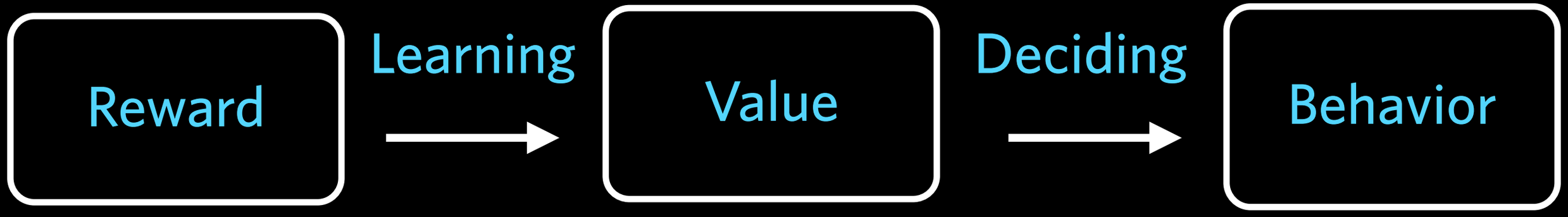
reinforcement learning
(learning what to do)

"Life is like playing a violin solo in public
while learning the instrument as one goes on"

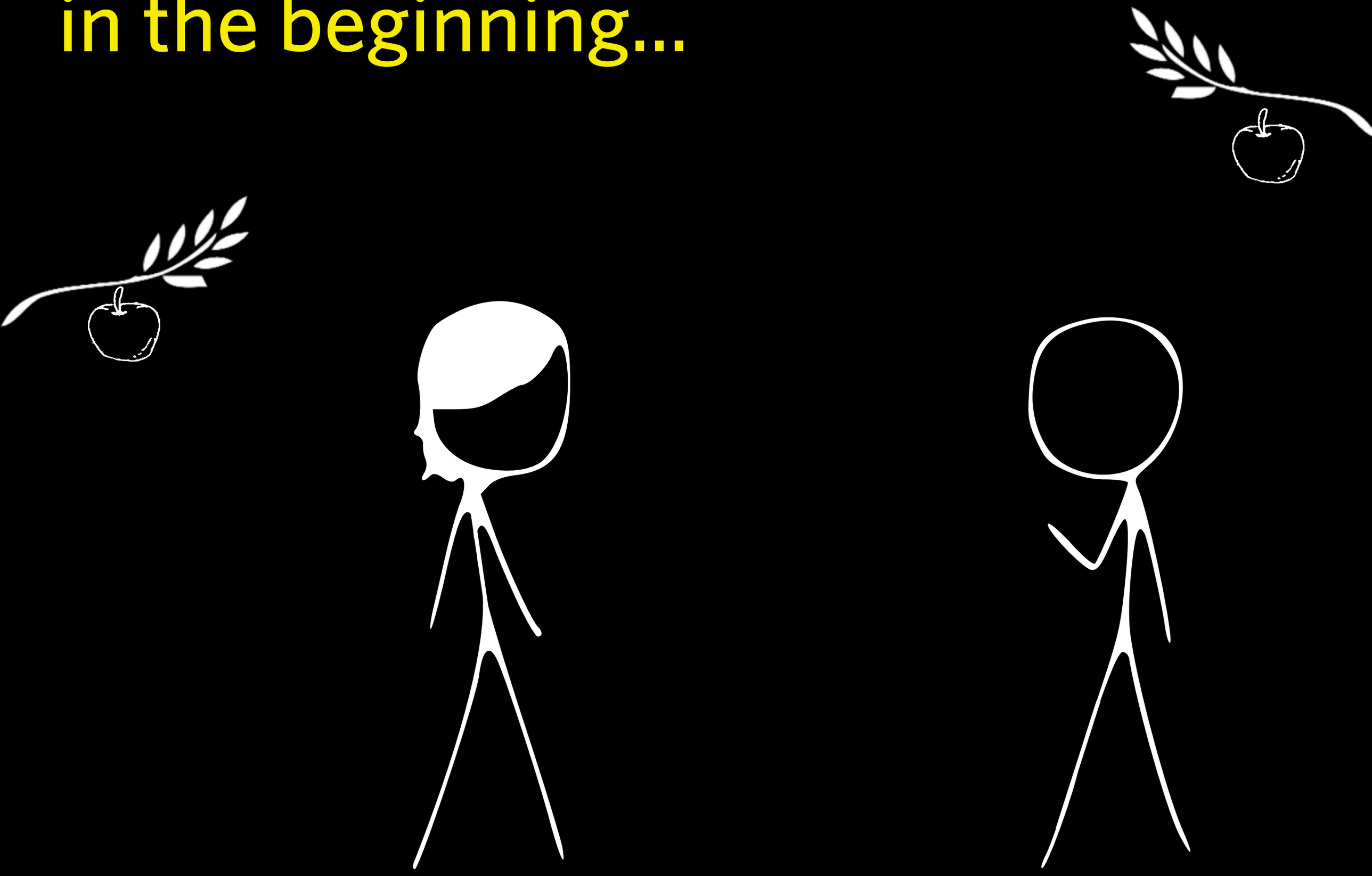


① understand challenges of decision-making

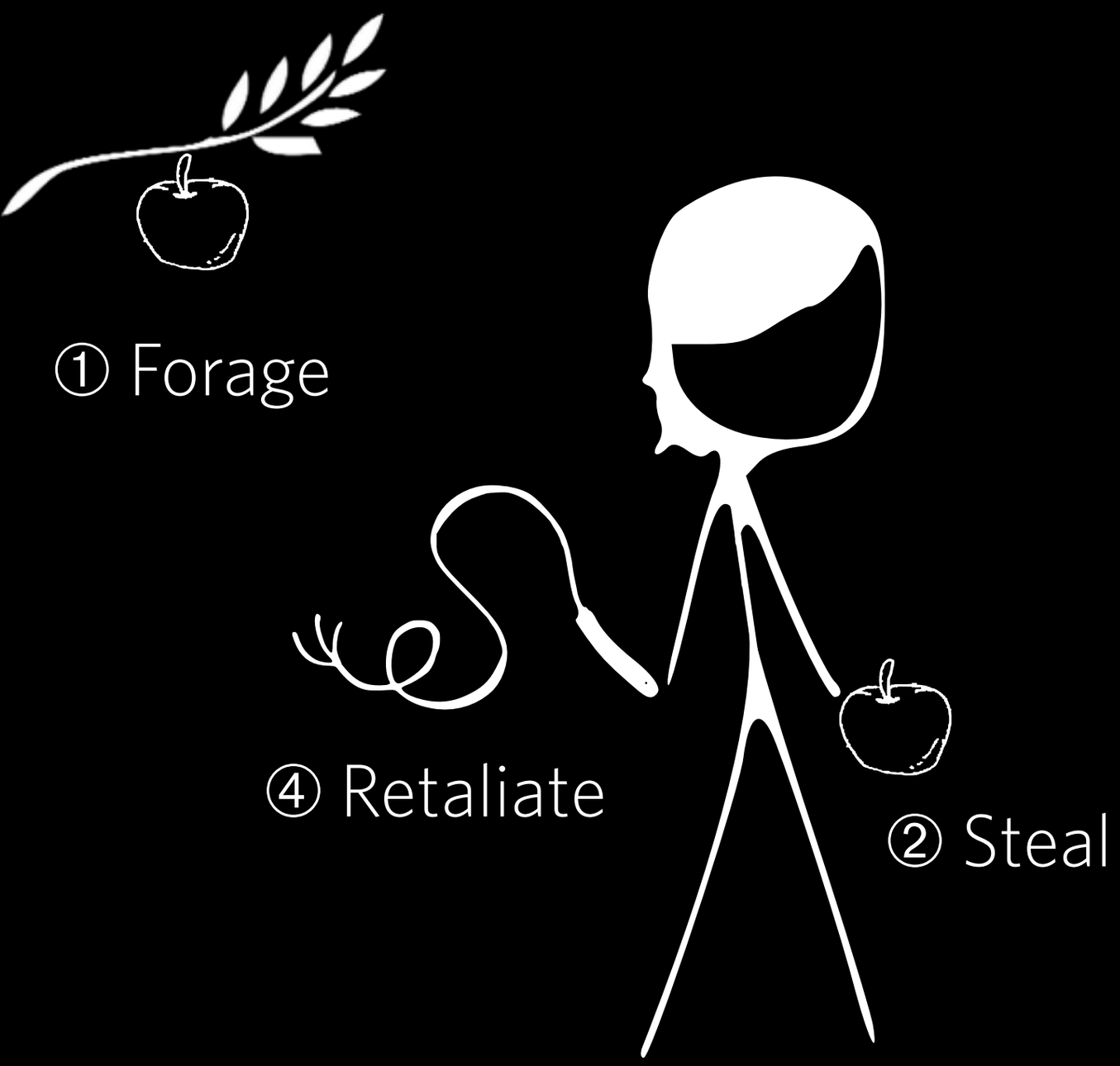
② learn basics of RL framework

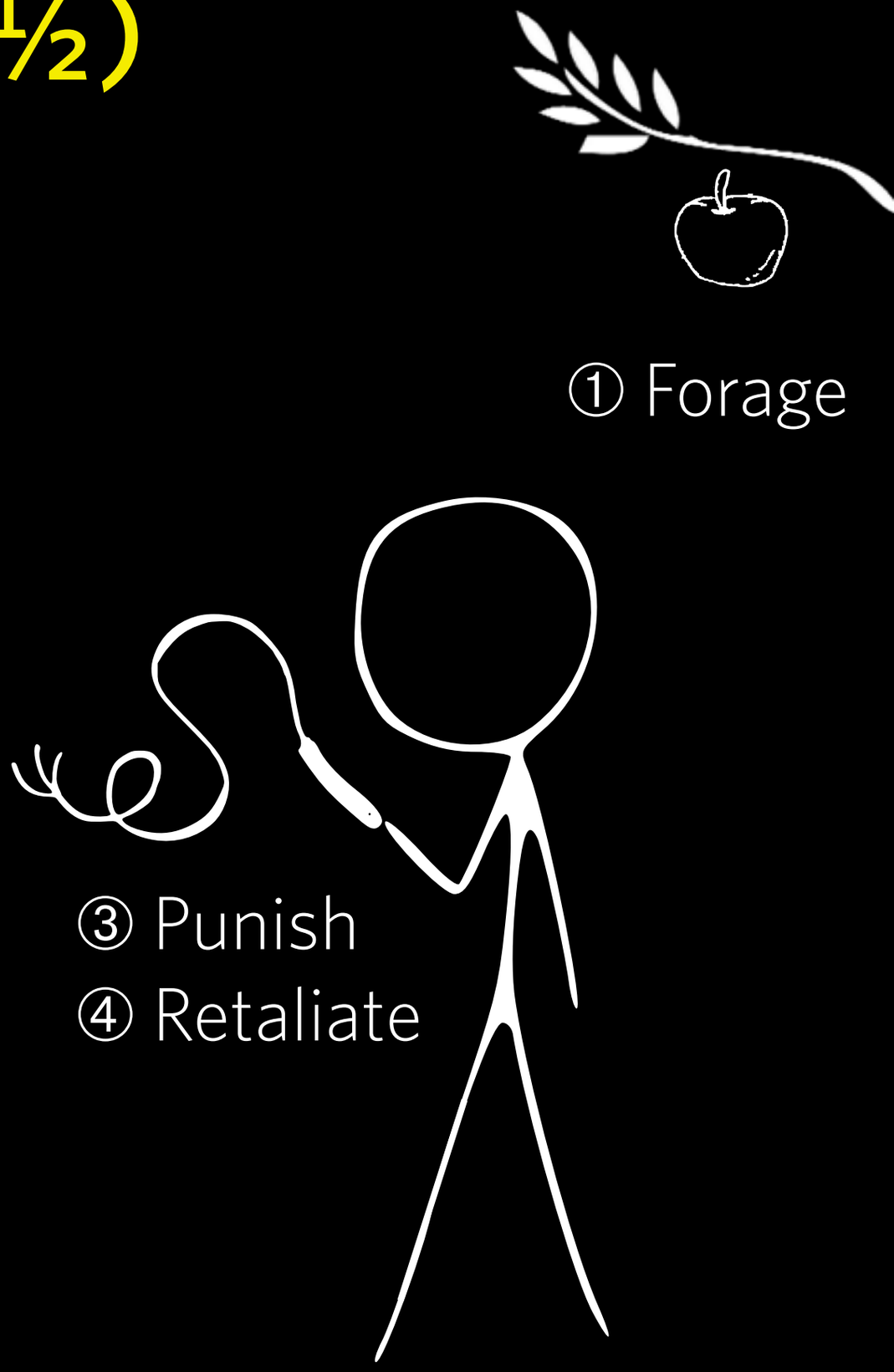
③ see applications to moral psychology

in the beginning...



