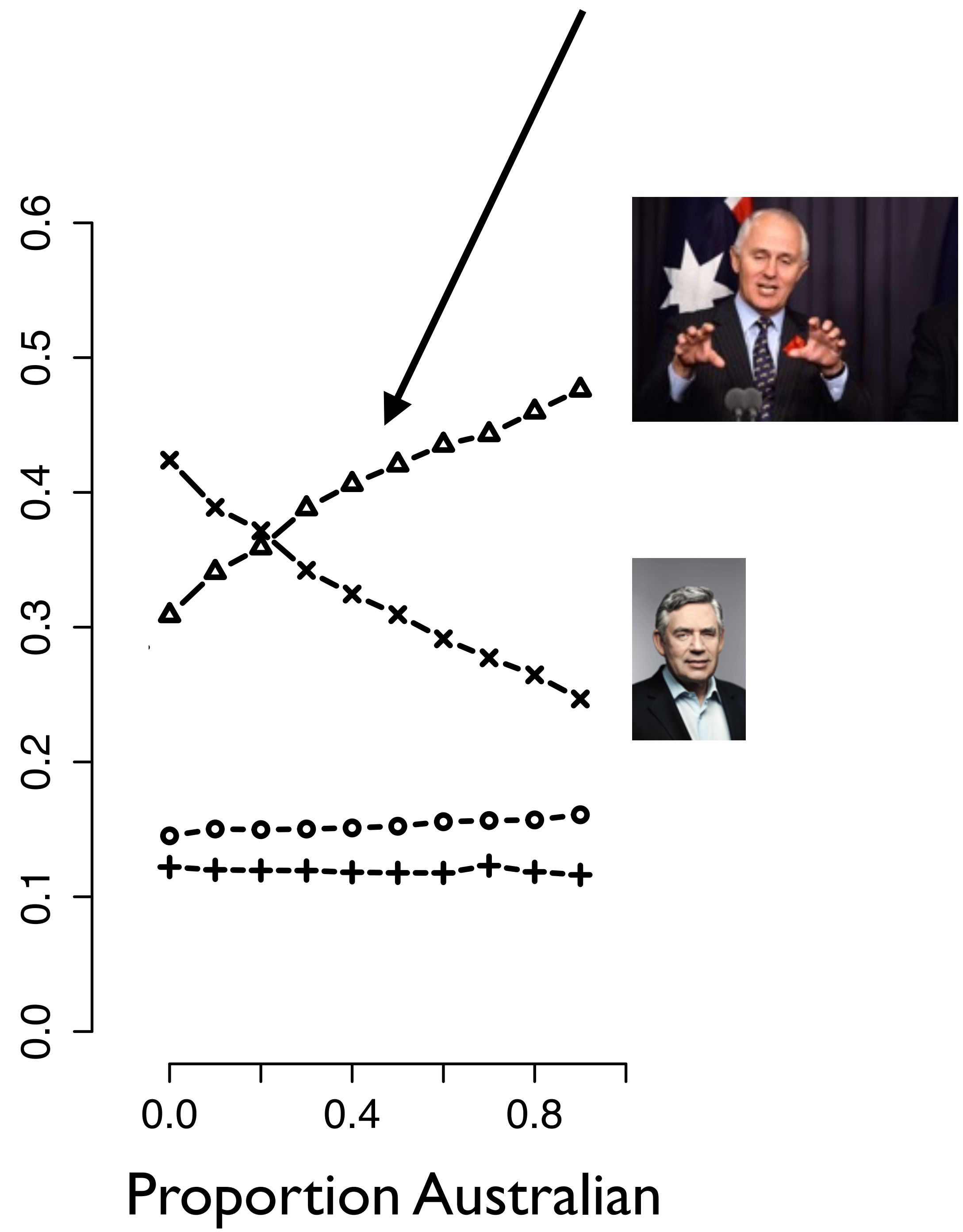


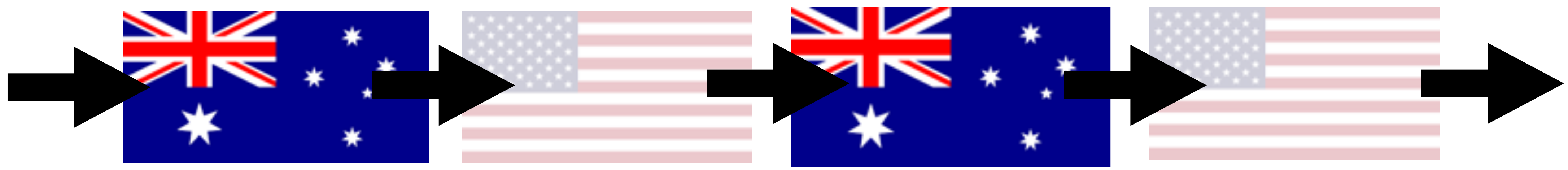
… and an all American iterated learning chain "reveals" a "preference" for *<u>Gordon Brown</u>* …



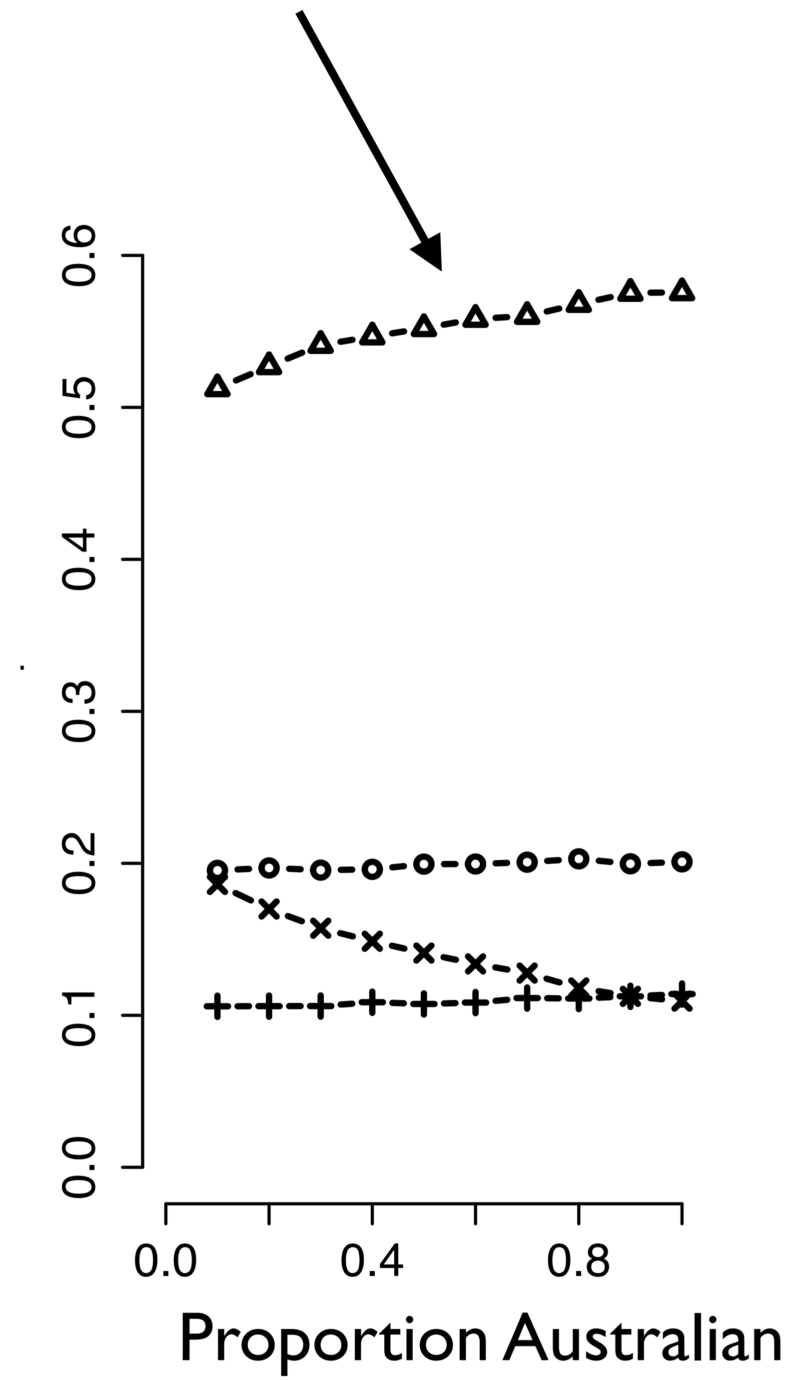


If we mix some Australians into the chain the Americans endorse *Malcolm Trunbull*





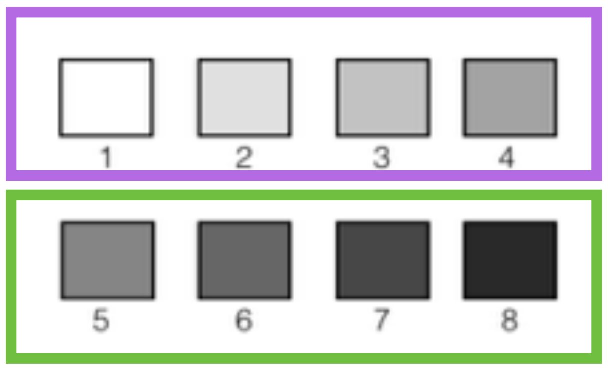
Australians choose Turnbull no matter how many Americans are included

