Reinforcement Learning II: Biology

Quentin Huys

Wellcome Trust Centre for Neuroimaging, UCL
Gatsby Computational Neuroscience Unit, UCL
Guy's and St. Thomas' Hospital NHS Trust

Psychiatrische Universitatsklinik, University of Zurich
Theoretical Neuromodeling Unit, ETH

Advanced Course in Computational Neuroscience, Bedlewo, Poland, August 2012

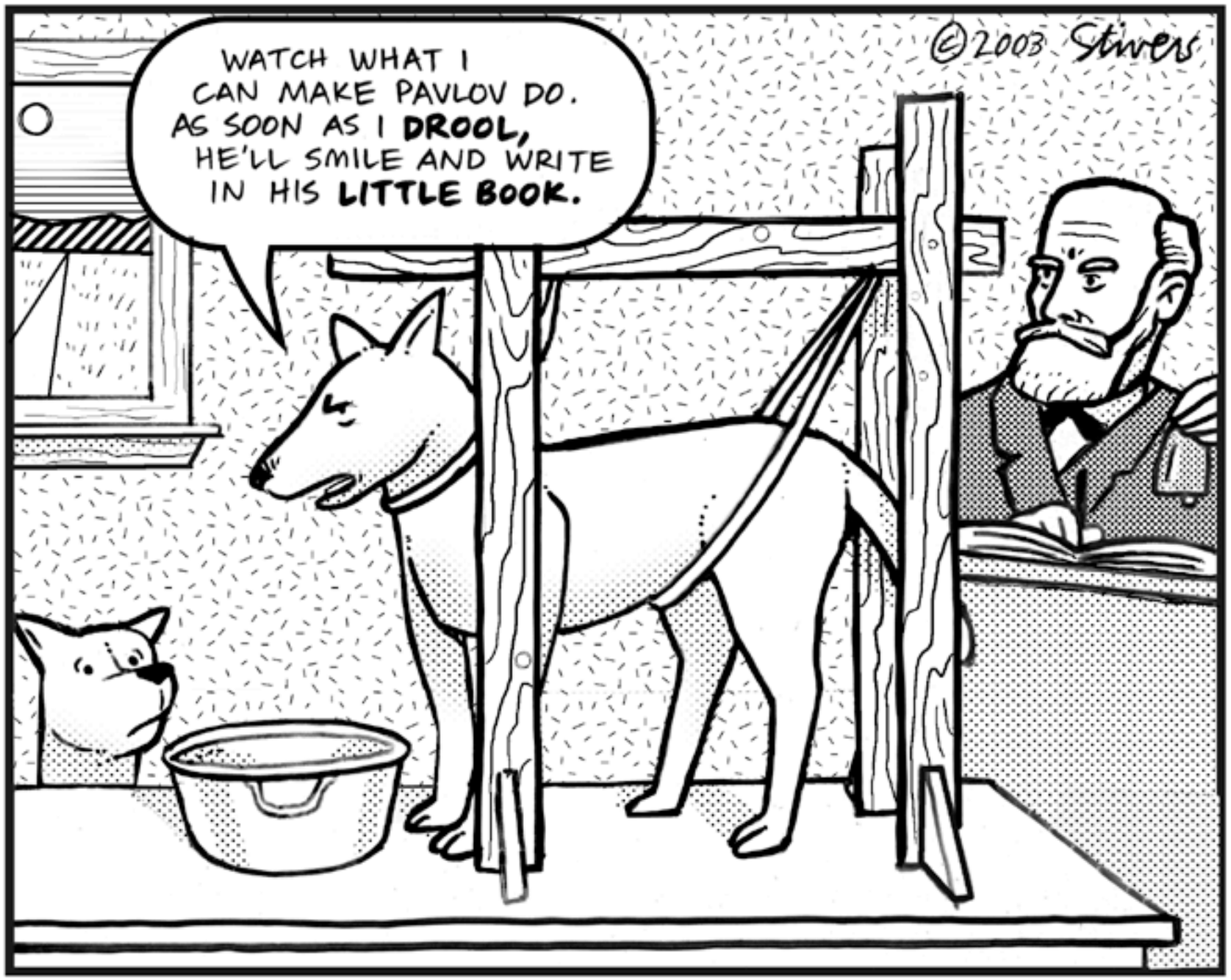
Recapitulation

- MDP: {s,a,T,R,pi}
- Multiple solution approaches
- “Global”: policy evaluation & iteration
- “Local”: sampling:
  - Monte Carlo
  - TD
- How do humans or animals solve RL problems?

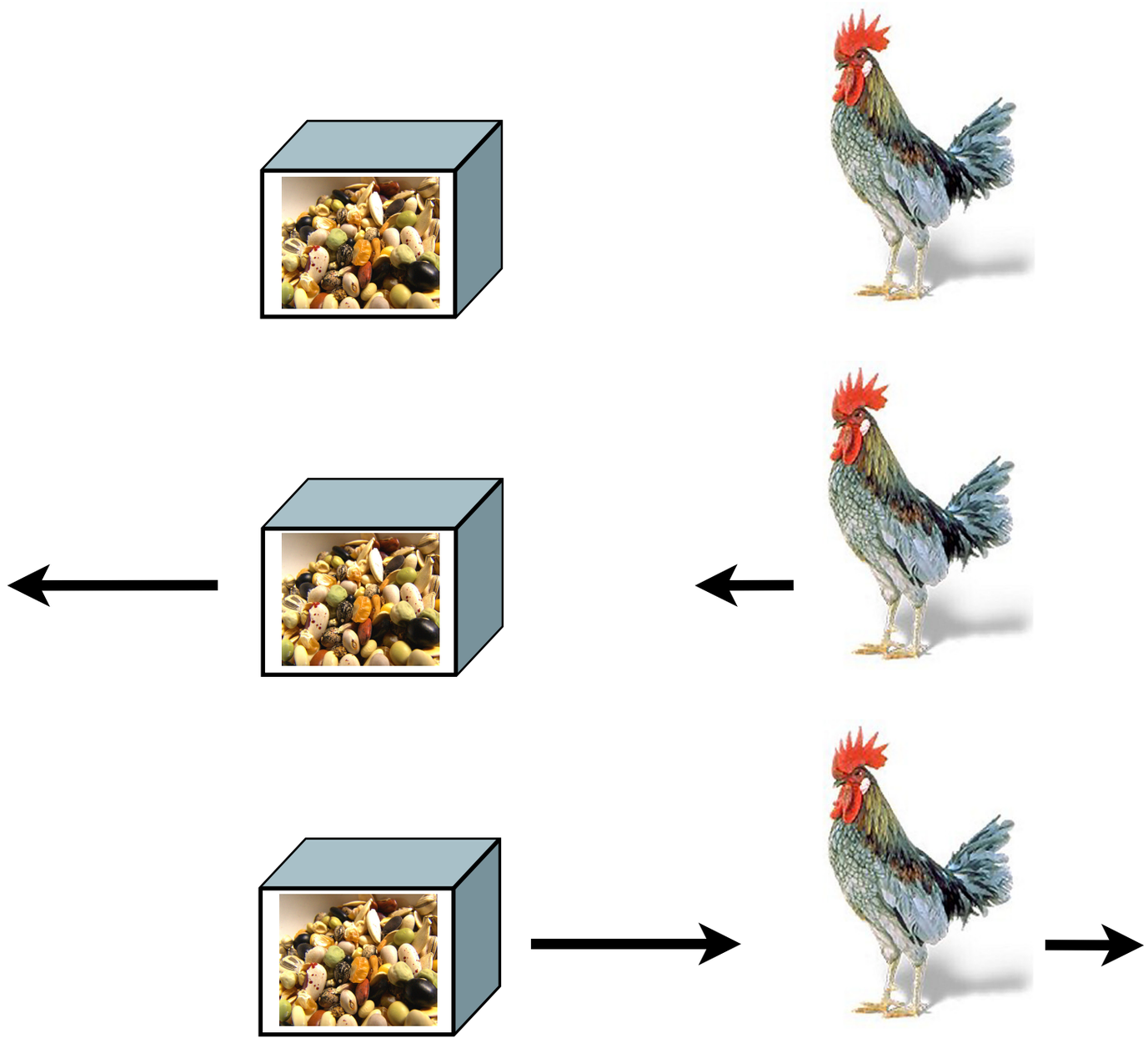
Policy

- Unconditioned responses
- Pavlovian conditioning
- Habits
- Goal-directed behaviour
- Doing tree search

Pavlovian conditioning



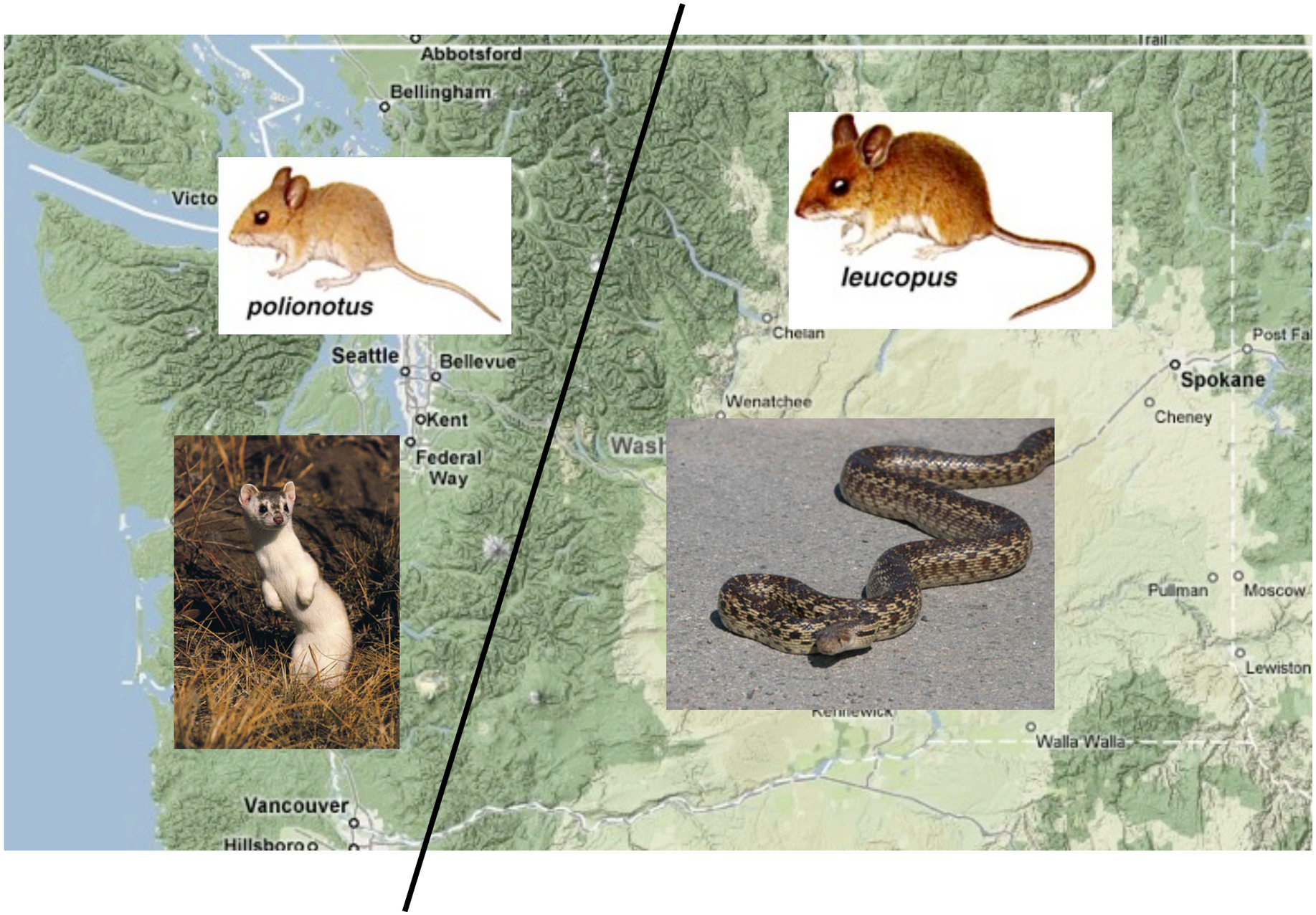
Unconditioned responses



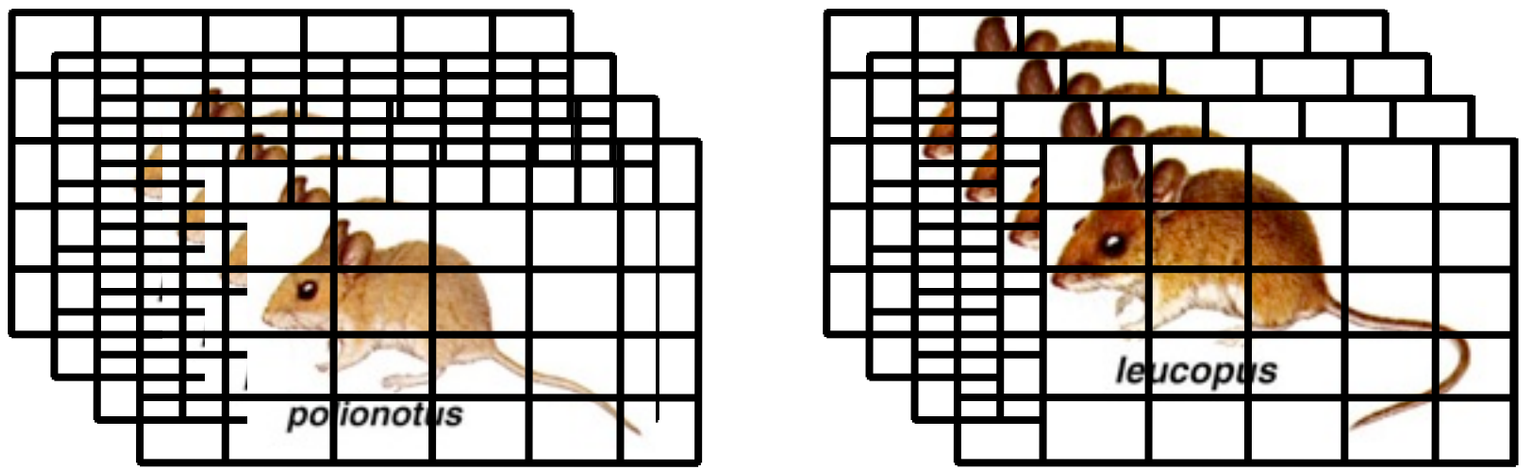
- powerful
- inflexible over short timescale
- adaptive on evolutionary scale

Hershberger 1986

are innate evolutionary strategies



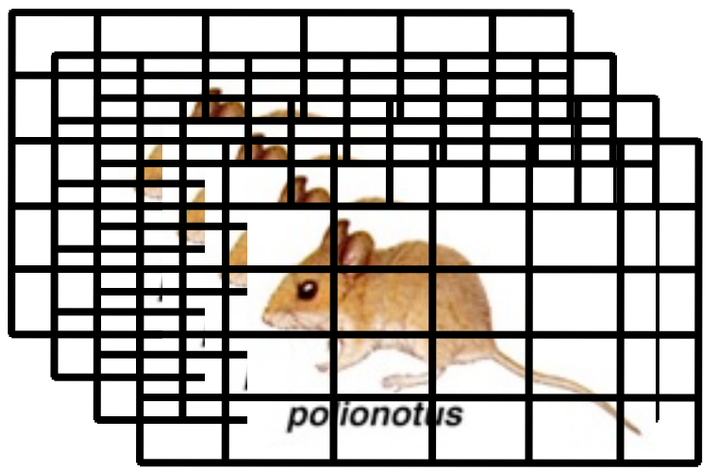
Innate evolutionary strategies

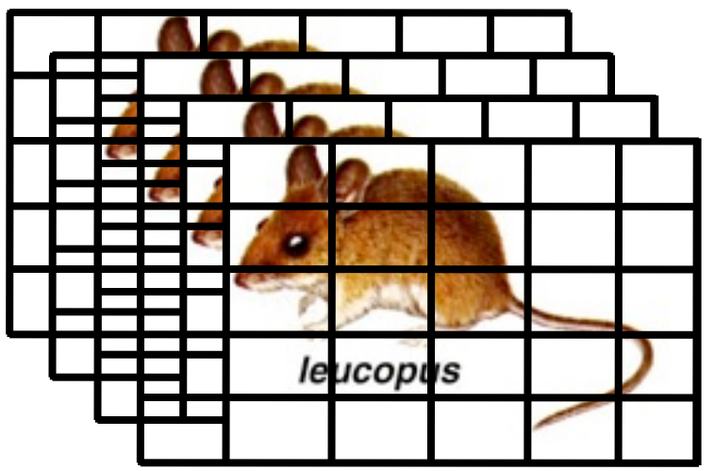


Hirsch and Bolles 1980

Innate evolutionary strategies

are quite sophisticated...





more survive

fewer survive

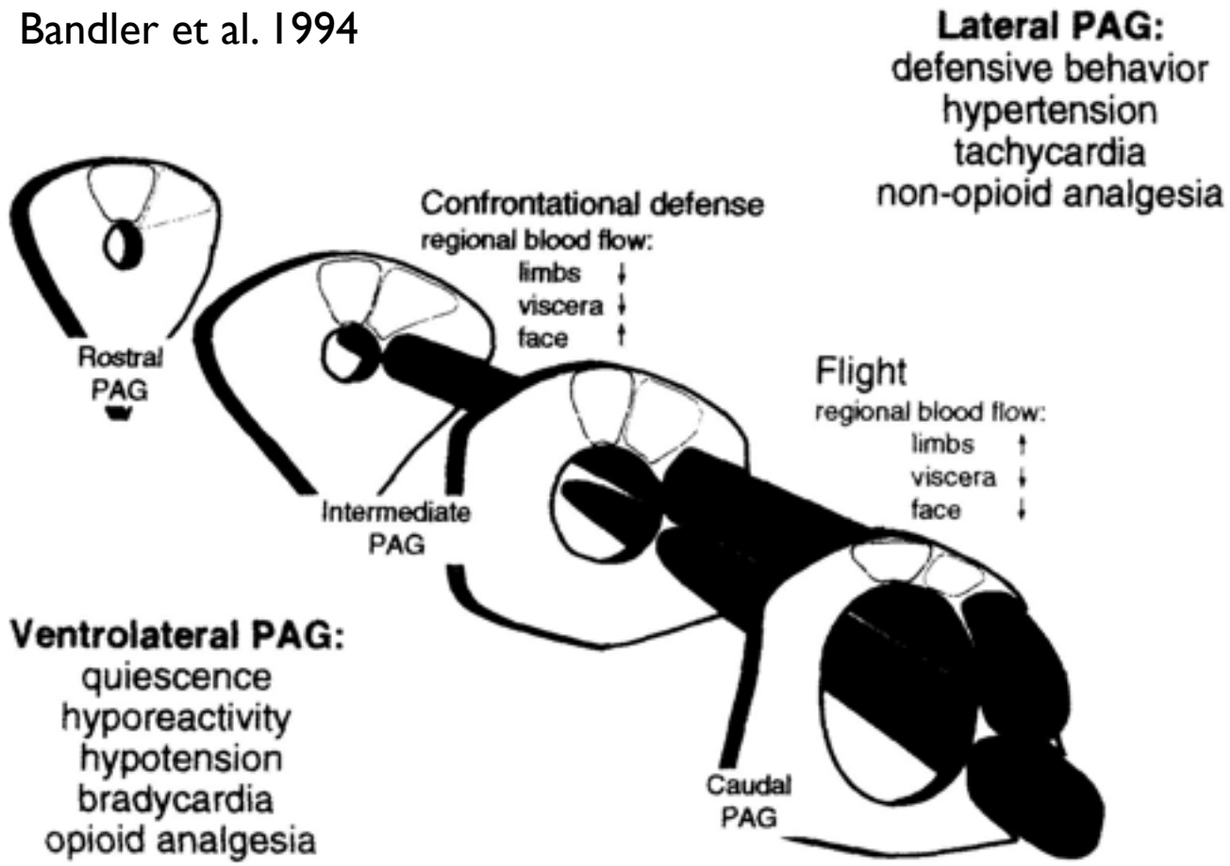
more survive

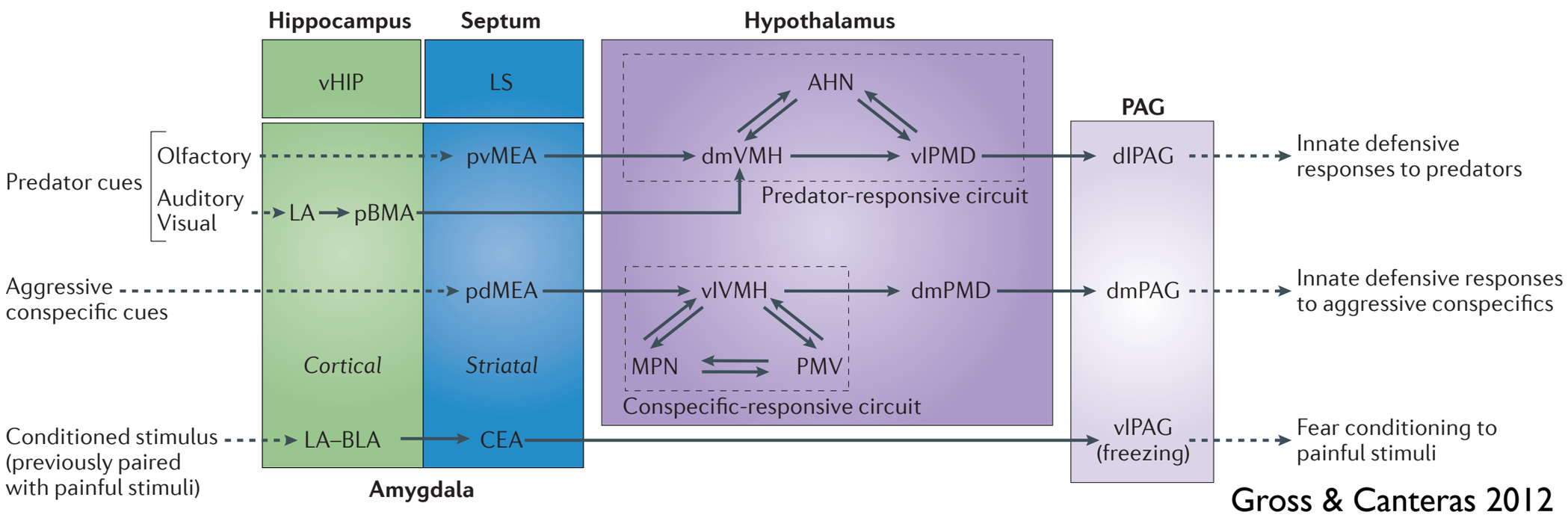
Hirsch and Bolles 1980

Specific biological substrates

- The dorsal periaquaeductal gray has a map of fear behaviours
- These may form part of three specific fear circuits to classes of dangerous stimuli

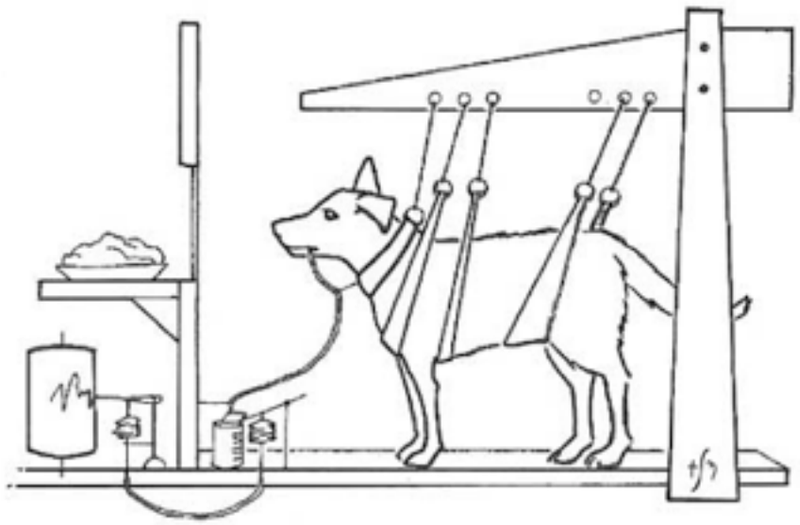
Bandler et al. 1994





Gross & Canteras 2012

Pavlovian conditioning

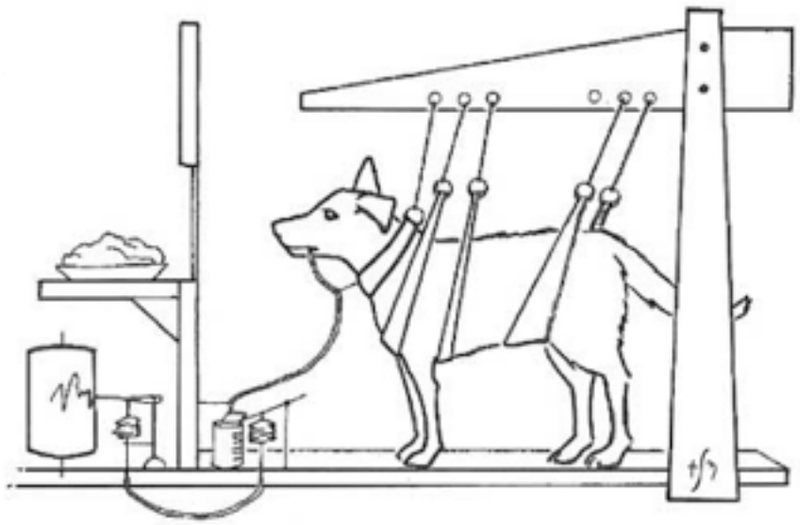


prior to training

CS: bell -> no response

US: food -> salivation & consumption

Pavlovian conditioning



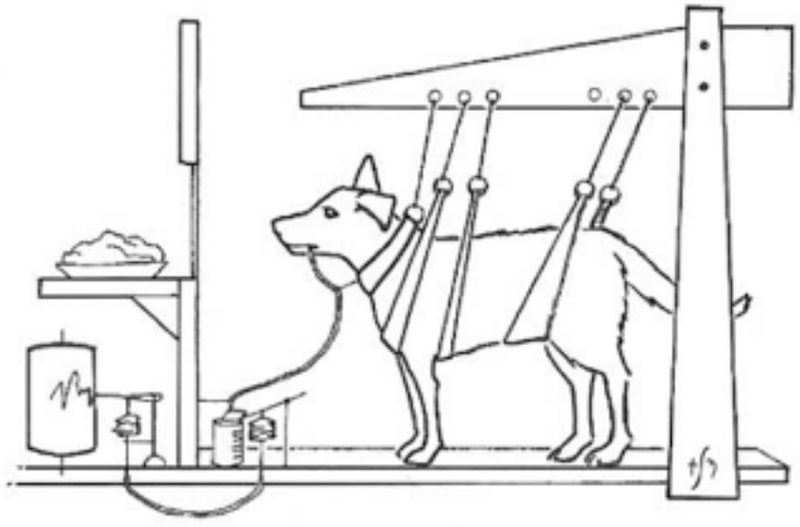
prior to training

CS: bell -> no response
US: food -> salivation & consumption

training

CS: bell ⟶ US: food

Pavlovian conditioning



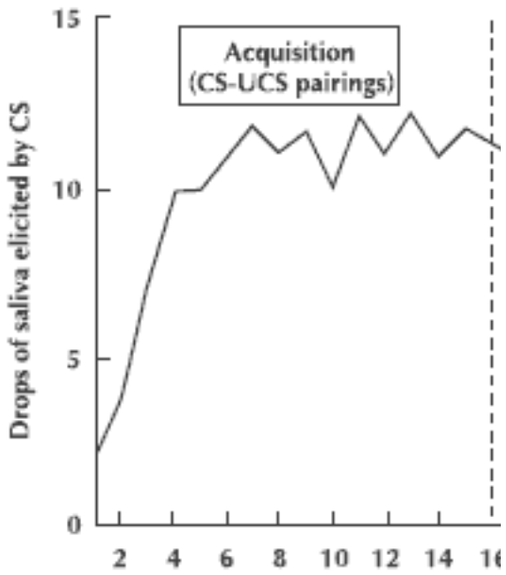
prior to training

CS: bell -> no response

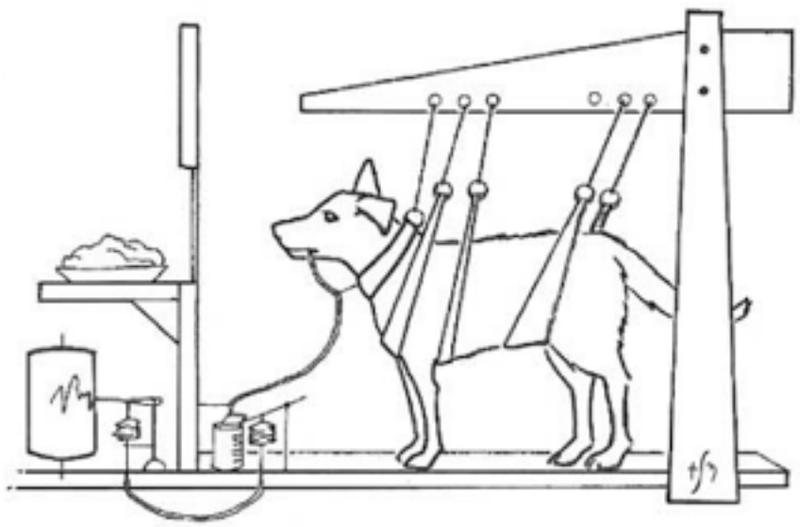
US: food -> salivation & consumption

training

CS: bell ⟶ US: food



Pavlovian conditioning



prior to training

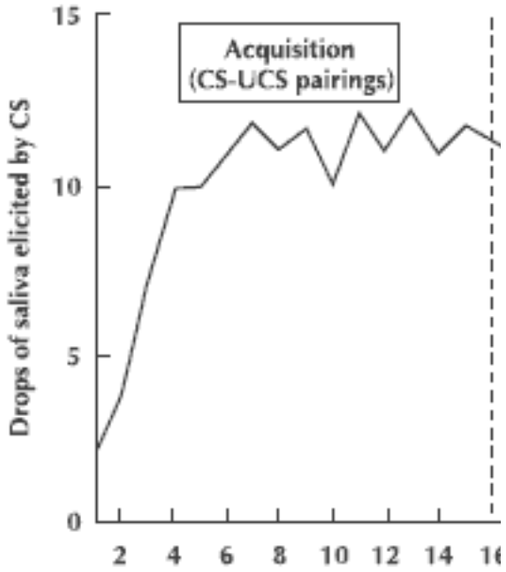
CS: bell -> no response
US: food -> salivation & consumption

training

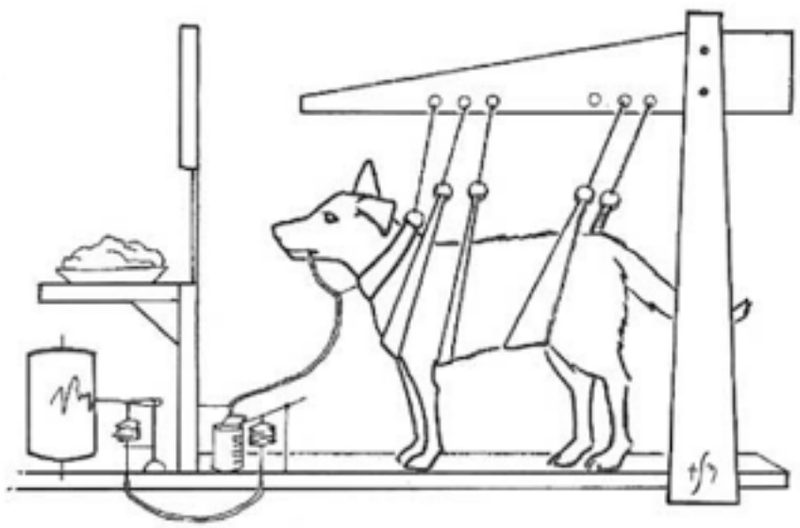
CS: bell ⟶ US: food

after training:

CS: bell -> salivation
US: food -> salivation & consumption



Pavlovian conditioning



prior to training

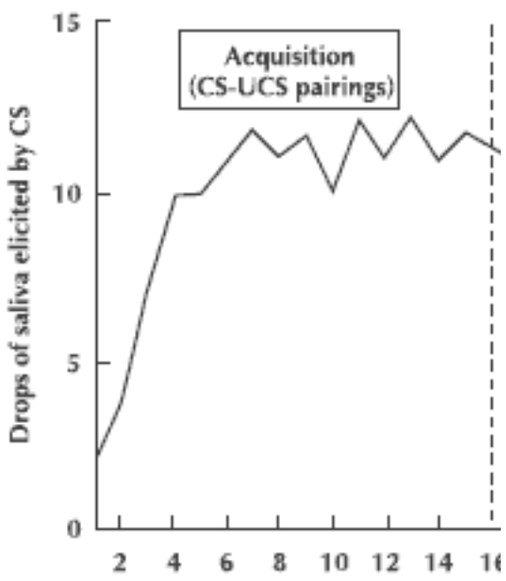
CS: bell -> no response
US: food -> salivation & consumption

training

CS: bell ⟶ US: food

after training:

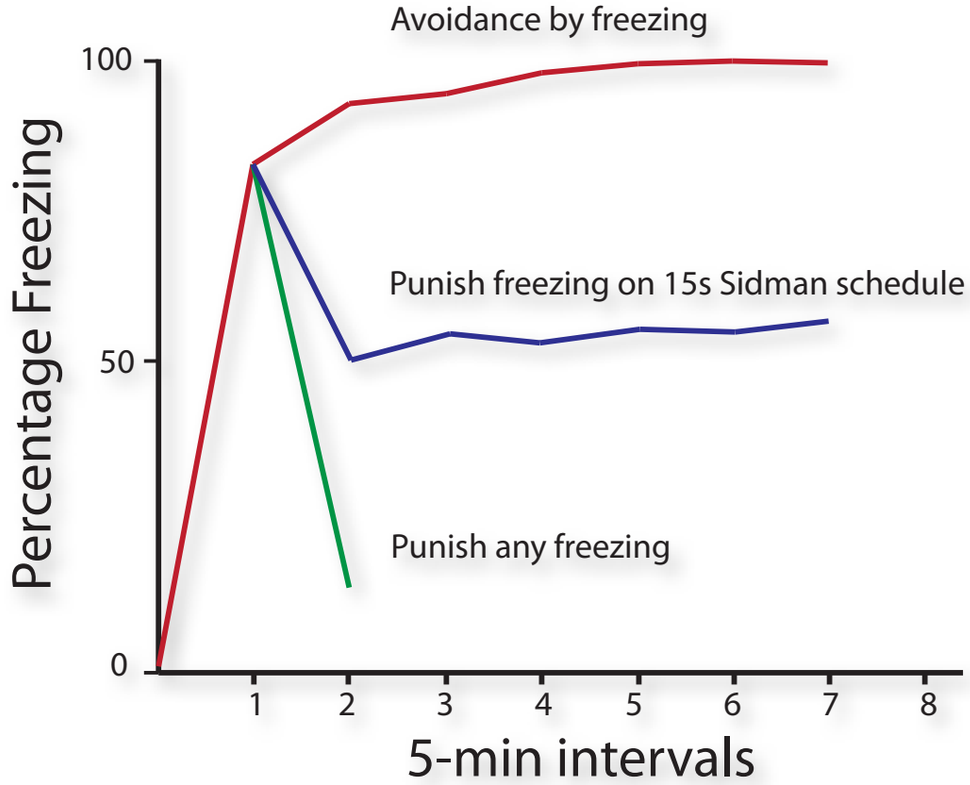
CS: bell -> salivation
US: food -> salivation & consumption



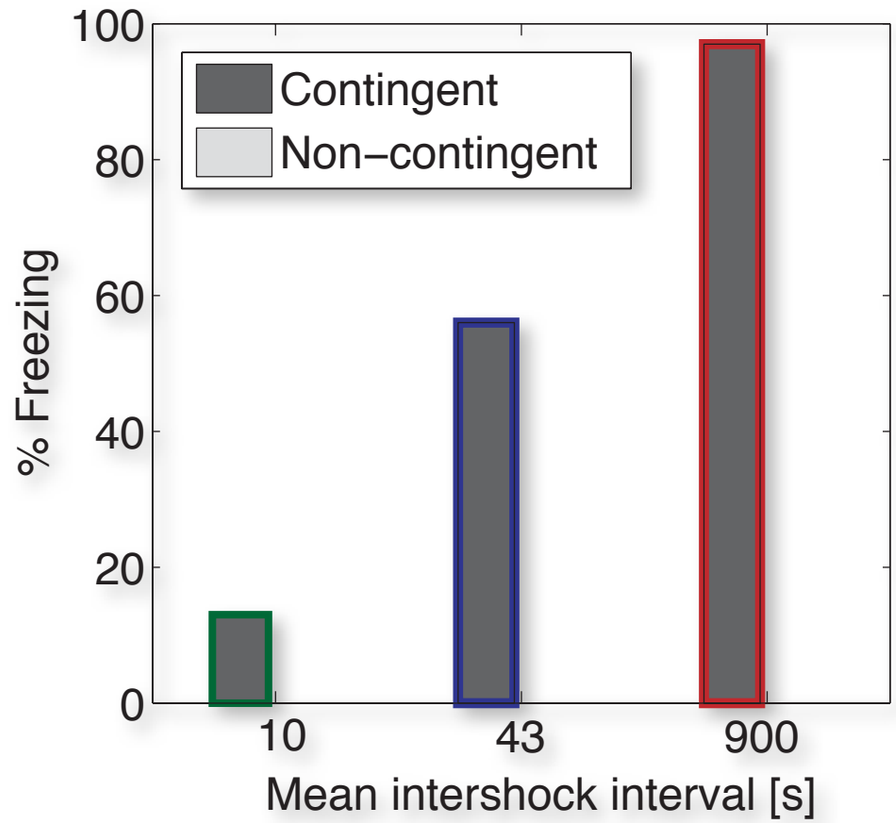
omission training:

CS: bell —salivation→ US: omitted
CS: bell —no salivation→ US: food

Aversive Pavlovian effects: freezing

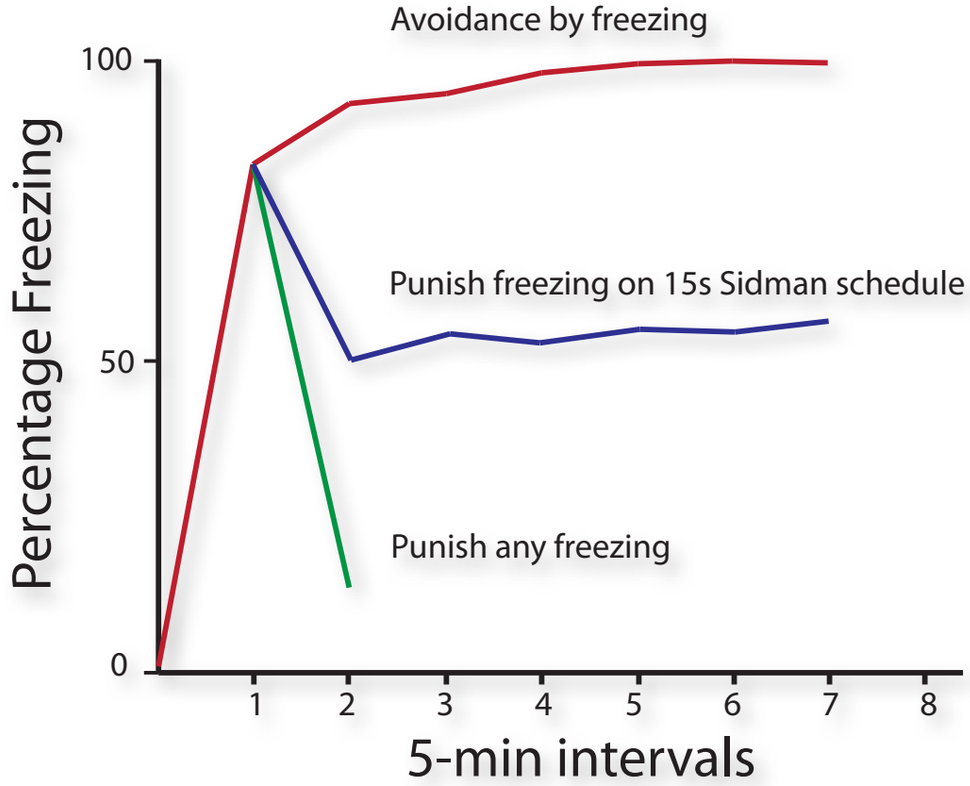


Bolles and Riley 1973

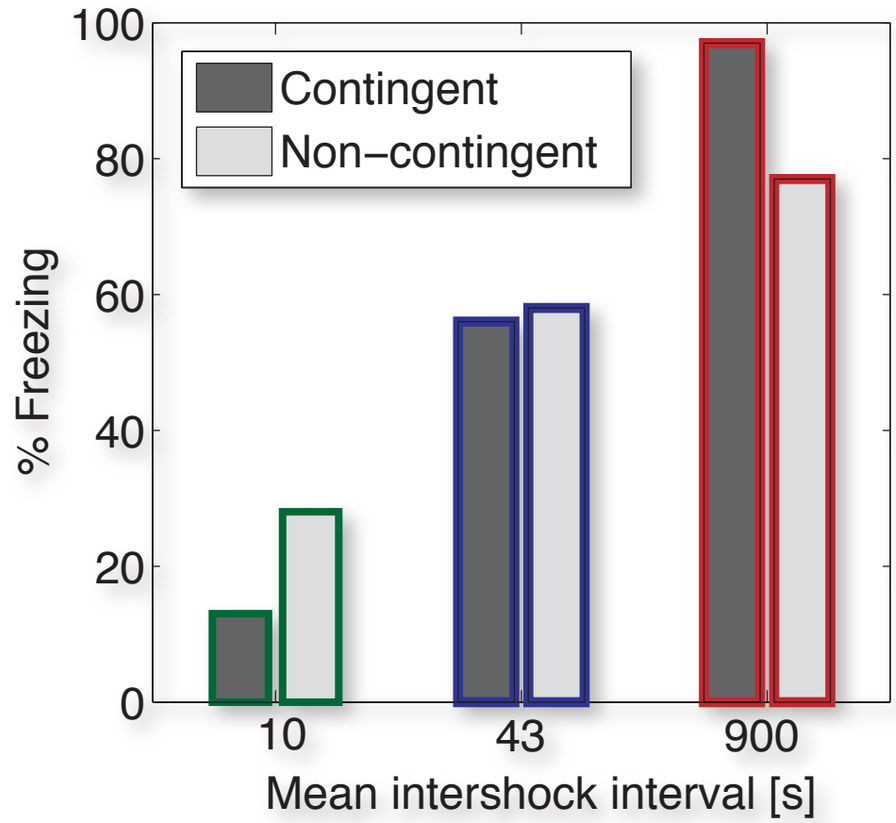


Bolles and Riley 1973

Aversive Pavlovian effects: freezing



Bolles and Riley 1973



Bolles and Riley 1973

A UK checkout

A UK checkout

A UK checkout

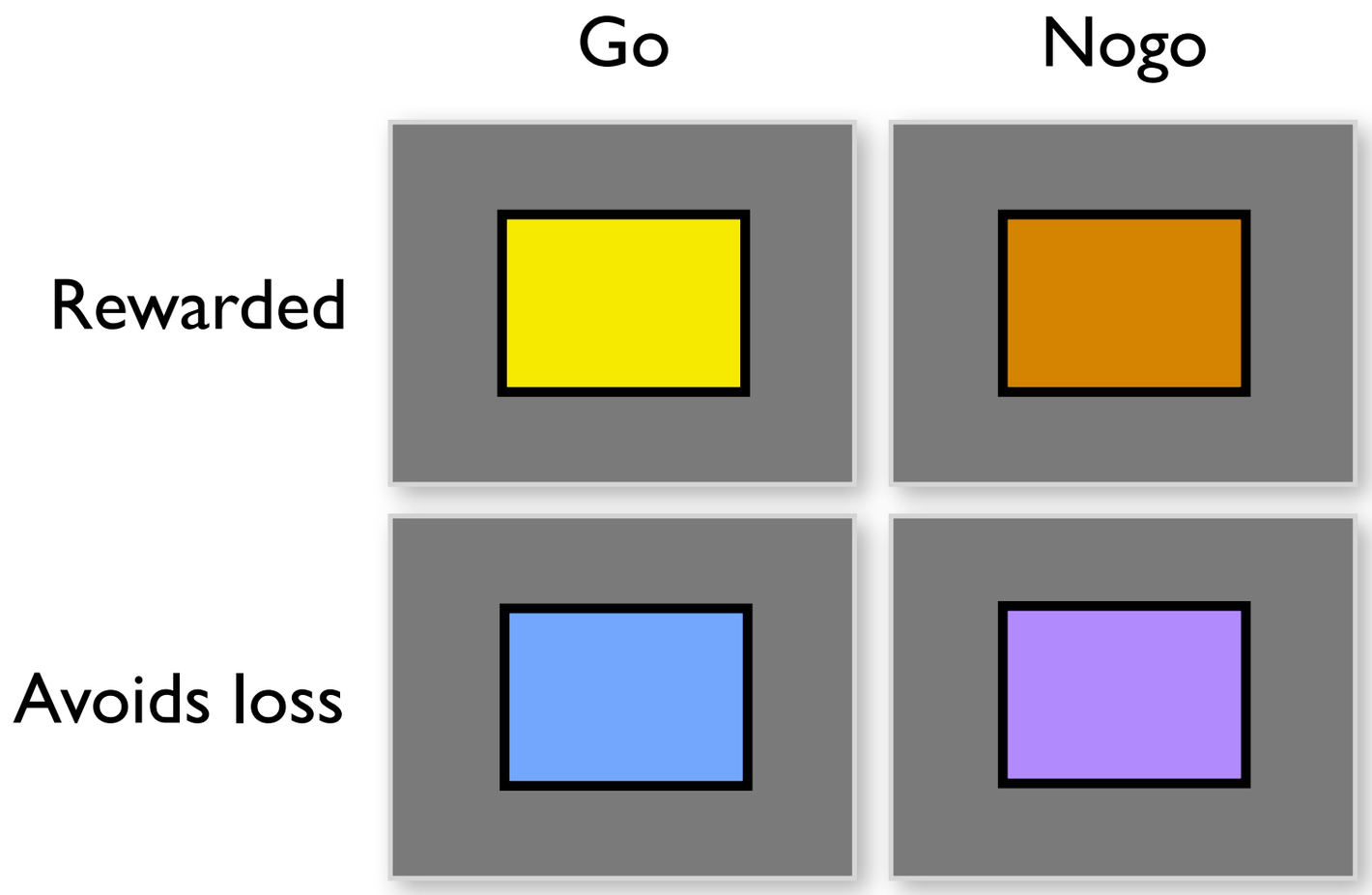
Tobacco firms accused of funding campaign to keep cigarettes on display

Jamie Doward and Alex Ascherson
guardian.co.uk, Saturday 26 February 2011 20.54 GMT
Article history

Cigarettes displayed for sale at a store in central London. Photograph: Andy Rain/EPA

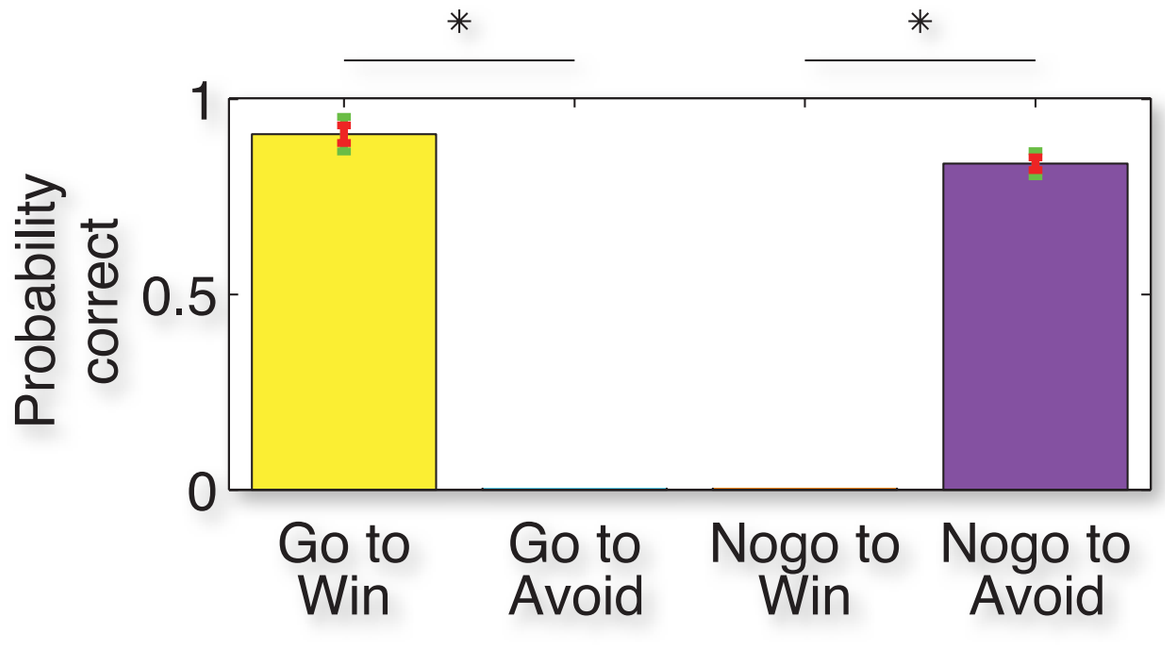
The Guardian, Sat Feb 26th 2011

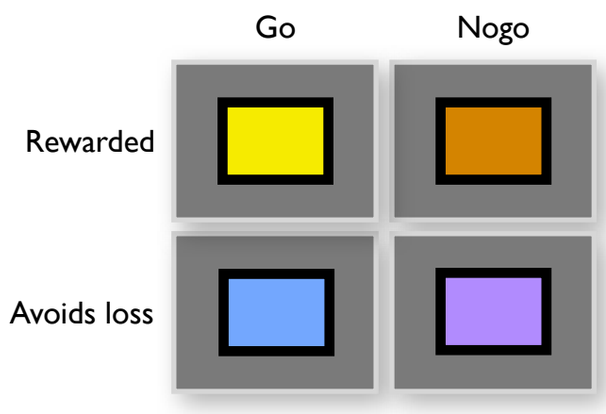
Pavlovian influences on instrumental learning

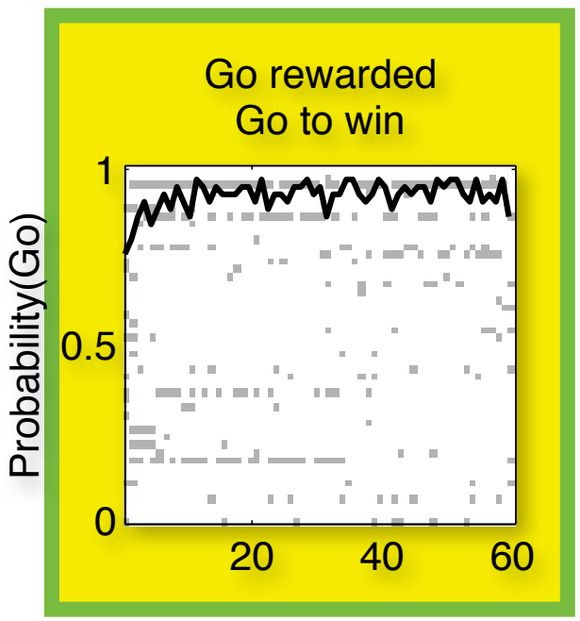


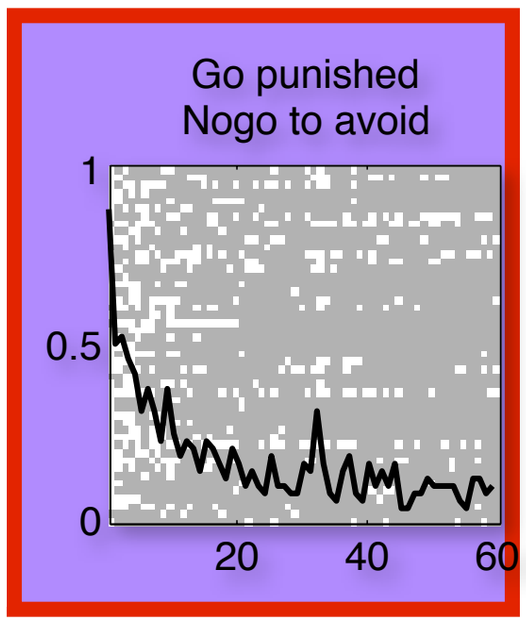
Guitart-Masip, Huys et al. 2011, 2012

Pavlovian influences on instrumental learning



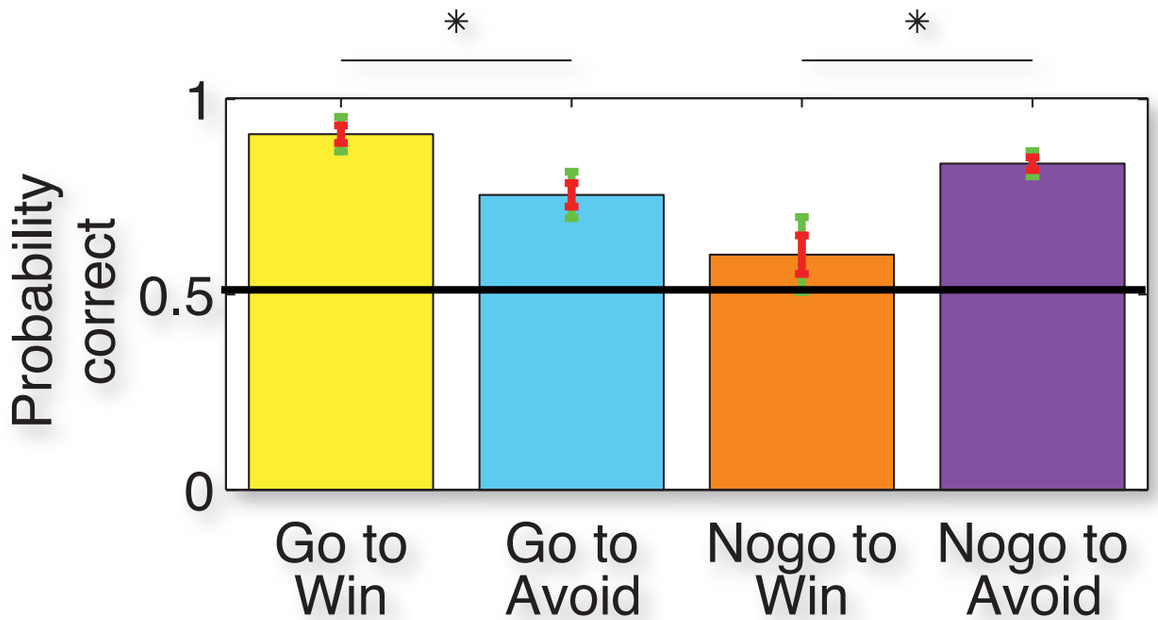


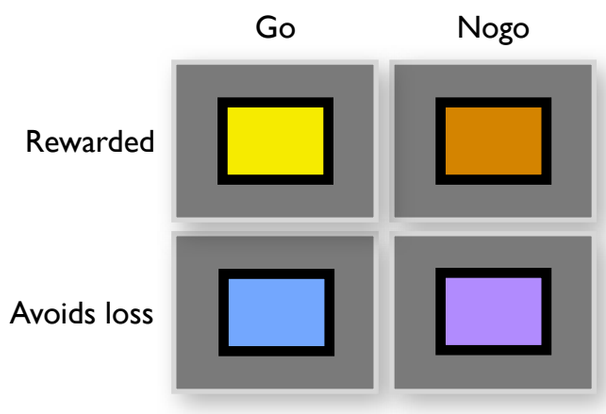


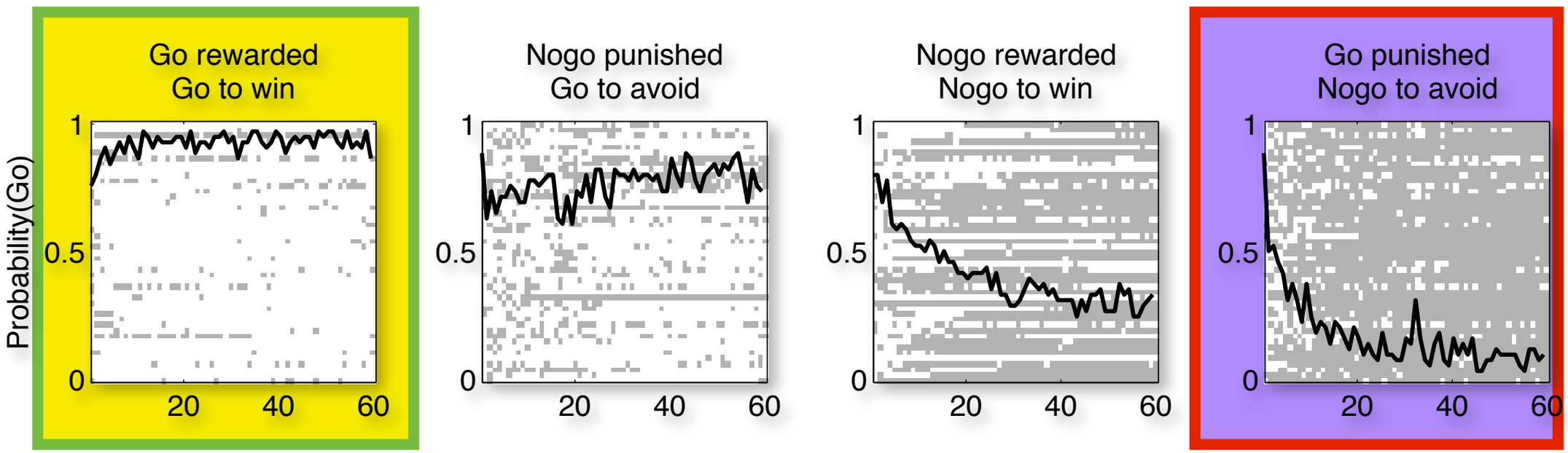


Guitart-Masip, Huys et al. 2011, 2012

Pavlovian influences on instrumental learning







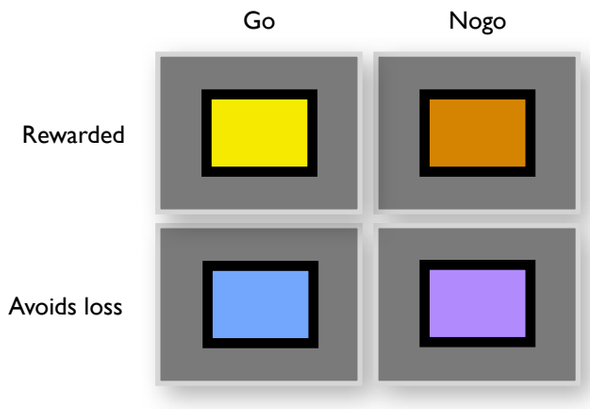
Guitart-Masip, Huys et al. 2011, 2012

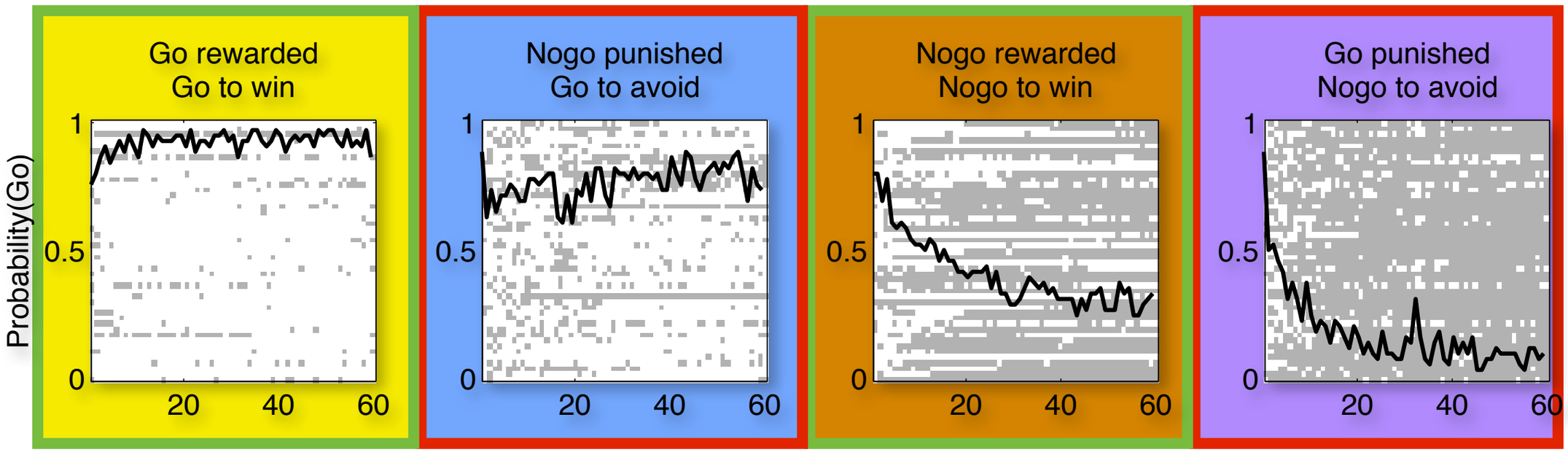
Pavlovian conditioning

- By being predictive of an affective outcome, a neutral stimulus can come to elicit the innate preparatory response usually evoked by the affective outcome.

- p(active response) ~ value(stimulus)?

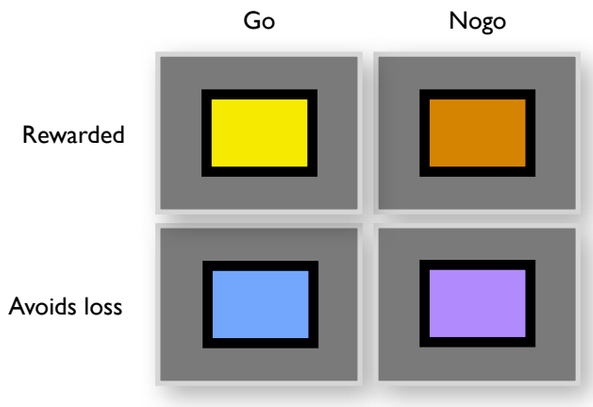
Model



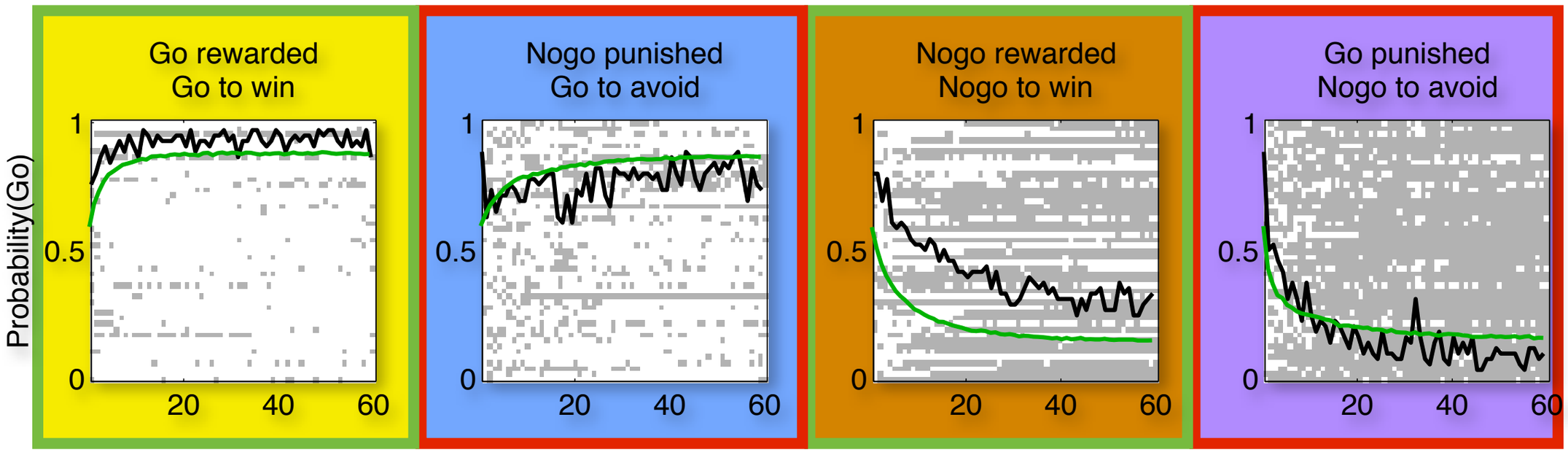


Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. *Submitted*

Model

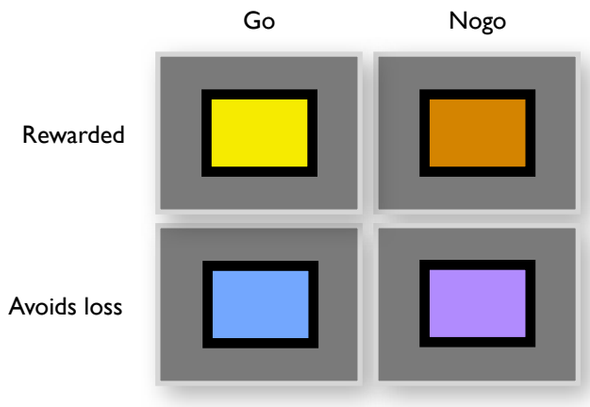


$$p(\text{go}|s_t) \quad \propto \quad \mathcal{Q}_t(\text{go}|s_t) + \text{bias}(\text{go})$$



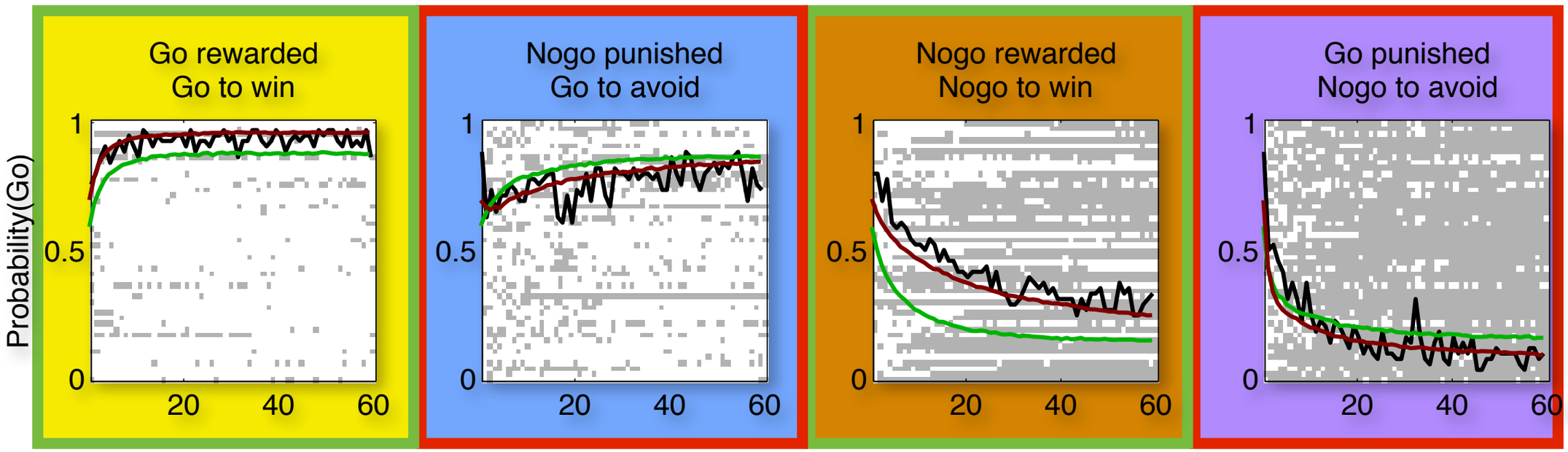
Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. *Submitted*

Model



$$p(\text{go}|s_t) \propto \mathcal{Q}_t(\text{go}|s_t) + \text{bias}(\text{go}) + \mathcal{V}_t(s_t)$$

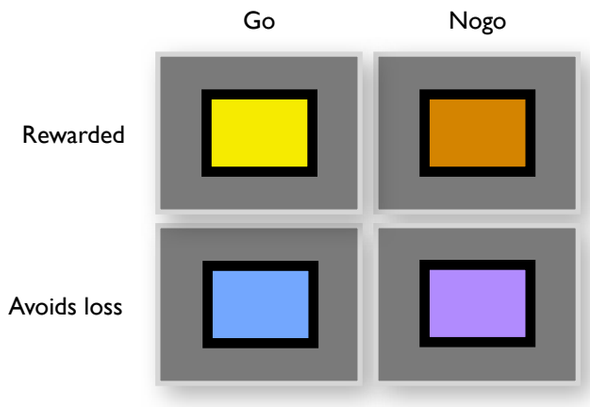
$$\mathcal{V}_t(s_t) = \mathcal{V}_{t-1}(s_t) + \epsilon(r_t - \mathcal{V}_{t-1}(s_t))$$



Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. *Submitted*

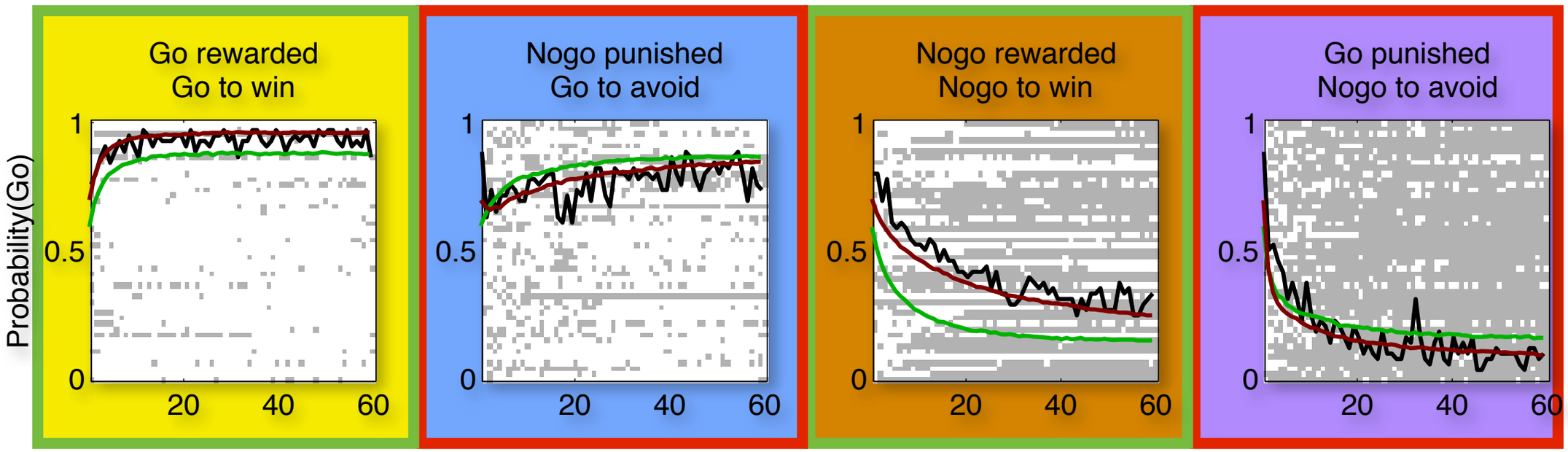
Model

P(go) $\propto$ value of stimulus



$$p(\text{go}|s_t) \propto \mathcal{Q}_t(\text{go}|s_t) + \text{bias}(\text{go}) + \mathcal{V}_t(s_t)$$

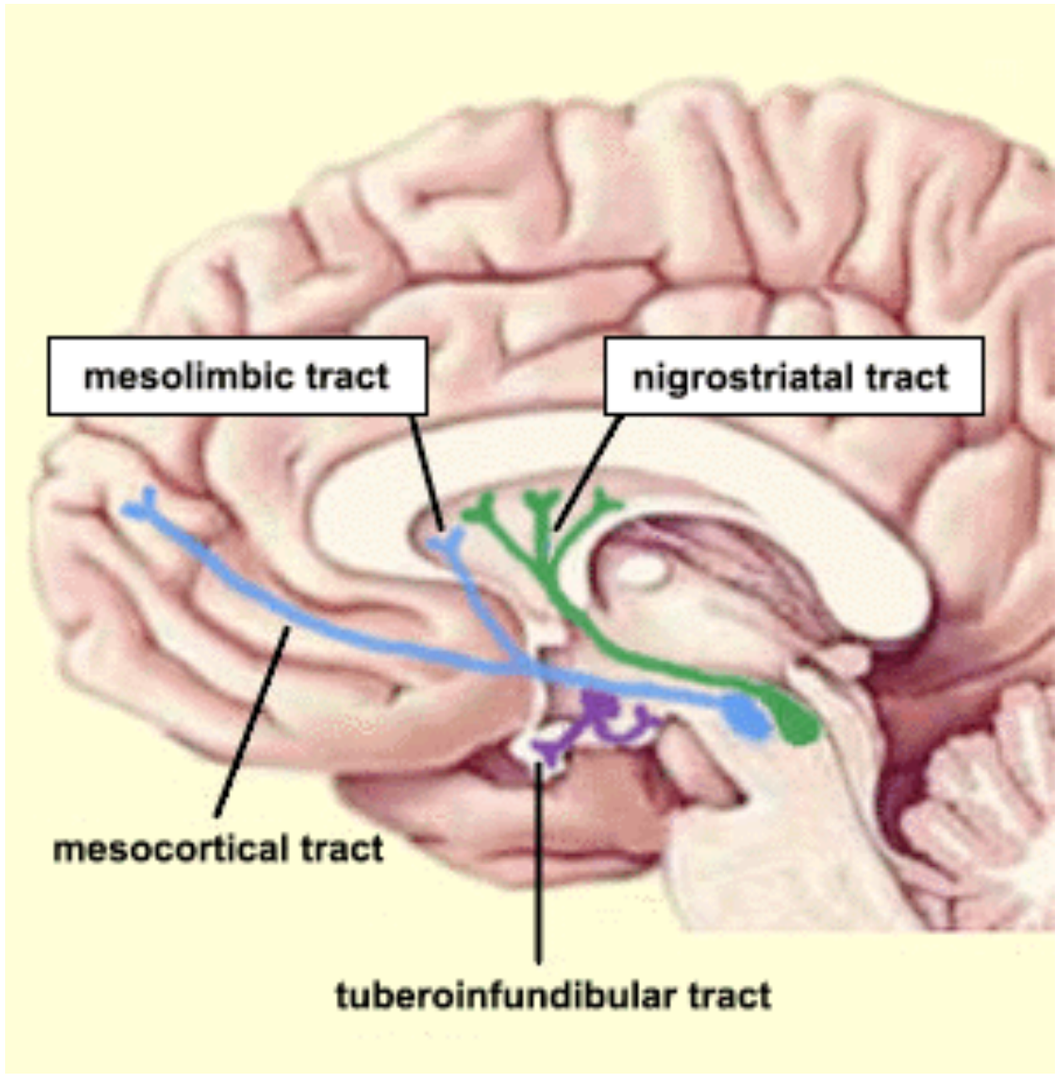
$$\mathcal{V}_t(s_t) = \mathcal{V}_{t-1}(s_t) + \epsilon(r_t - \mathcal{V}_{t-1}(s_t))$$



Guitart-Masip et al. 2011, Guitart-Masip, Huys et al. *Submitted*

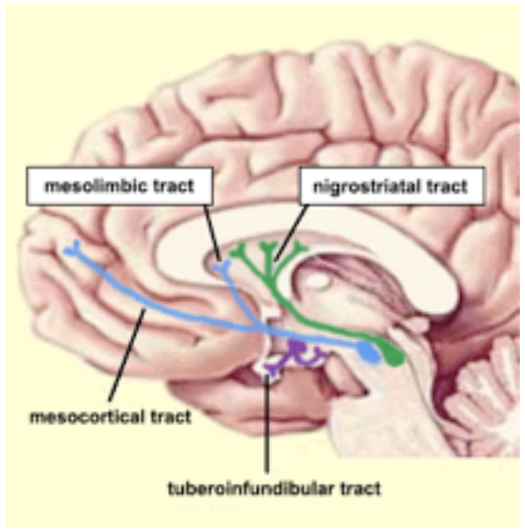
Pavlovian values in the brain?