the moral universe (½)





the moral universe (½)

THE EVOLUTION OF RECIPROCAL ALTRUISM

By Robert L. Trivers

articles

Altruistic punishment in humans

Ernst Fehr* & Simon Gächter†

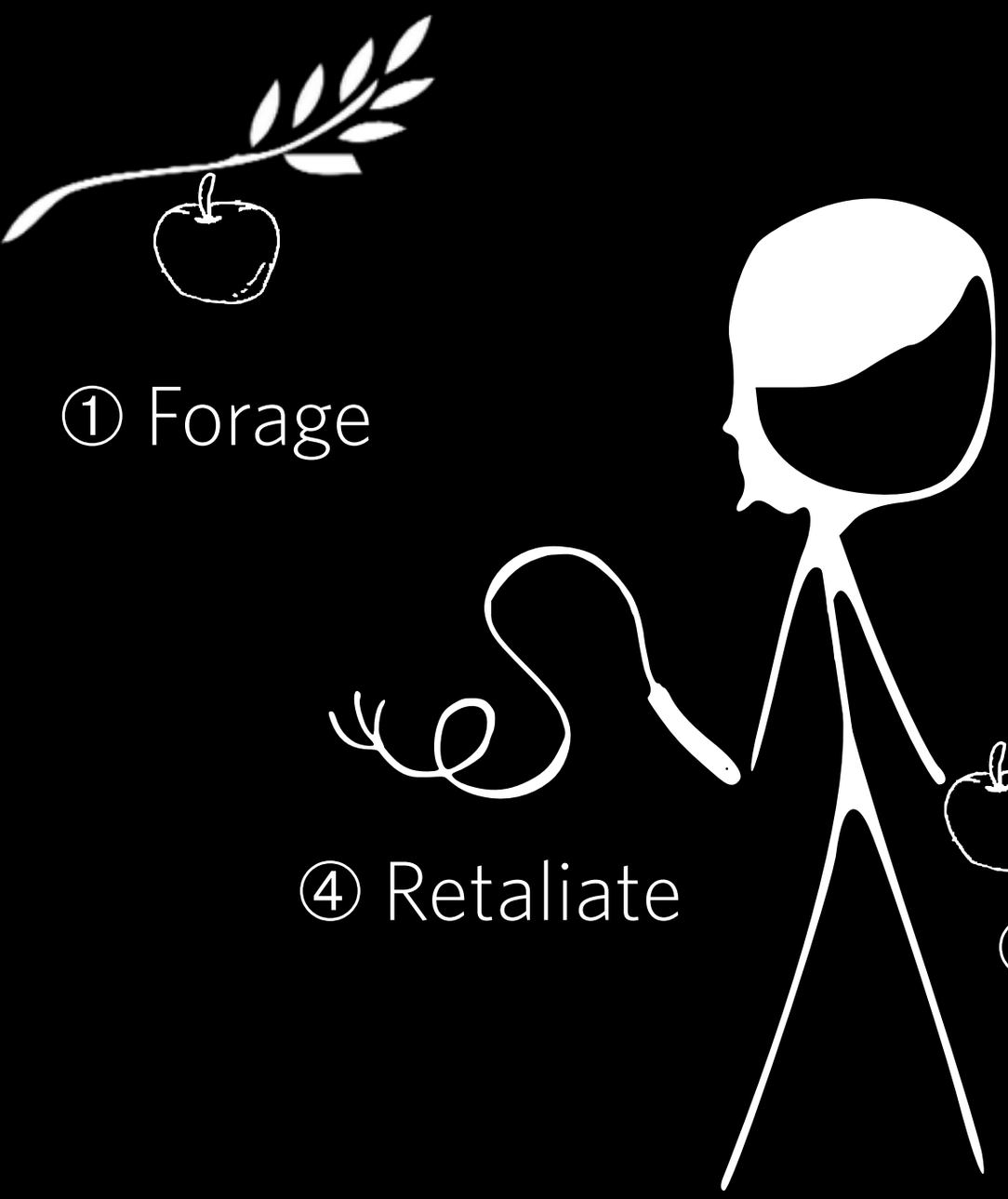
* *University of Zürich, Institute for Empirical Research in Economics, Blümlisalpstrasse 10, CH-8006 Zürich, Switzerland*
† *University of St Gallen, FEW-HSG, Varnbüelstrasse 14, CH-9000 St Gallen, Switzerland*