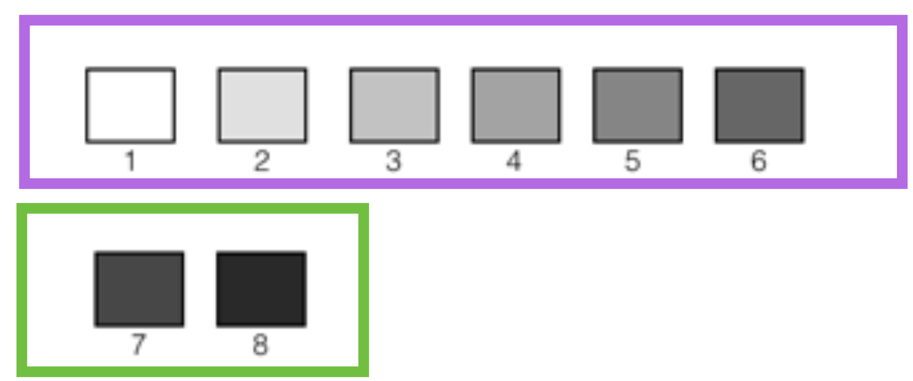


Case study 4:
A non-Bayesian example

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

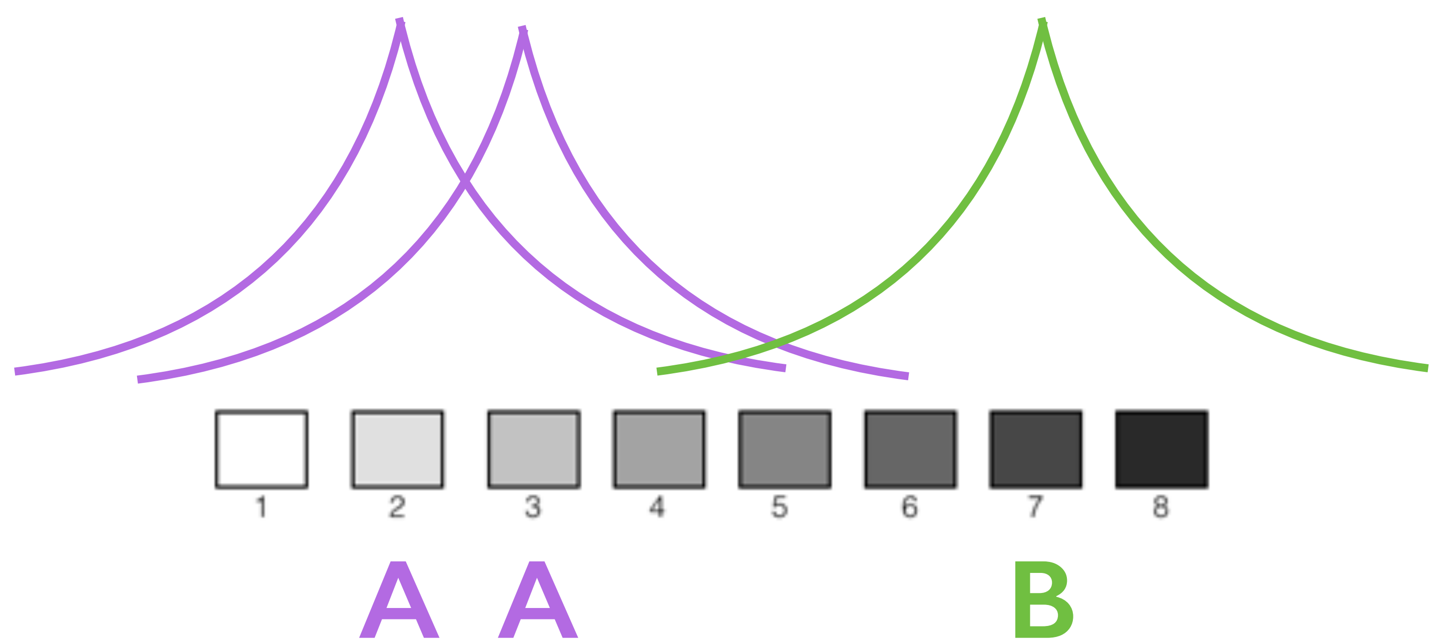
(e.g., Austerweil 2014)





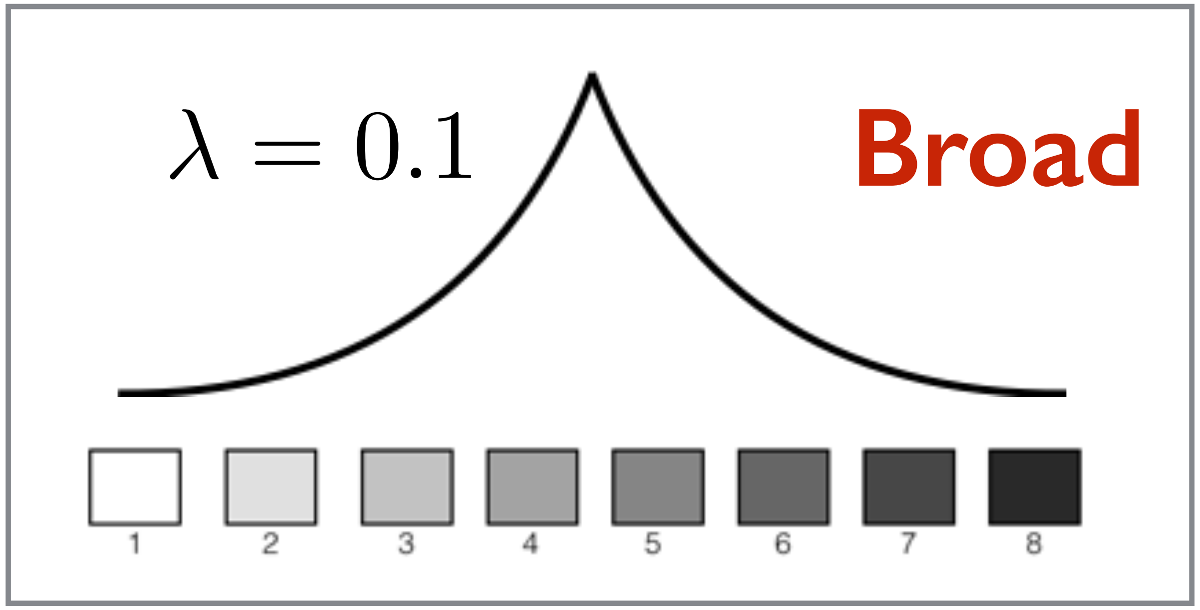
Exemplar model of categorisation

(Nosofsky 1986; Pothos & Bailey 2009)

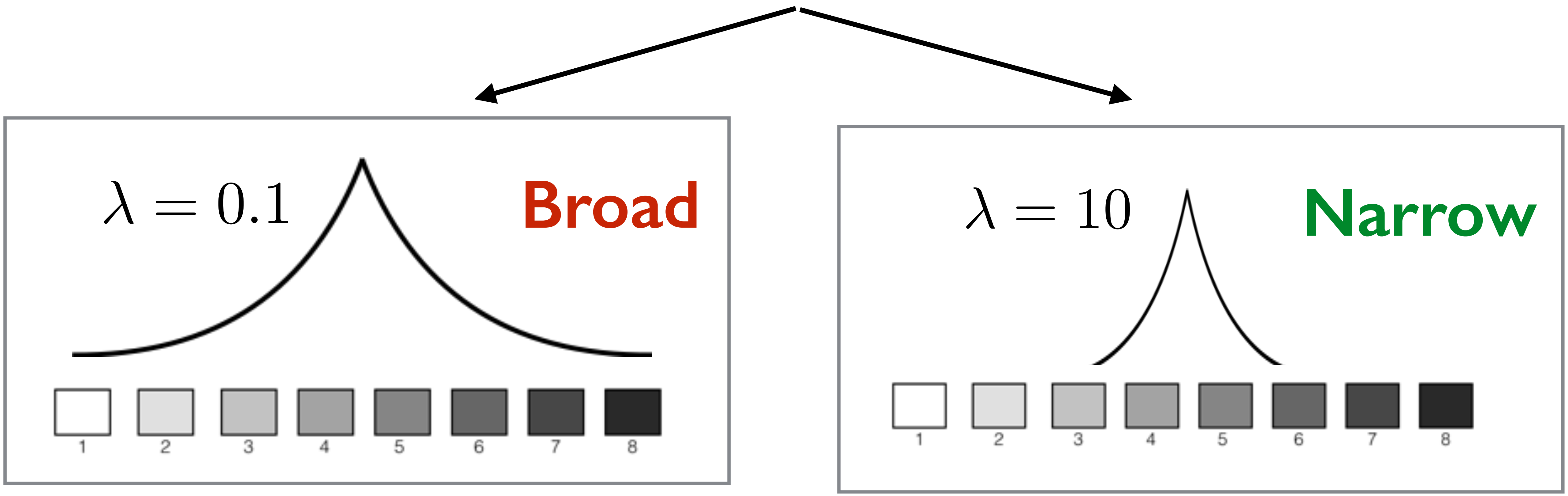


GCM: categorisation probability is proportional to sum similarity

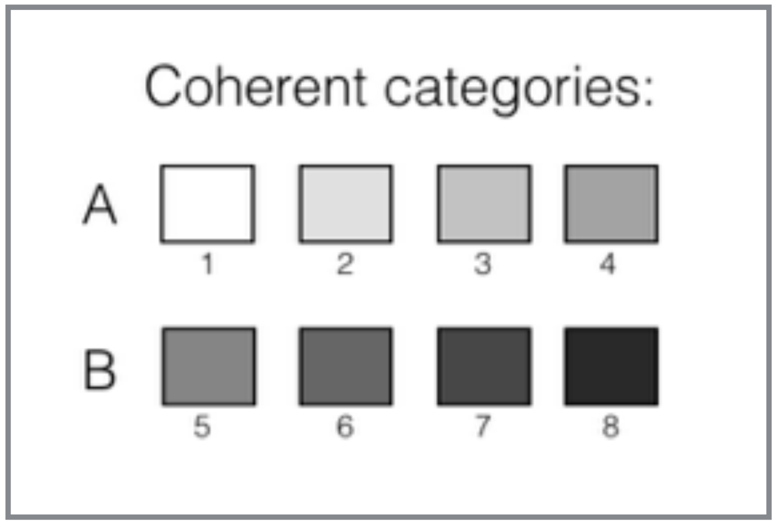
GCM allows learners to vary in how broadly they generalise from a stimulus



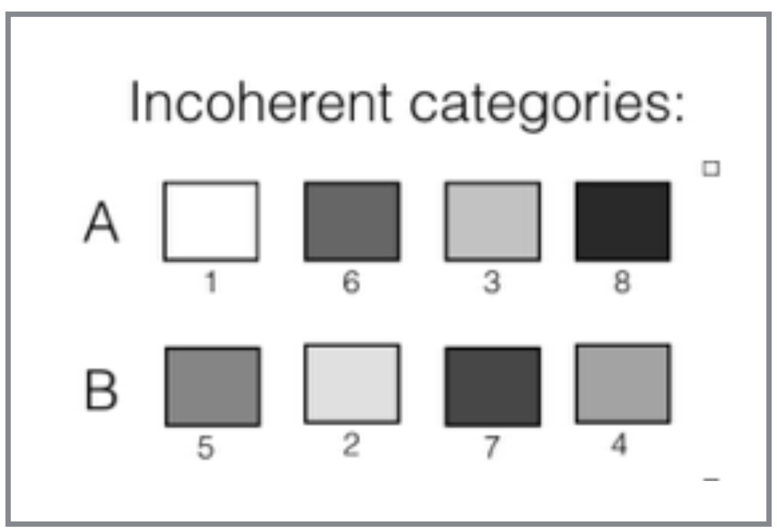
GCM allows learners to vary in how broadly they generalise from a stimulus