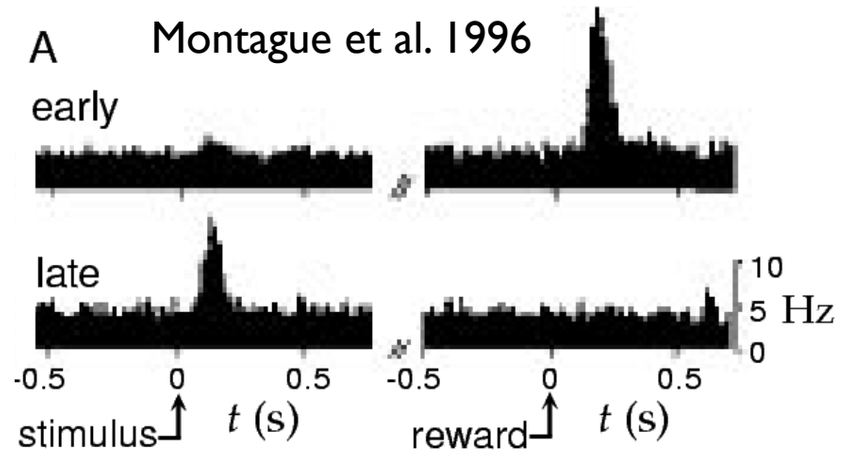
- Where the values V(s) are is not so clear
- Midbrain dopamine neurons seem to report a TD error



Pavlovian values in the brain?

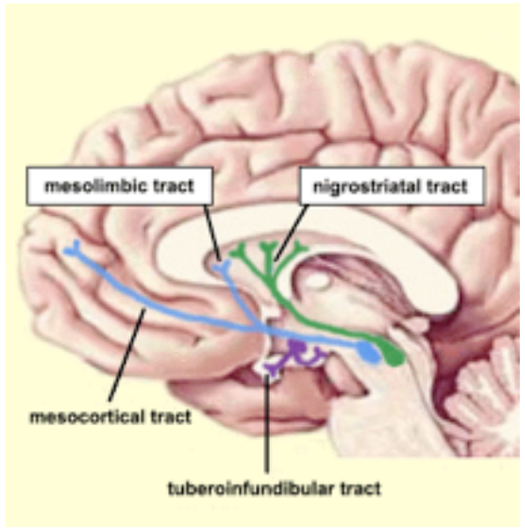
- Where the values V(s) are is not so clear
- Midbrain dopamine neurons seem to report a TD error

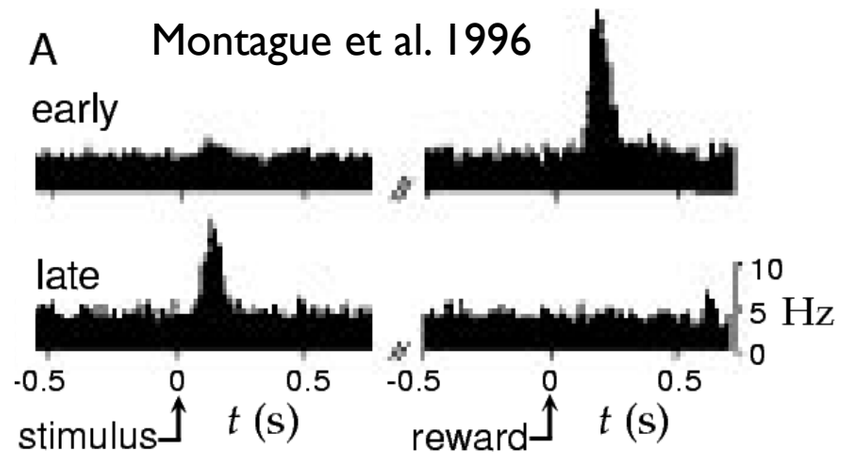


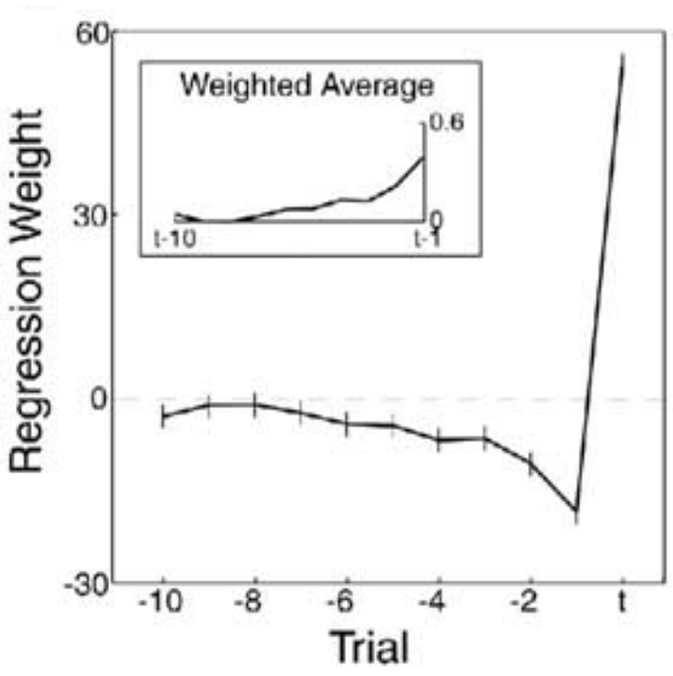


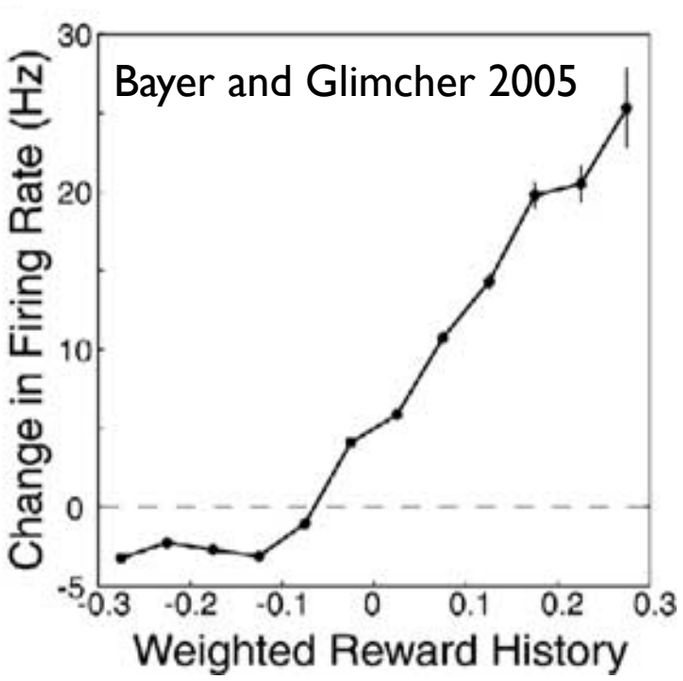
Pavlovian values in the brain?

- Where the values V(s) are is not so clear
- Midbrain dopamine neurons seem to report a TD error



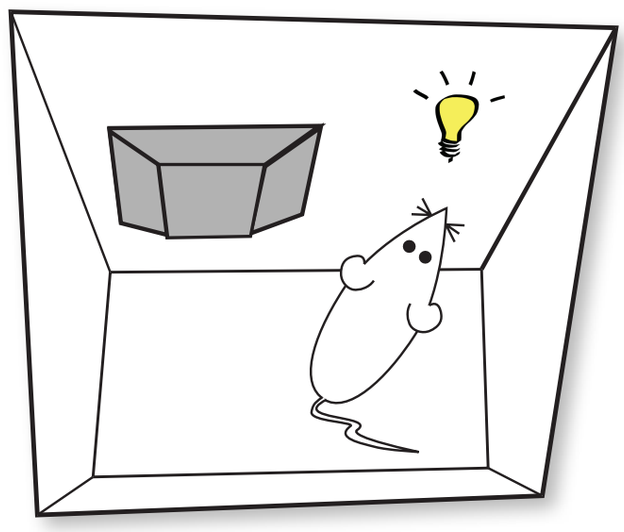




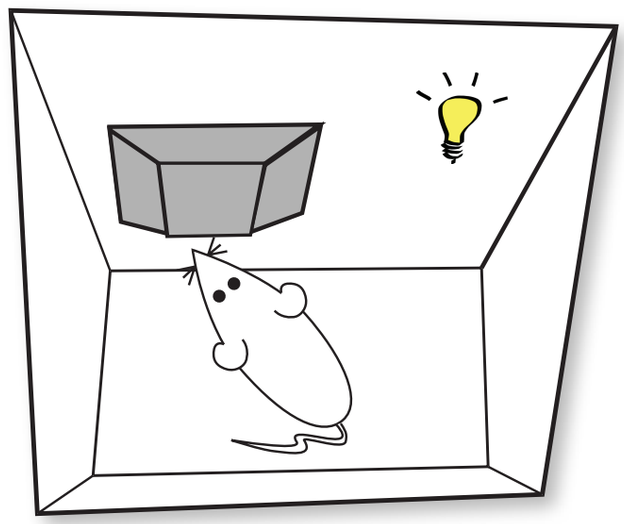


Addictive Pavlovian values

A



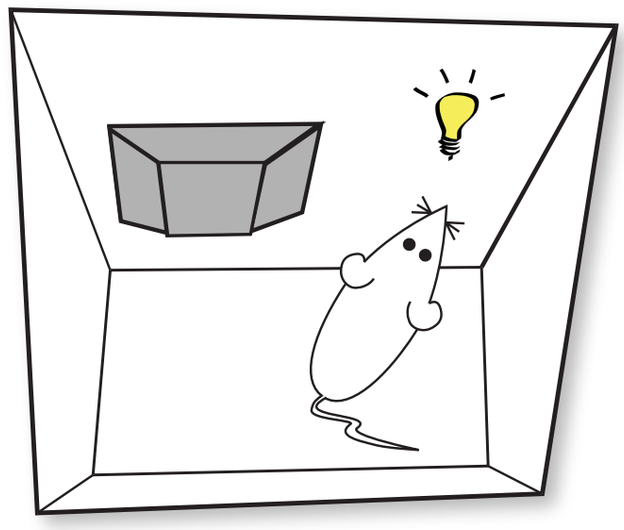
D



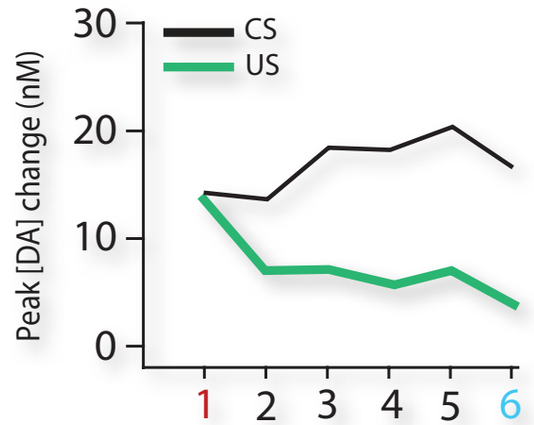
Huys et al. 2012 after Flagel et al. 2009, 2011

Addictive Pavlovian values

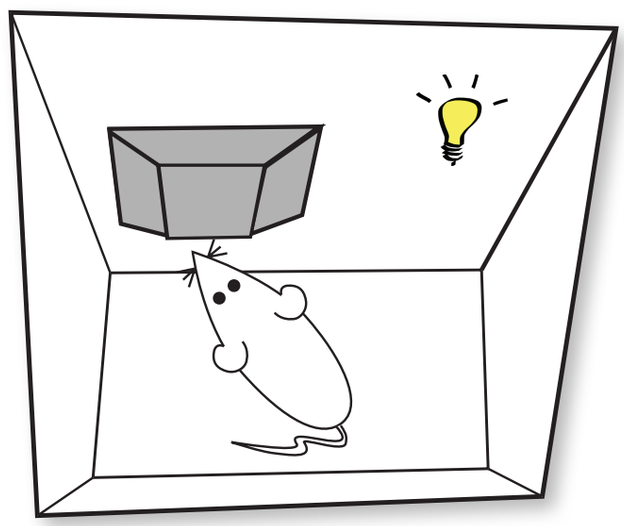
A



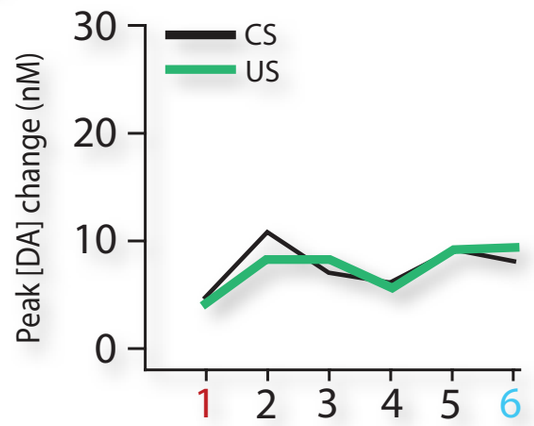
C



D



F

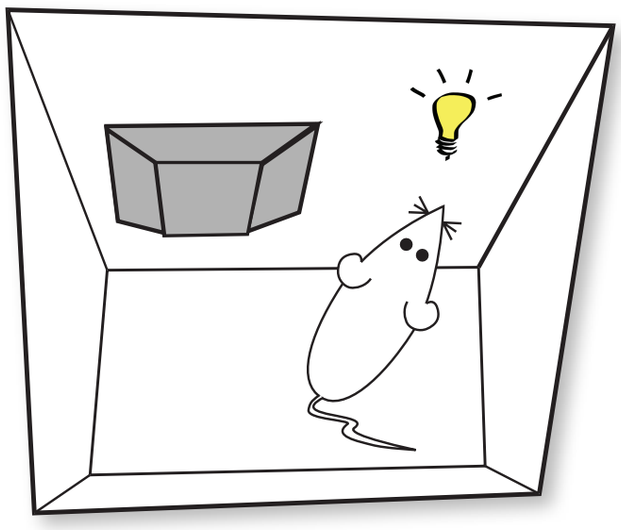


Huys et al. 2012 after Flagel et al. 2009, 2011

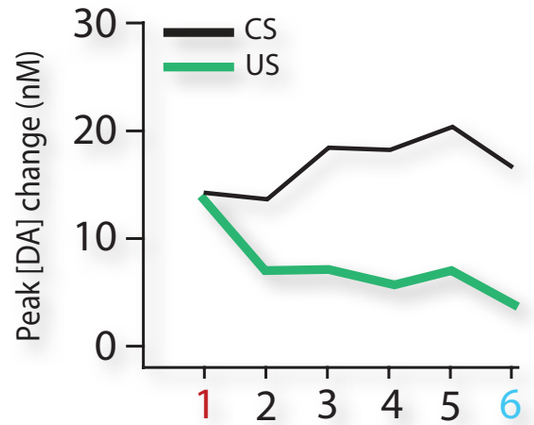
Addictive Pavlovian values

Pavlovian state value acquisition

A

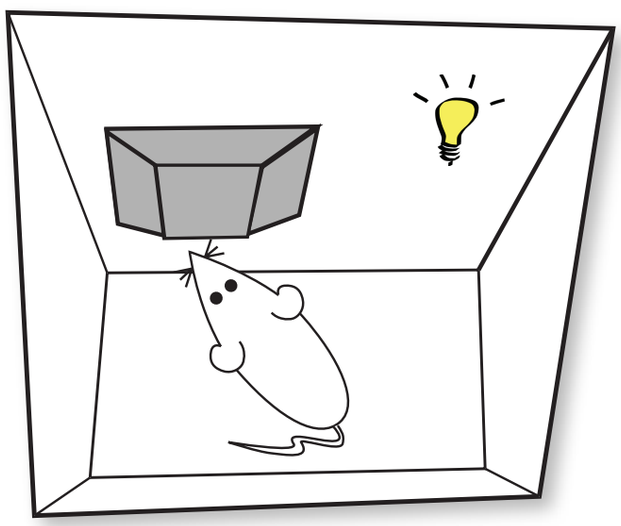


C

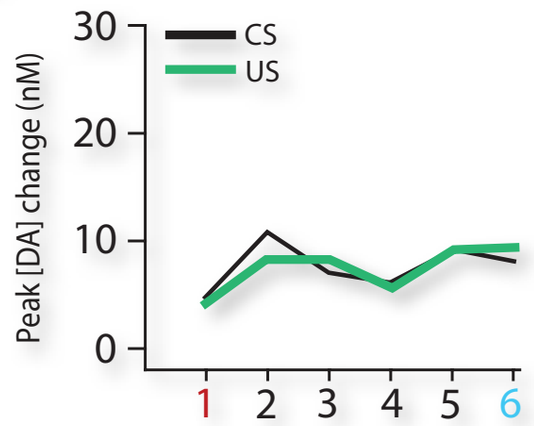


No Pavlovian state learning

D



F

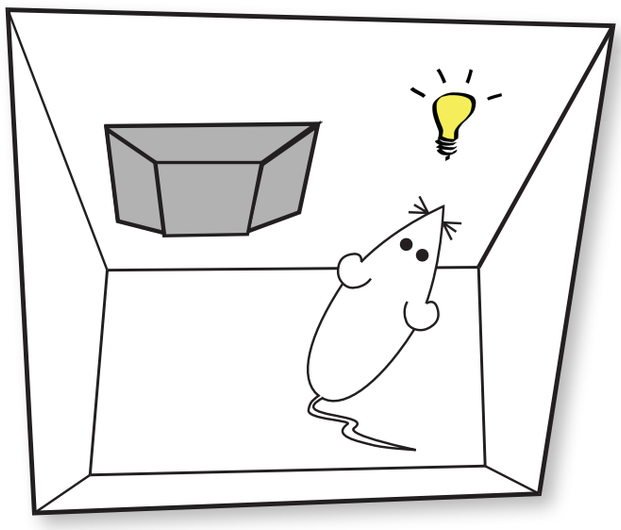


Huys et al. 2012 after Flagel et al. 2009, 2011

Addictive Pavlovian values

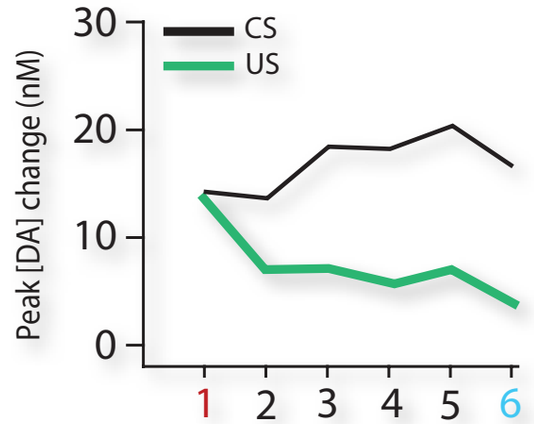
Pavlovian state value acquisition

A



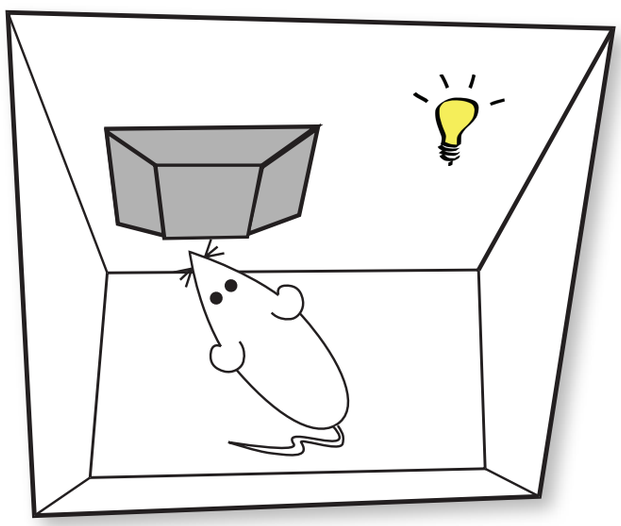
Impulsive
Amphetamine
DA antagonism
D2R<->phasic DA

C

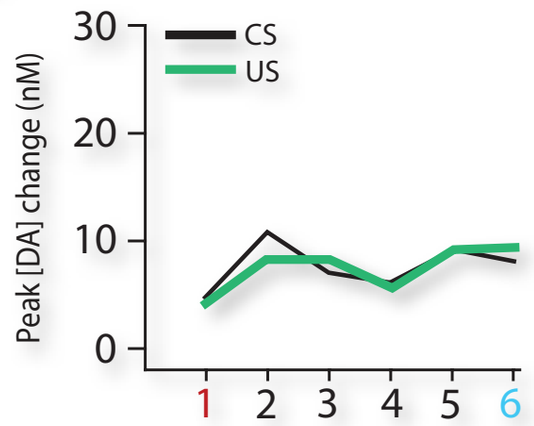


No Pavlovian state learning

D



F



Huys et al. 2012 after Flagel et al. 2009, 2011

‘Sign-tracking’

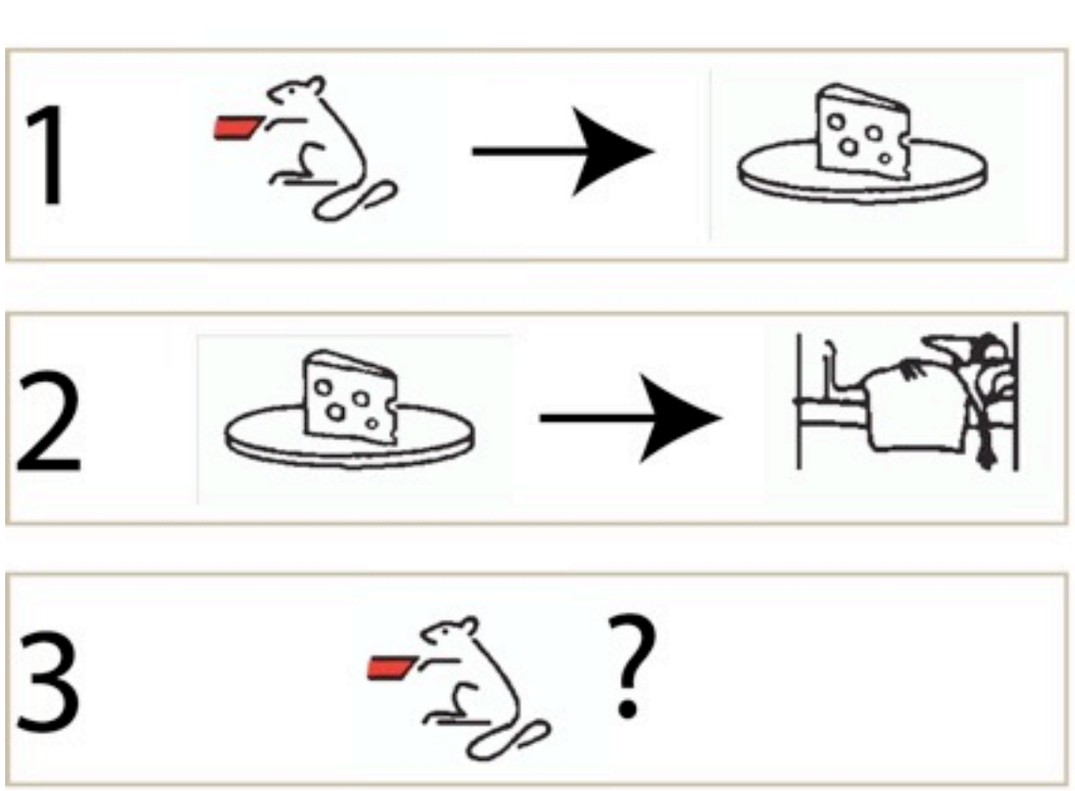
‘Sign-tracking’

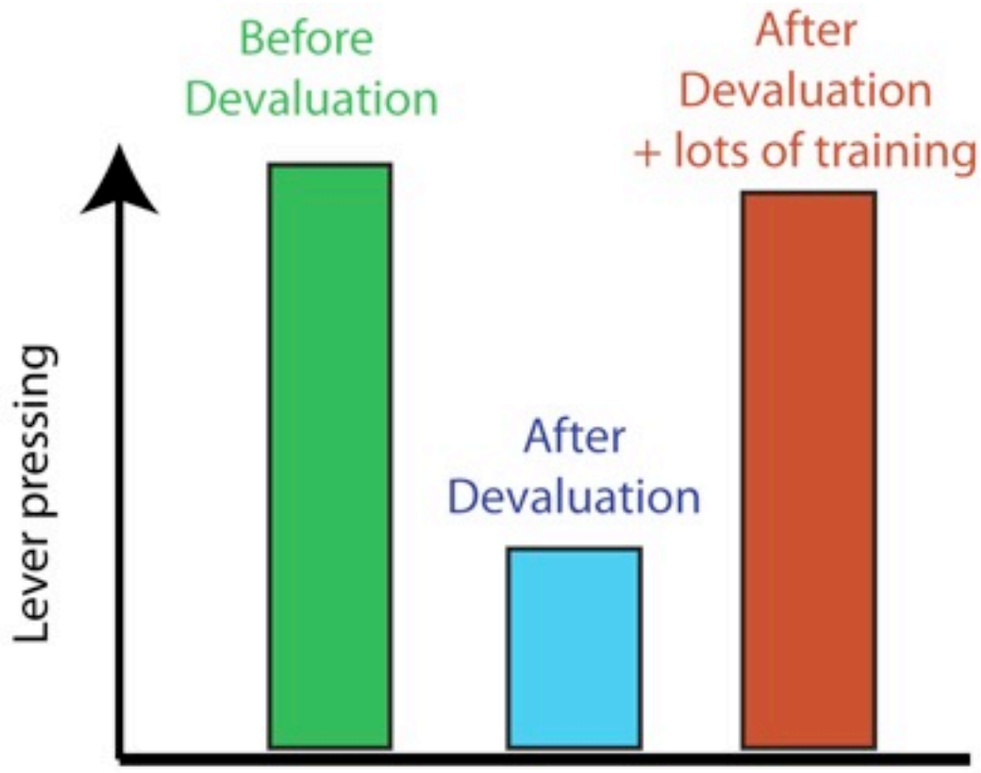
- Wiers et al. 2010: Attentional training reduces relapse
- Heinze et al. 2009: NAcc DBS reduces relapse & craving

Depression: aversive stimulus values

- dot probe task
- emotional stroop
- recollection bias

Devaluation





Goal-directed vs. habitual behaviour
mix and match

Habits

- Instrumental stimulus-response tendencies
- Lever pressing: not innate
- SARSA:

$$\mathcal{Q}(s_t, a_t) \leftarrow \mathcal{Q}(s_t, a_t) + \alpha[r_t + \gamma \mathcal{Q}(s_{t+1}, a_{t+1}) - \mathcal{Q}(s_t, a_t)]$$

- any response
- slow acquisition
- slow ‘extinction’ if reward changes

Decision systems in parallel

- Devaluation
  - Train briefly / overtrain
  - Devalue outcome
  - Immediate behaviour?

early: g-d

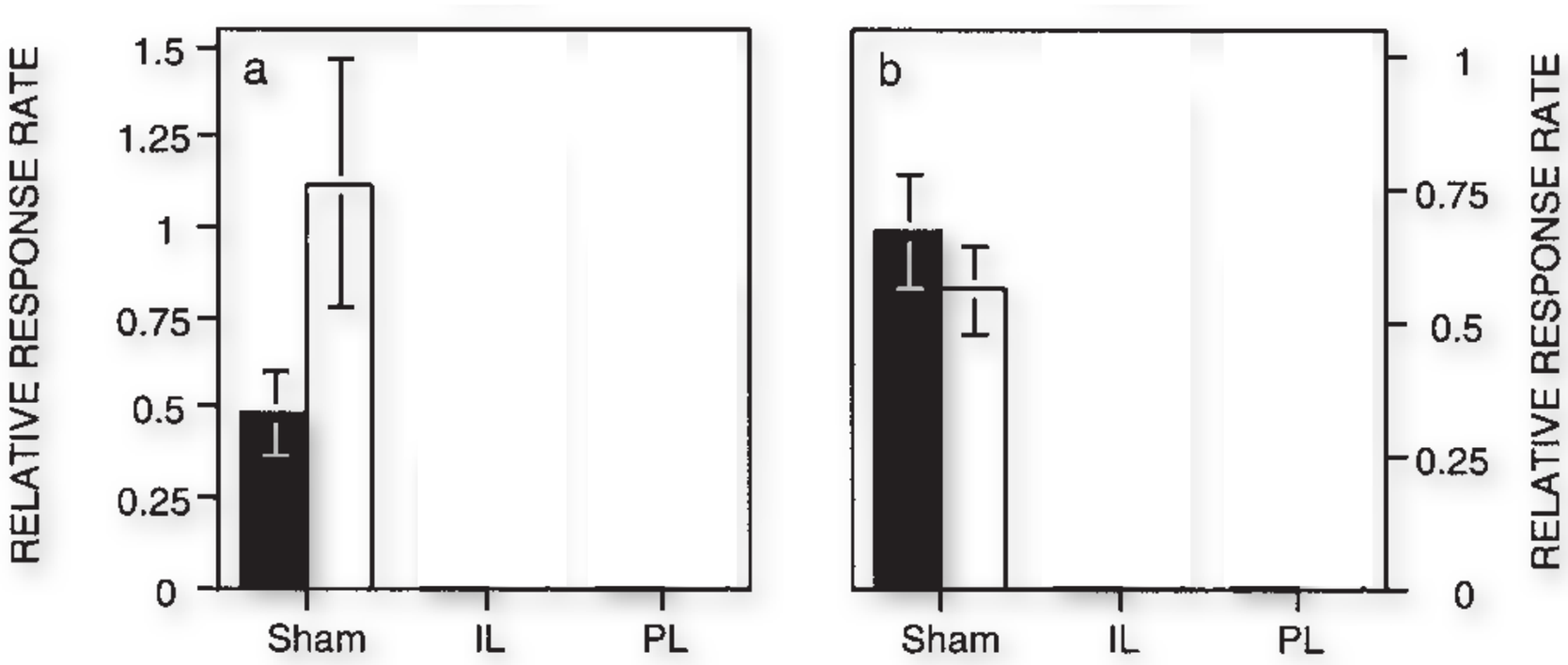
late: habits

Decision systems in parallel

- Devaluation
  - Train briefly / overtrain
  - Devalue outcome
  - Immediate behaviour?

early: g-d

late: habits



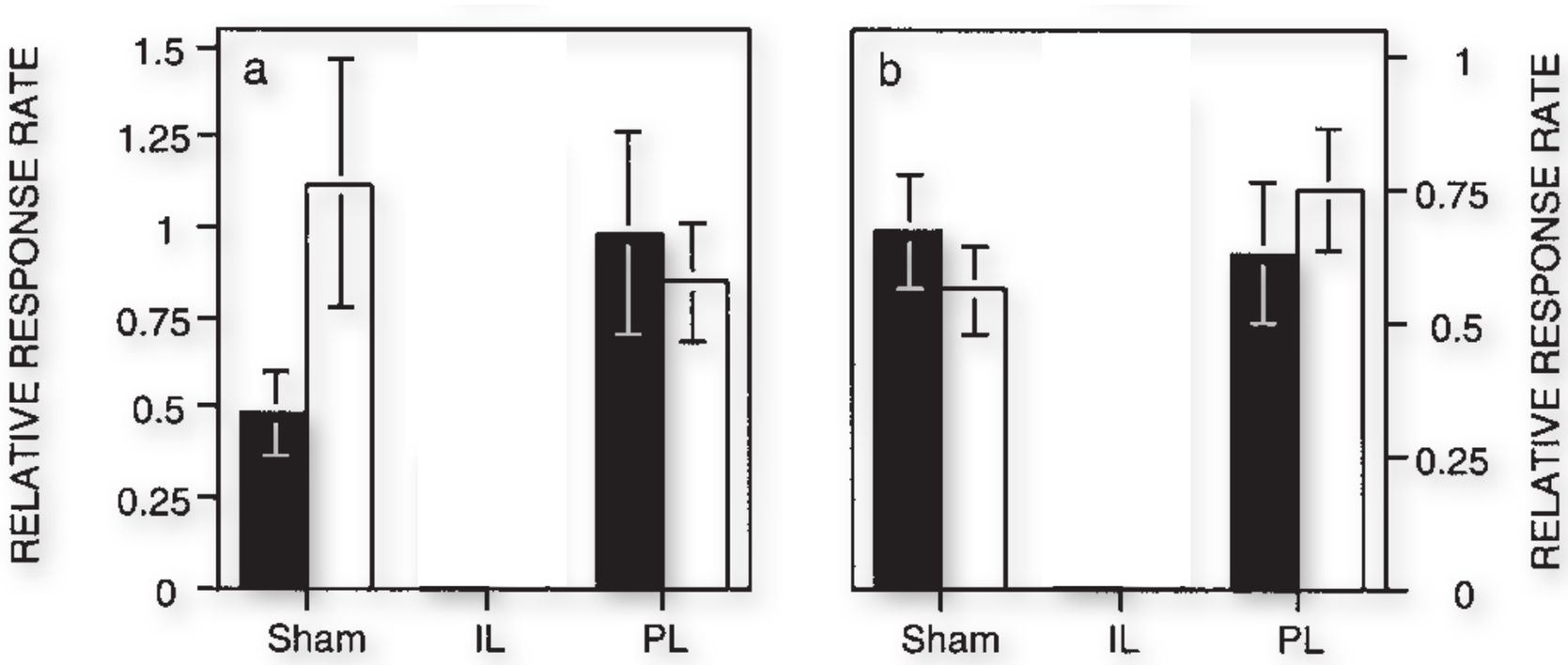
Killcross and Coutureau 2003

Decision systems in parallel

- Devaluation
  - Train briefly / overtrain
  - Devalue outcome
  - Immediate behaviour?

early: g-d

late: habits



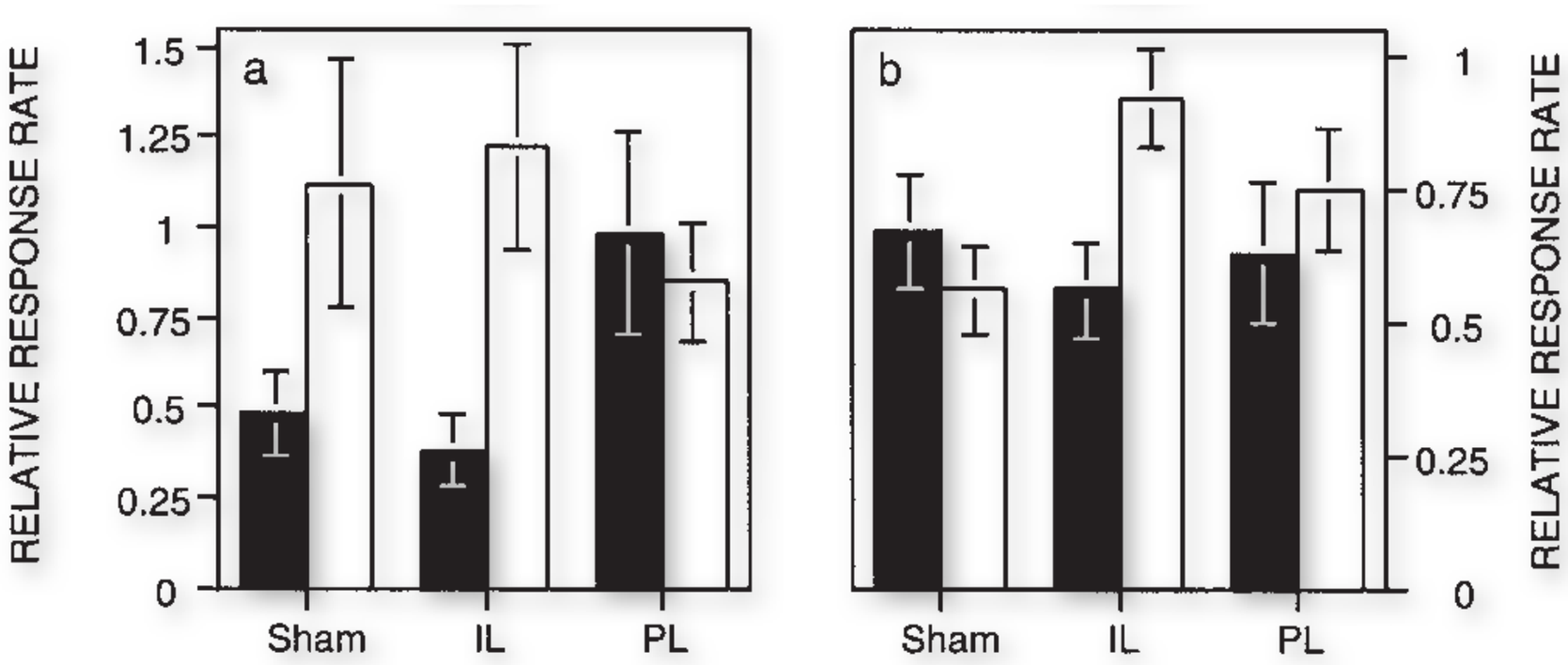
Killcross and Coutureau 2003

Decision systems in parallel

- Devaluation
  - Train briefly / overtrain
  - Devalue outcome
  - Immediate behaviour?