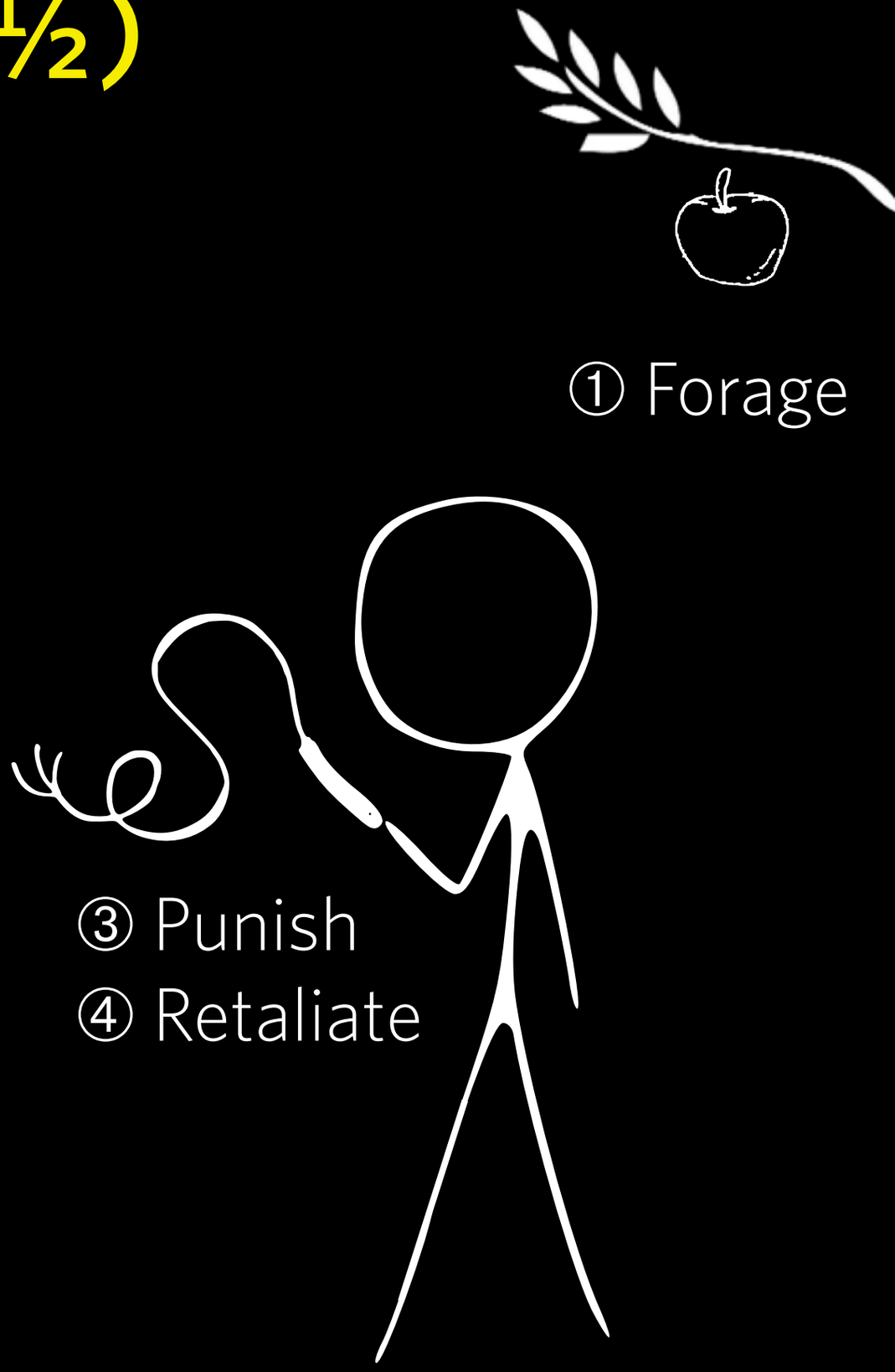
REVIEW ARTICLE

Punishment in animal societies

T. H. Clutton-Brock & G. A. Parker

the moral universe (½)





learning + deciding

thief

punisher

Expected Value = Reward × Probability

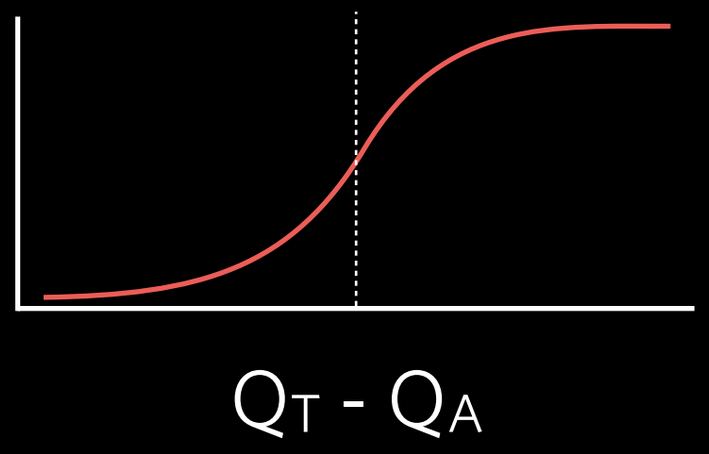
$$Q_T = B + C \times p(P|T)$$

$$Q_A = 0$$

$$p(T) = \frac{e^{Q_T}}{\Sigma e^{Q}}$$

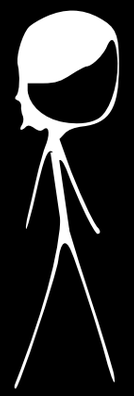
$$p(T) = \text{softmax}(Q_t)$$

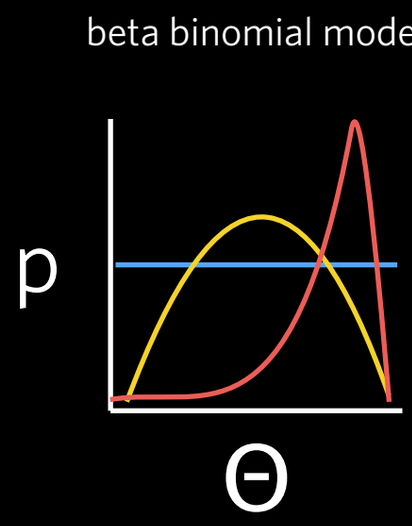
$\Theta = p(P|T)$



the bayesian solution

thief





punisher

allow to evolve

Θ = p(P|T)

$$p(\Theta) = \frac{N_P + \alpha}{N_P + \alpha + \beta}$$

Q_T =  - $\int \Theta \cdot$ 

$Q_T = B - \int \Theta \cdot C$

$p(T) = \text{softmax}(Q_T)$

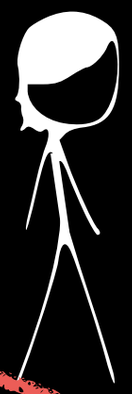
so what happens?