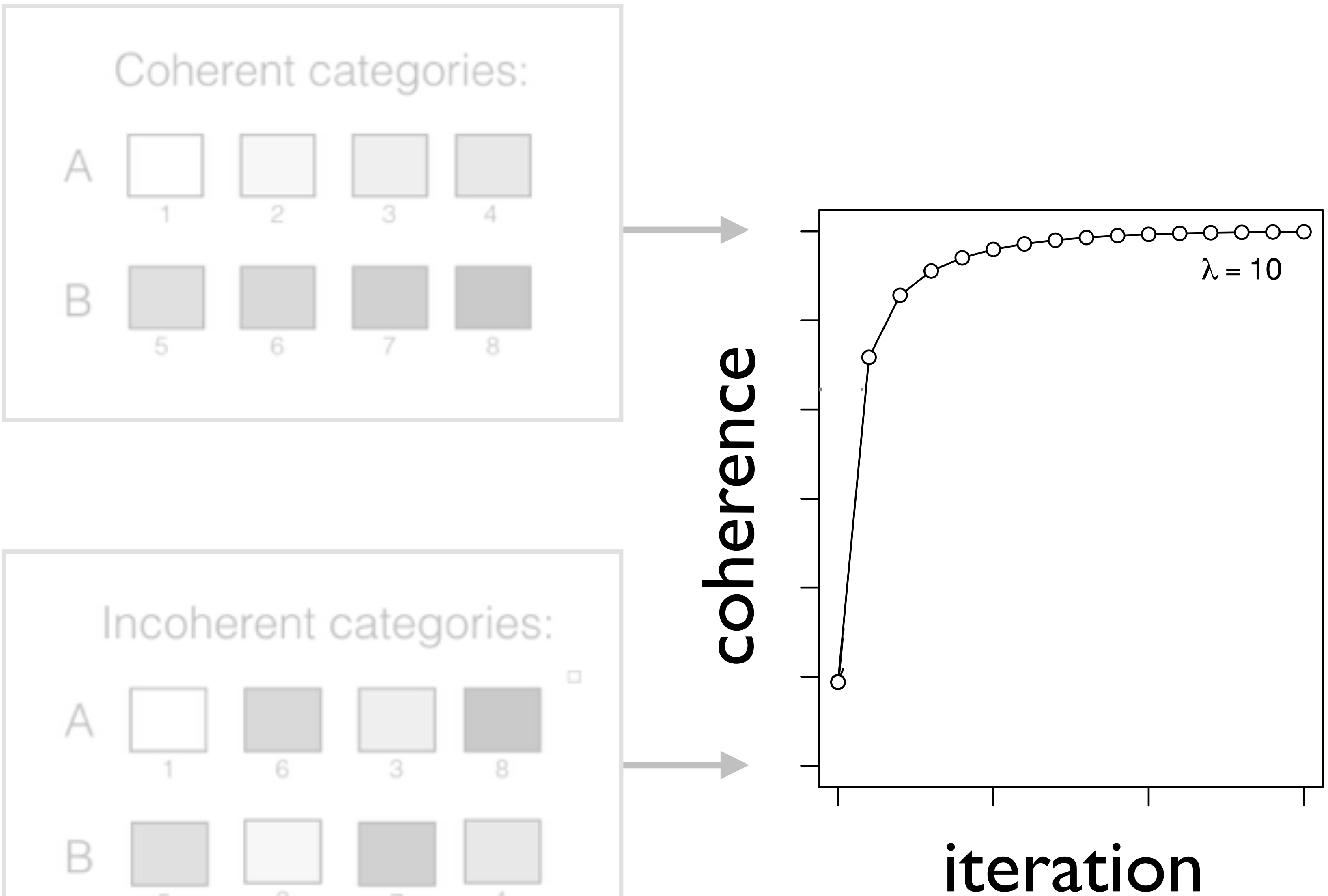
Categorisation bias #1



<u>Coherent</u> systems
assign similar items
to the same category

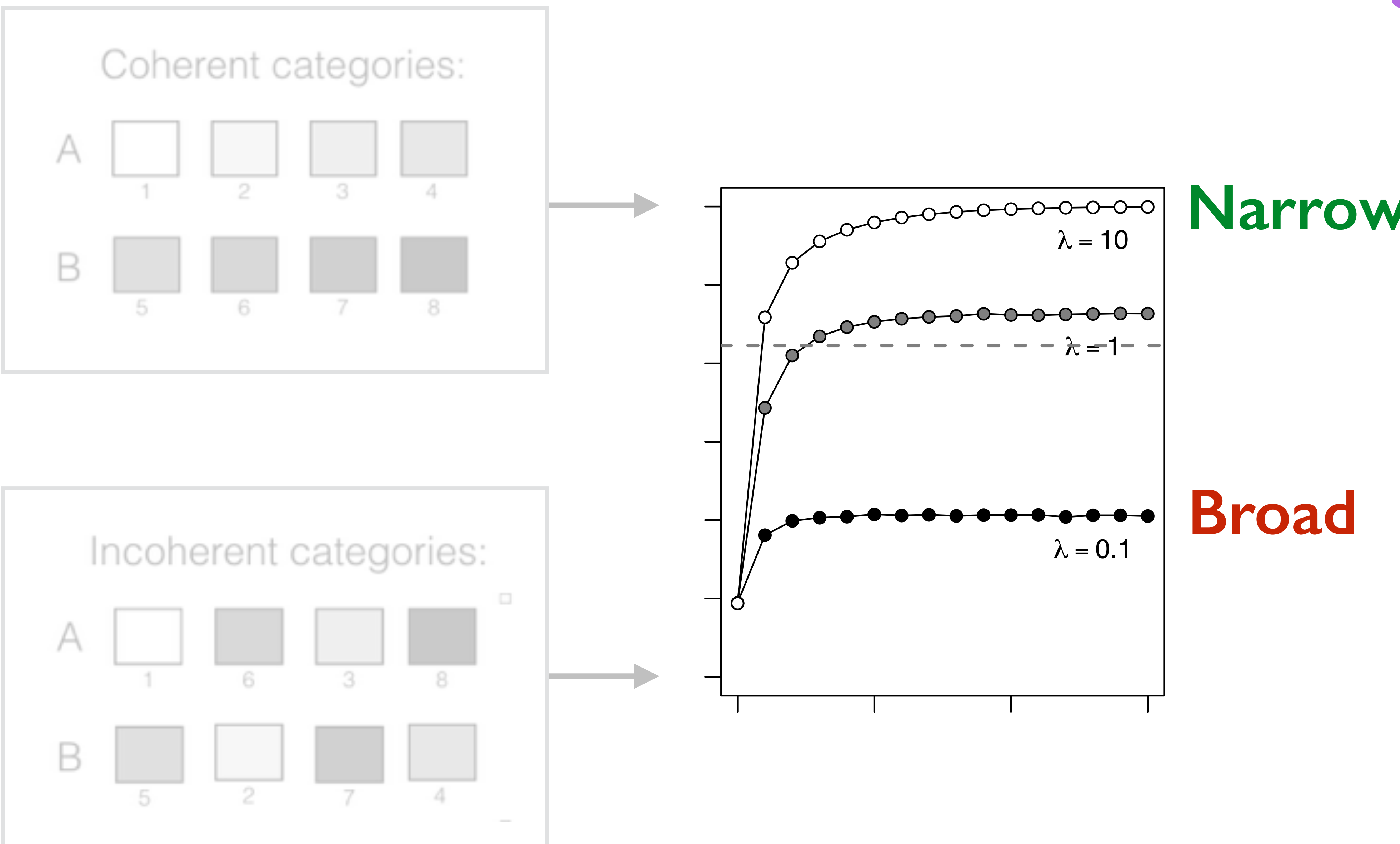


Iterated learning with GCM when learners are homogenous

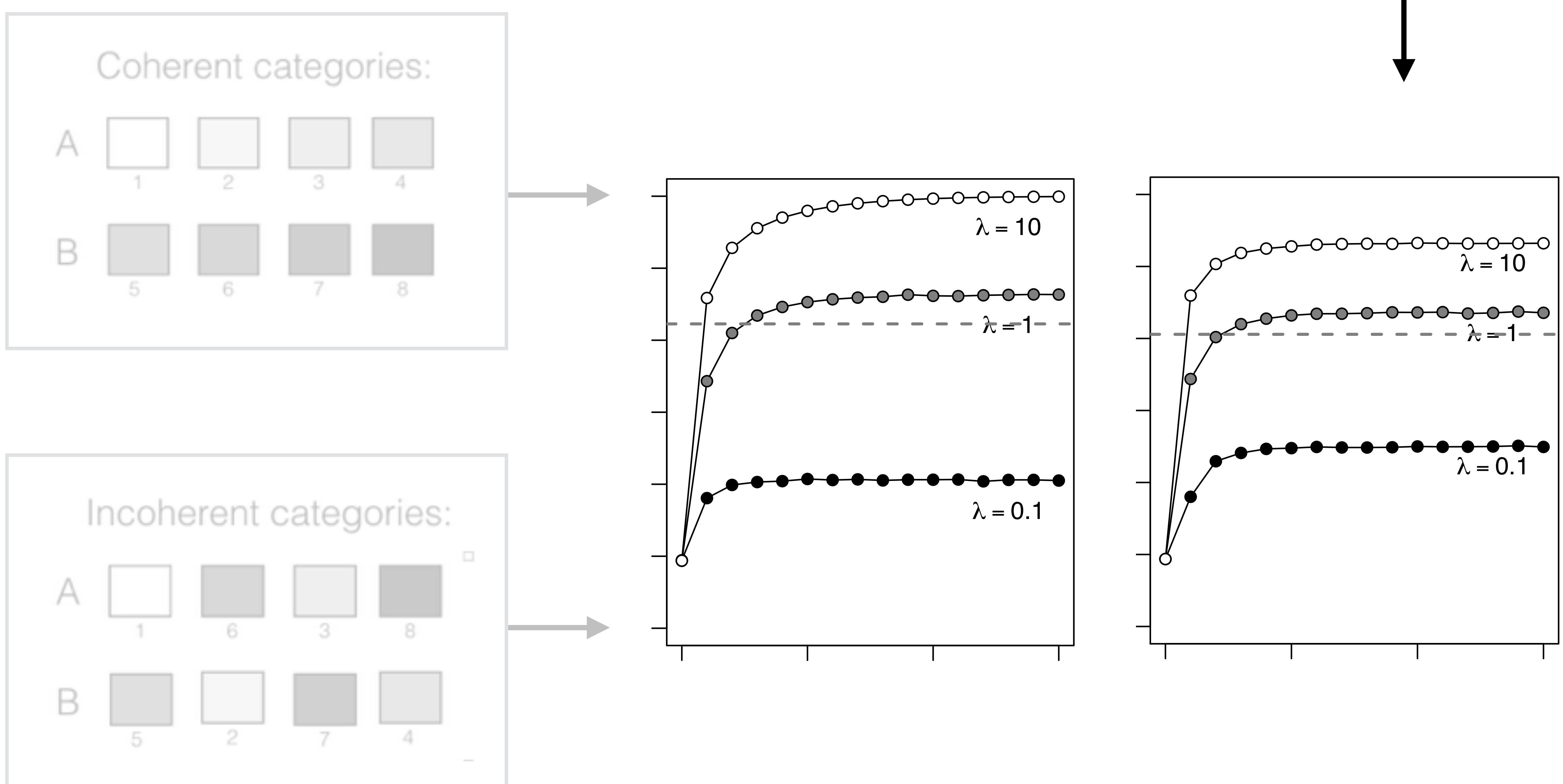


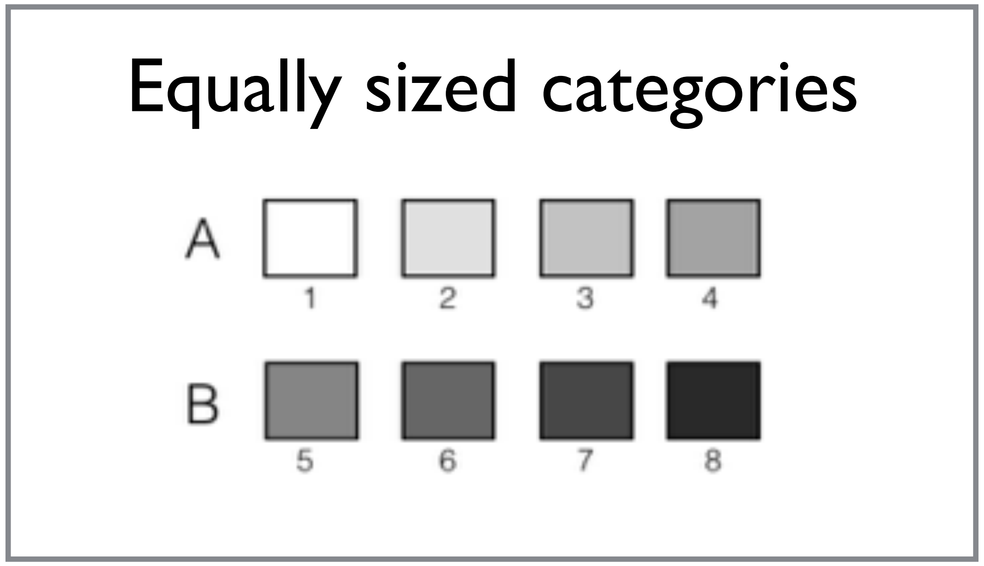
Narrow generalisation implies strong coherence bias

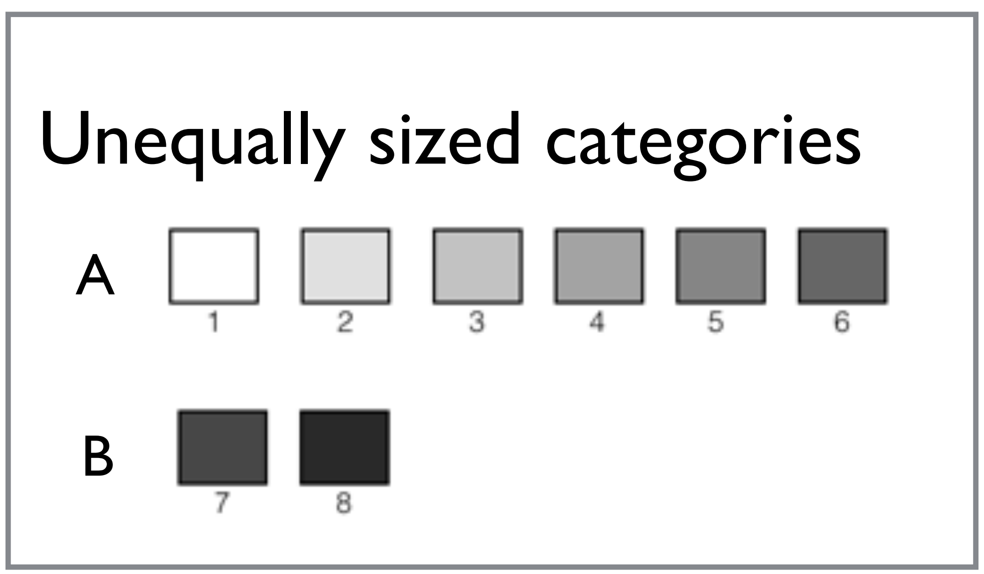
Iterated learning with GCM when learners are homogenous