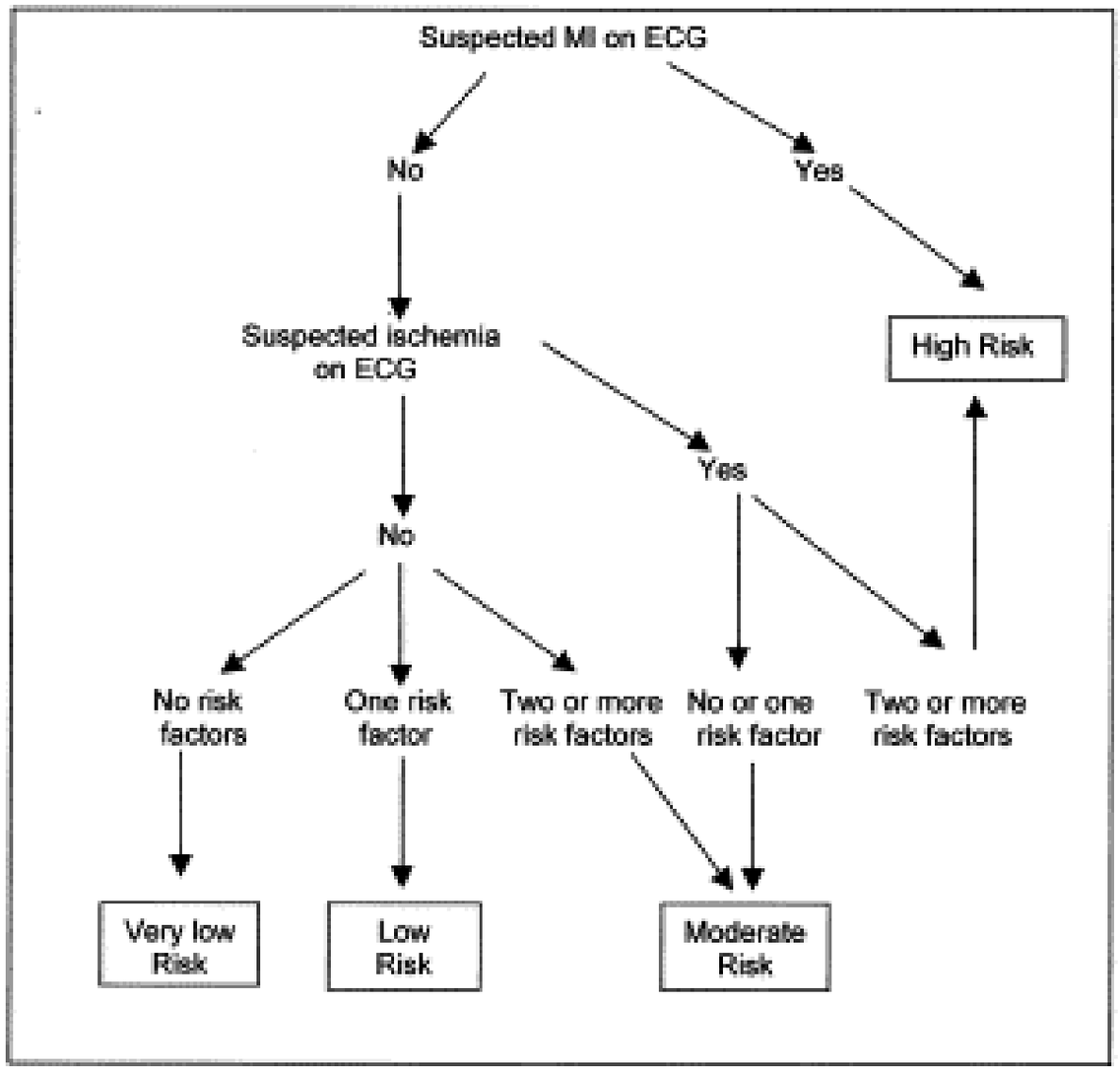
early: g-d

late: habits



Killcross and Coutureau 2003

Simple is better at times: doctors

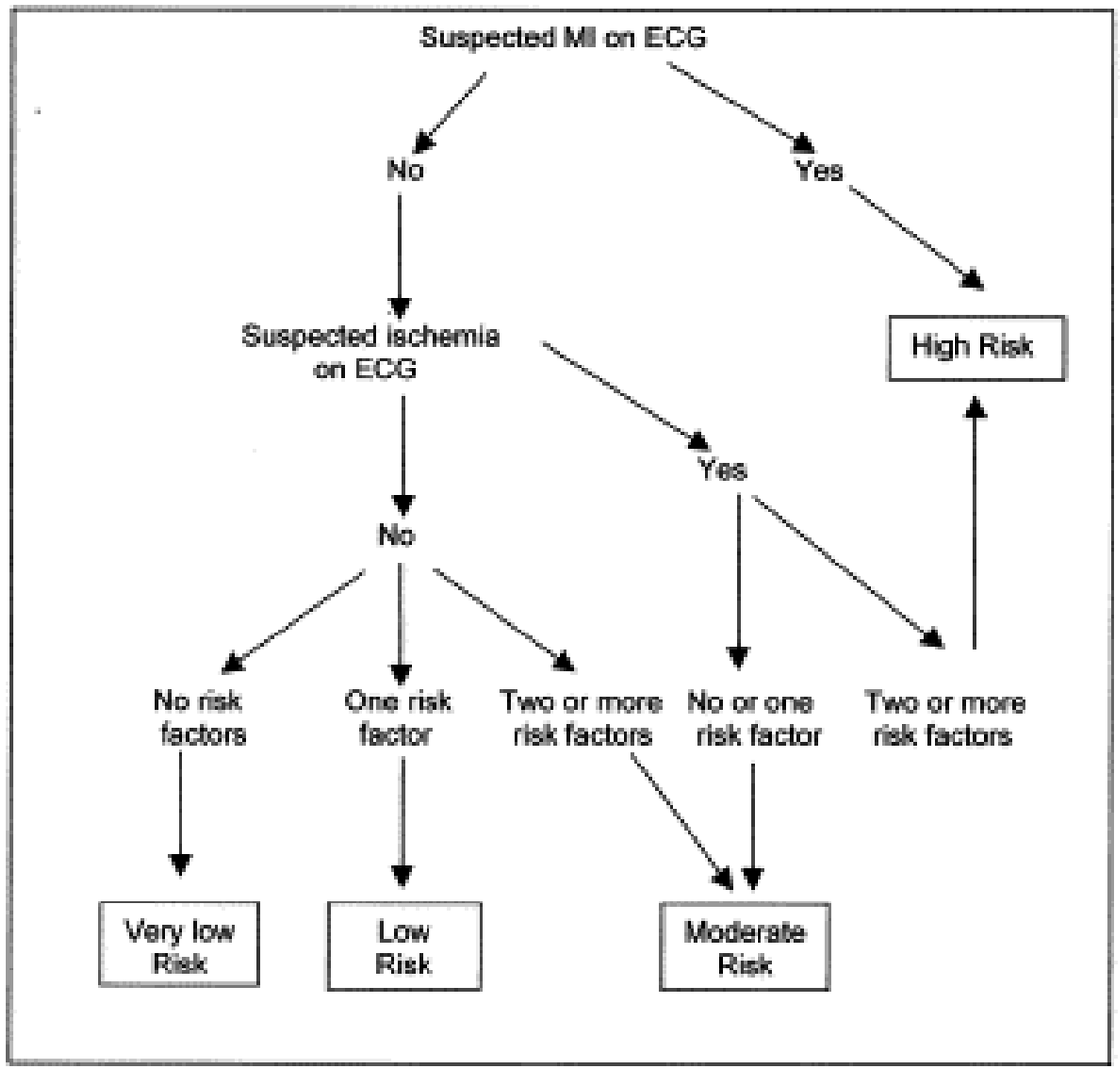


20 cases for which truth known

Cardiologists
General physicians
A&E physicians

Melly et al. 2002

Simple is better at times: doctors



20 cases for which truth known

Cardiologists
General physicians
A&E physicians

Physicians overly cautious, but still miss many -> complications

Melly et al. 2002

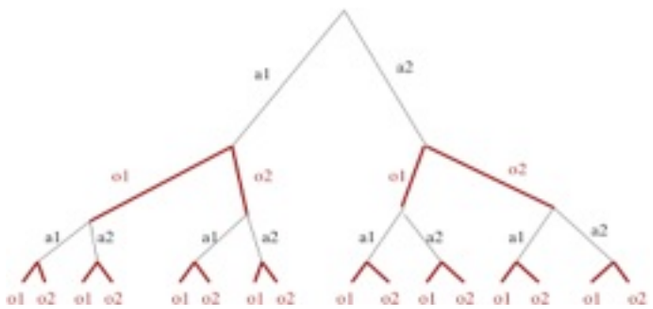
Dopamine and habits

- Also see phasic DA responses in instrumental learning

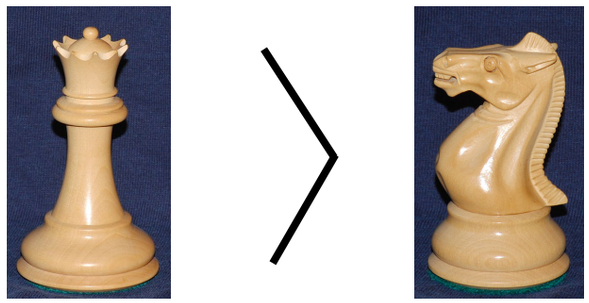
- DA seems to be involved in learning with prediction errors
  - Sum of prediction errors = "cached" value (Daw et al. 2005)

Many decision systems in parallel

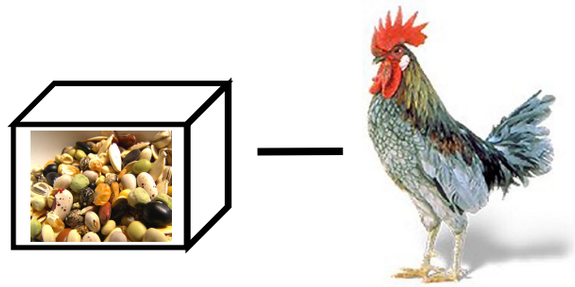
Goal-directed system
Tree search



Habit system
Experience average



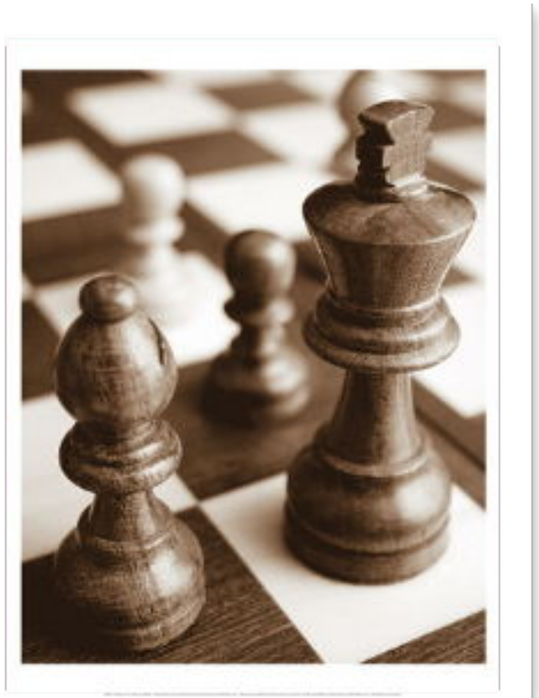
Innate system
Evolutionary strategy

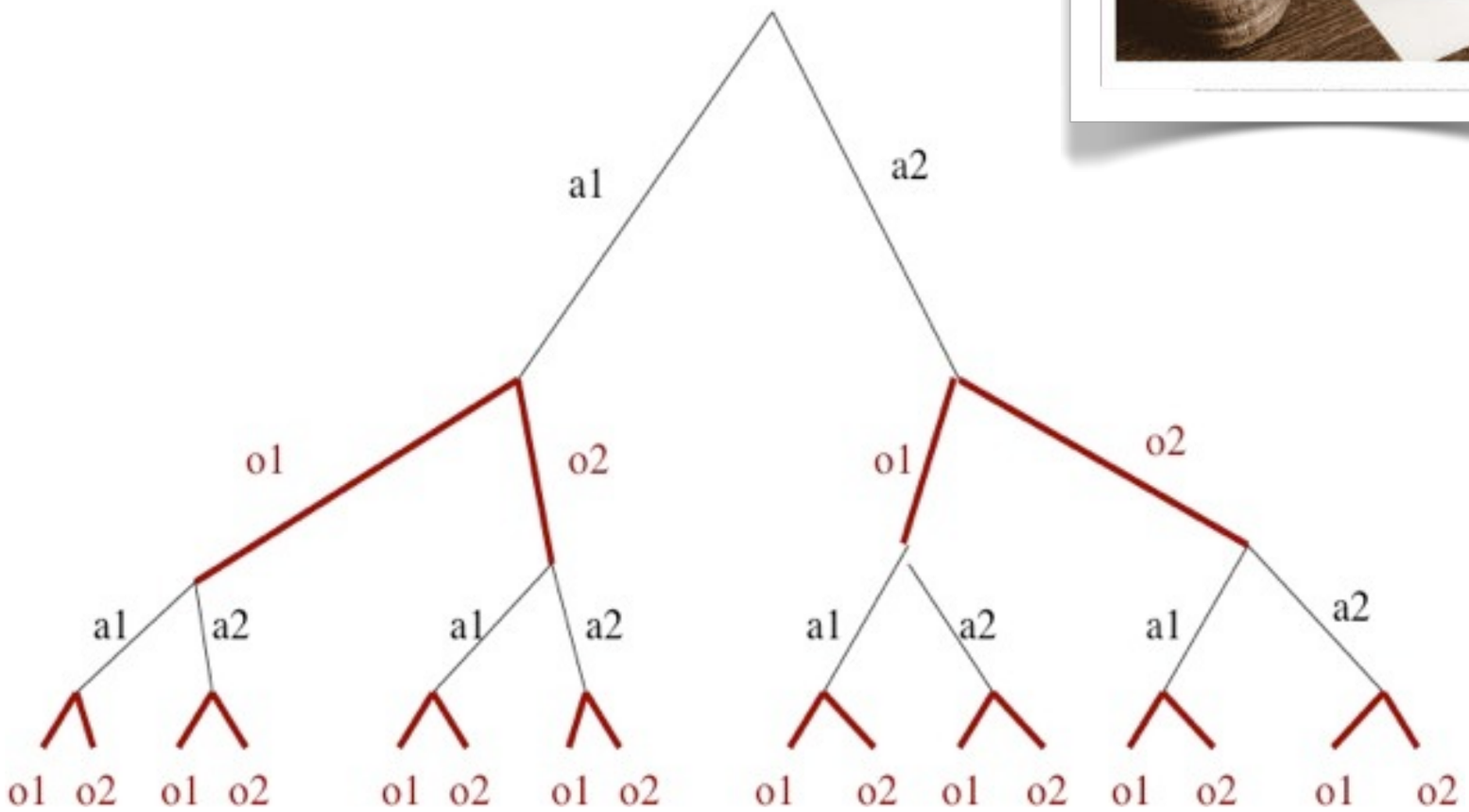


- behaviourally and neurobiologically distinct and identifiable
  - How to arbitrate? according to advantages and disadvantages?
- interactions
  - Pavlovian pruning of decision trees
  - model-based teaching of habits

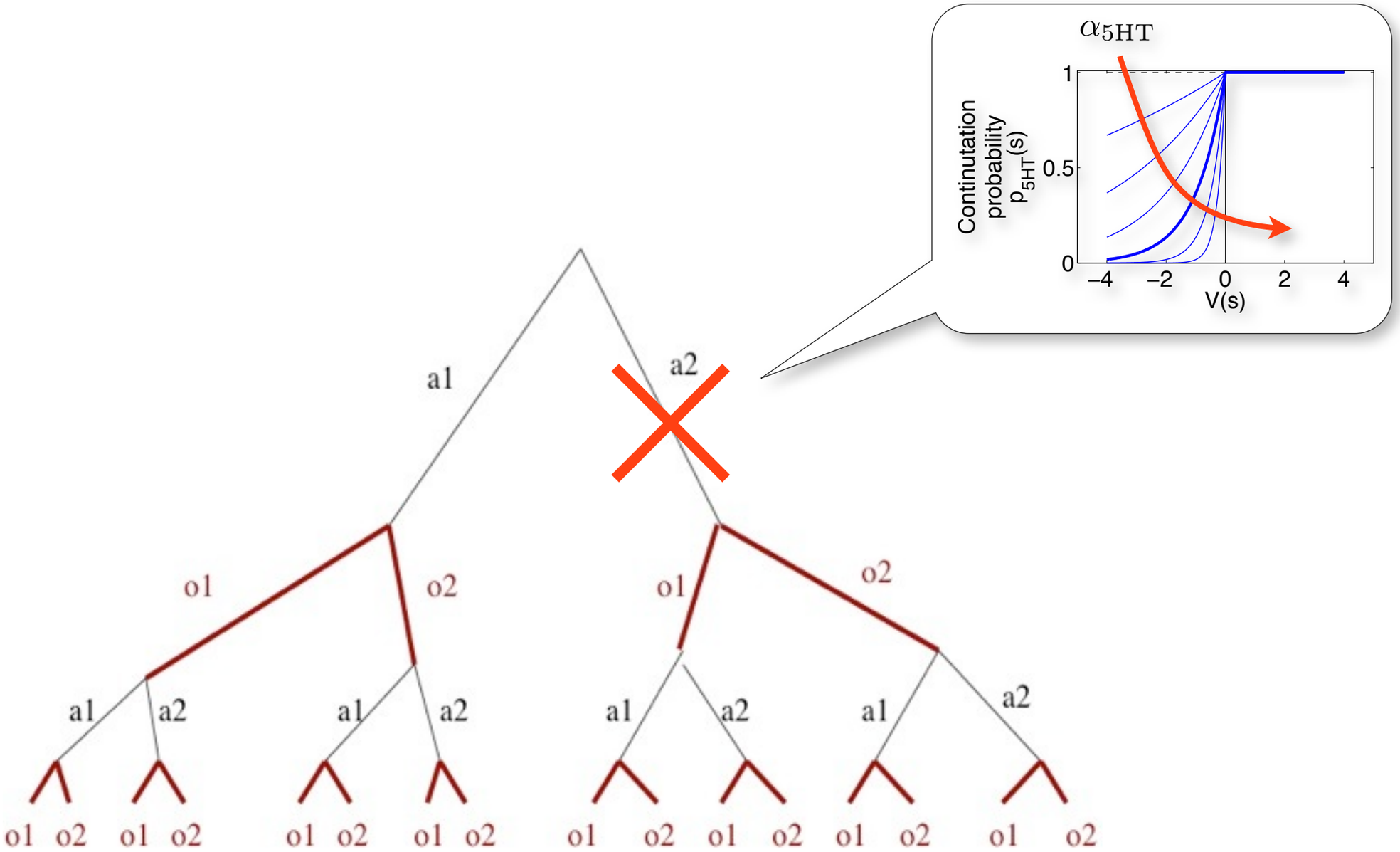
Goal-directed decision making

- Each move: 30 odd options
- 30^{40}?
- Legal boards ~10^{123}
- Can't just do full tree search.

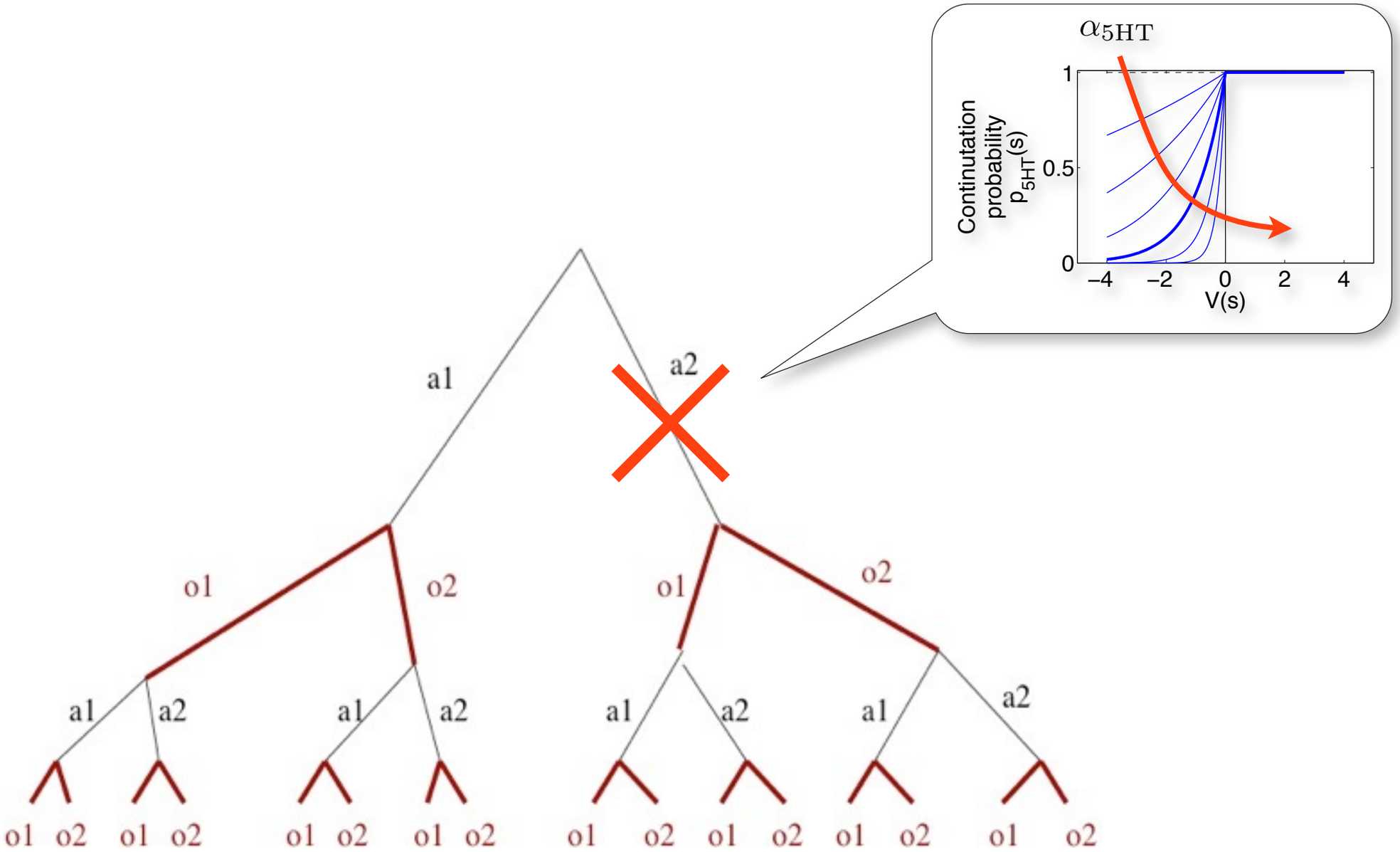




Pruning a decision tree



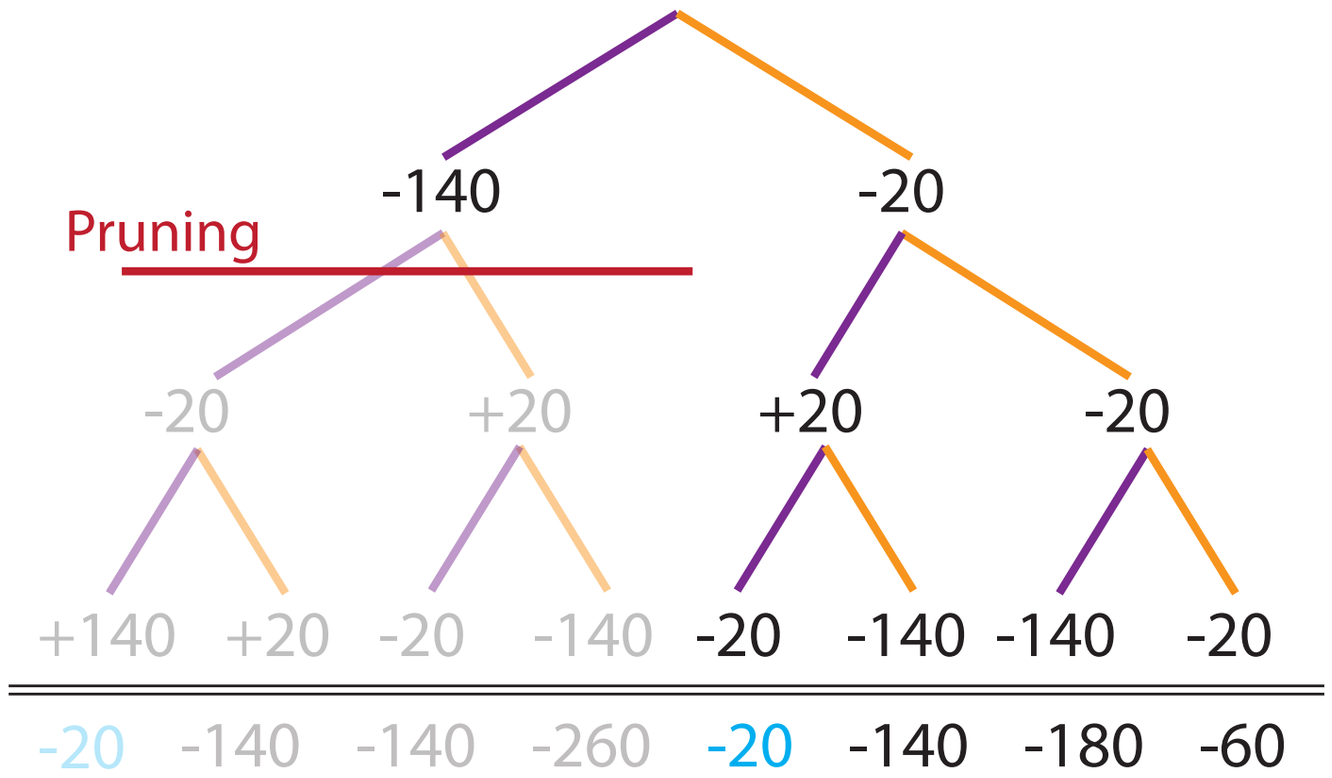
Pruning a decision tree



Don’t go there… Don’t think it either?

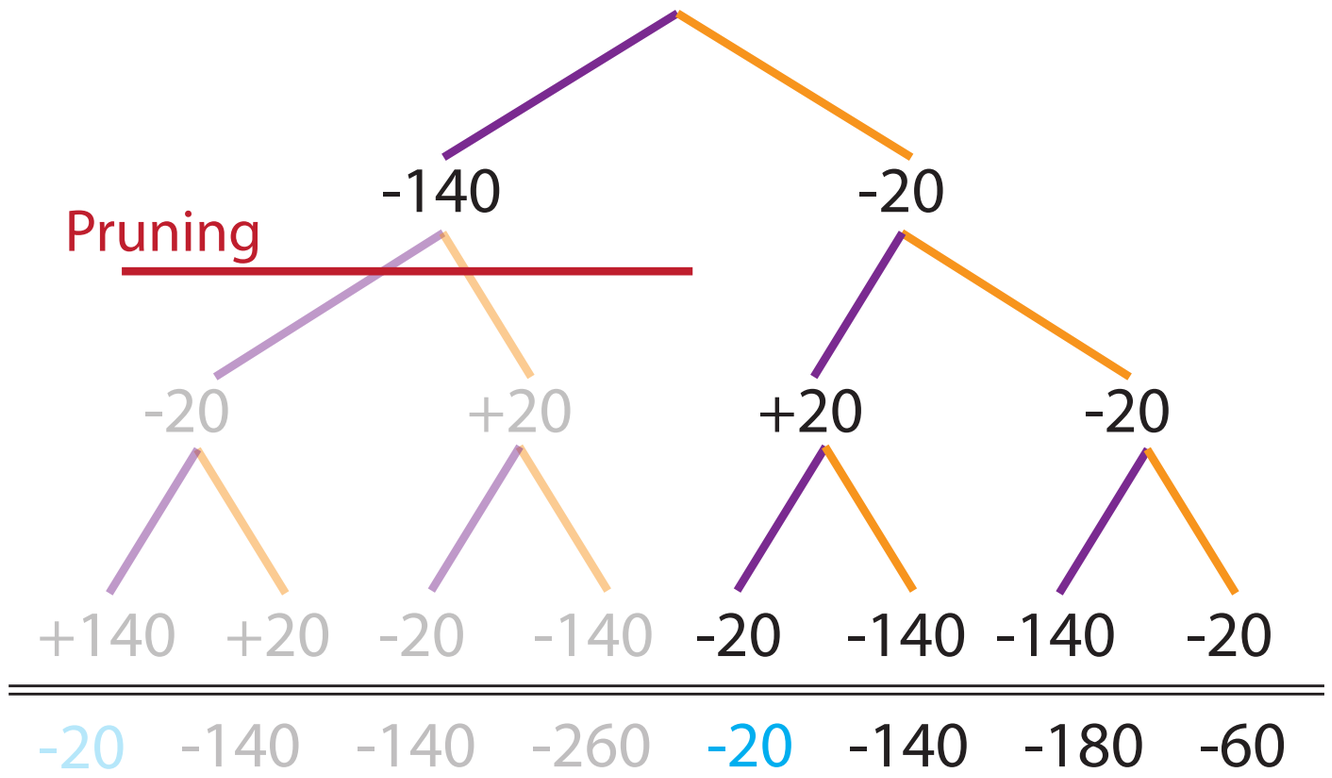
Pavlovian pruning

- Optimality
  - conserve guarantees
  - difficult & computationally expensive
- Approximate
  - trade optimality for speed



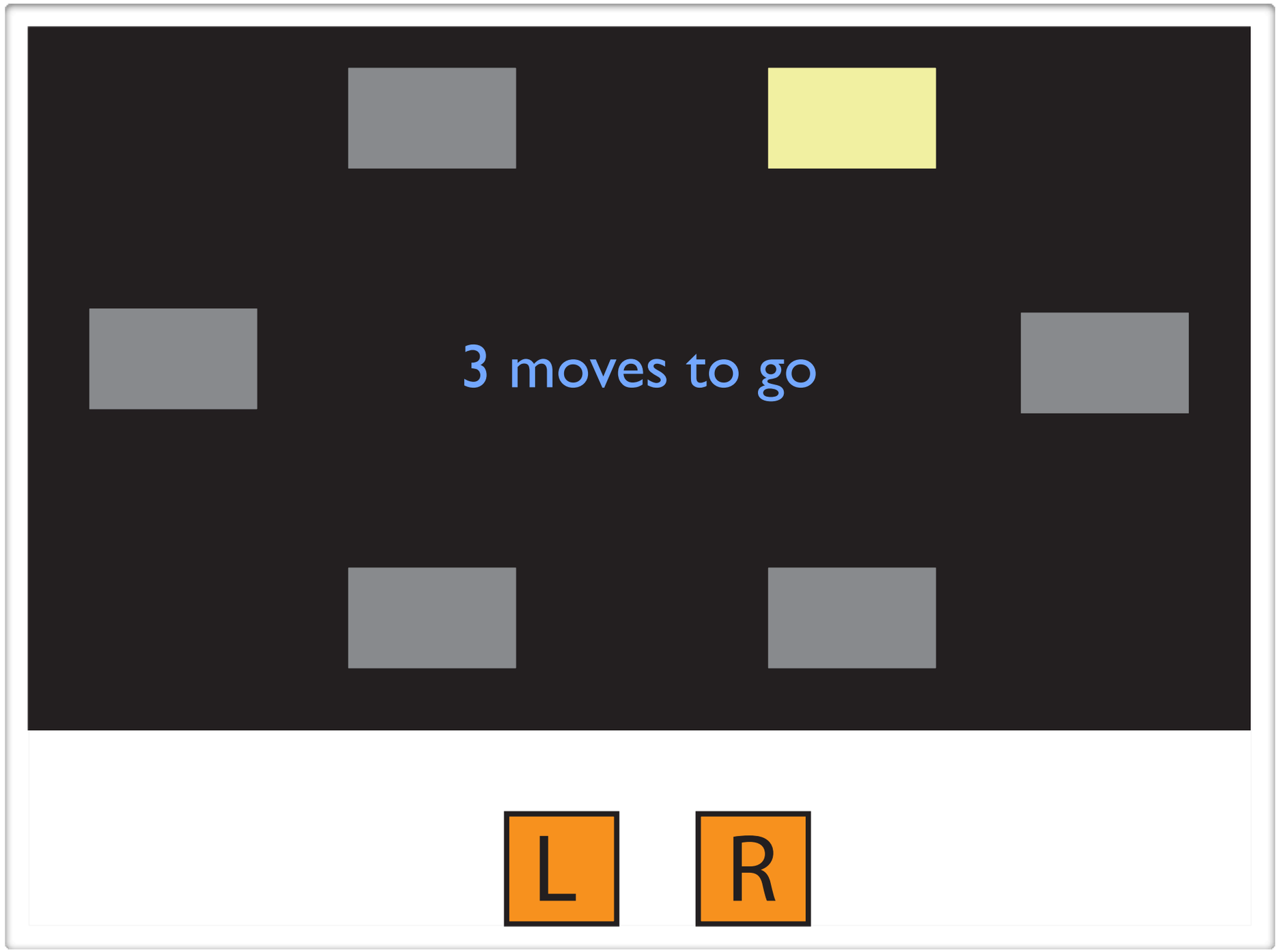
Pavlovian pruning

- Optimality
  - conserve guarantees
  - difficult & computationally expensive
- Approximate
  - trade optimality for speed