① Punishers evolve punishment

② Thieves learn restraint

the bayesian solution

thief



allow to evolve

punisher

allow to evolve

Θ = p(P|T)

$$p(\Theta) = \frac{N_P + \alpha}{N_P + \alpha + \beta}$$

$$Q_T = \text{🍎} - \int \Theta \cdot \text{〰}$$

$$Q_T = B - \int \Theta \cdot C$$

$$p(T) = \text{softmax}(Q_T)$$

test your intuitions

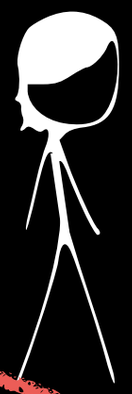
What values would ensure theft?

What values would prevent theft?

What values maintain learnability?

the bayesian solution

thief



allow to evolve

punisher

allow to evolve

$\Theta = p(P|T)$

$$p(\Theta) = \frac{N_P + \alpha}{N_P + \alpha + \beta}$$

$$Q_T = \text{apple} - \int \Theta \cdot \text{squiggle}$$

$$Q_T = B - \int \Theta \cdot C$$

$$p(T) = \text{softmax}(Q_T)$$

so what happens?

① Punishers evolve punishment

② Thieves learn restraint

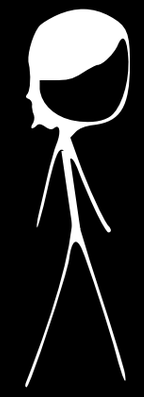
the bayesian solution

PROBLEMS

① an awful lot of mental accounting to do

② a fear of god or a taste for justice?

thief



steal

0

1.5

3

1.8

2

Q_T 2.9

4

$p(T) = \text{softmax}(Q_t)$

abstain

0

.5

1

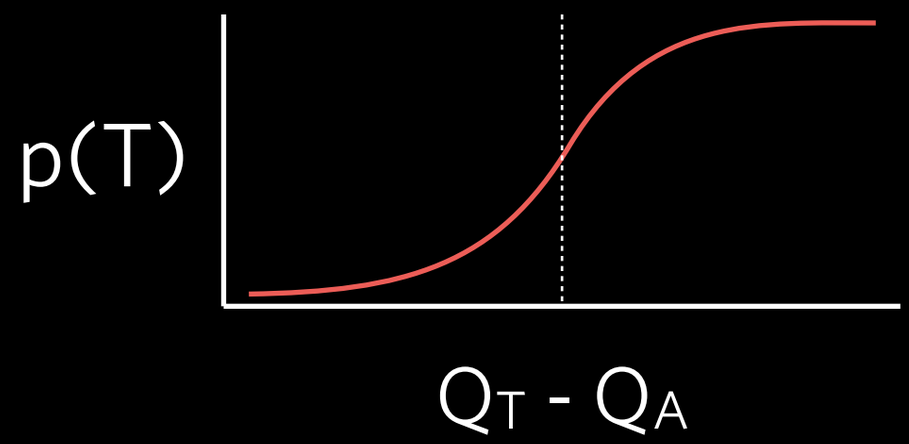
.8

2

.9

1

Q_A

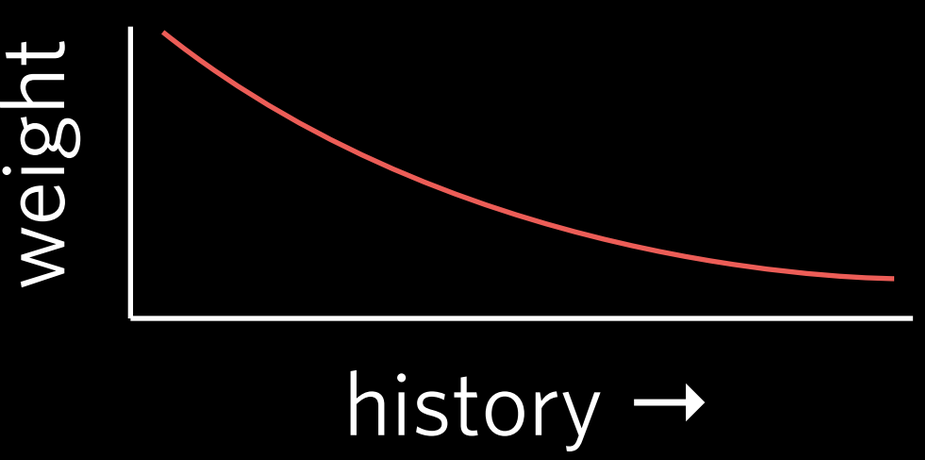


the reinforcement learning solution

prediction error learning
(c.f. rescorla-wagner)

$PE = R - Q_t$ $\quad$ $R = B - C$

$Q_t \leftarrow Q_t + \alpha(PE)$



the bayesian solution

PROBLEMS

① an awful lot of mental accounting to do

② a fear of god or a taste for justice?

Reinforcement Learning

Reward

② Steal

③ Punish

Learning

Value

① Forage

② Steal

③ Punish

④ Retaliate

Deciding

Behavior

① Forage

② Steal

③ Punish

④ Retaliate

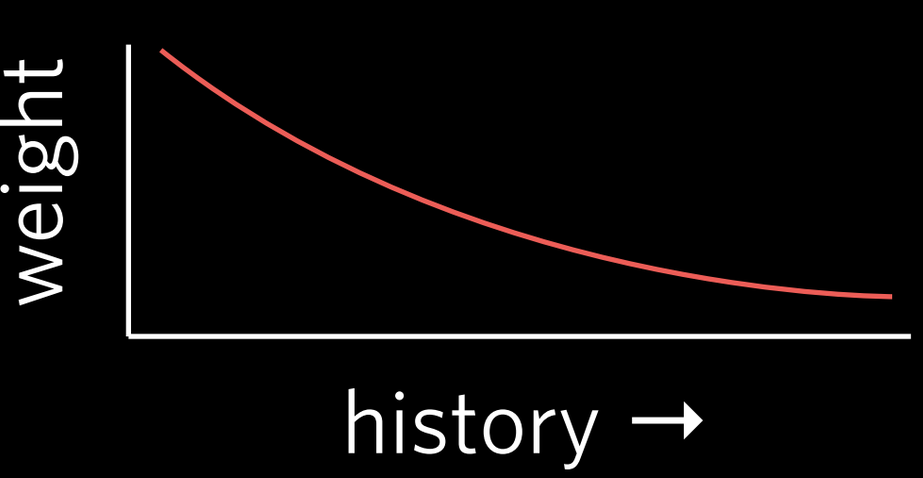
Evolution

the reinforcement learning solution

prediction error learning
(c.f. rescorla-wagner)

$PE = R - Q_{t-1}$ $R = B - C$

$Q_t \leftarrow Q_t + \alpha(PE)$



thief

punisher

allow the reward
function to evolve

theft / punishment

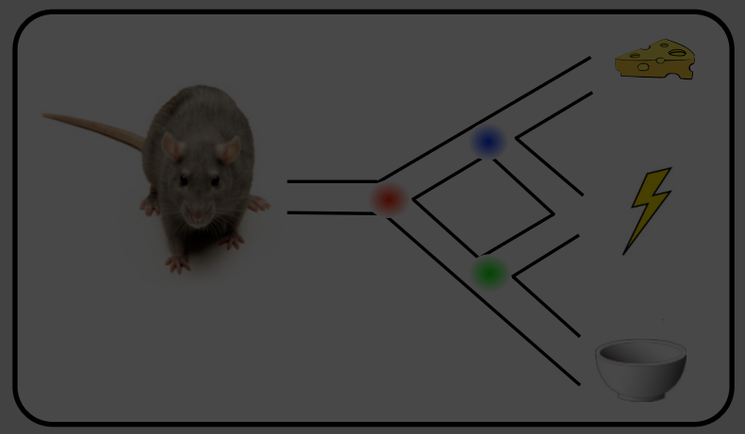
so what happens?