Heterogeneity isn’t much of a problem here

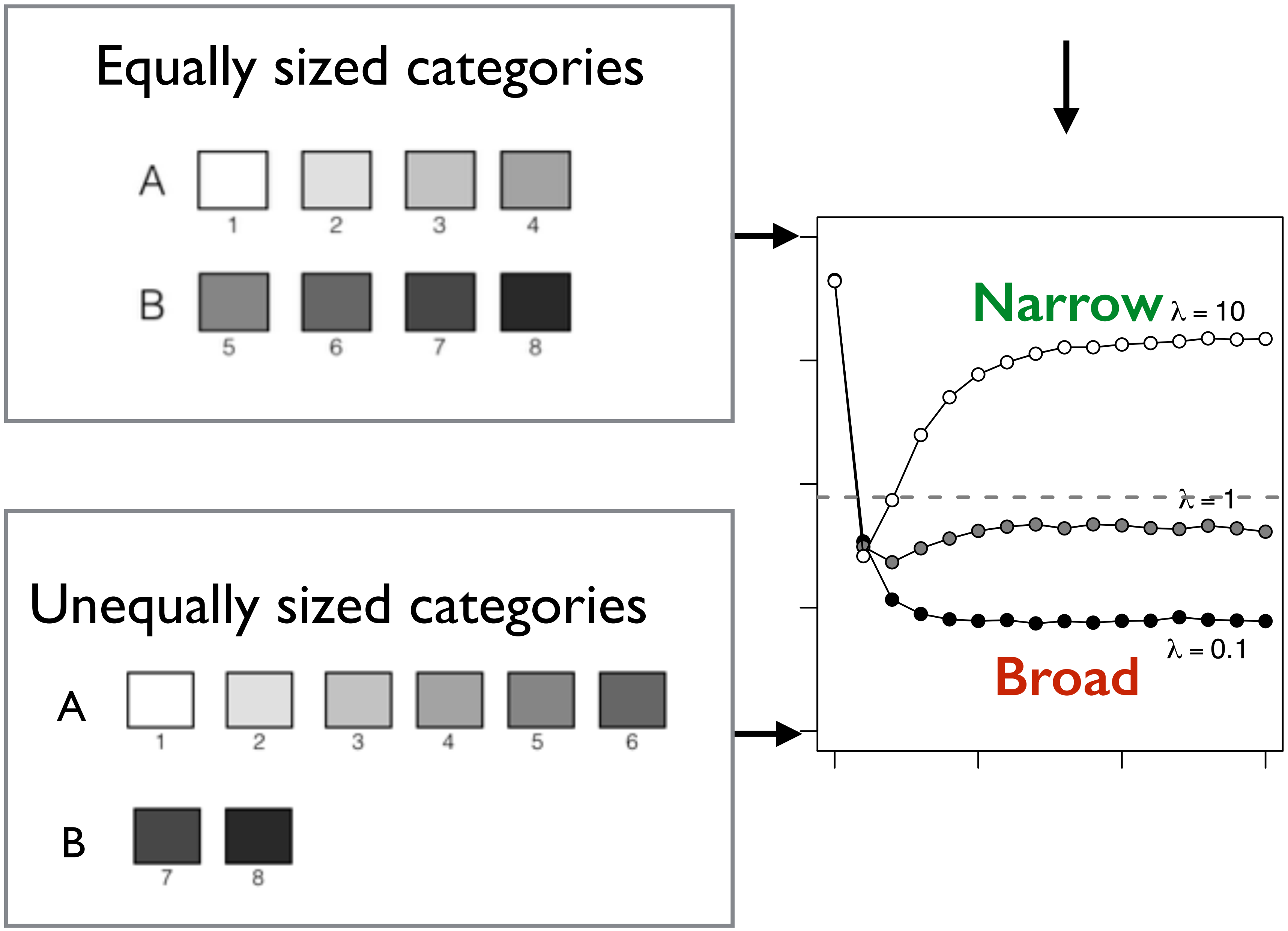




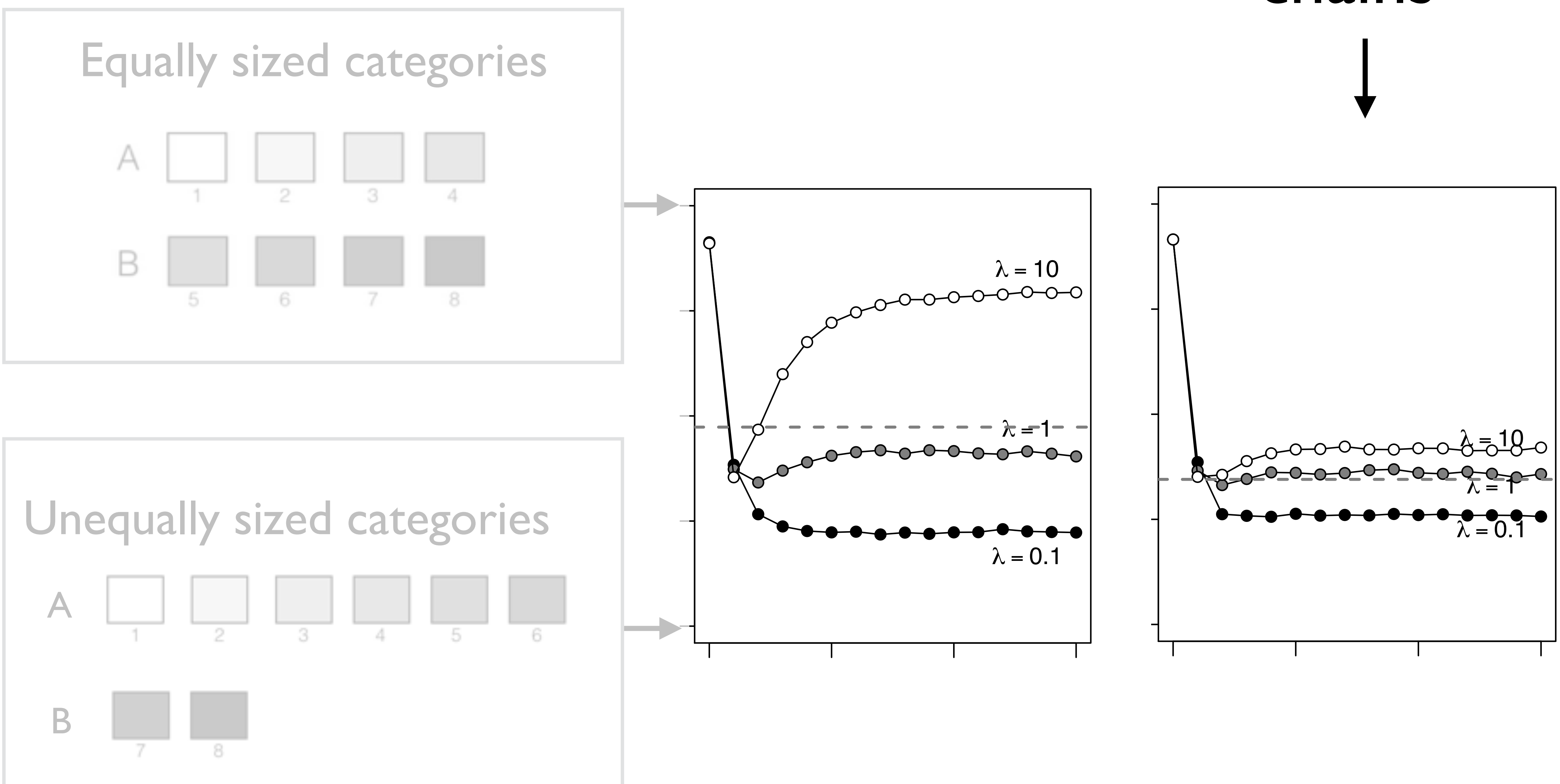


Categorisation bias #2

Homogenous chains

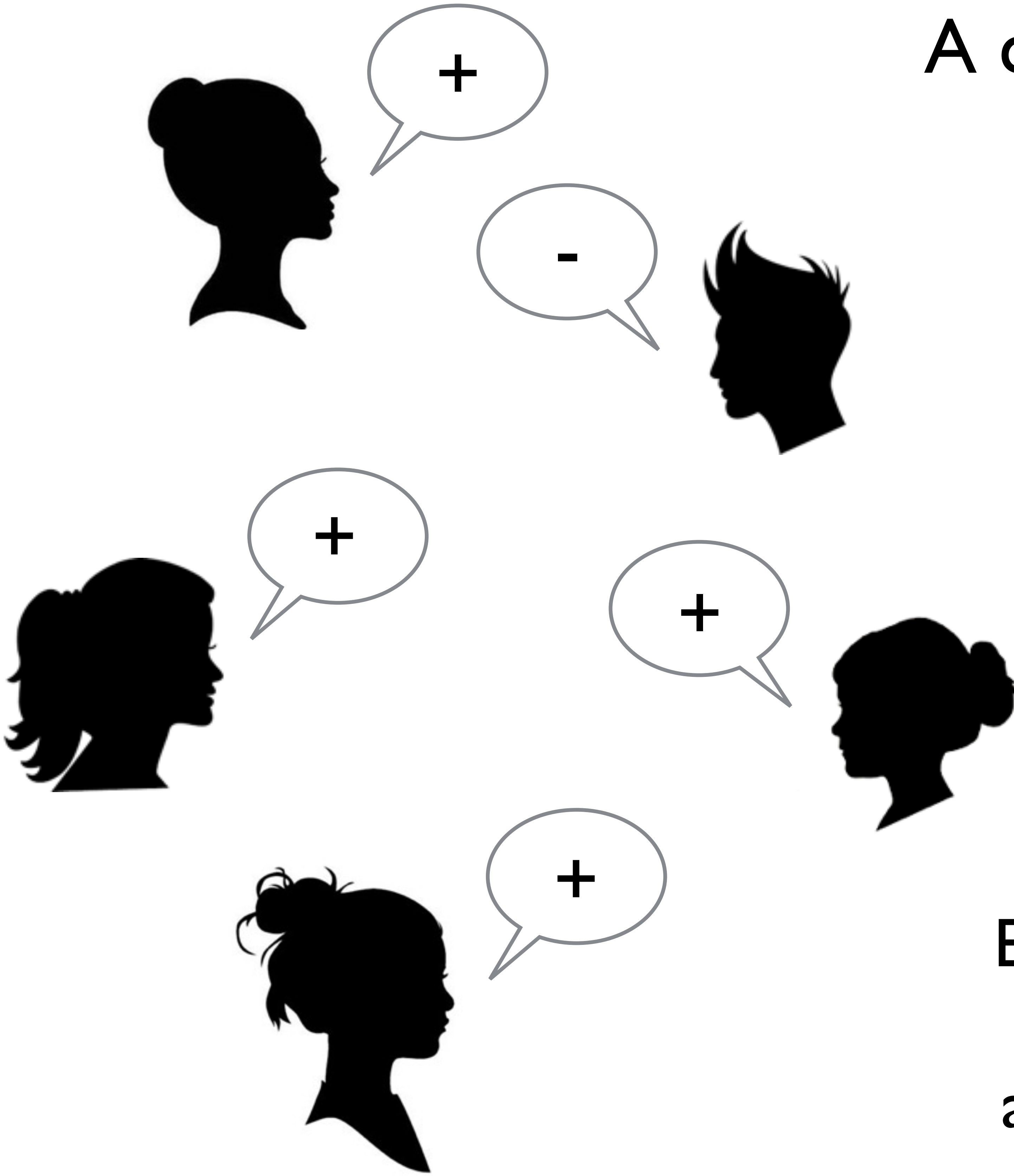


Heterogenous chains

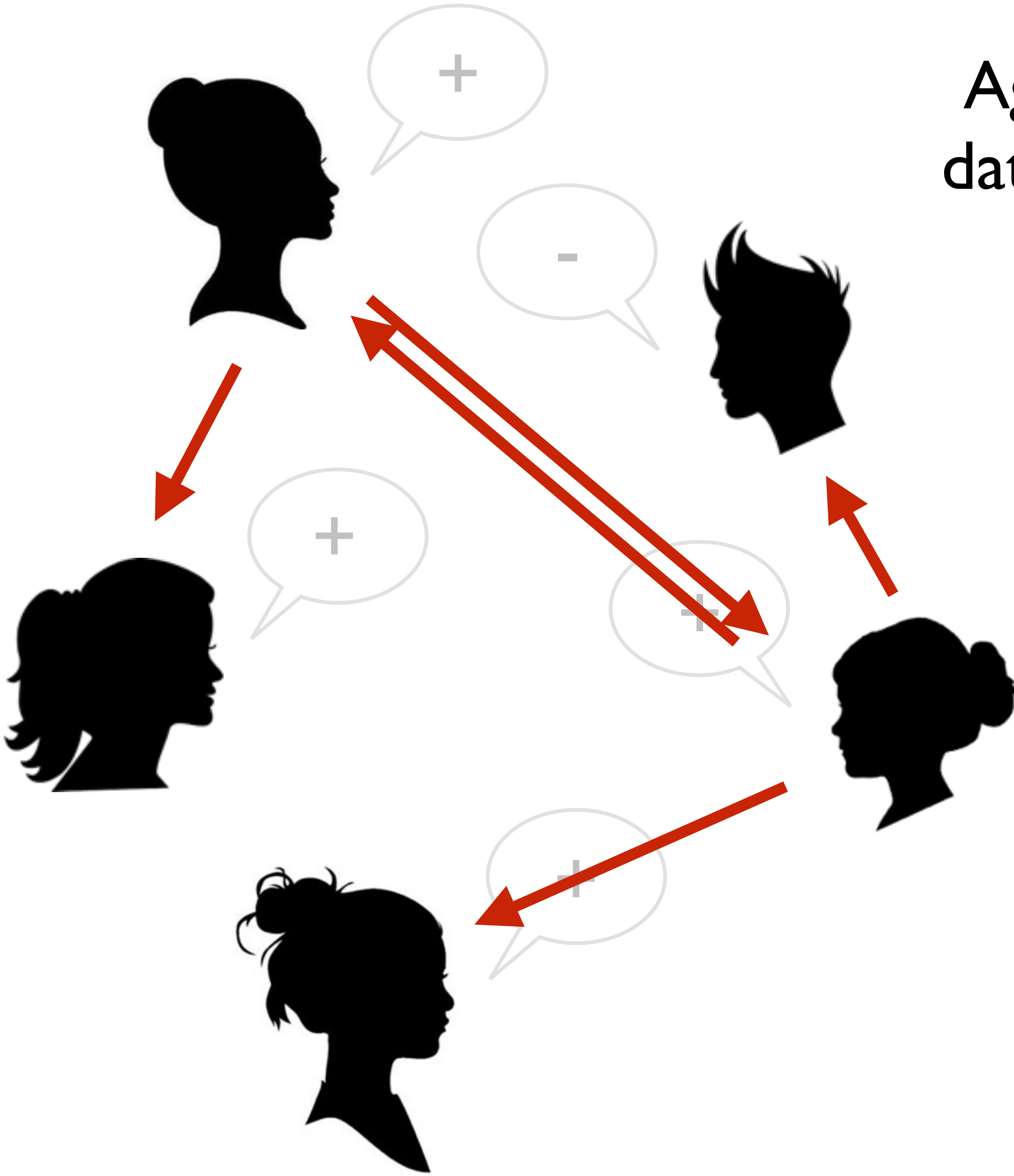


Case study 5:
Belief evolution in a self-organising Bayesian social network

A crowd of Bayesian speakers

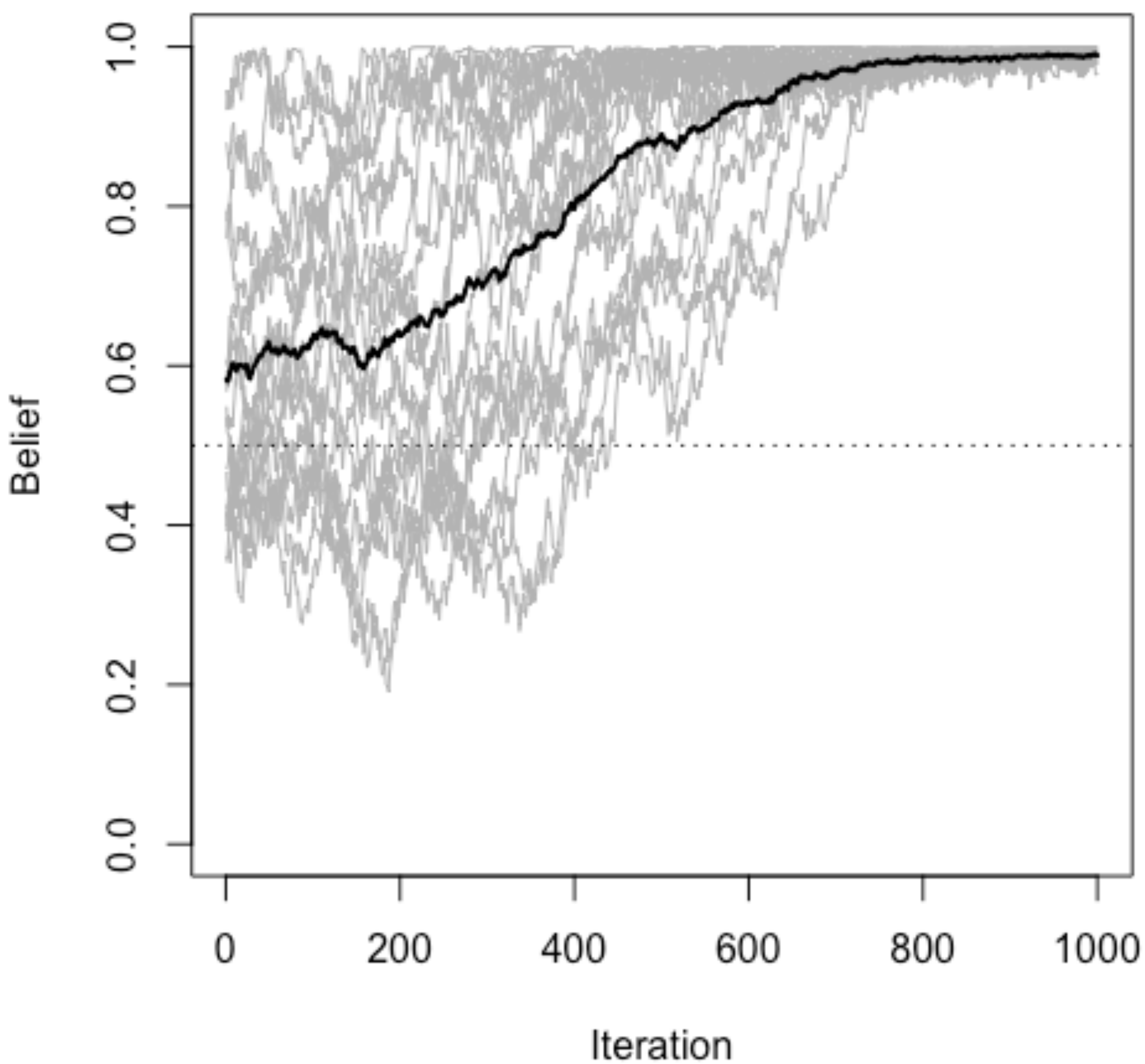


Each agent maintains belief about the rate of + and about the trustworthiness of other agents