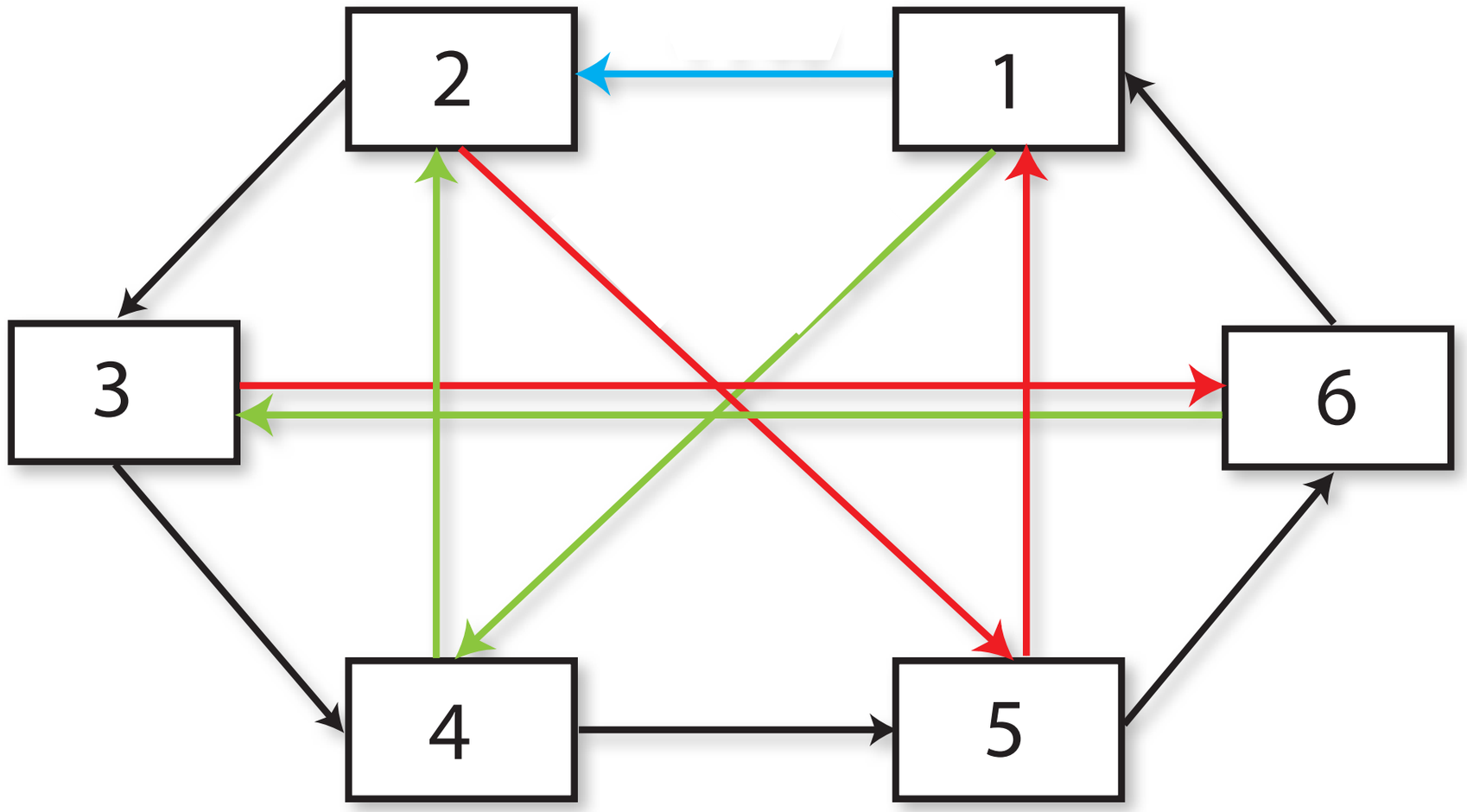


- Pavlovian
  - reflexively prune on encountering a punishment

Psychochess



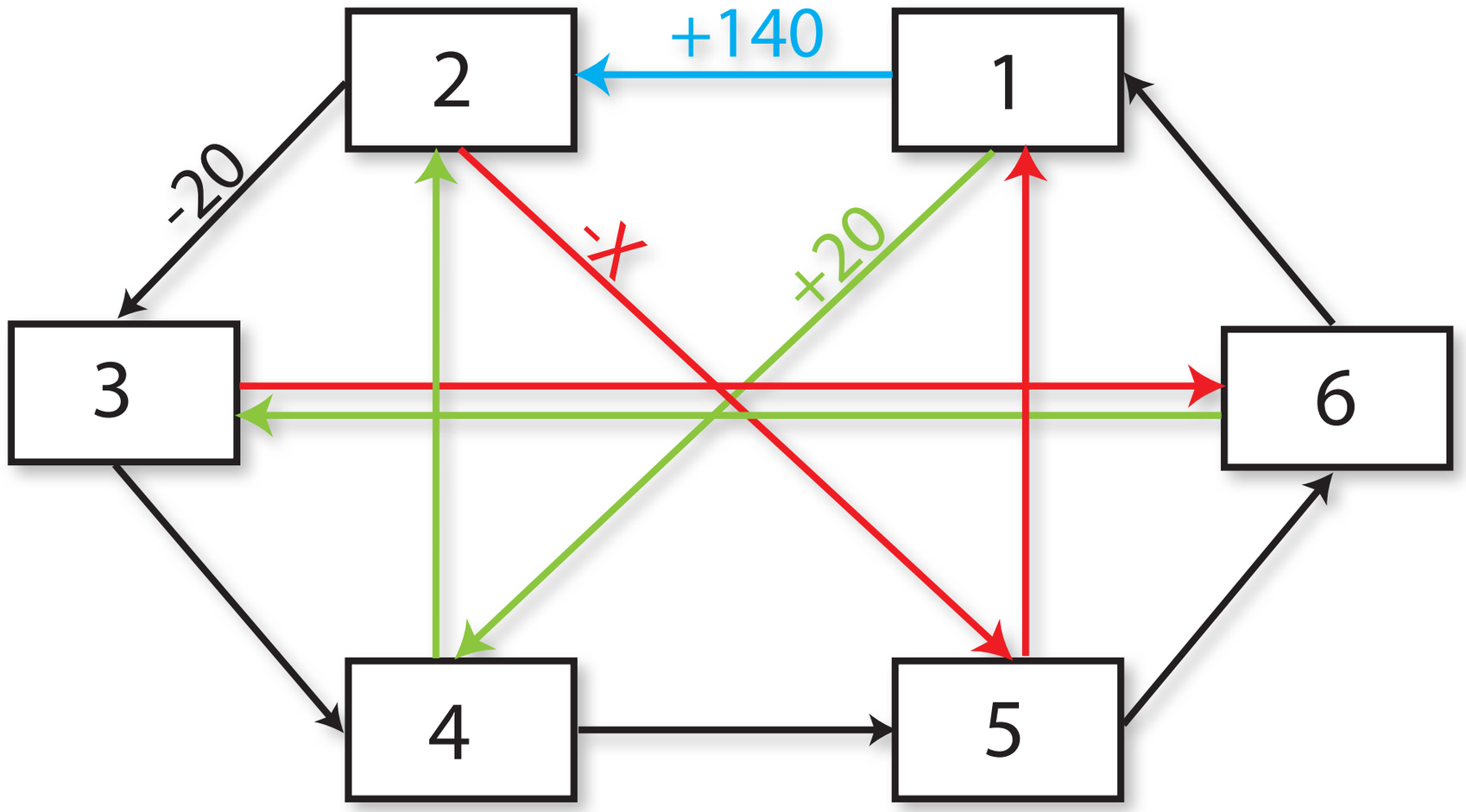
A poor experimental psychologist's version of chess



-X = -140

A tree search task

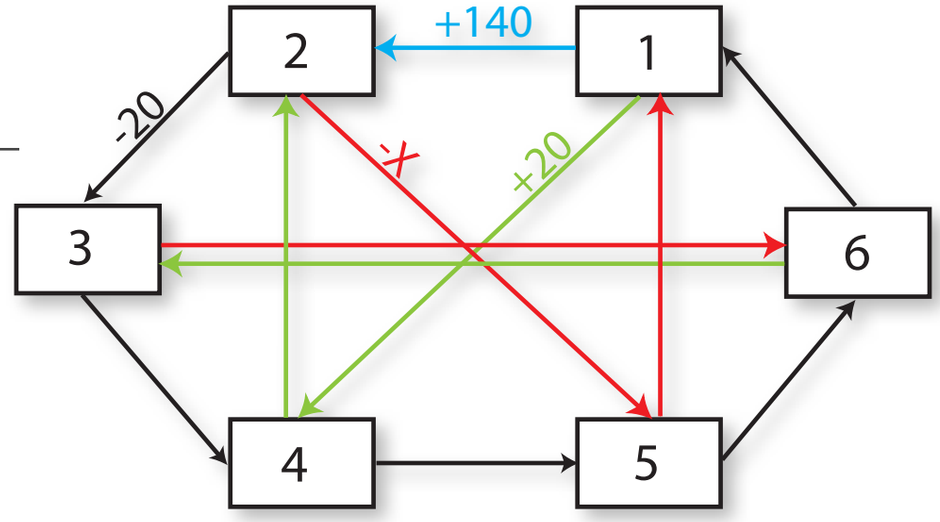
A poor experimental psychologist's version of chess

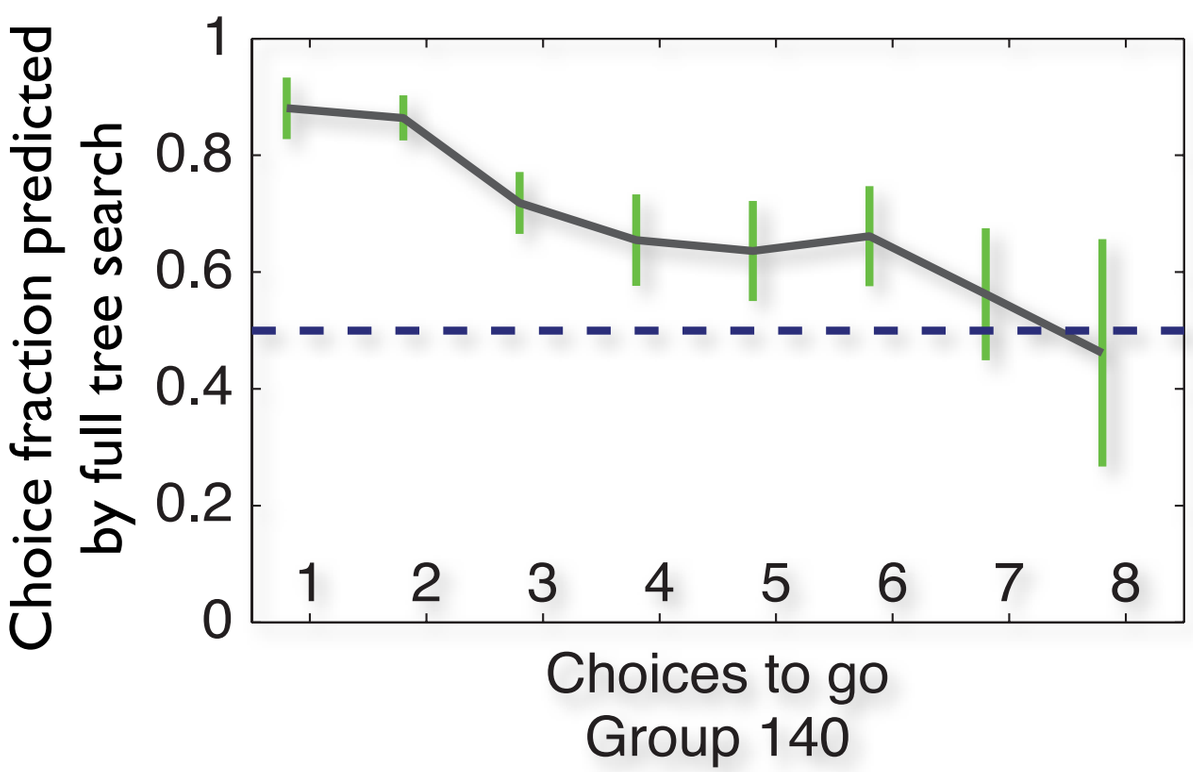


-X = -140

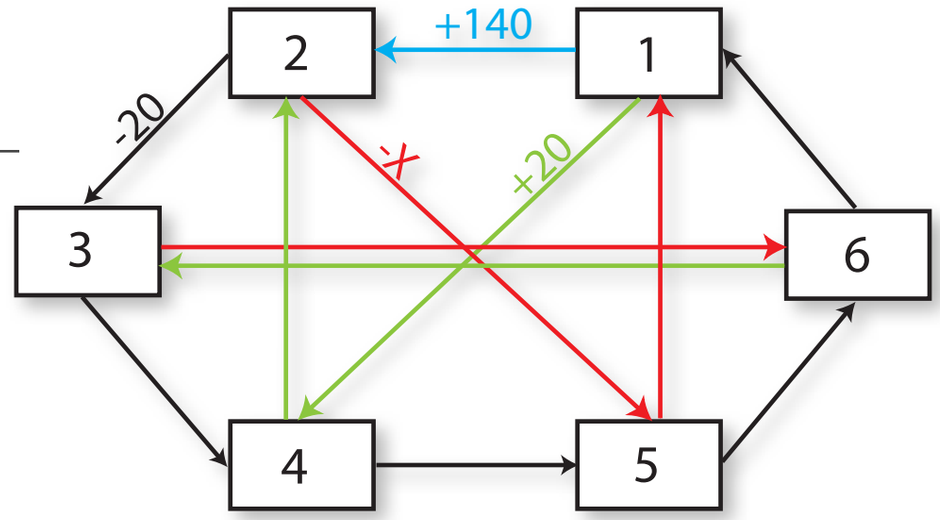
A tree search task

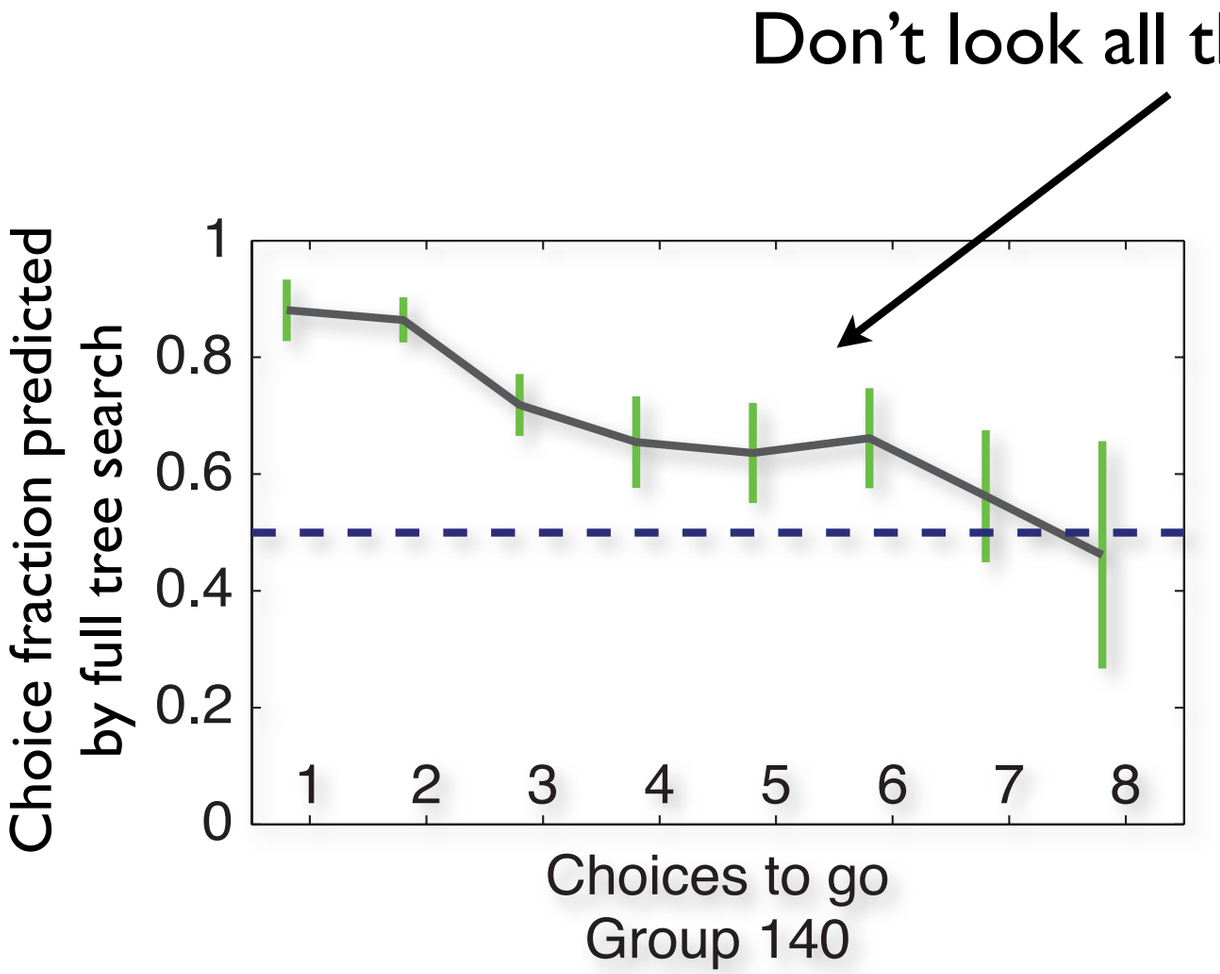
Model 1: full lookahead



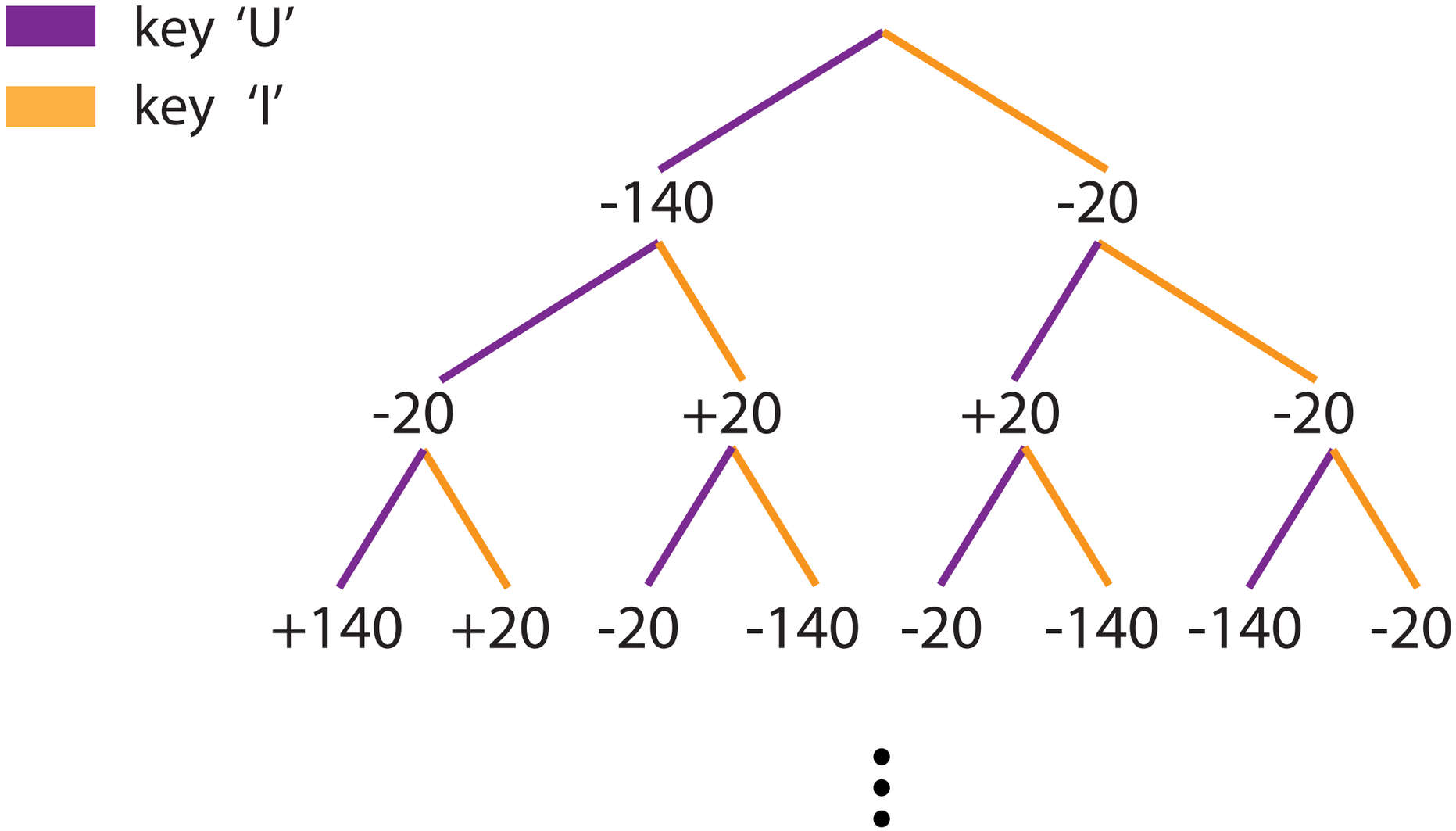


Model 1: full lookahead





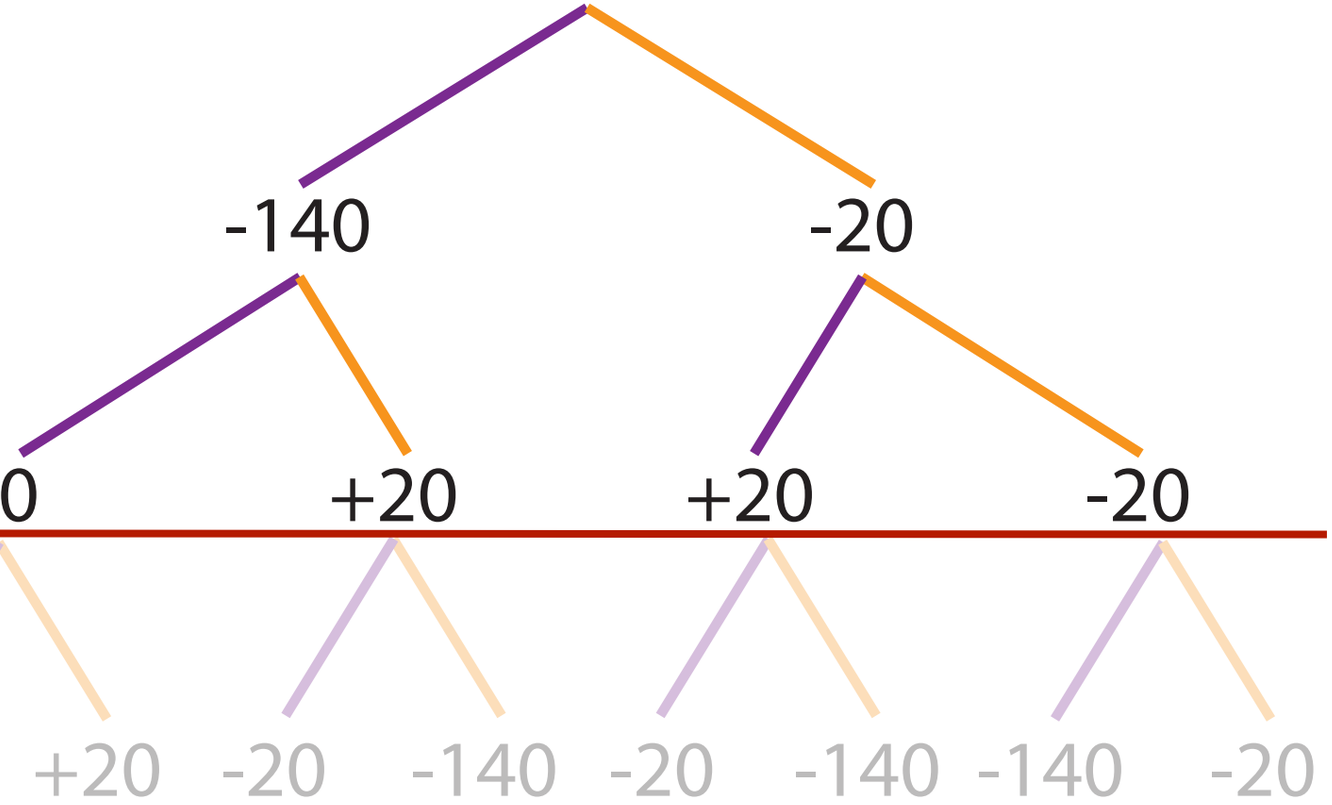
Pruning



Pruning

key ‘U’

key ‘I’



Flat pruning

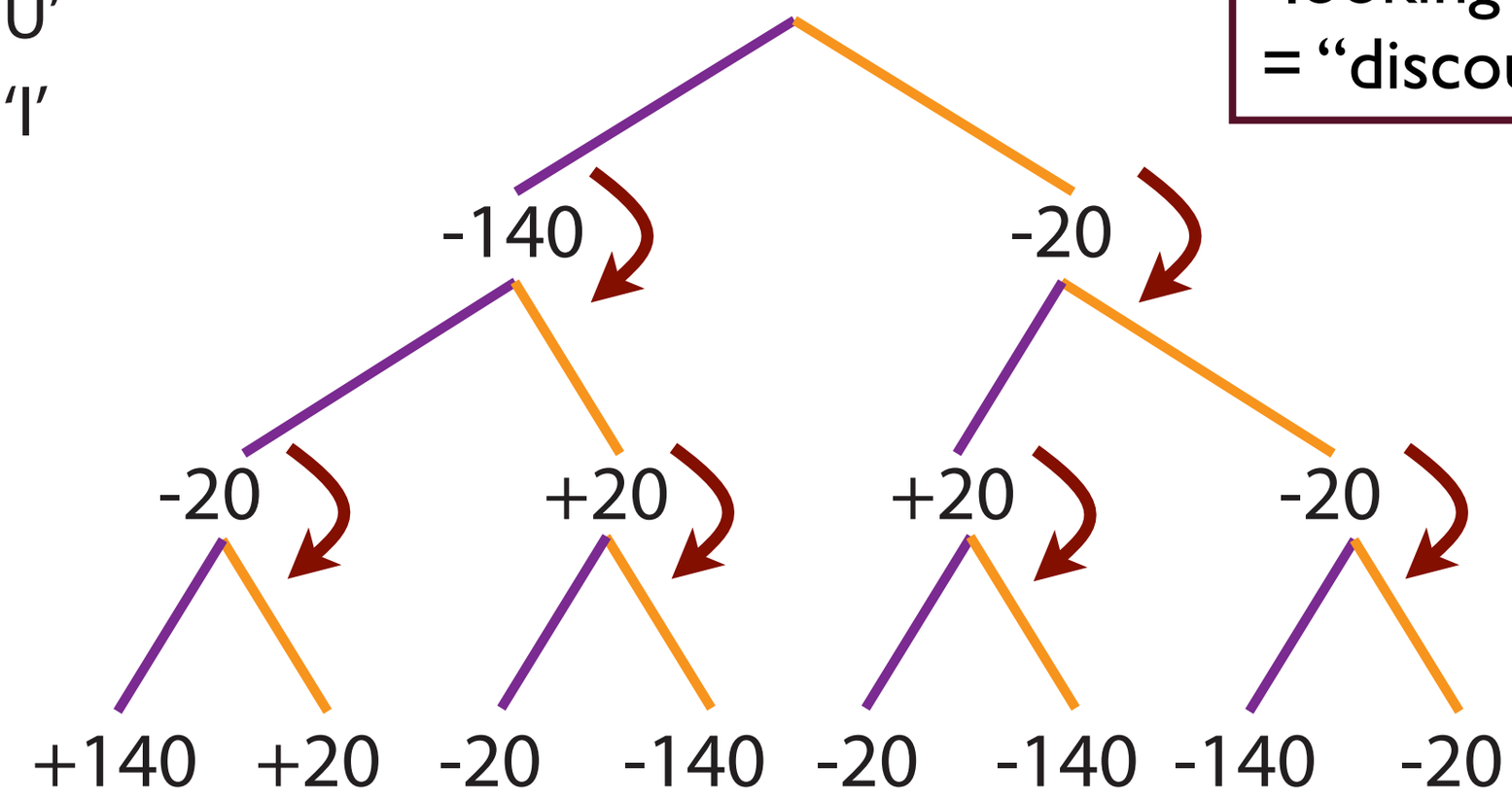
Pruning

key ‘U’

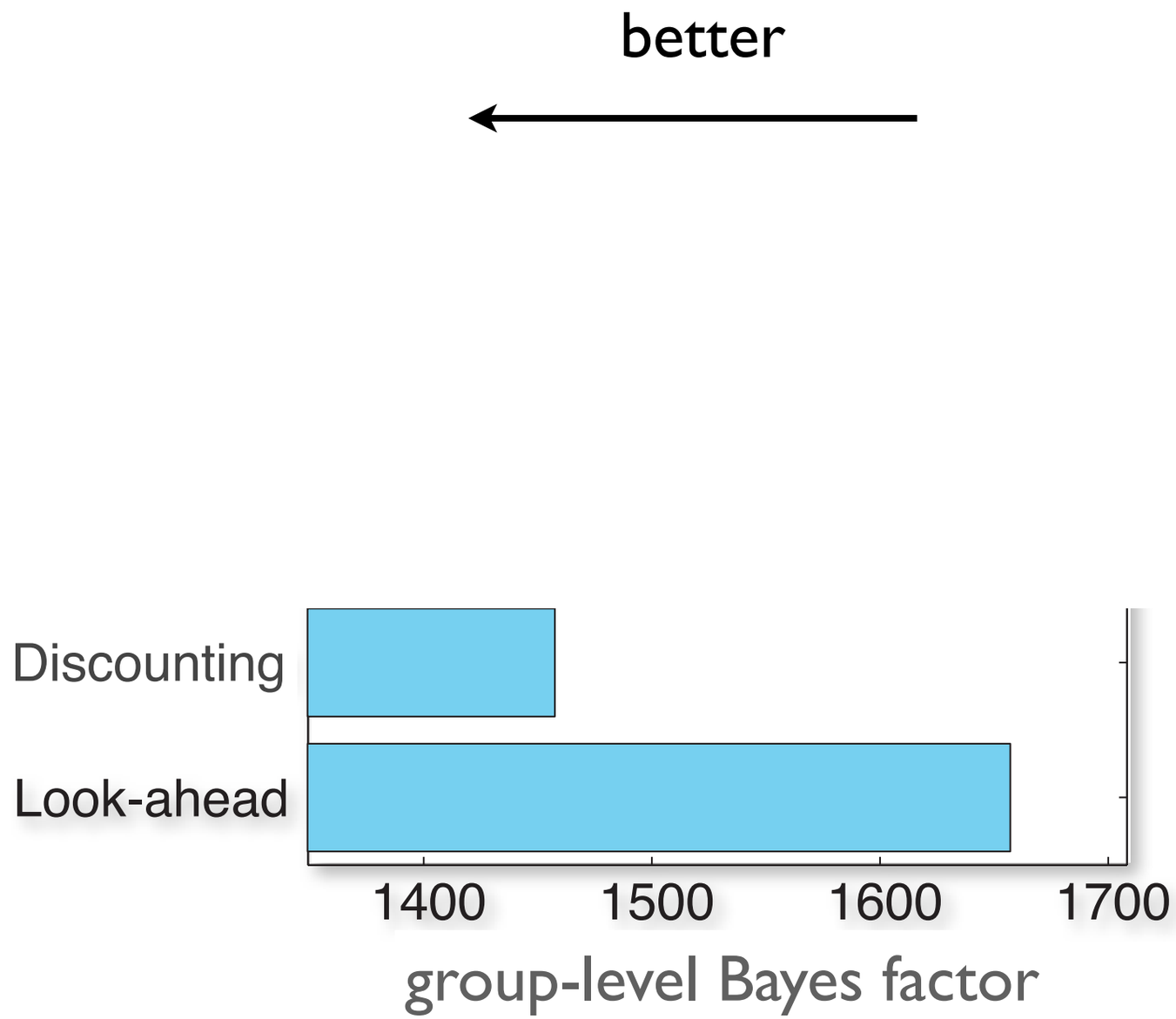
key ‘I’

Probability of looking ahead = “discounting”

Flat pruning



People don’t look to the end...

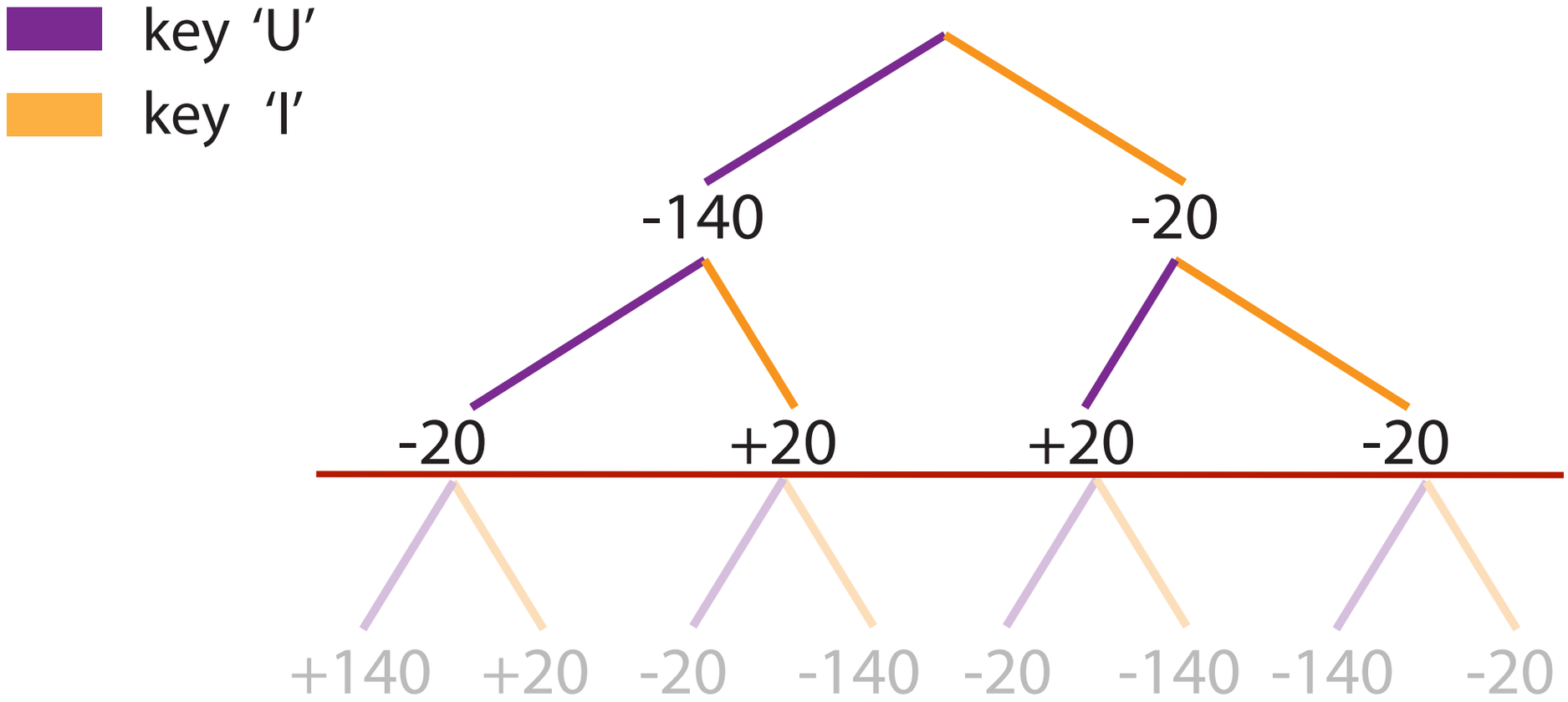


Huys et al. (2011) Submitted

Chess

Selective pruning!
Pavlovian pruning???