① Resolute punishment (innately rewarding)

② Flexible theft (“objective” rewards only)

Temporal Difference Learning

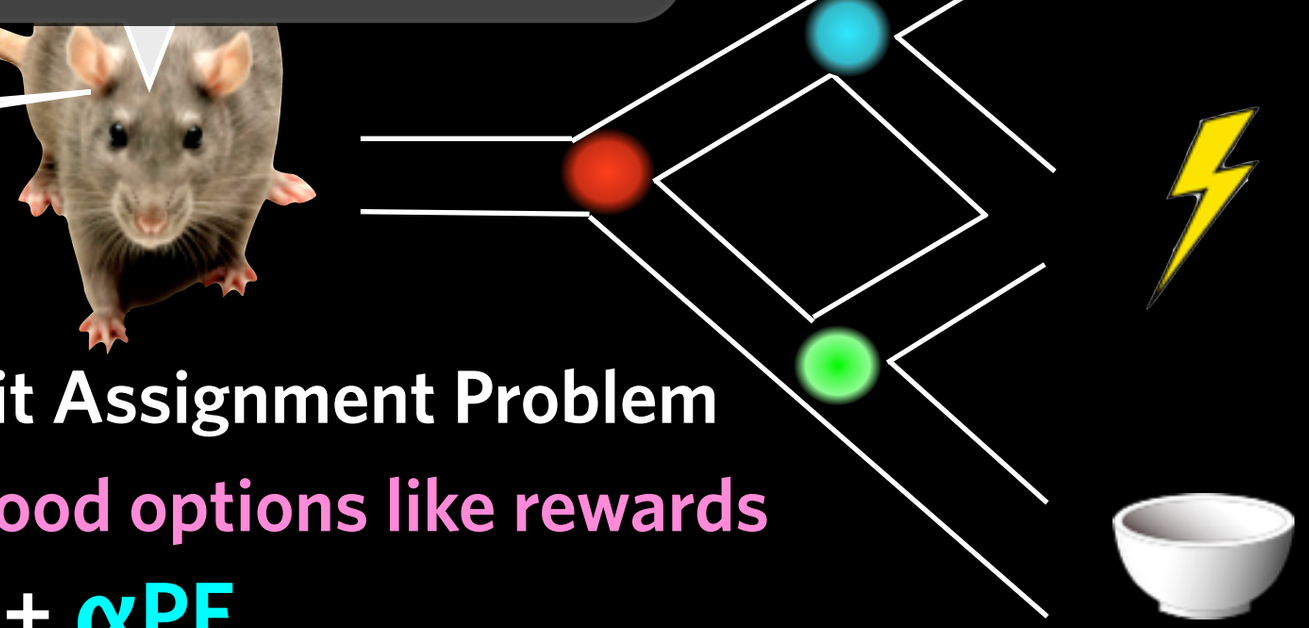
L =
R =
L =
R =
L =
R =



What should I do?

1. L L +
2. L R -
3. R L -
4. R R 0

Model Based Reinforcement Learning



L = +
R = 0

L = +
R = -

L = -
R = 0

Model-Free Reinforcement Learning

Temporal Credit Assignment Problem

Solution: Treat good options like rewards

$$Q_{(s,a)} \leftarrow Q_{(s,a)} + \alpha PE$$

$$Q_{(s,a)} \leftarrow Q_{(s,a)} + \alpha(R_{s'} + Q_{(s',a')} - Q_{(s,a)})$$

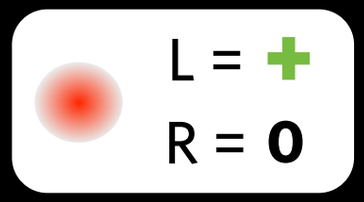
$$0 = Q_{(\bullet,L)} \leftarrow Q_{(\bullet,L)} + \tfrac{1}{2}(R_{\bullet} + Q_{(\bullet,L)} - Q_{(\bullet,L)})$$

$$2 = Q_{(\bullet,L)} \leftarrow Q_{(\bullet,L)} + \tfrac{1}{2}(0 + Q_{(s',a')} - Q_{(\bullet,L)})$$