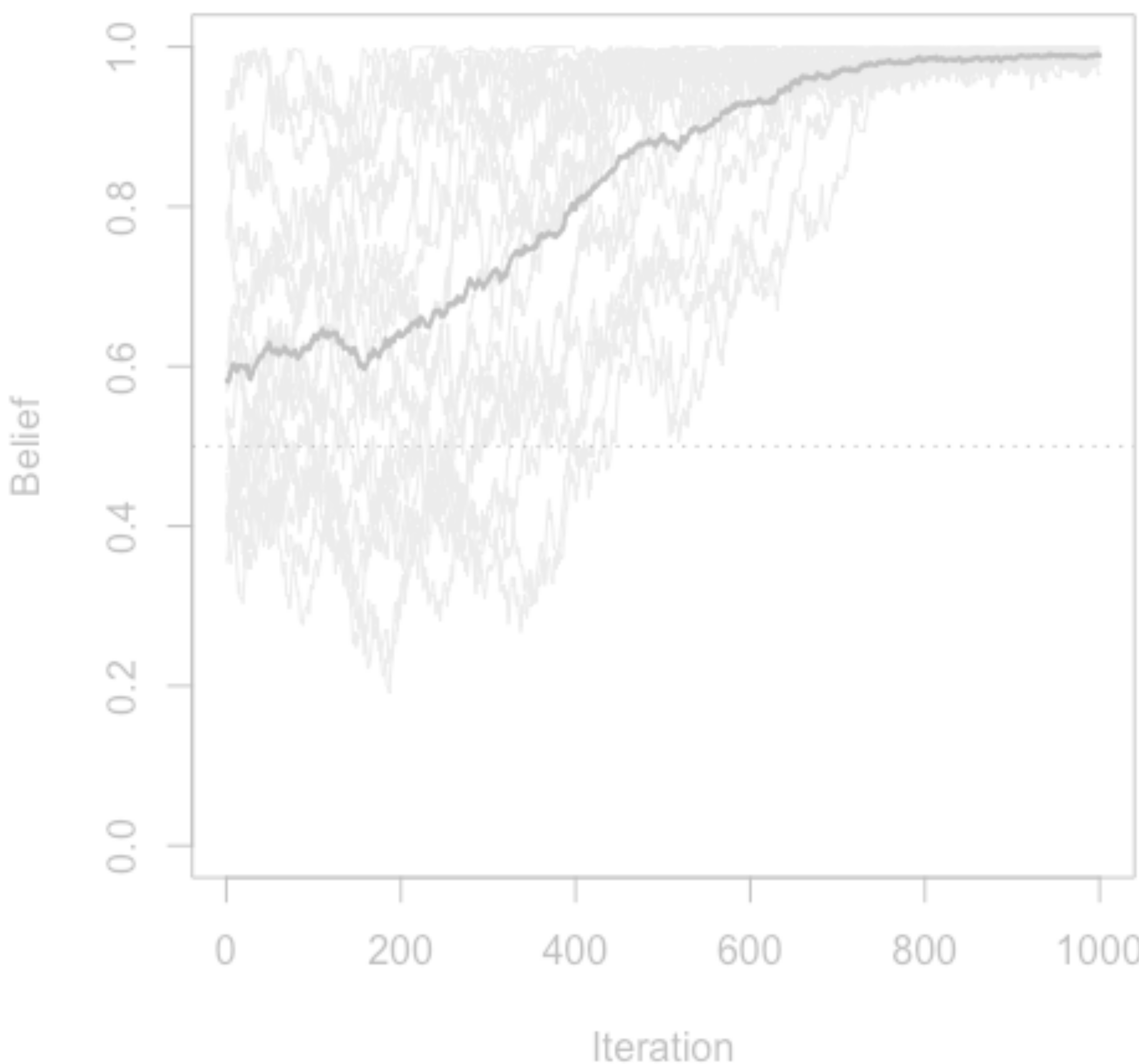
Agents prefer to receive data from trusted sources

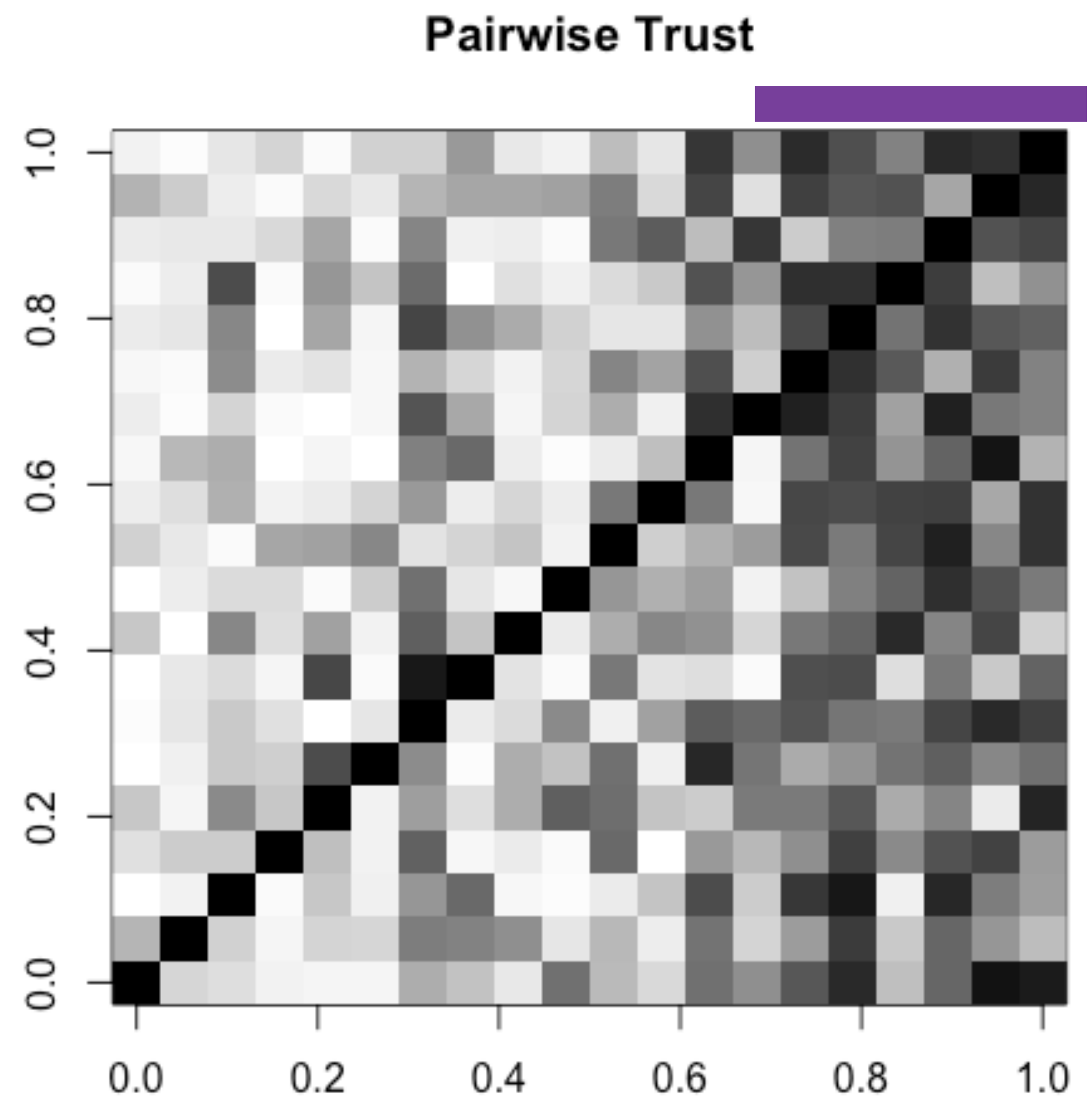
What could possibly go wrong?