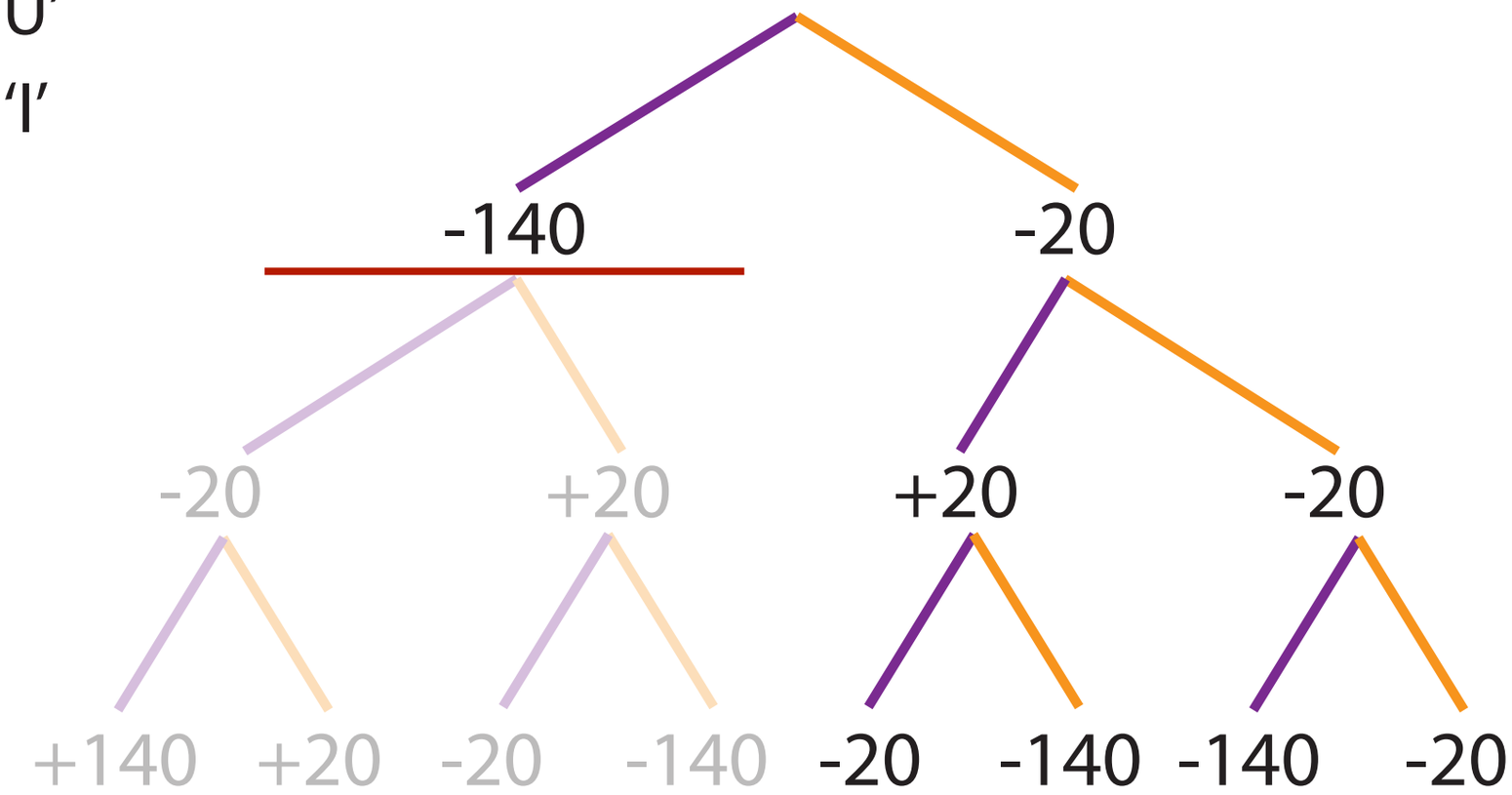
Pruning



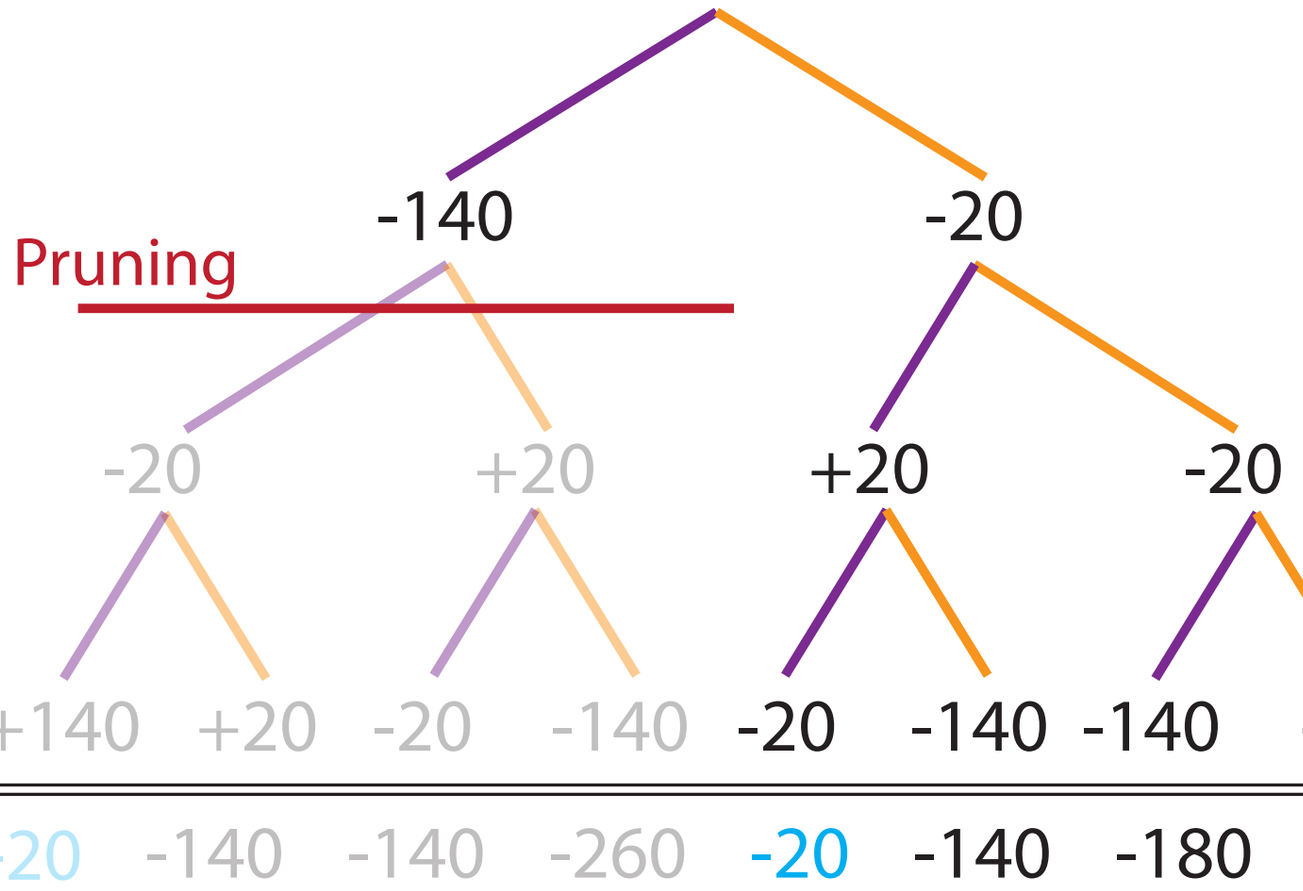
Pruning

key ‘U’

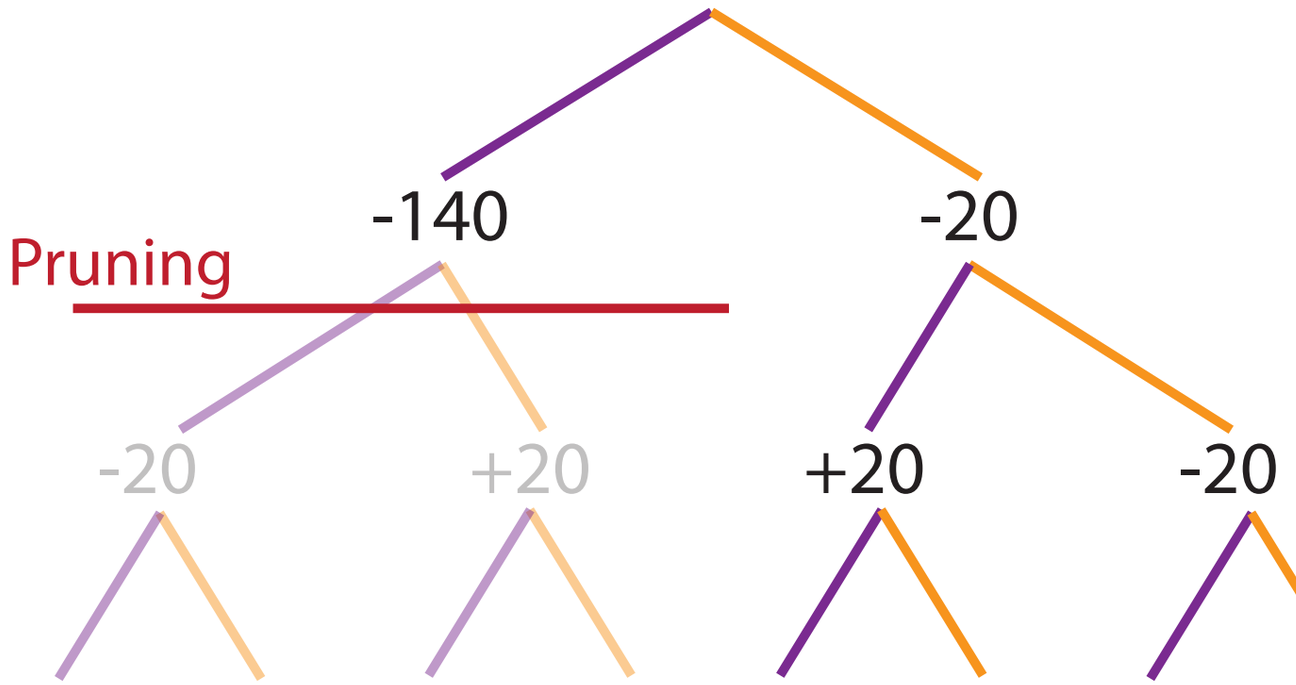
key ‘I’



Pavlovian?

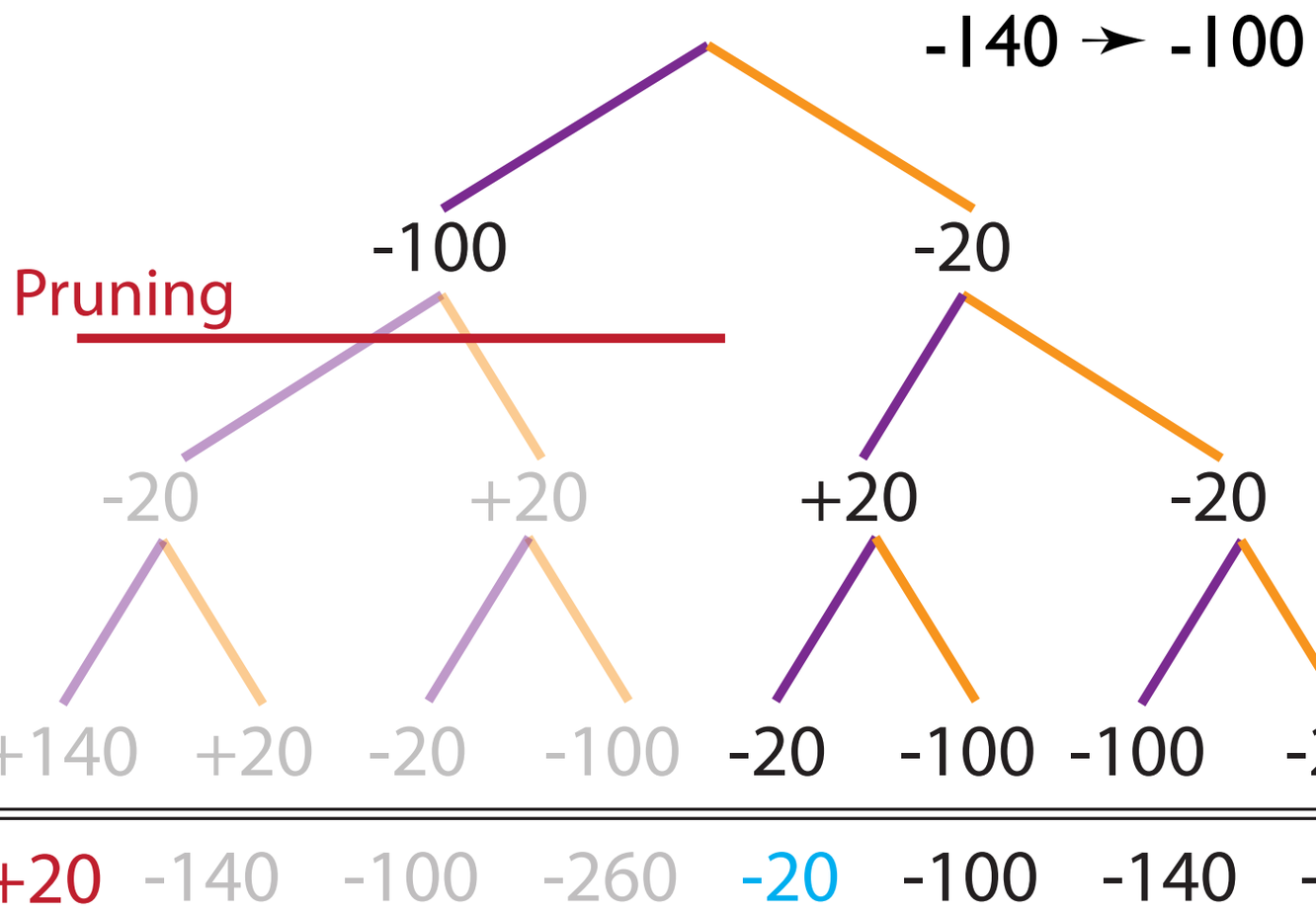

-140
-20
Pruning
-20
+20
+20
-20
+140
+20
-20
-140
-20
-140
-140
-20
-20
-140
-140
-260
-20
-140
-180
-60

Pavlovian?

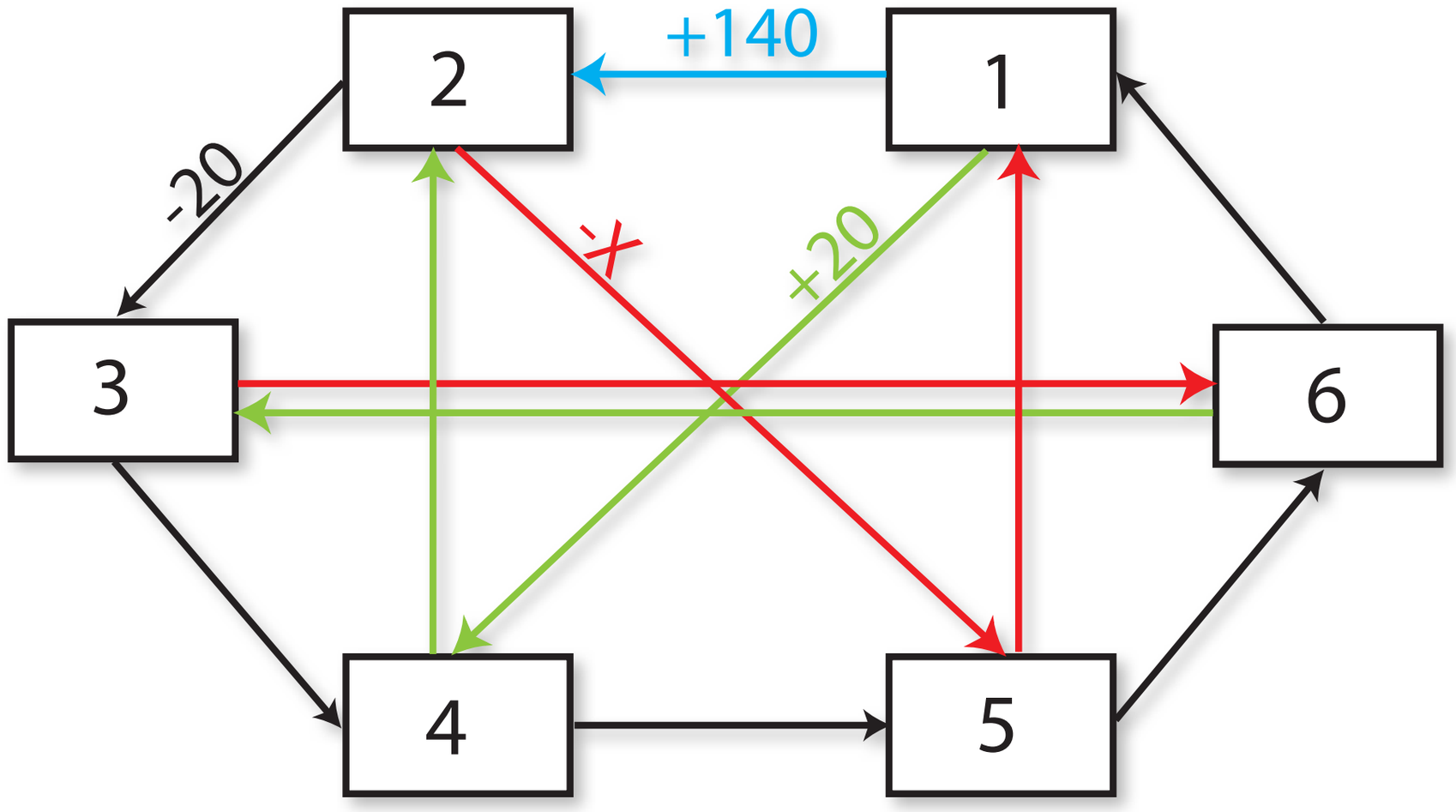


-20	-140	-140	-260	-20	-140	-180	-60

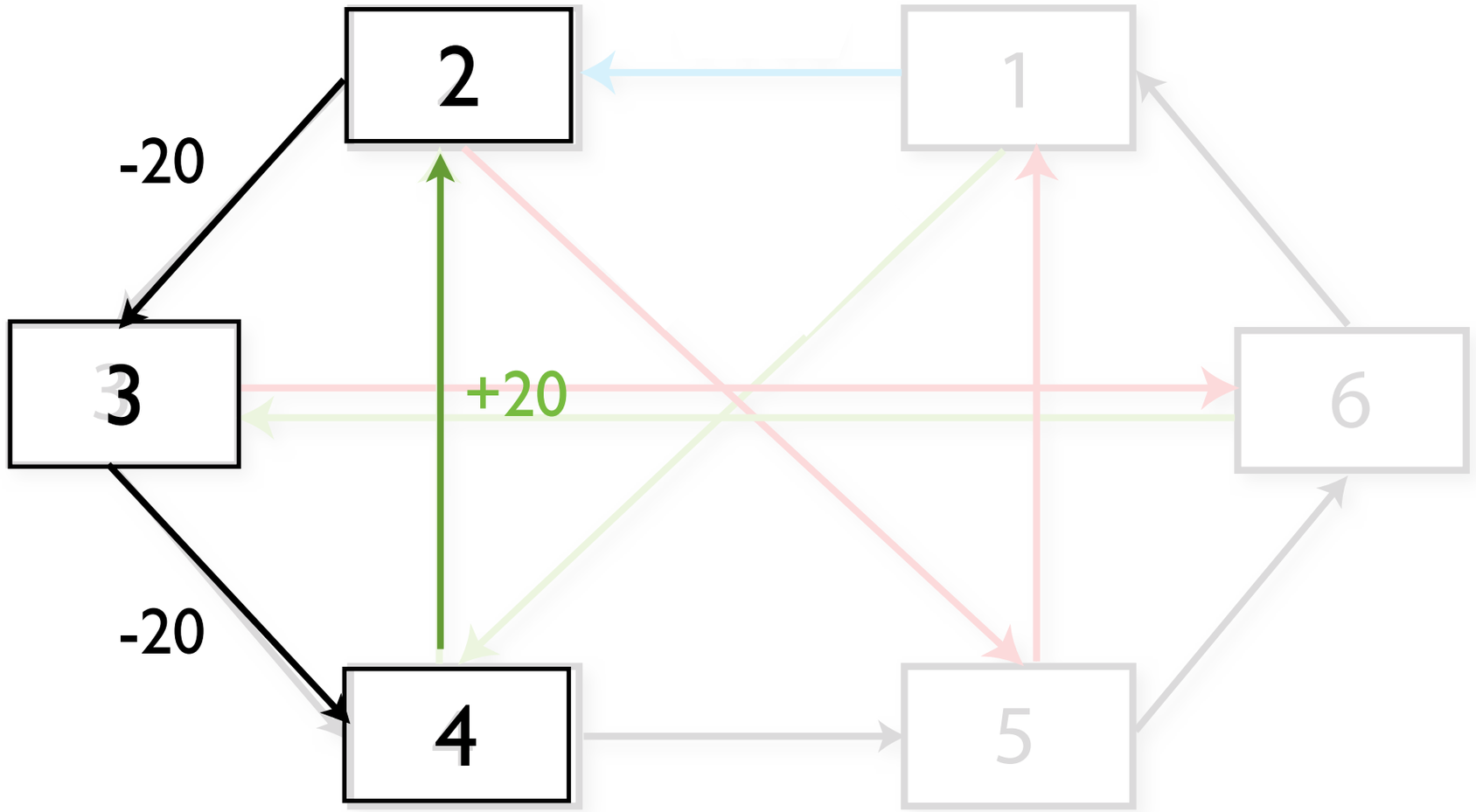
Pavlovian?


-140 ➔ -100
-100
-20
Pruning
-20
+20
+20
-20
+140
+20
-20
-100
-20
-100
-100
-20
+20
-140
-100
-260
-20
-100
-140
-60

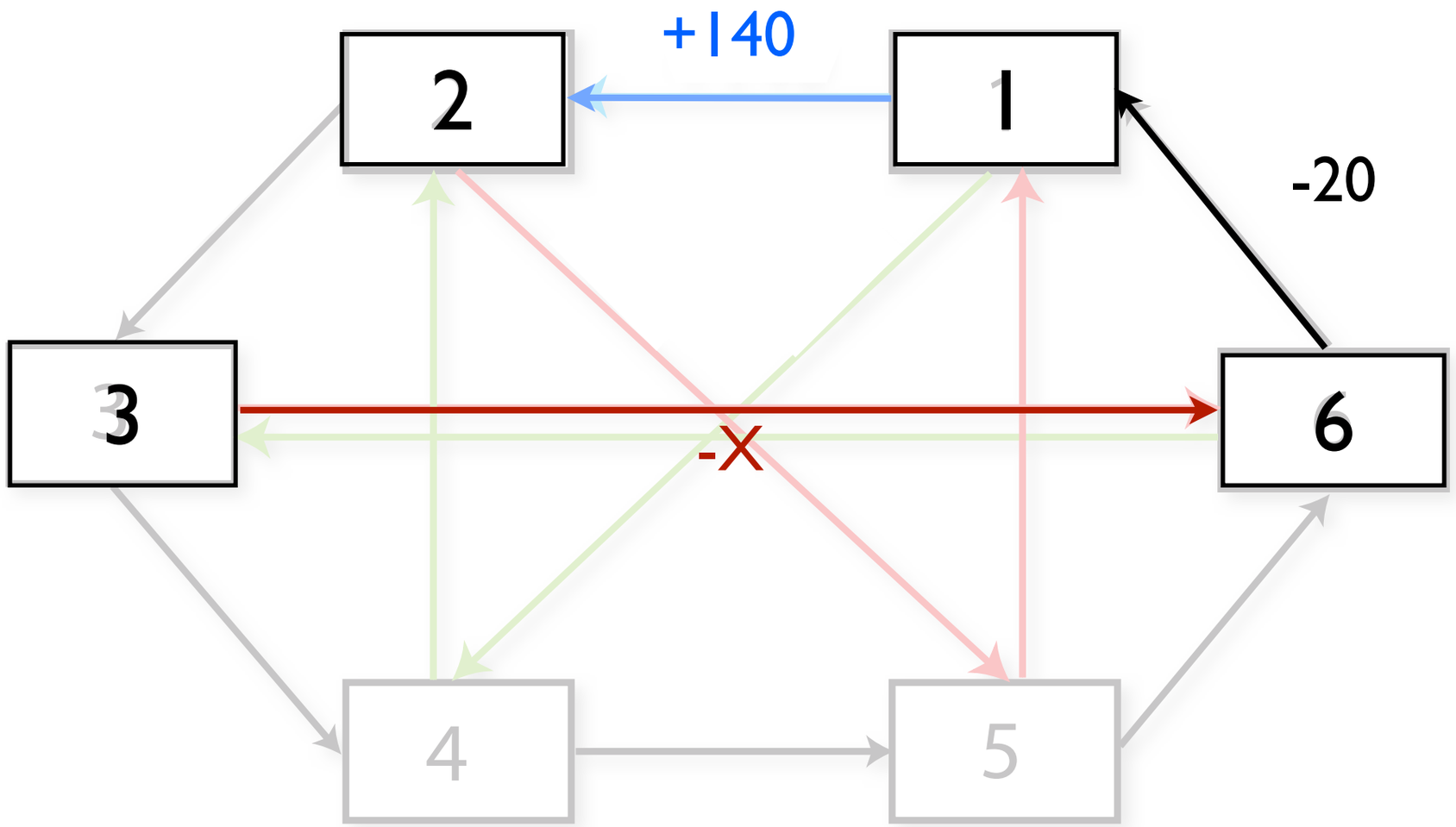
Stop at losses?



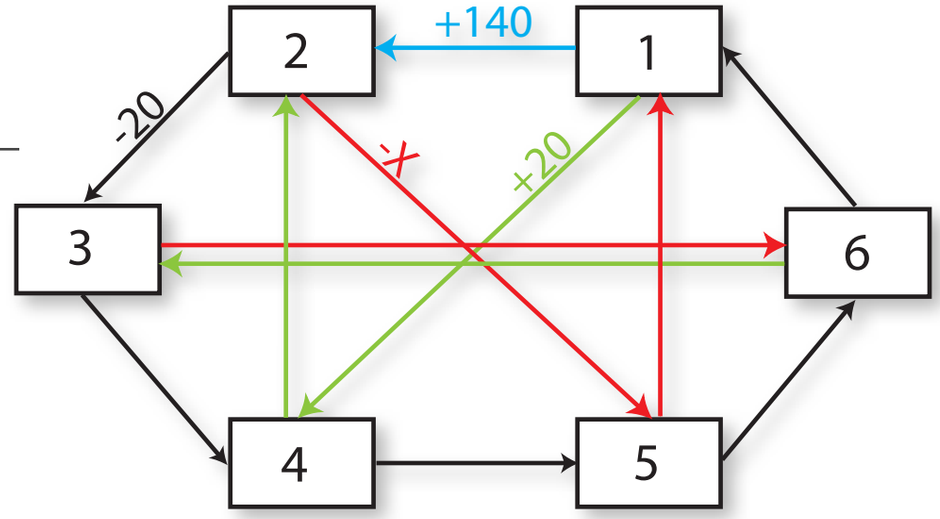
Stop at losses?



Stop at losses?