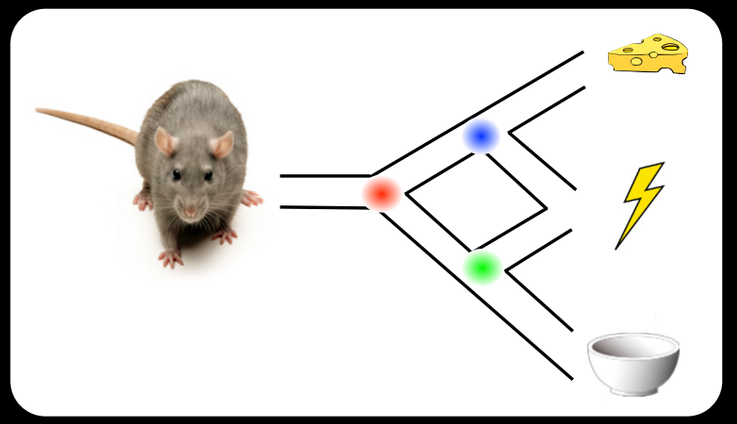
0 4 0

From Association to Action

e.g. Dickinson et al 1995

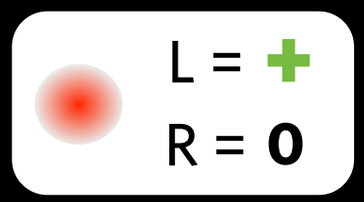


Model-Free
Reinforcement Learning

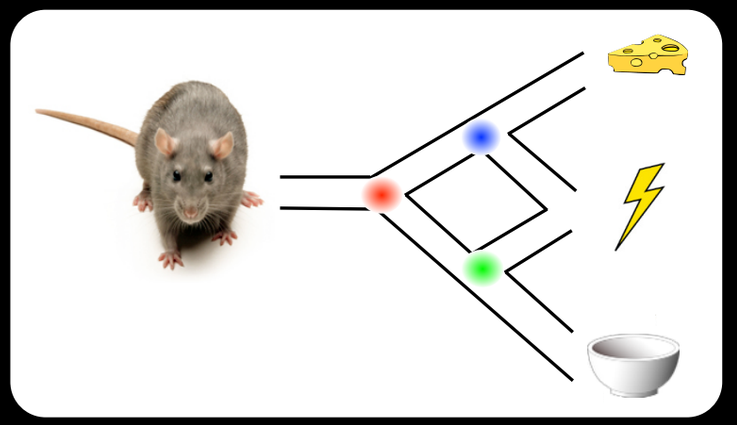


Model
Based
Reinforcement
Learning

From Association to Action



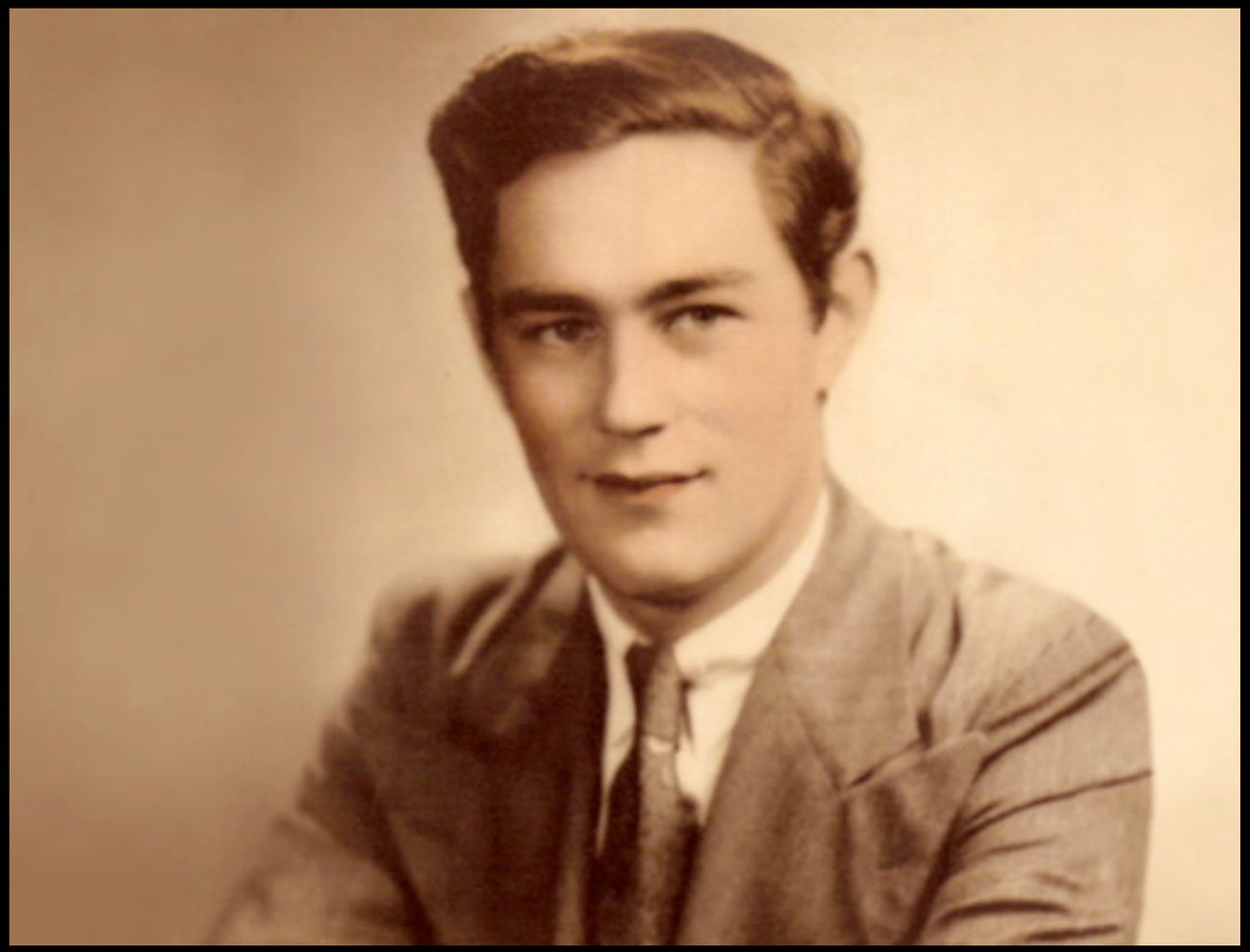
Model-Free
Reinforcement Learning

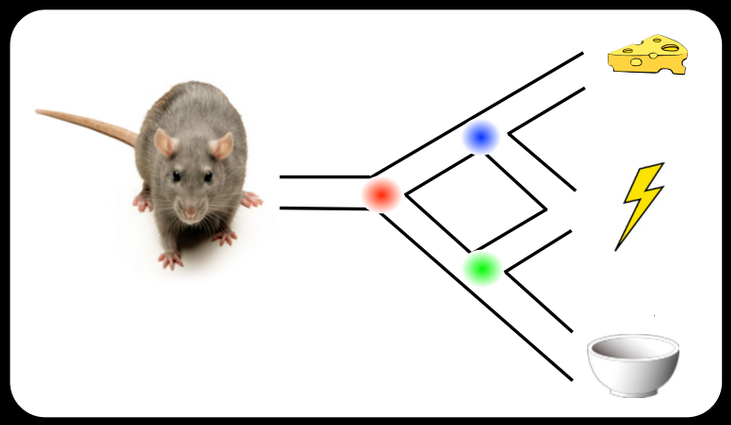


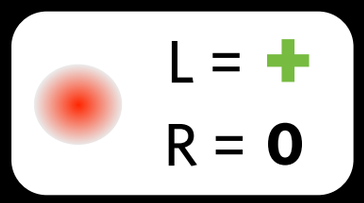
Model
Based
Reinforcement
Learning

Prediction of value, but no model of outcome

Memory: Patient HM

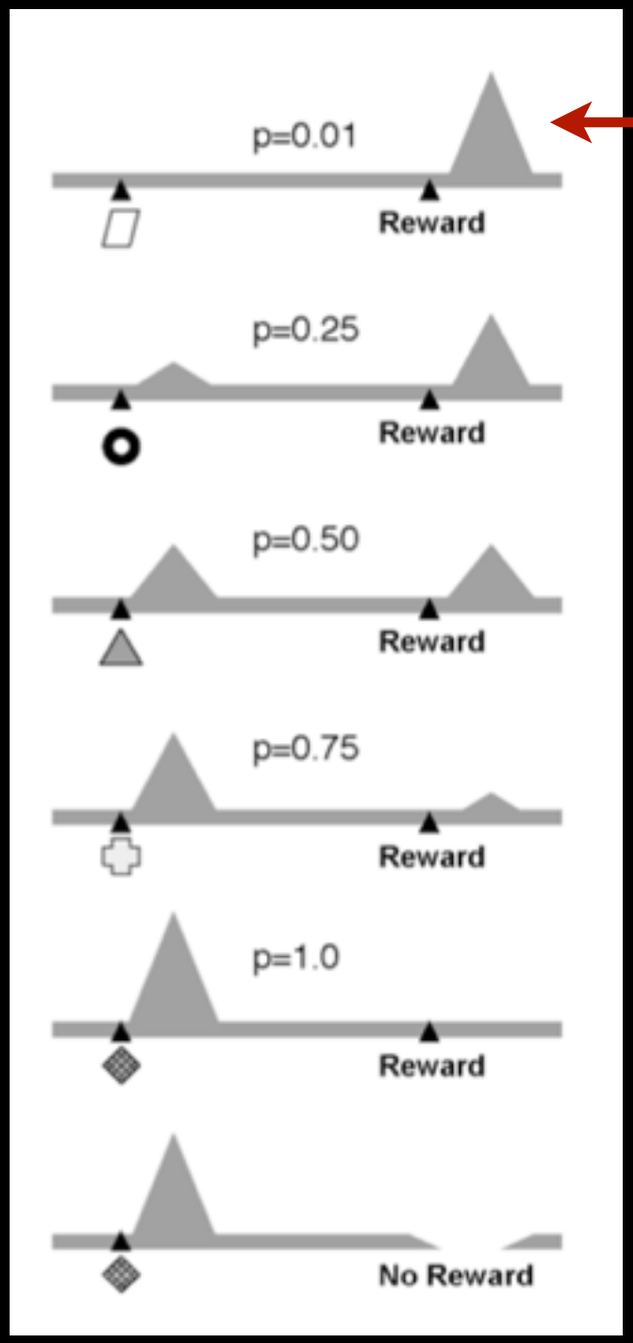




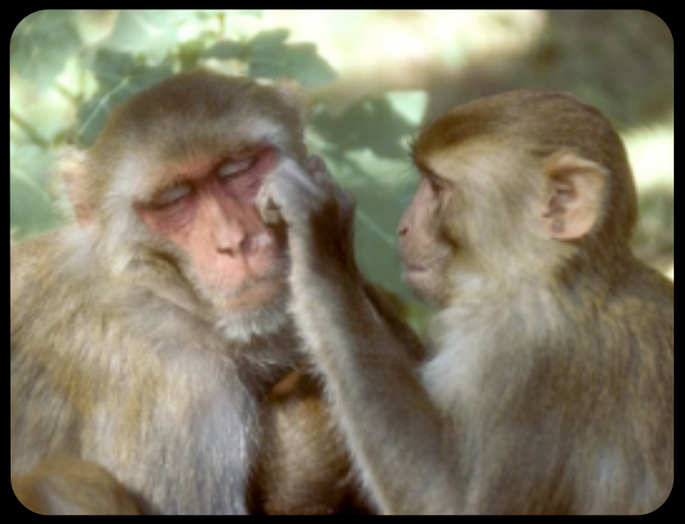


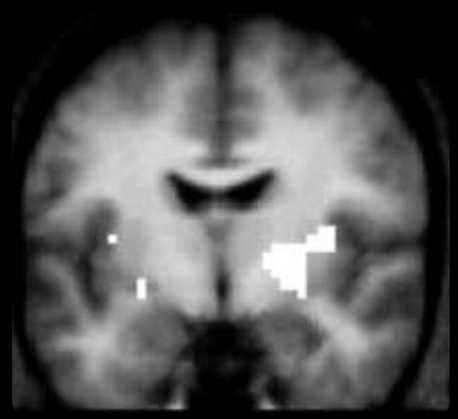
Neural Mechanisms of Model-Free Learning