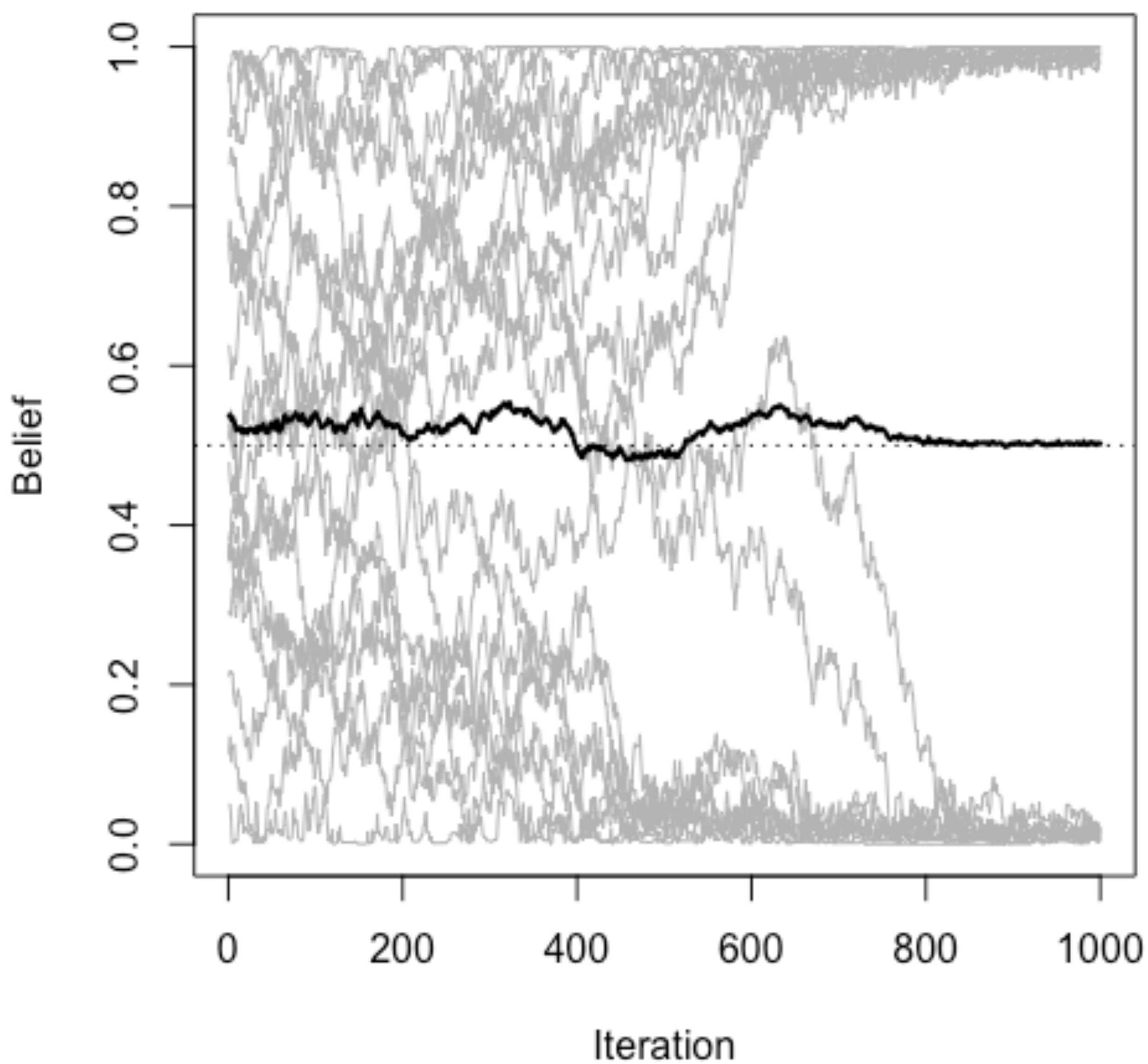


They might ratchet themselves into extremism?