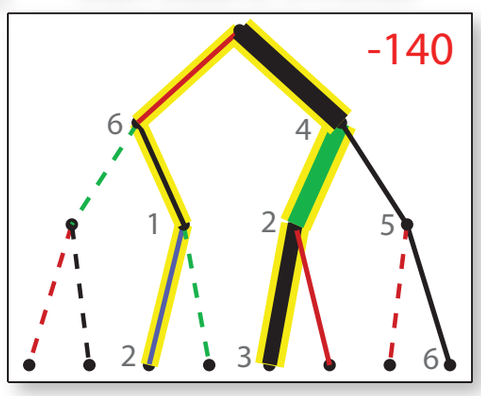


Avoiding losses



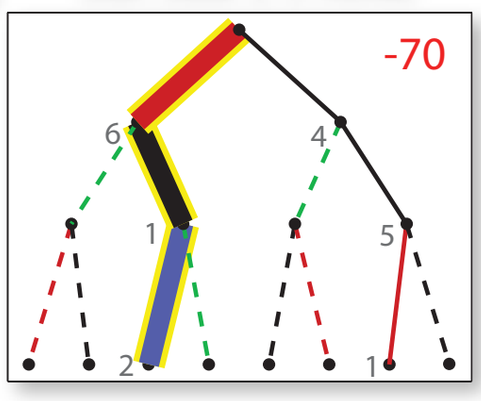
A

Start state 3, 3 moves

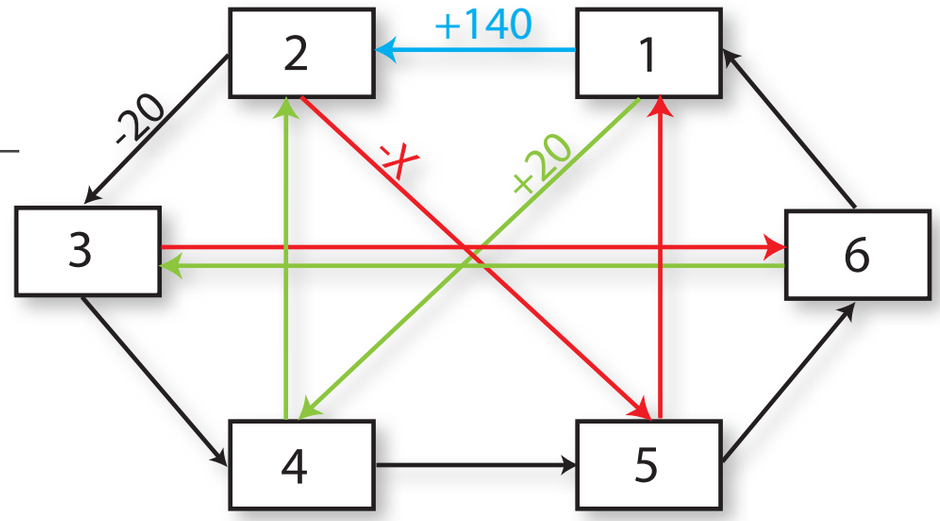


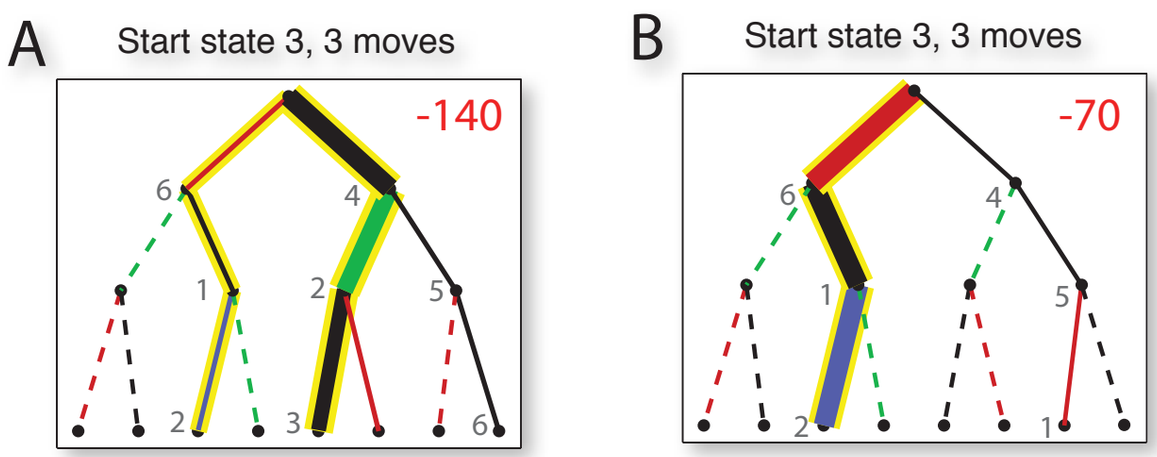
B

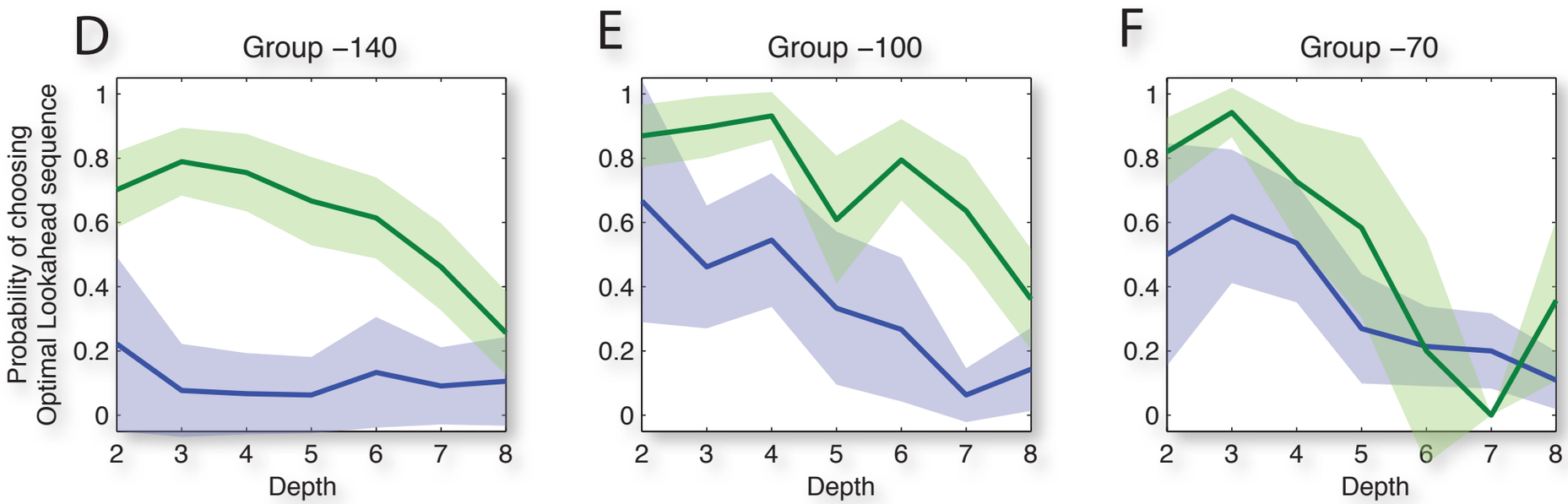
Start state 3, 3 moves



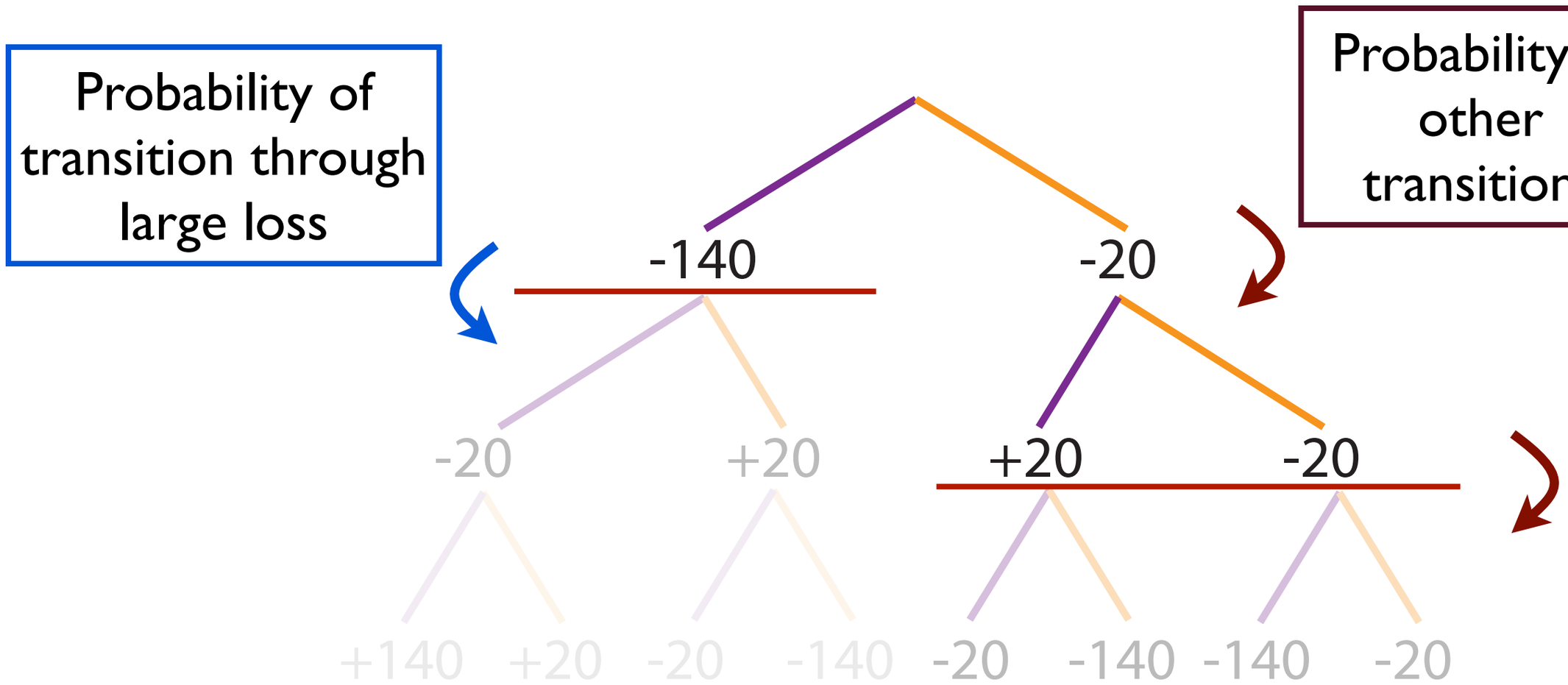
Avoiding losses



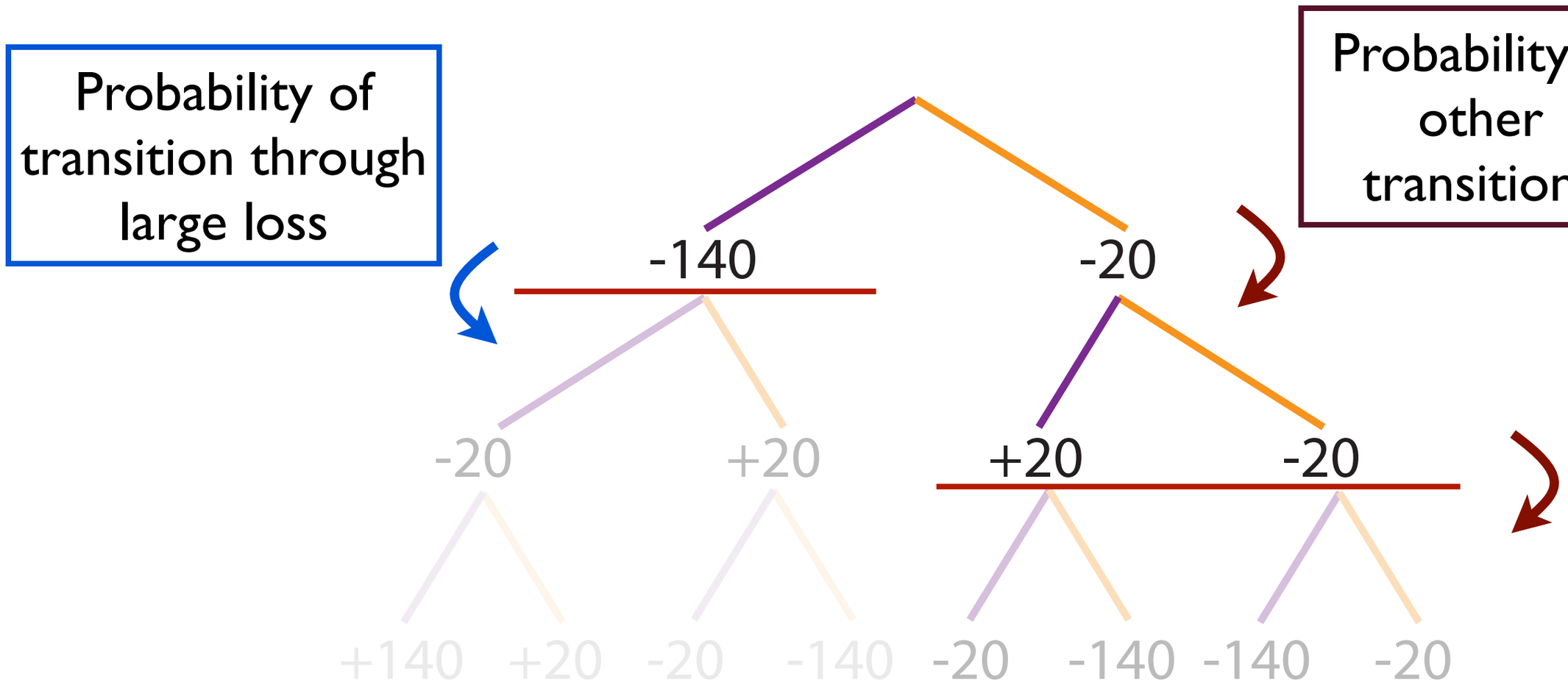




Adaptive pruning model

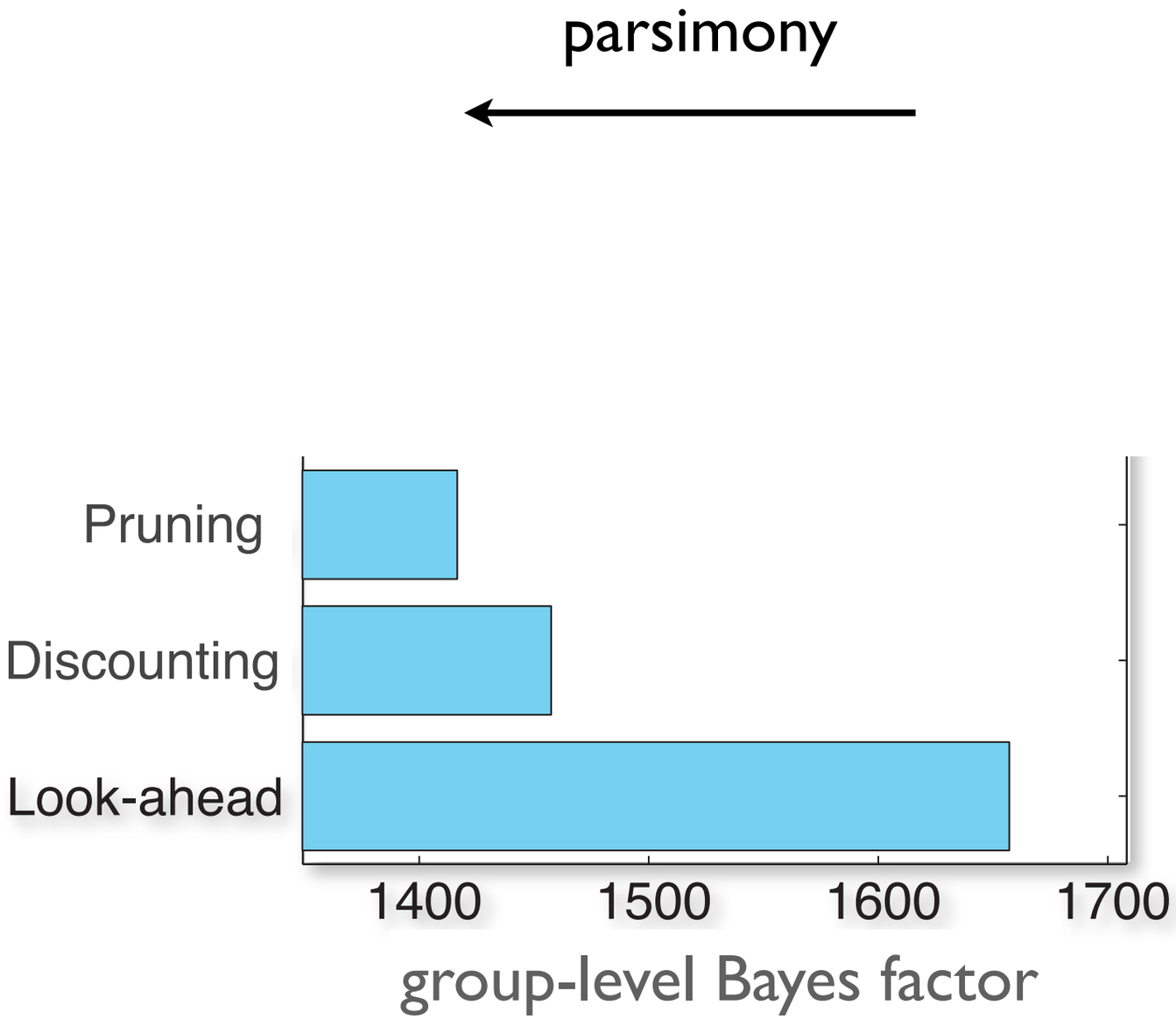


Adaptive pruning model



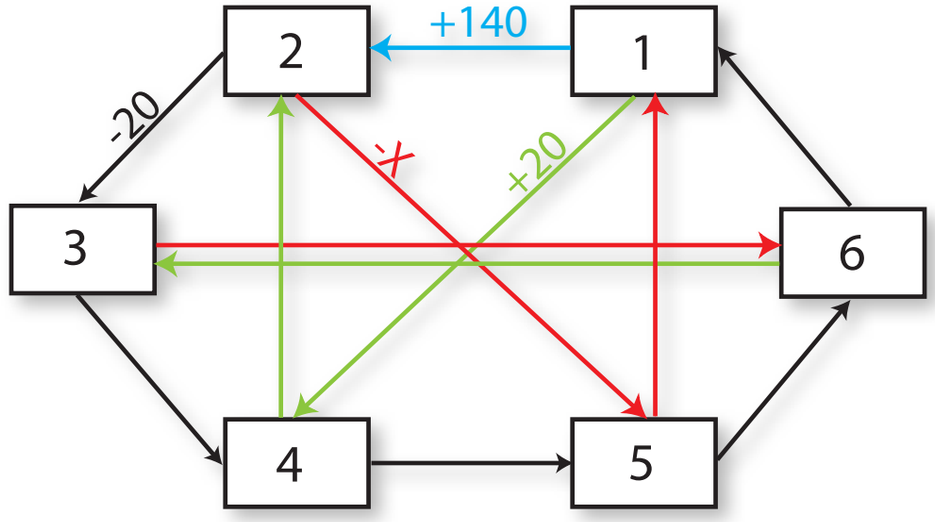
$$p(X|\mathcal{M}) = c \int d\theta \, p(X|\theta)$$

Adaptive pruning wins

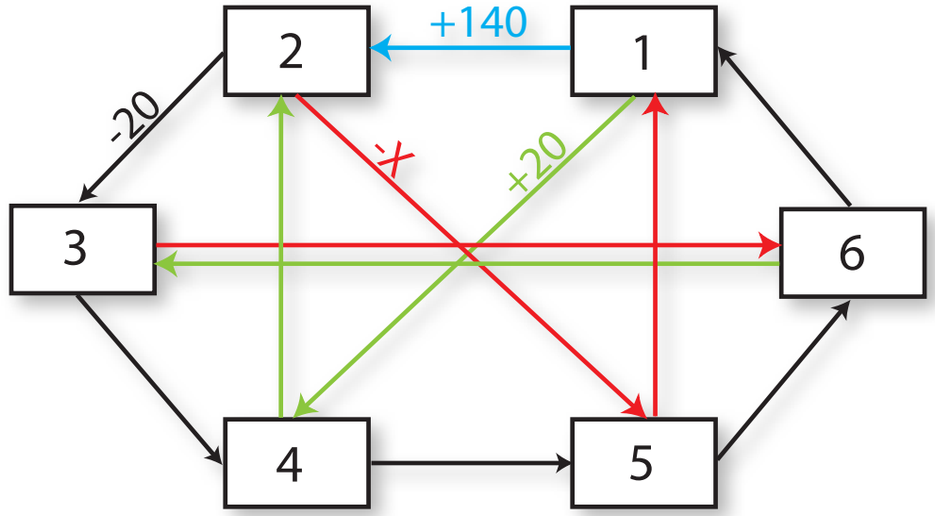


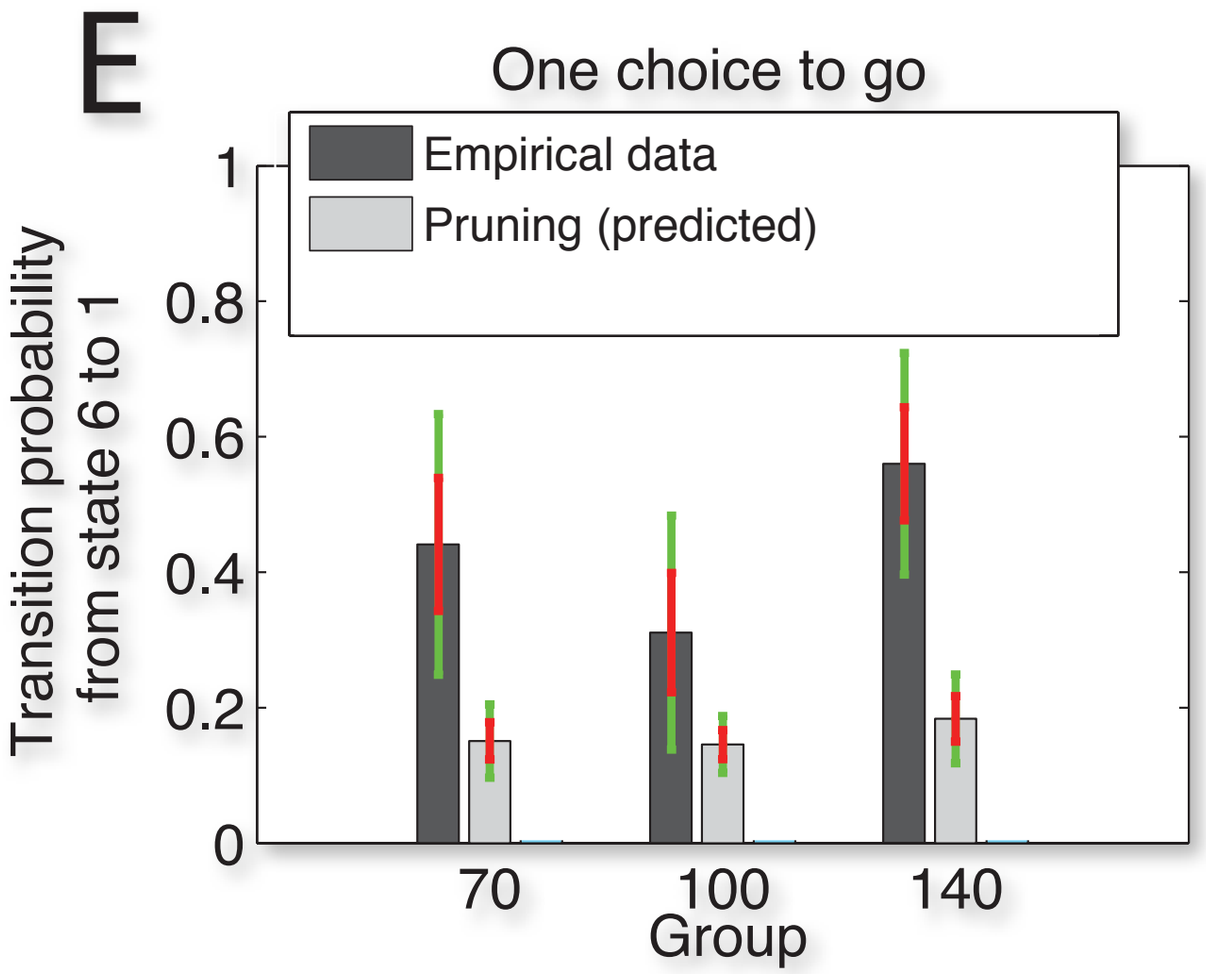
Huys et al. (2011) Submitted

Generate some data

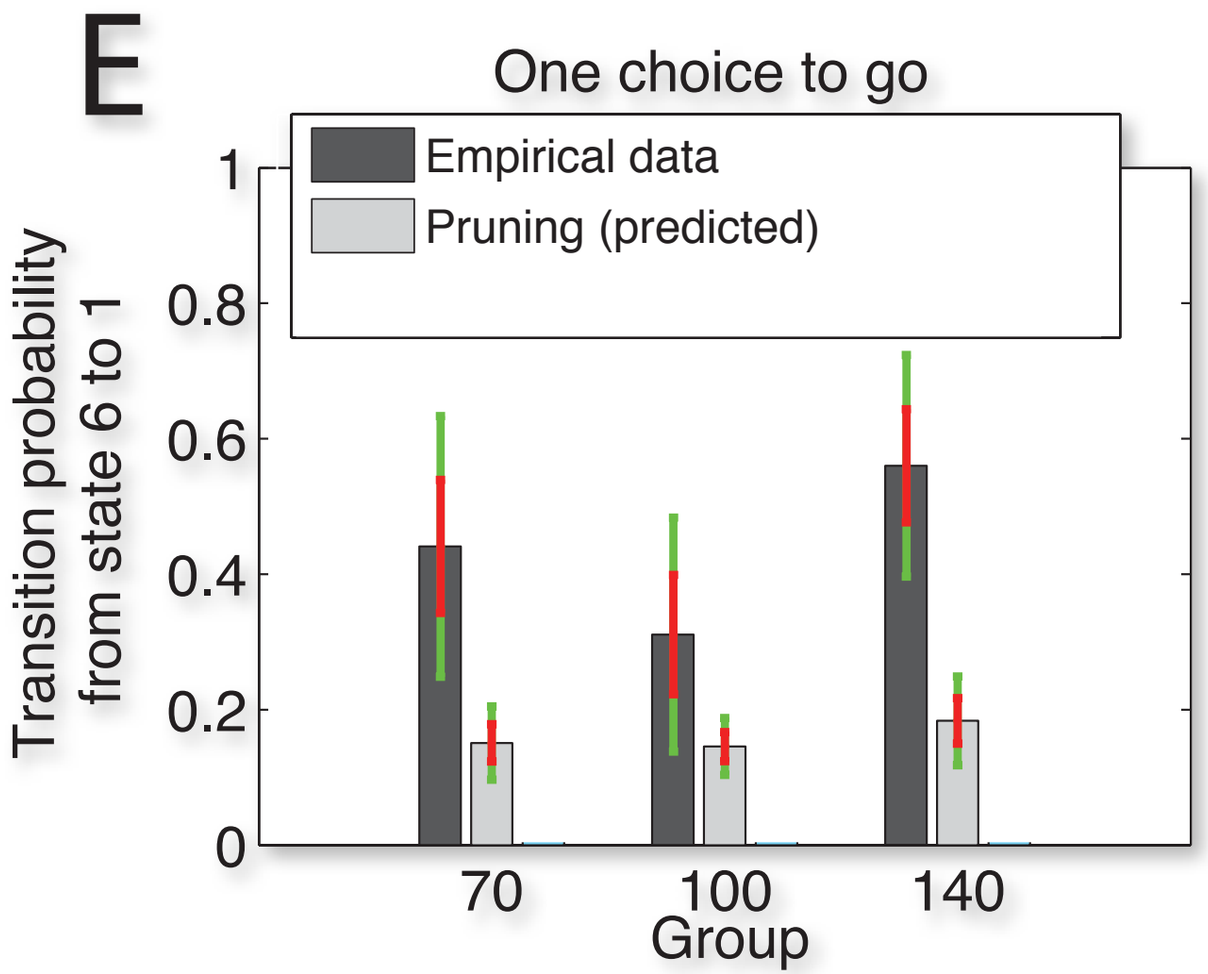


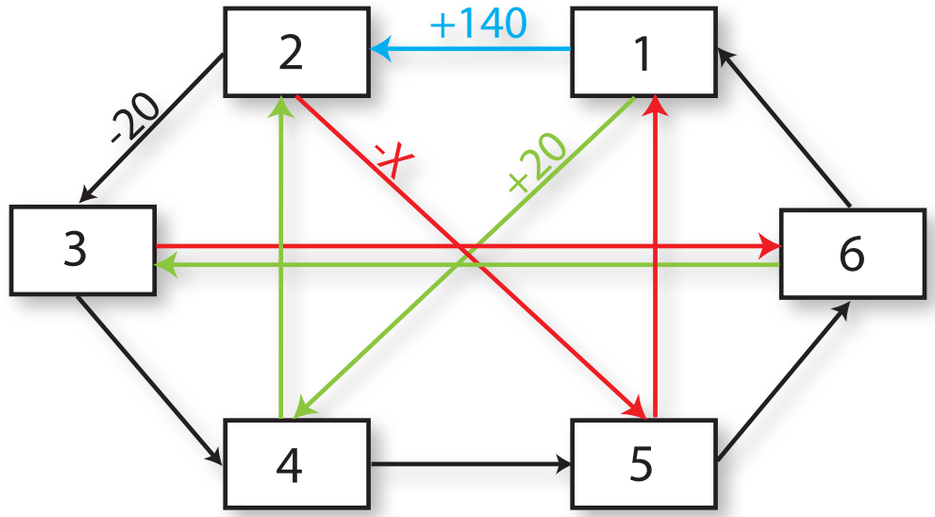
Generate some data

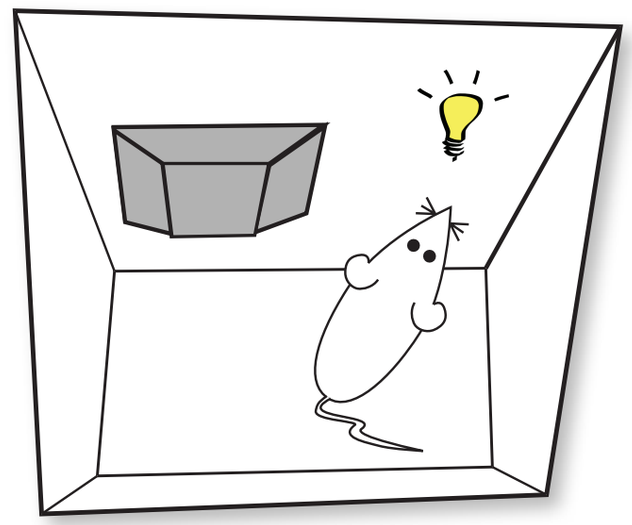




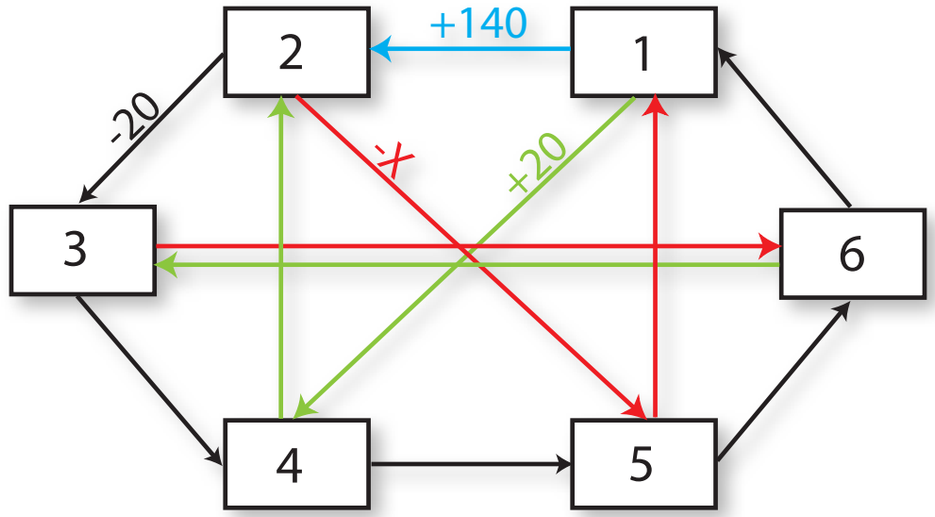
Generate some data

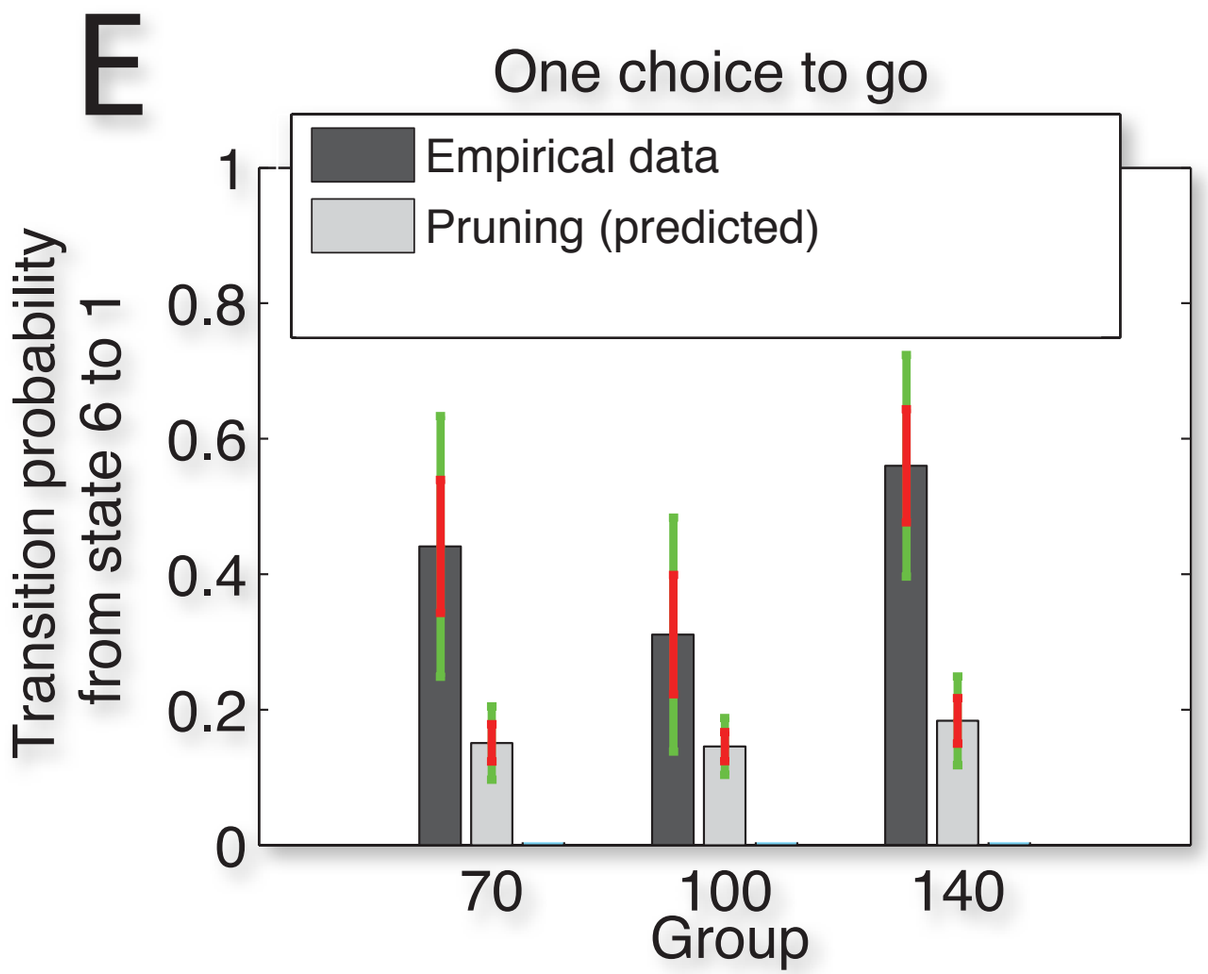


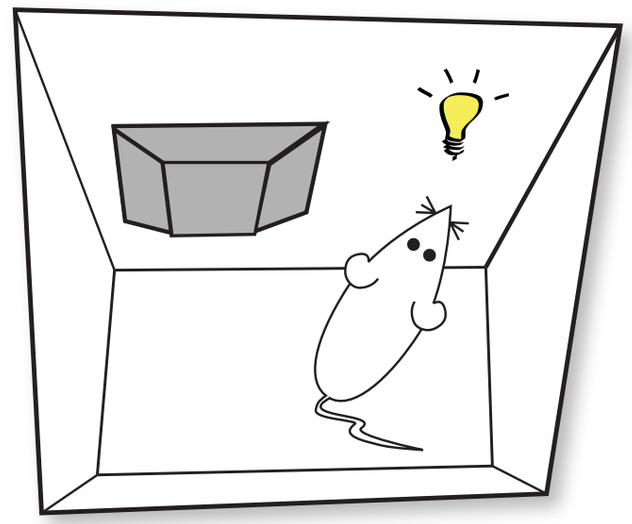




Generate some data

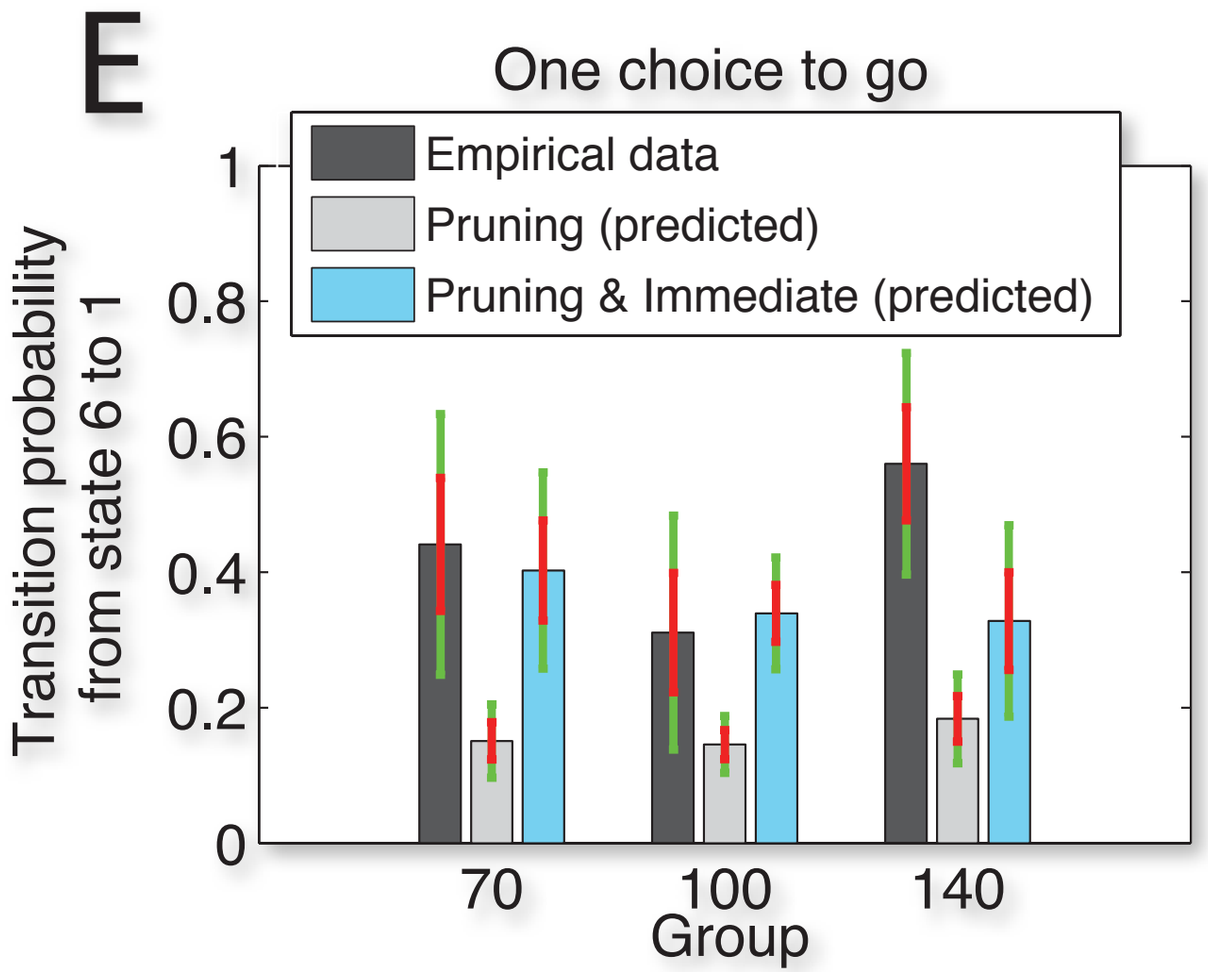


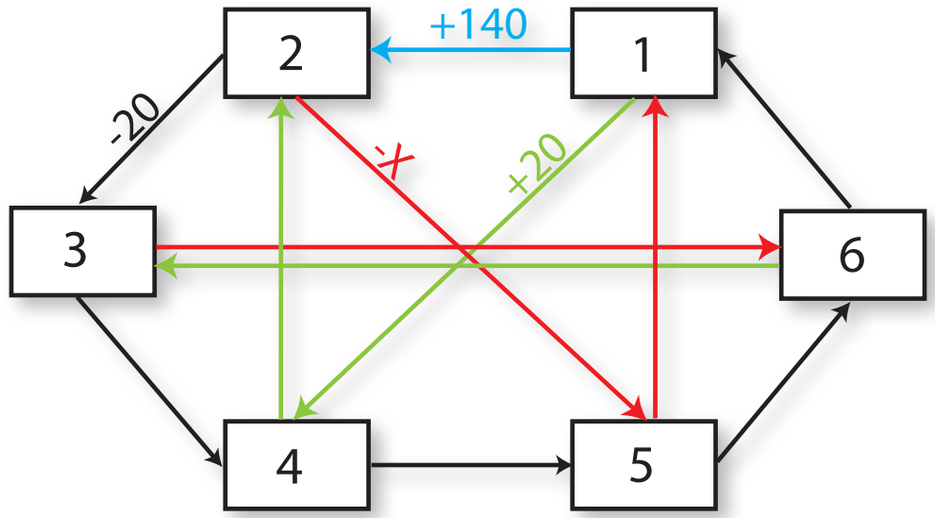


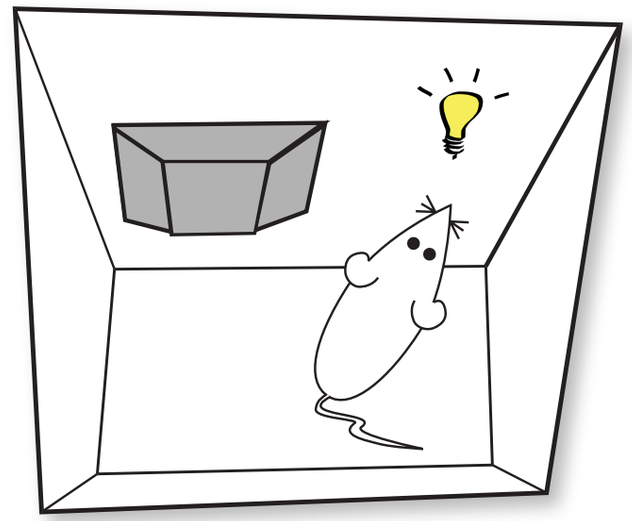


$$\mathcal{V}(s) \leftarrow \mathcal{V}(s) + \epsilon(\mathcal{V}(s') + r_t - \mathcal{V}(s))$$

Generate some data

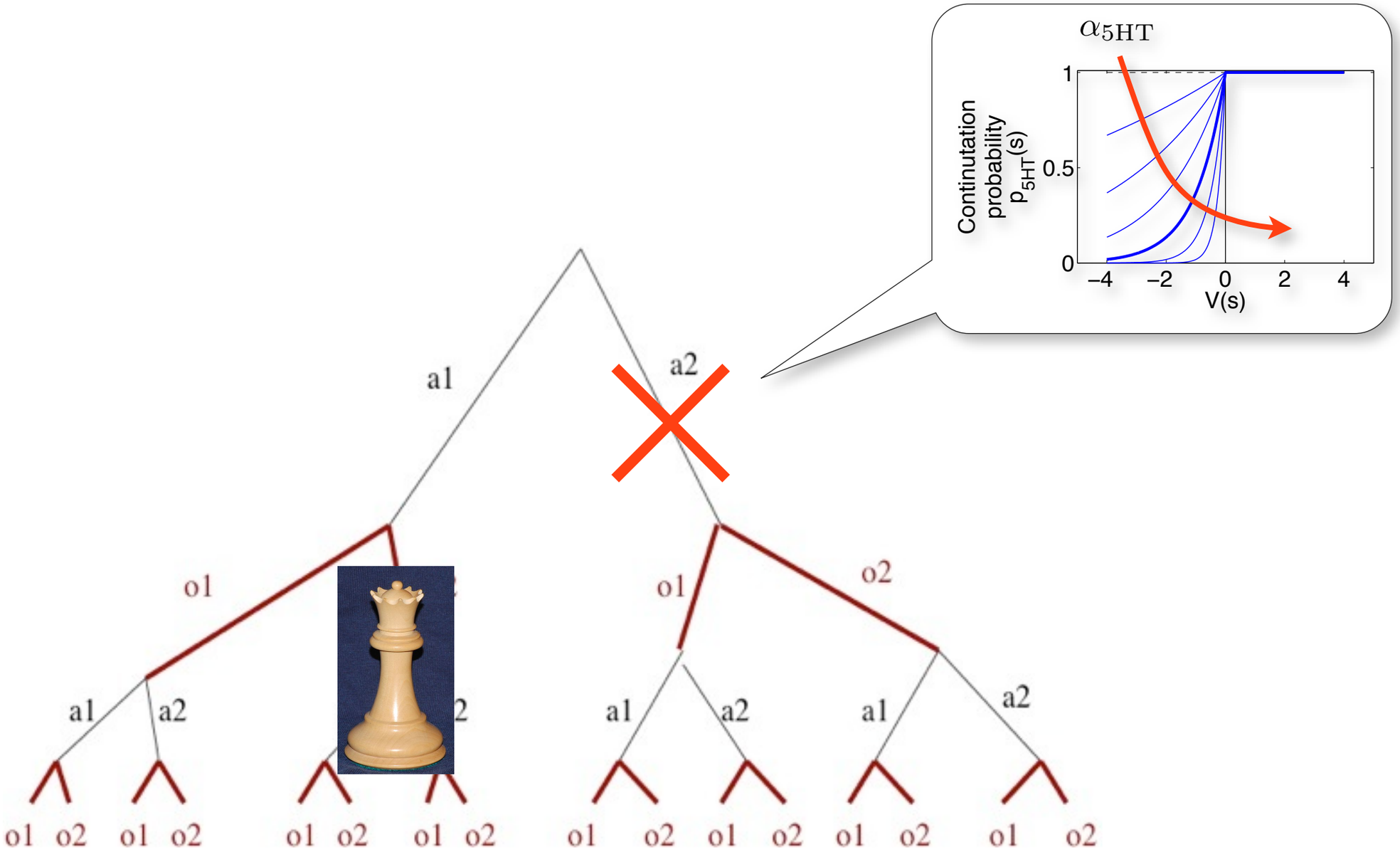




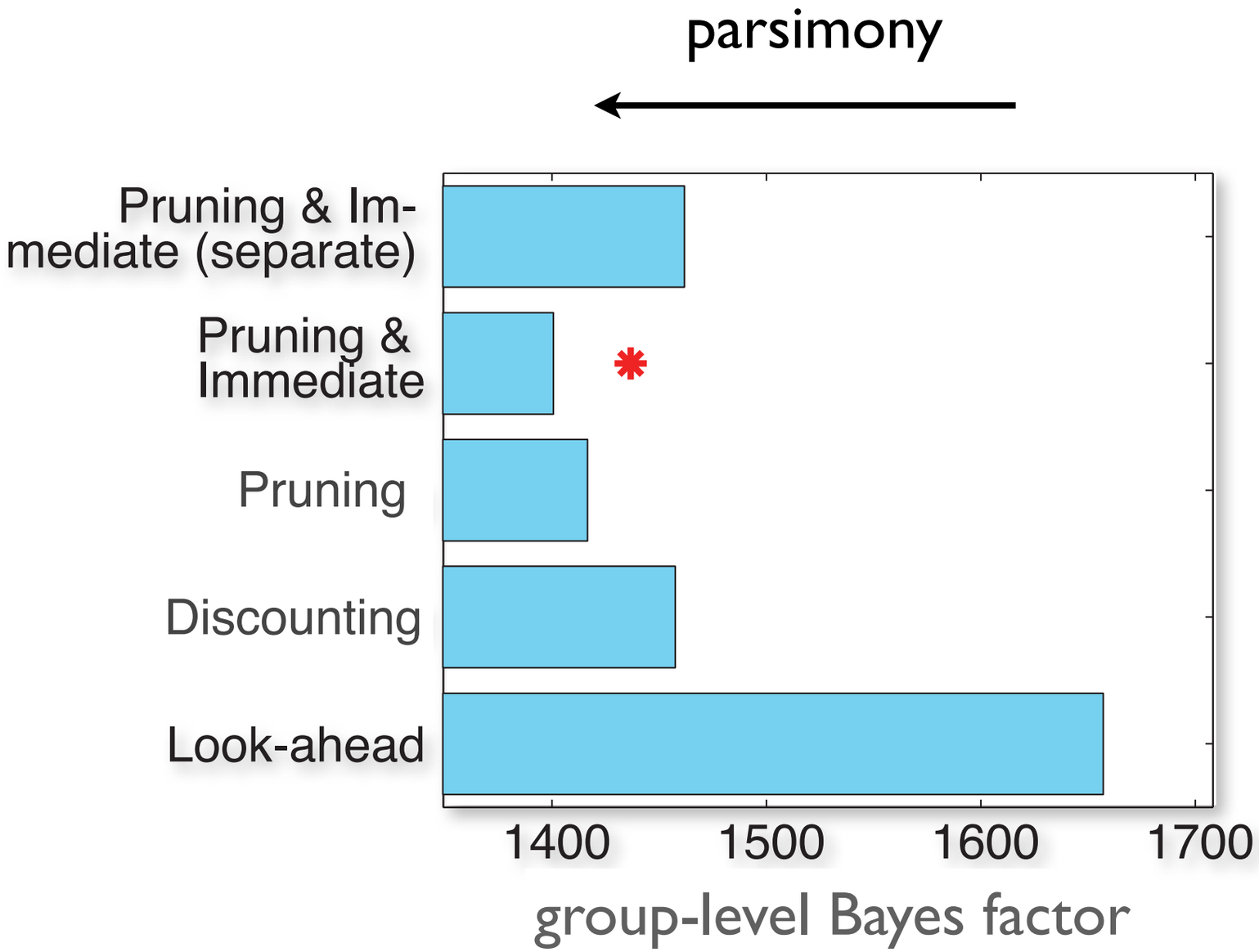


$$\mathcal{V}(s) \leftarrow \mathcal{V}(s) + \epsilon(\mathcal{V}(s') + r_t - \mathcal{V}(s))$$