Graphic borrowed from Daw & Shohamy 2008
For experiments, see e.g. Shultz et al 1996

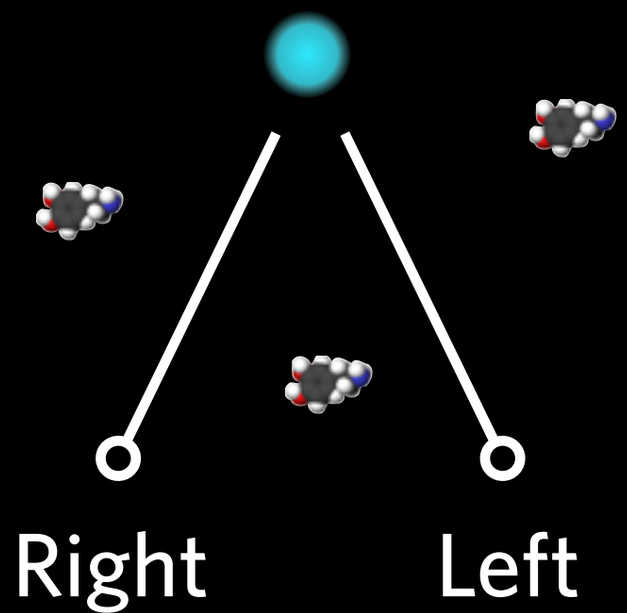


Dopamine = prediction error signal



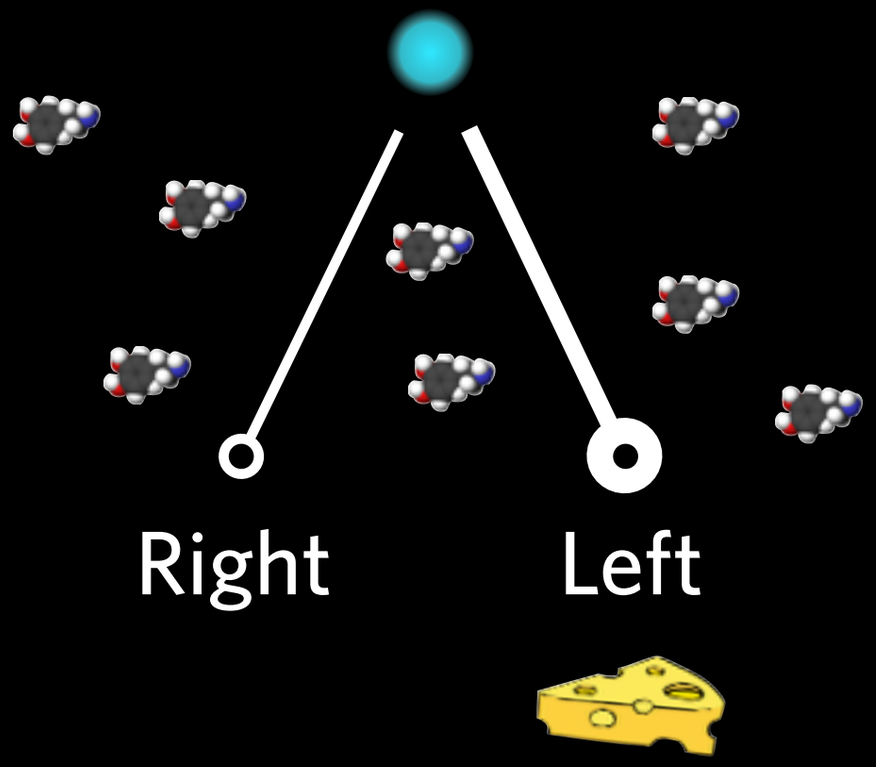


Model-Free Learning



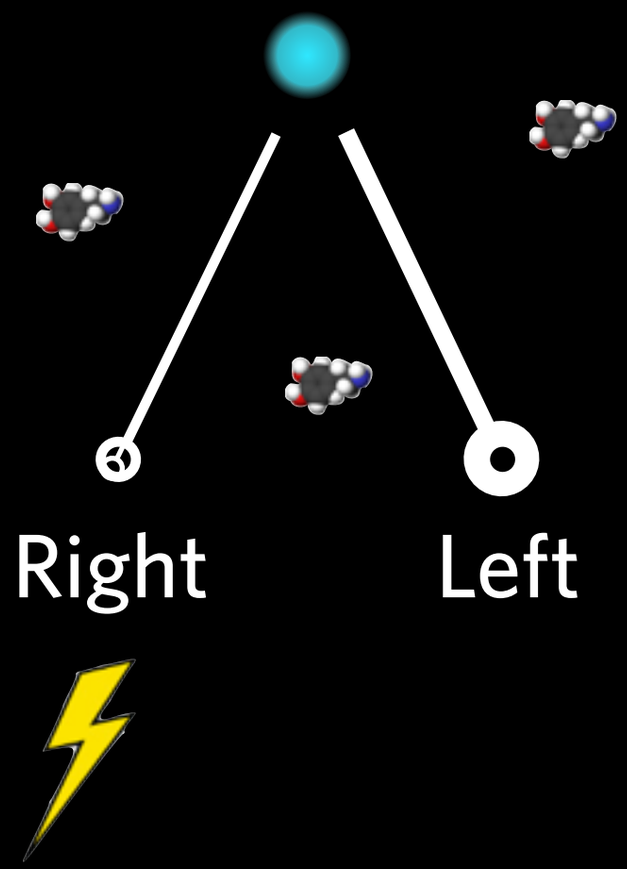
Nucleus Accumbens
of the Striatum

Model-Free Learning



Nucleus Accumbens
of the Striatum

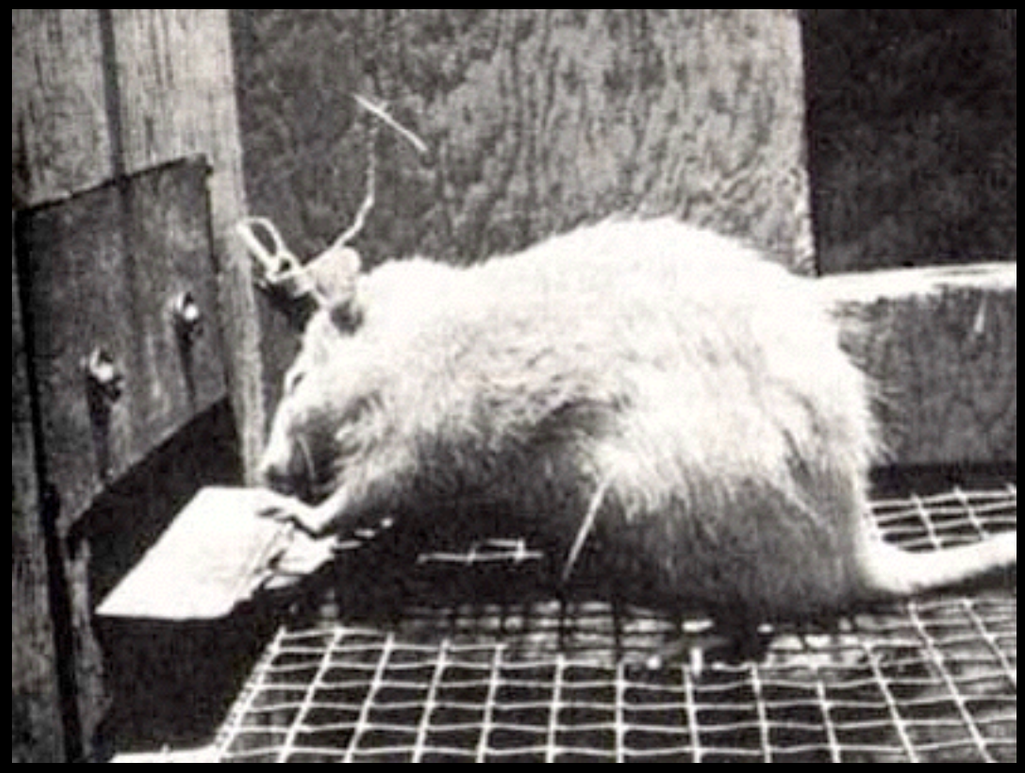
Model-Free Learning



Nucleus Accumbens
of the Striatum

Application: Why is cocaine addictiQe?