…with the biggest **extremists** being the most trusted agents

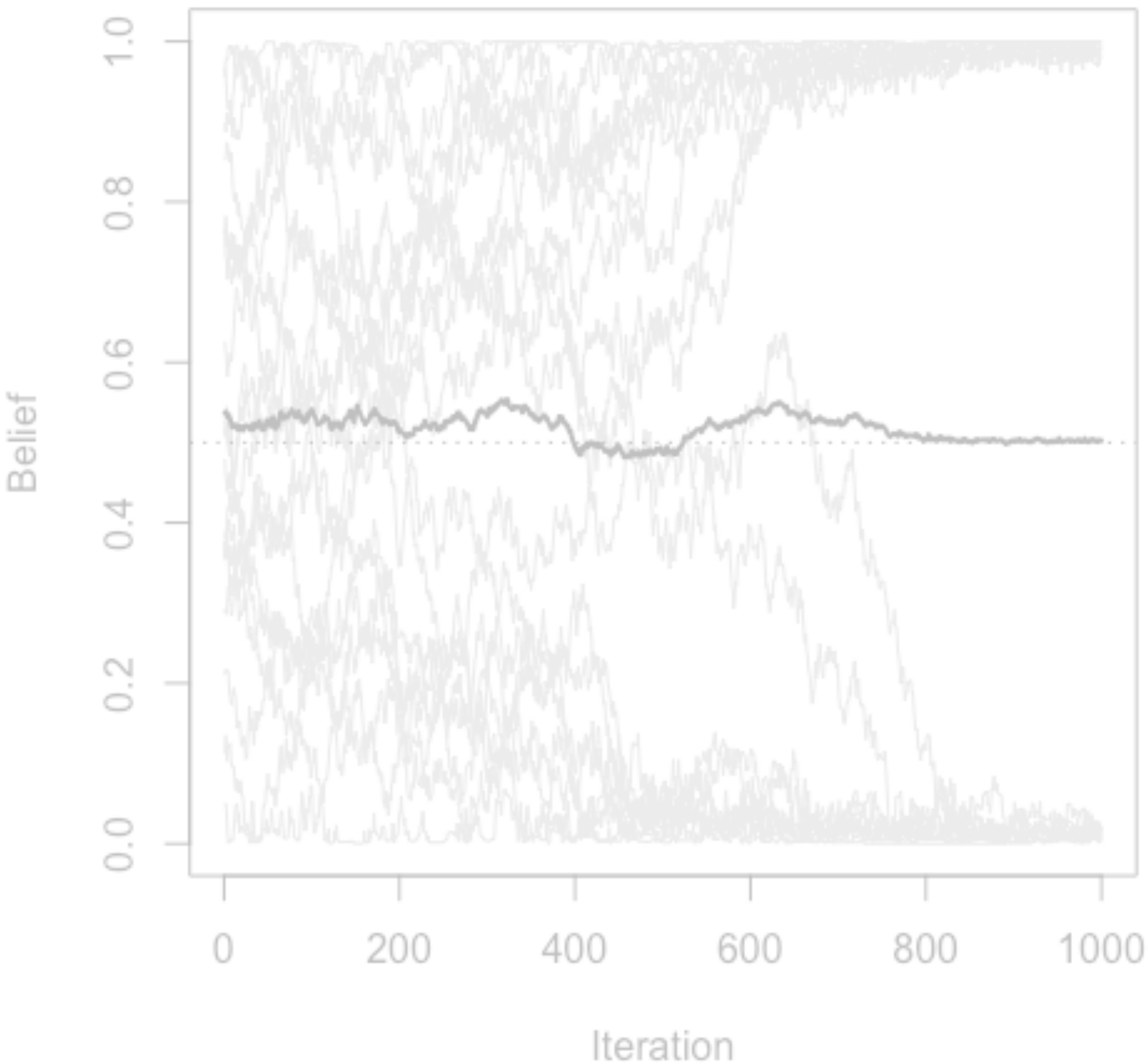


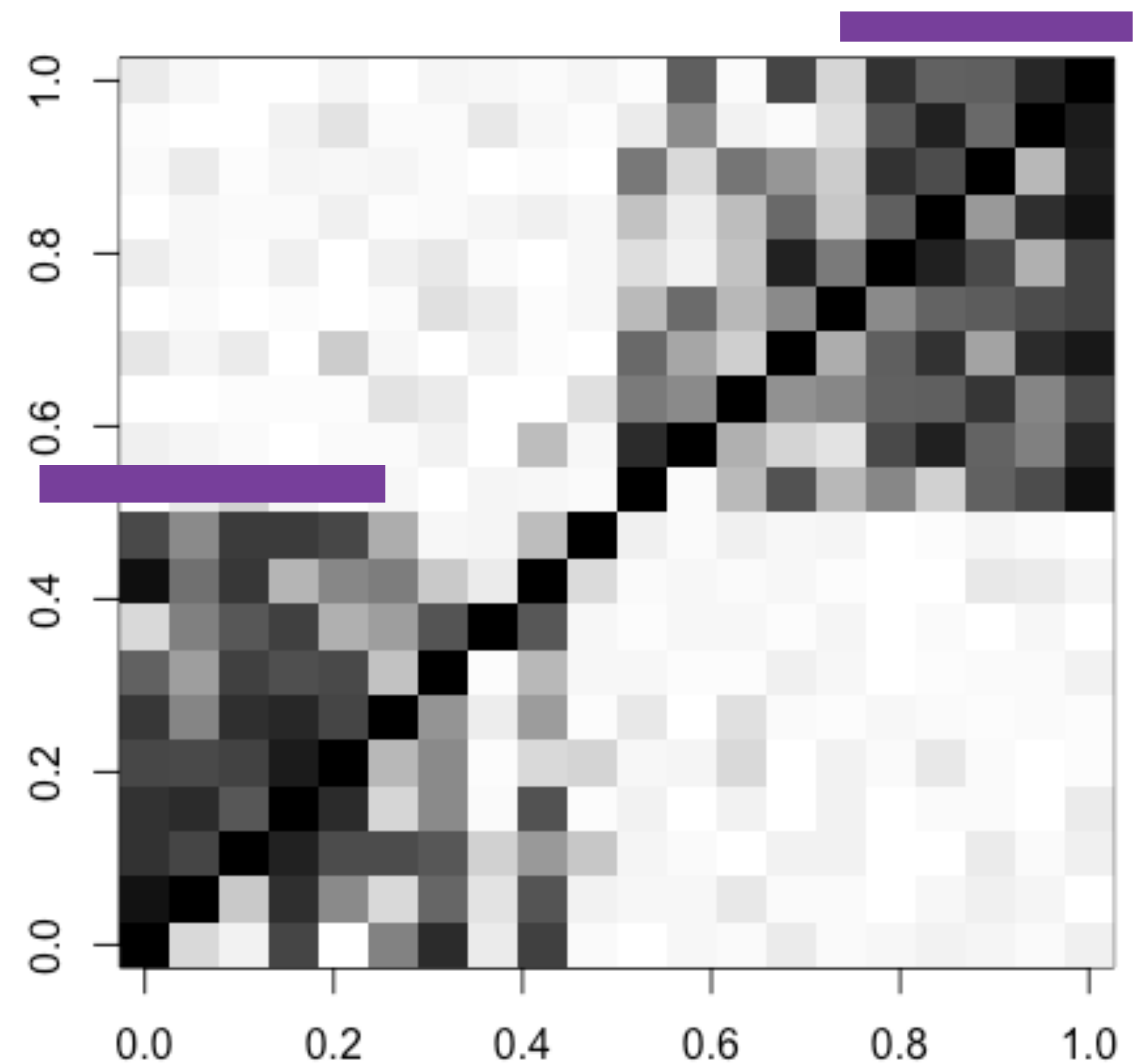




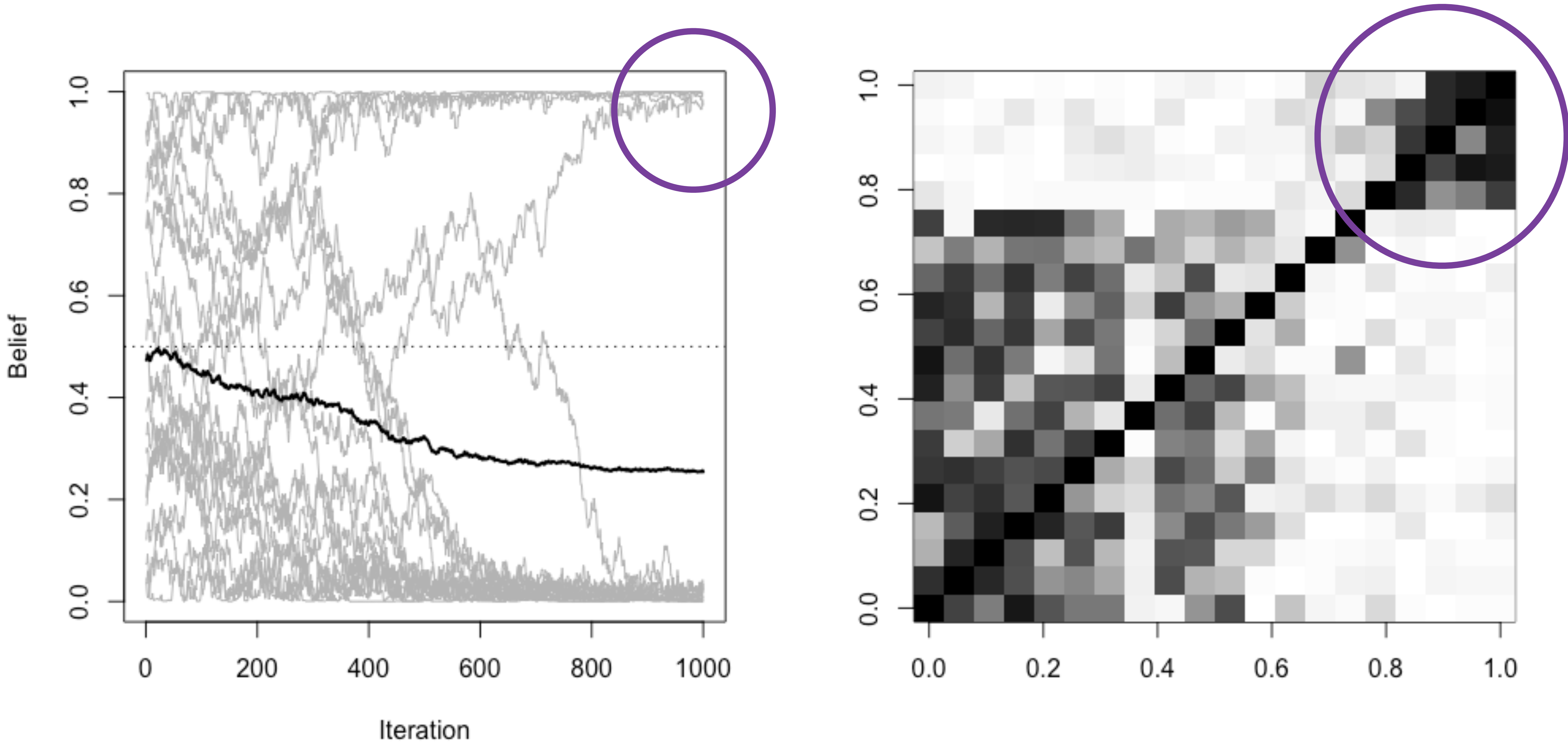
They might polarise into warring factions

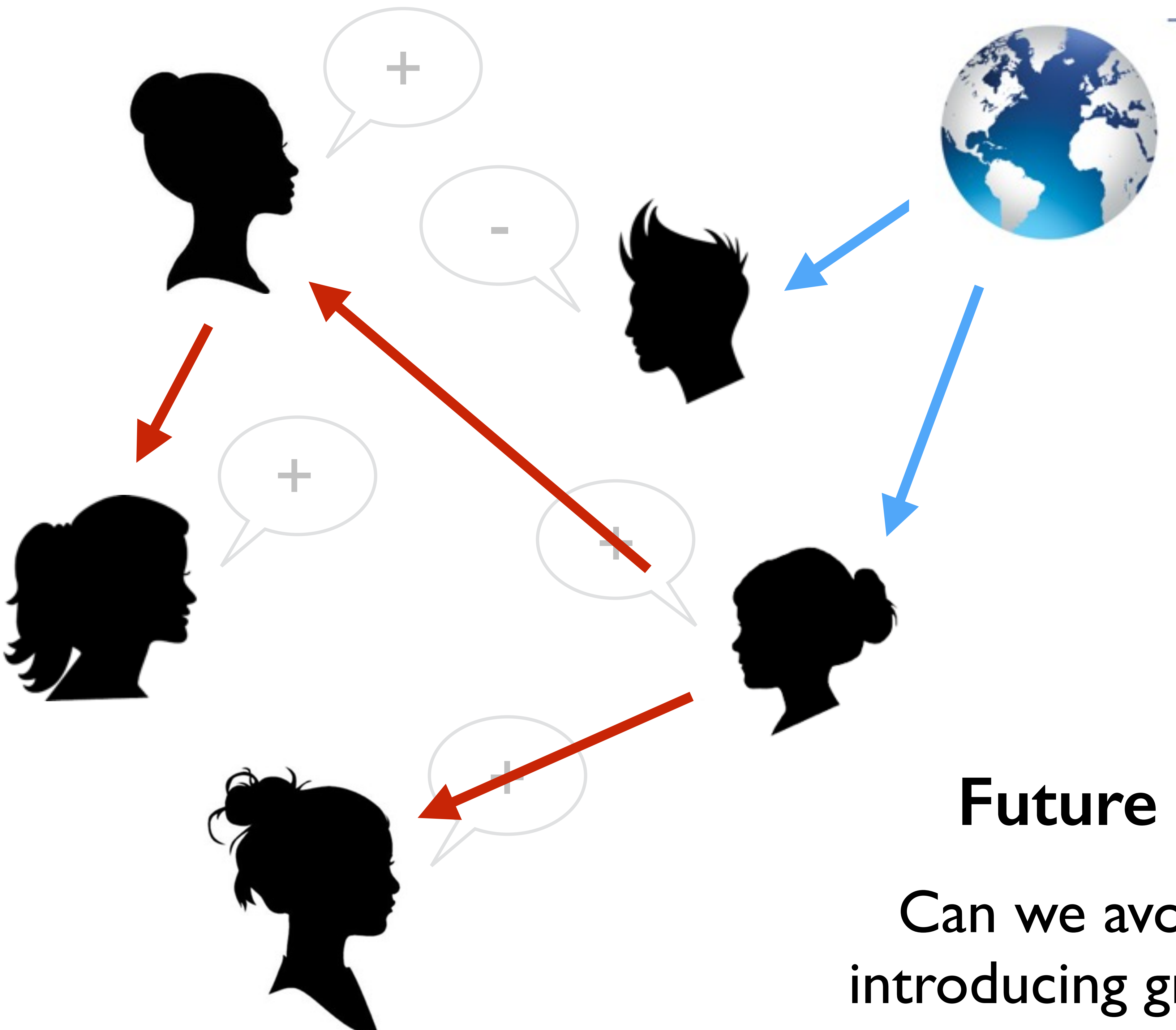
…with the **extremists** being most trusted within group; and no between-group trust





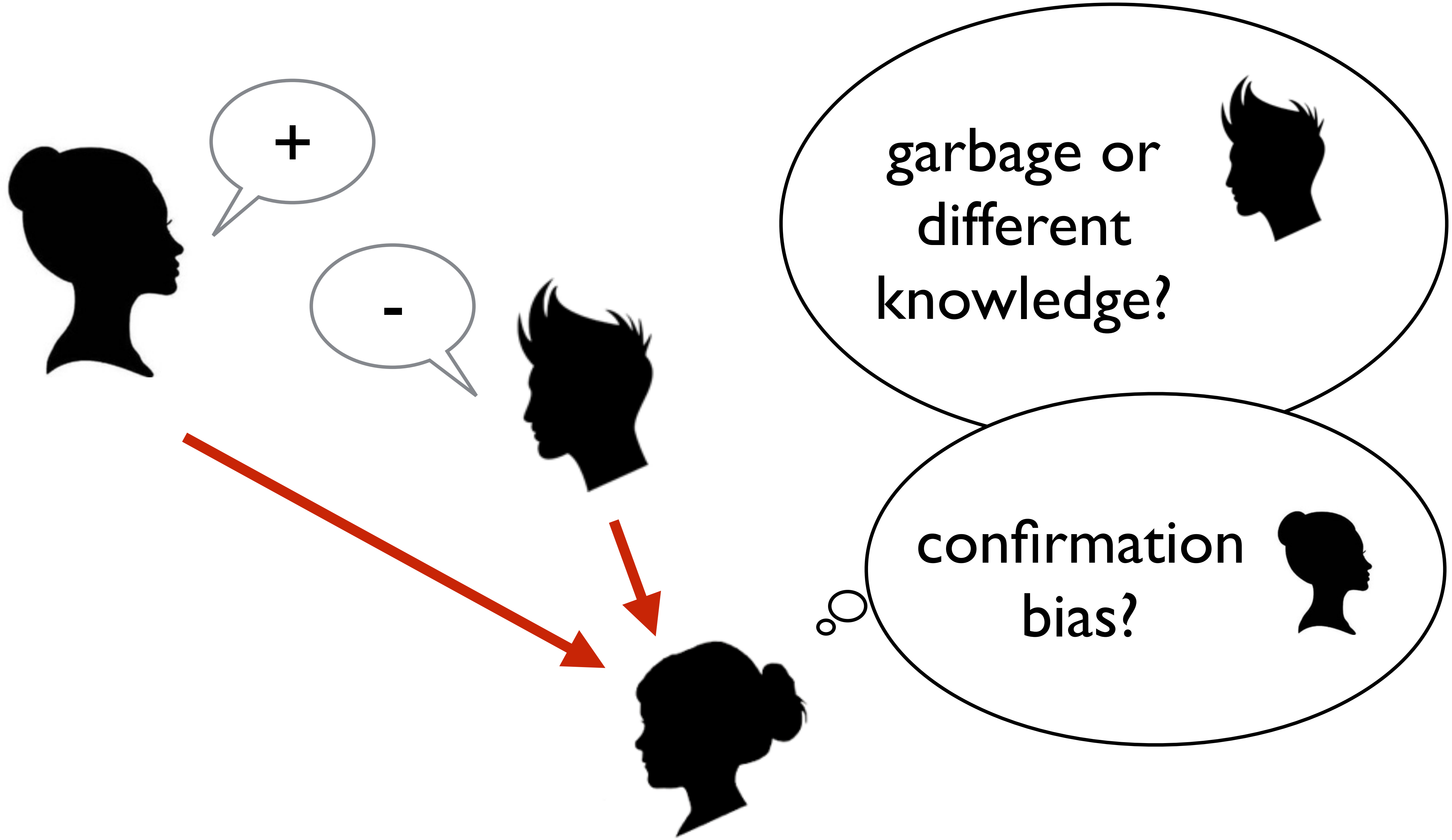
And small "rogue" groups might form their own isolated world.





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?

- <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 strong biases
    - The magnitude of the distortion is variable
    - Social structure, theory of mind, the link to the world… they all matter

- <u>Implications</u>:
    - IL is limited as a tool for "revealing inductive priors"
    - IL is potentially useful for studying "distortions" in cultural and linguistic evolution

Thanks!