Pavlovian influences inside a decision tree



Adaptive pruning wins

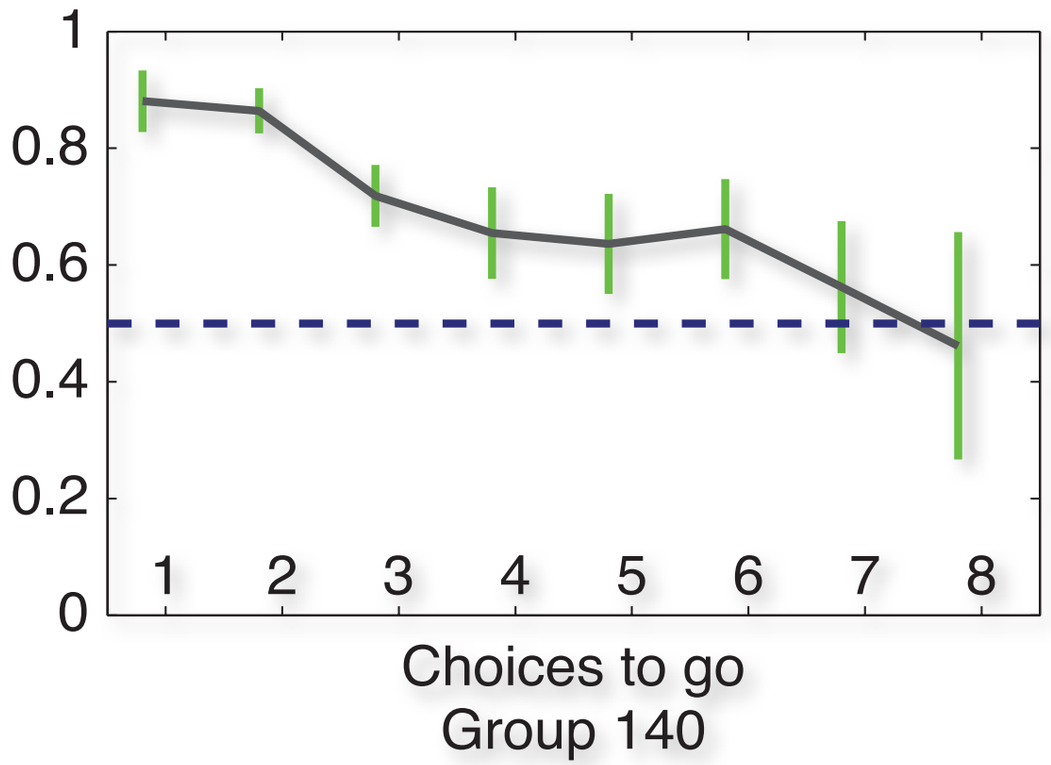


Huys et al. (2011) Submitted

Further alternatives

- Loss aversion
  - maybe subjects dislike large loss disproportionately?
- Cached terminal values
- Cached learning only
- Choices of entire sequences
- Compare models' abilities to explain the ENTIRE set of data

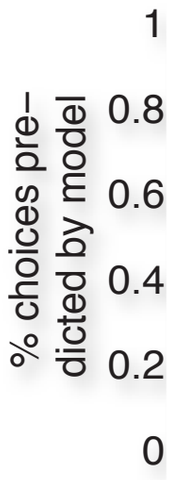
Pruning is Pavlovian

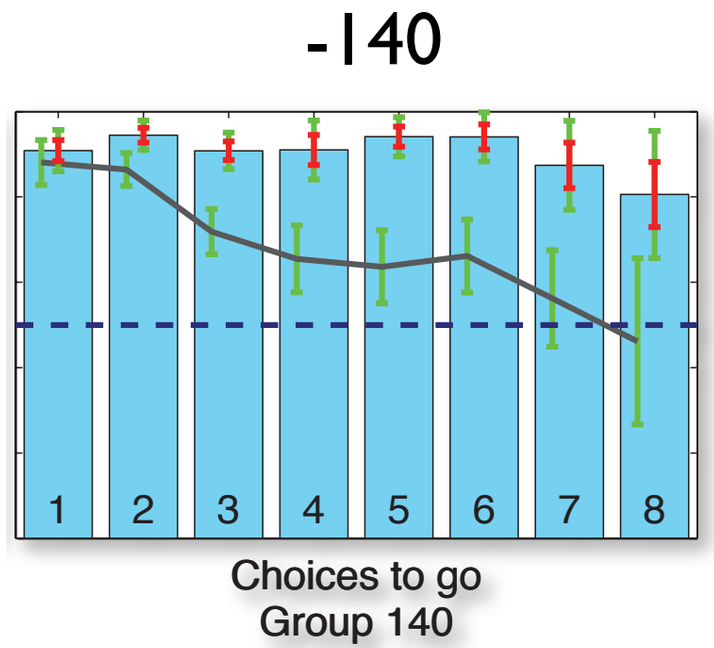


Huys et al. (2011) Submitted

Pruning is Pavlovian

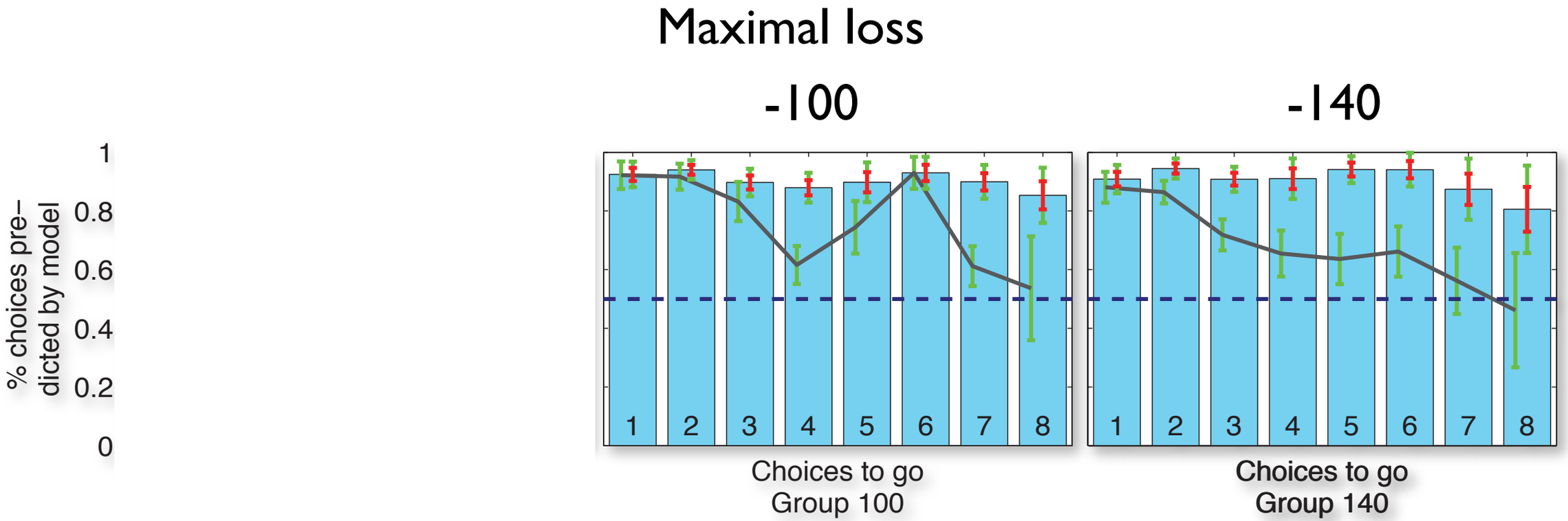
Maximal loss





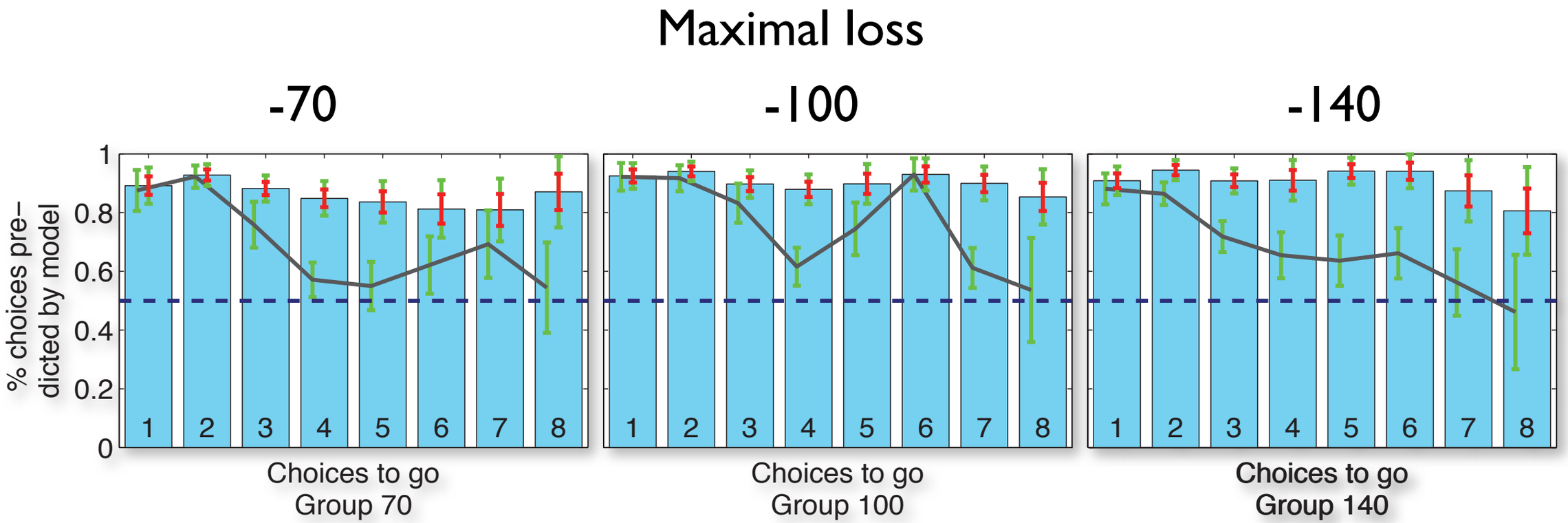
Huys et al. (2011) Submitted

Pruning is Pavlovian



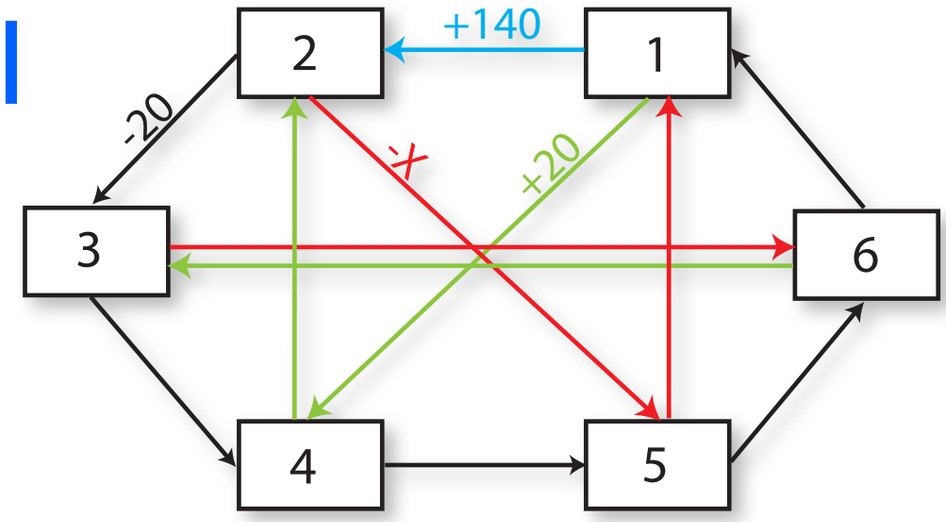
Huys et al. (2011) Submitted

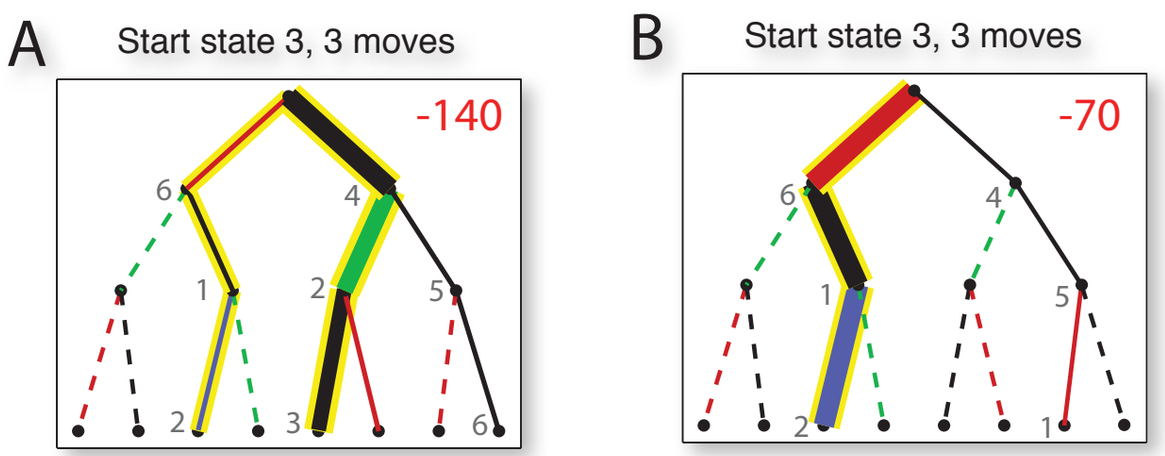
Pruning is Pavlovian

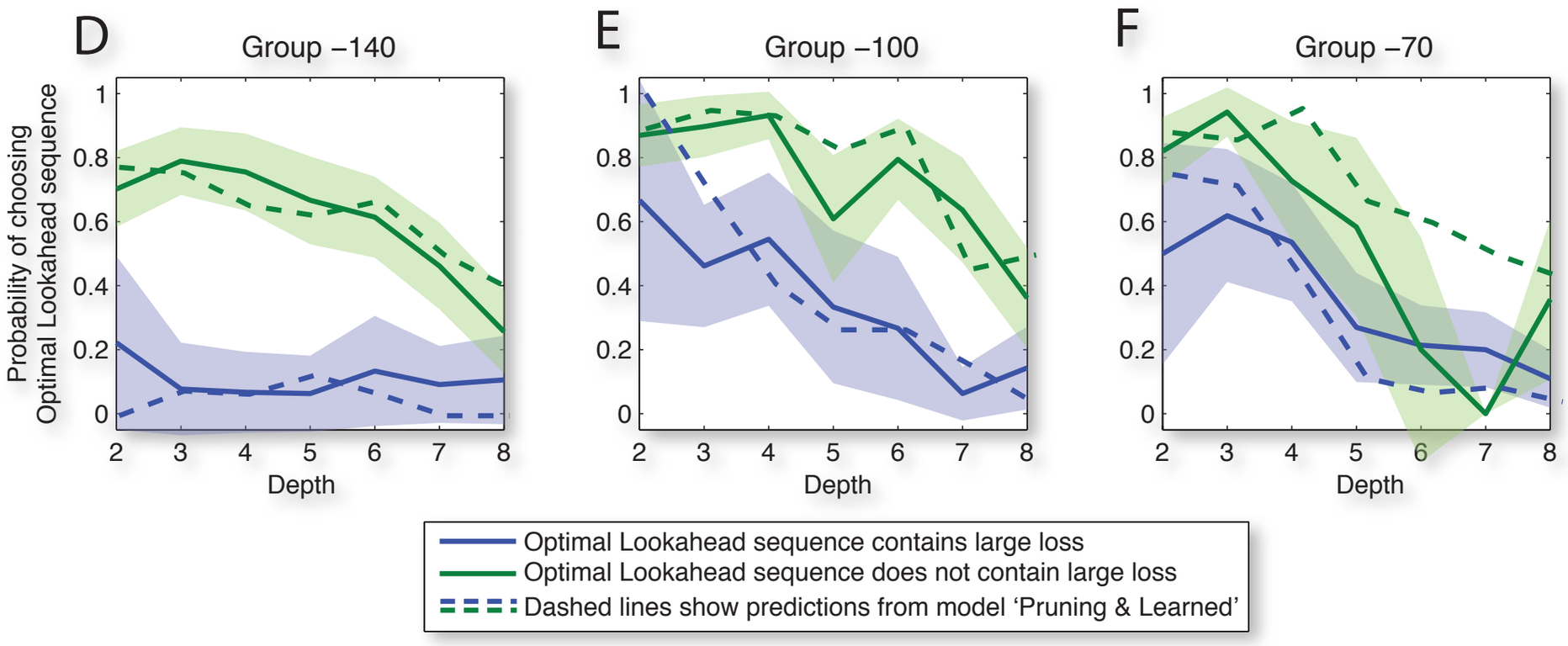


Huys et al. (2011) Submitted

Generating data from the model

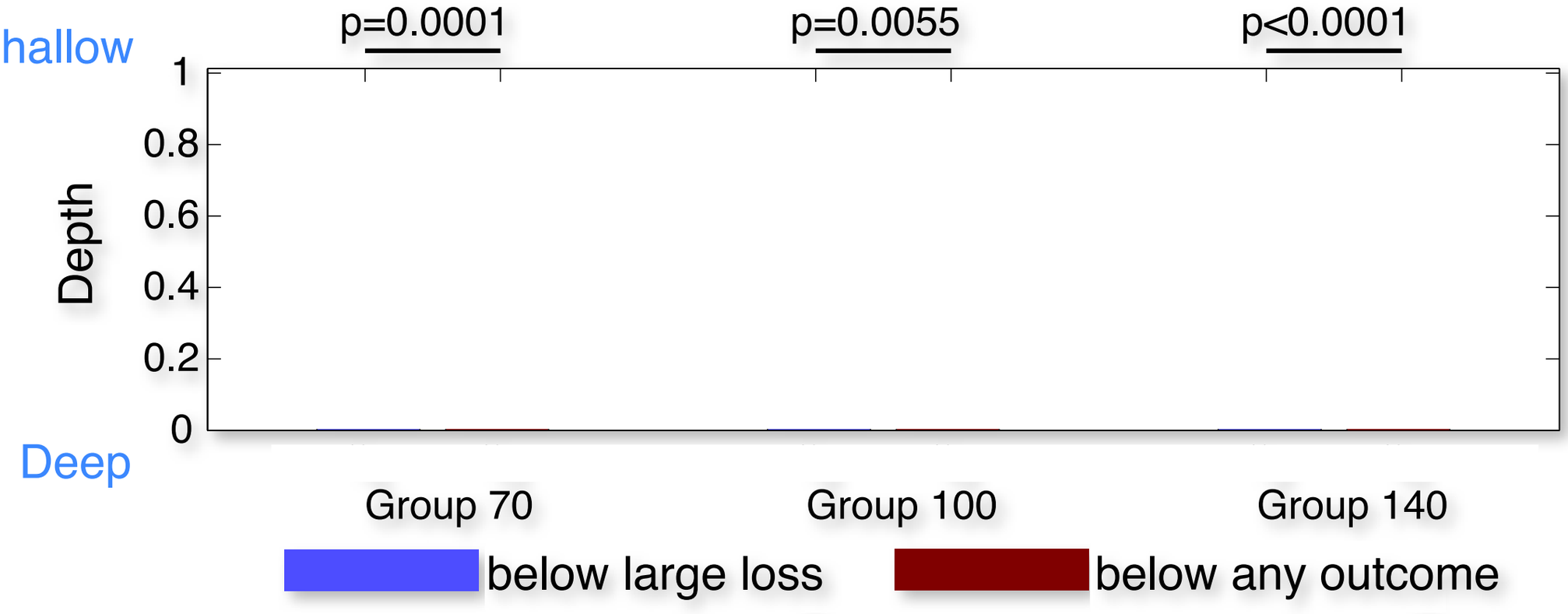






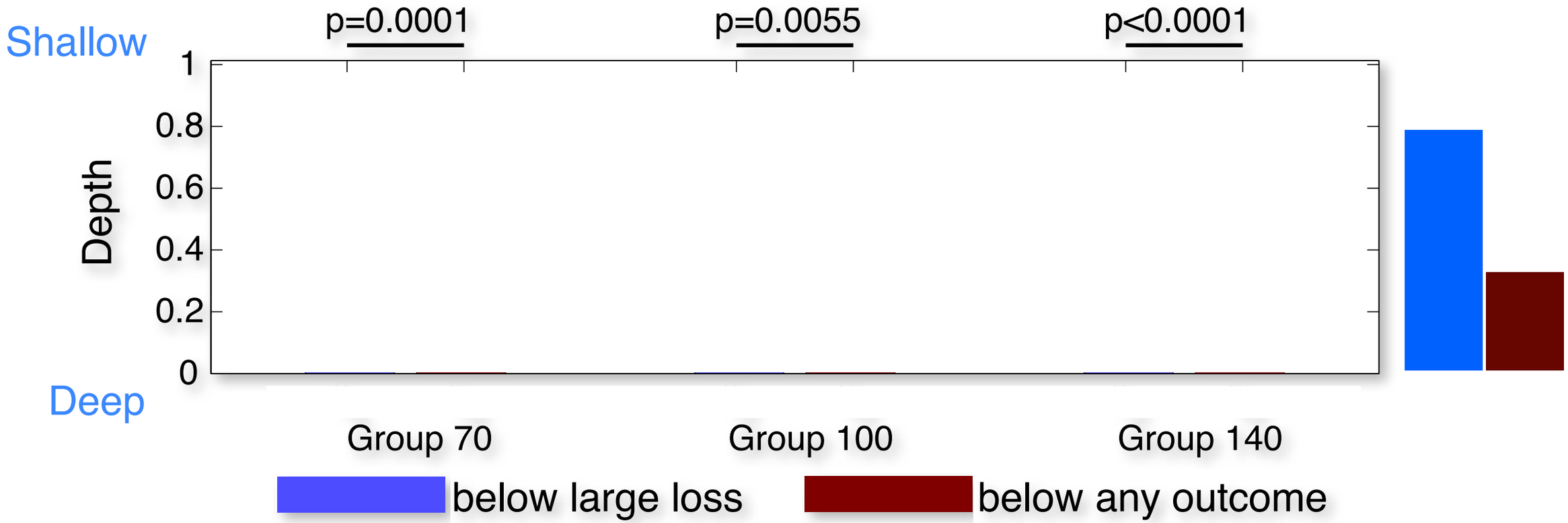
Pruning parameters

- Given the model, can now look at parameters
  - Ensure they are meaningful



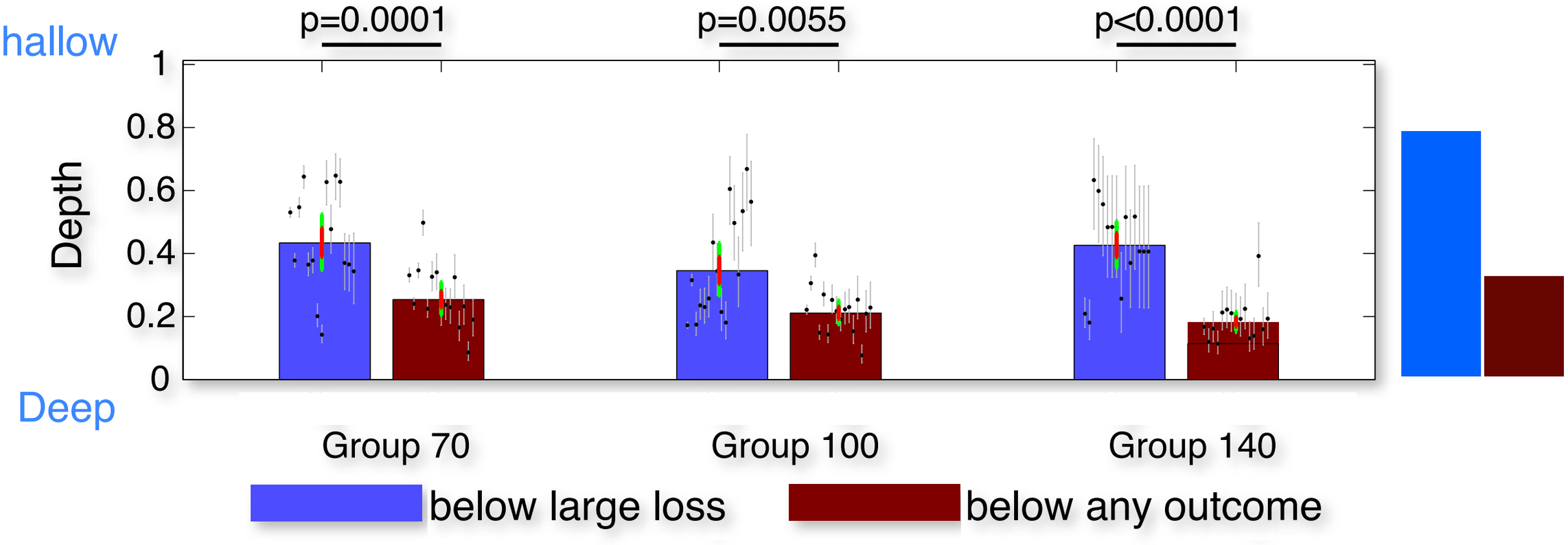
Pruning parameters

- Given the model, can now look at parameters
  - Ensure they are meaningful

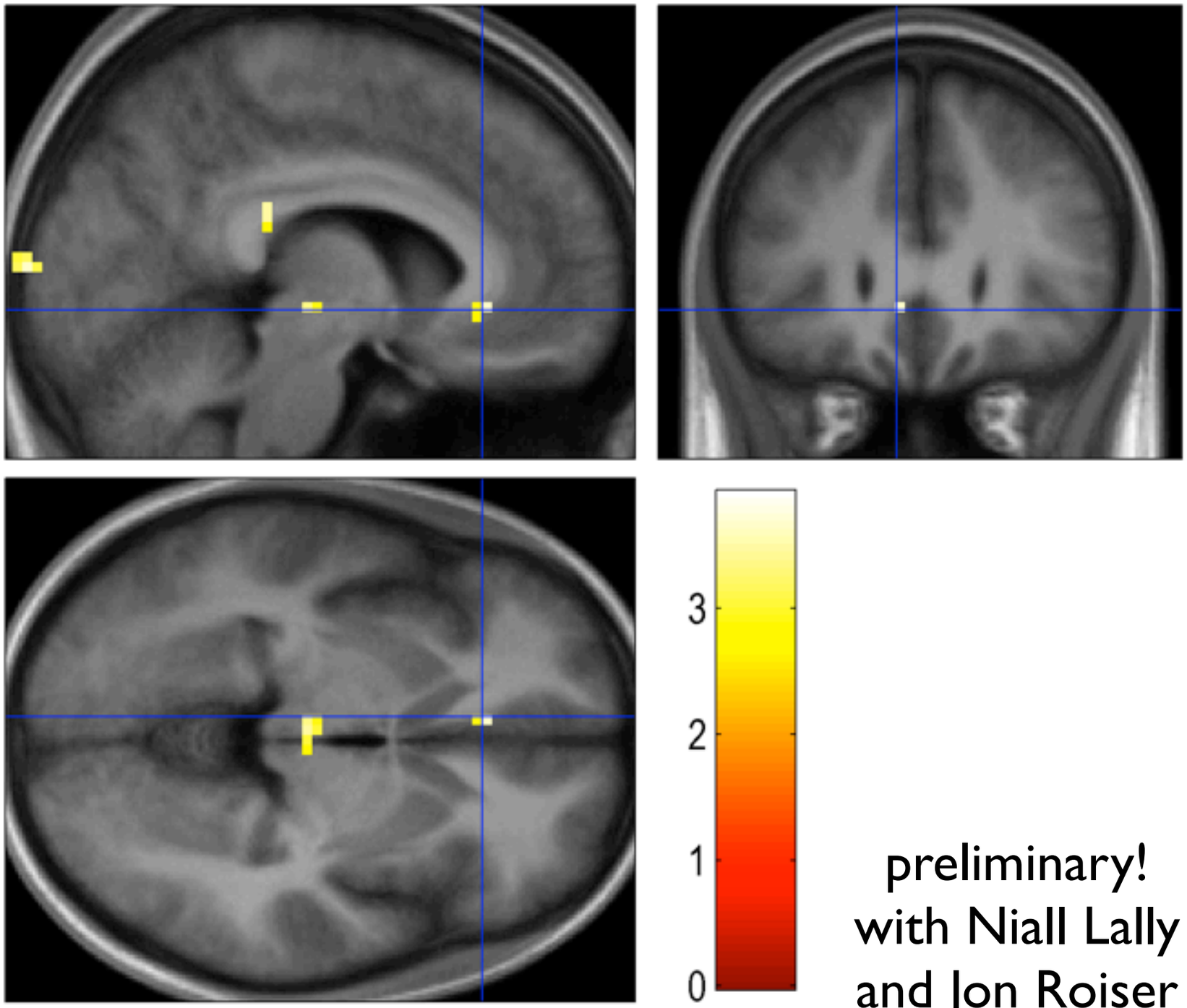


Pruning parameters

- Given the model, can now look at parameters
  - Ensure they are meaningful



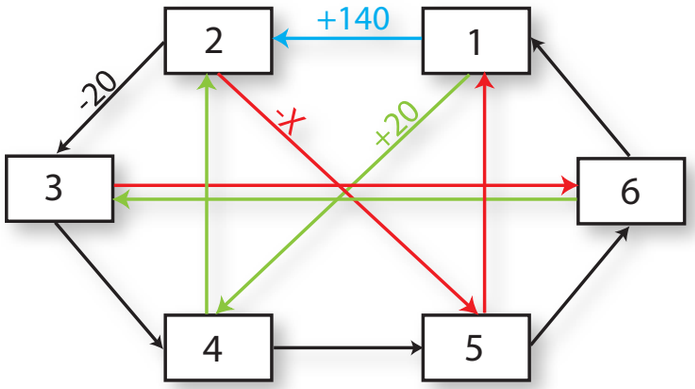
Pruning in the brain: PAG and sgPFC

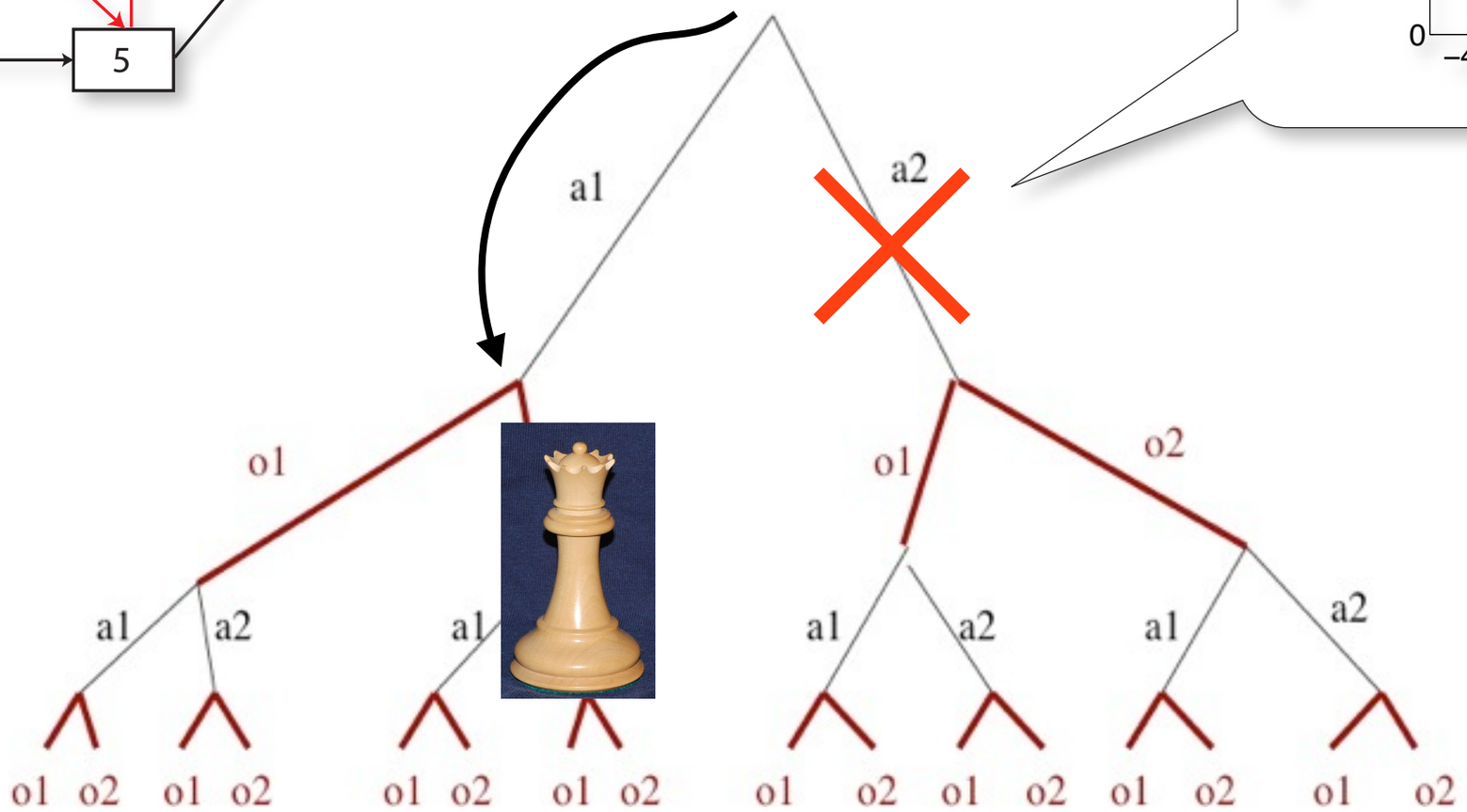


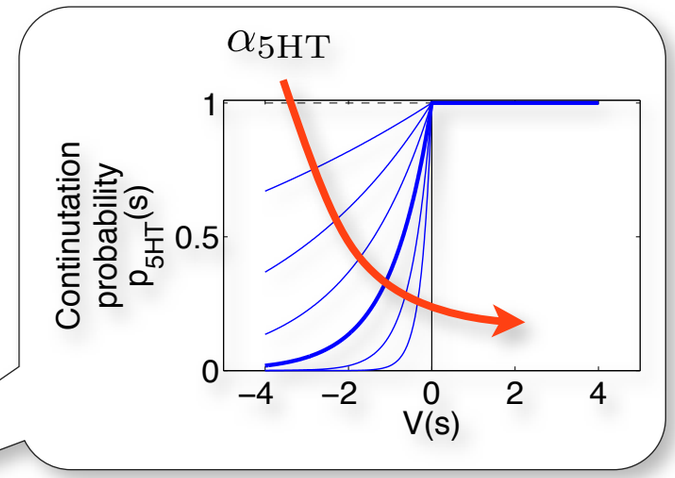
preliminary!
with Niall Lally
and Jon Roiser

Two Pavlovian influences

(i.e. state values directly linked to actions thorughout learning)







Recap

- Multiple decision systems
- Multiple values
- Multiple action mechanisms
- Interactions
  - Override
  - Uncertainty
- Complex problem
- Identification via critical features