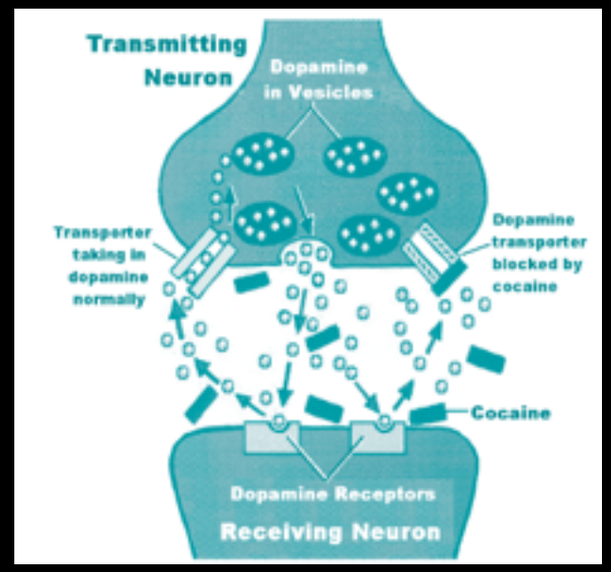
Redish, 2004

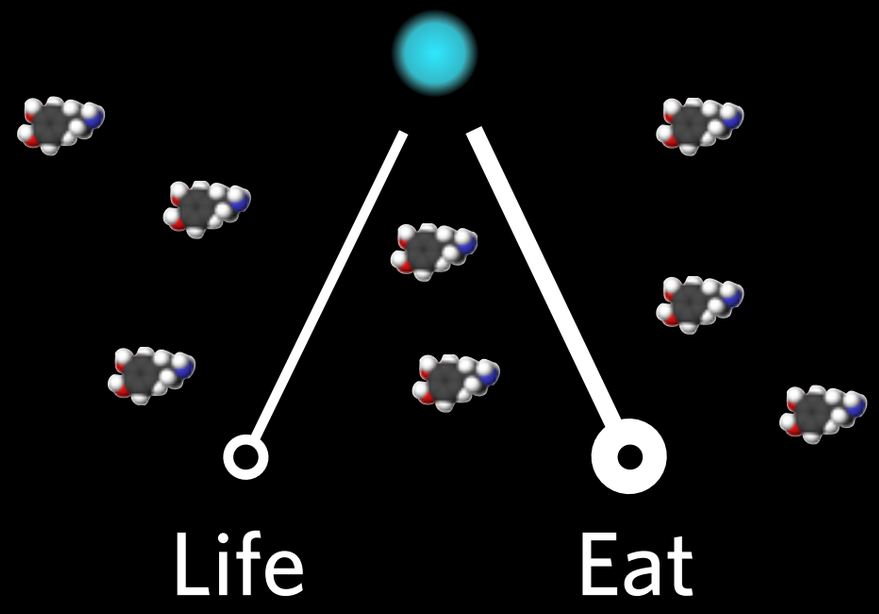


Olds & Milner 1964

Application: Why is cocaine addictiQe?

Redish, 2004

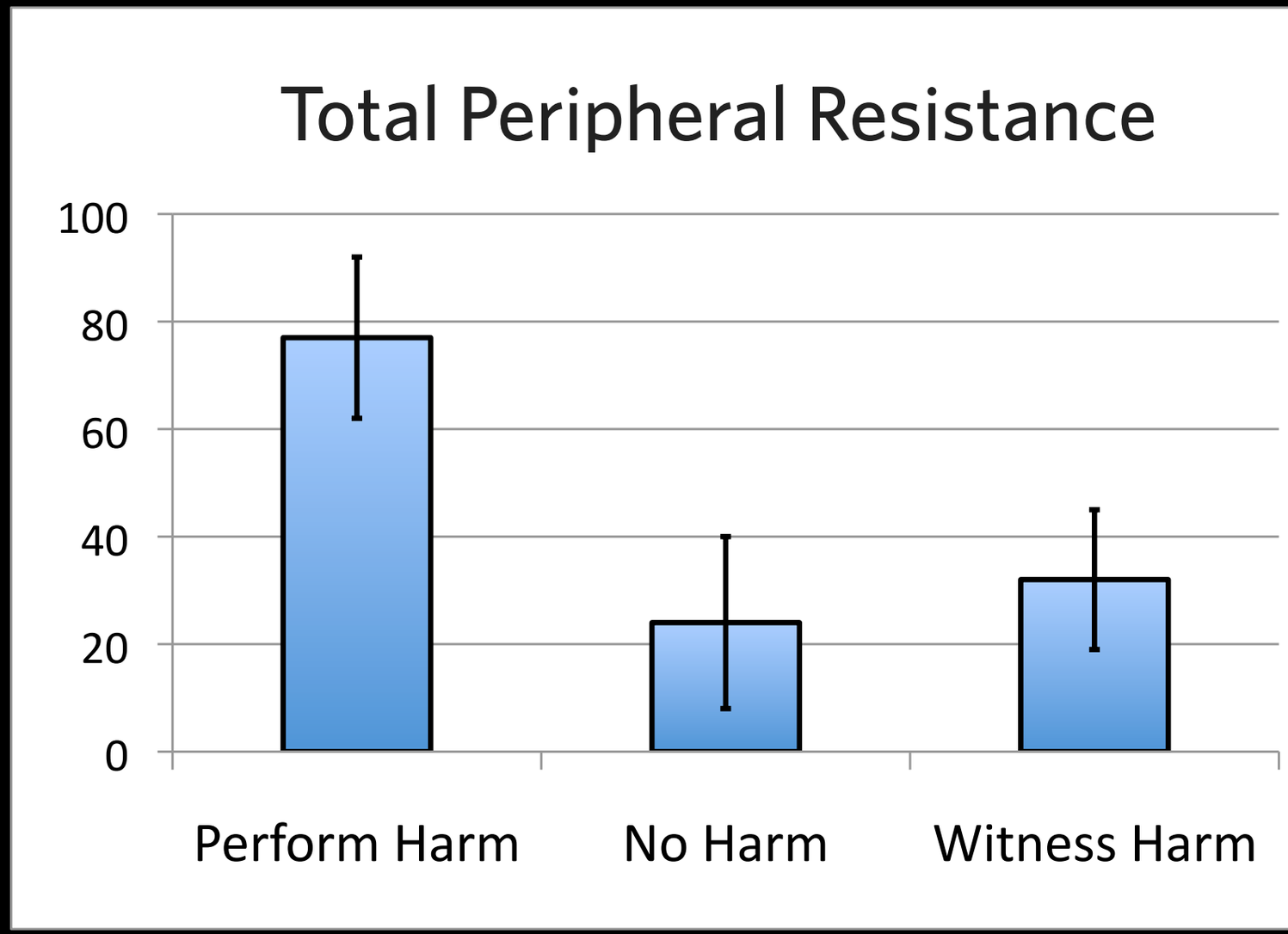


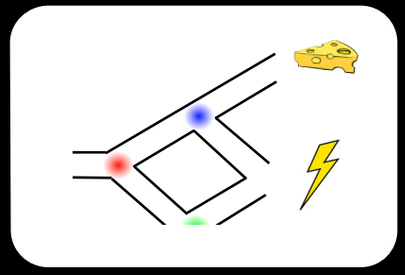


Prediction Error Learning

The AQersion to Harm

Wendy Mendes





Cushman, Gray, Gaffey & Mendes, 2012

Dual Process Morality

Josh Greene