Neuronal reward mechanisms underlying reinforcement learning

Wolfram Schultz

University of Cambridge

www.wolframschultz.org

What are rewards?

Overall function: keep gene carriers (agents) alive and ensure propagation of their genes into the next generation.

Daily function: provide essential substances for survival and activities for gene propagation.

Rewards are all attractive stimuli, events, objects, situations and activities that are evolutionary beneficial.

Thus, rewards are not defined by their physical and chemical properties but by their usefulness for the survival and gene propagation of biological agents.


当店の人気盛り込み
Food
Liquid, alcohol, taste, aesthetics
Ice cream
Addicted
Art - aesthetics
Sex
South Africa
Foraging albatrosses
Risk
Mental rewards
Rosetta Stone
J-F Champollion
Prince Edouard Is
Crozet Is
Kerguelen Is
Heard Is
60°S
Antarctica
NYSE
New York Stock Exchange
August Macke
Friends

Rewards have three principal behavioural functions

Learning (positive reinforcement)

Testable using experimental psychology:
Pavlovian and operant conditioning,
based on animal learning theory.

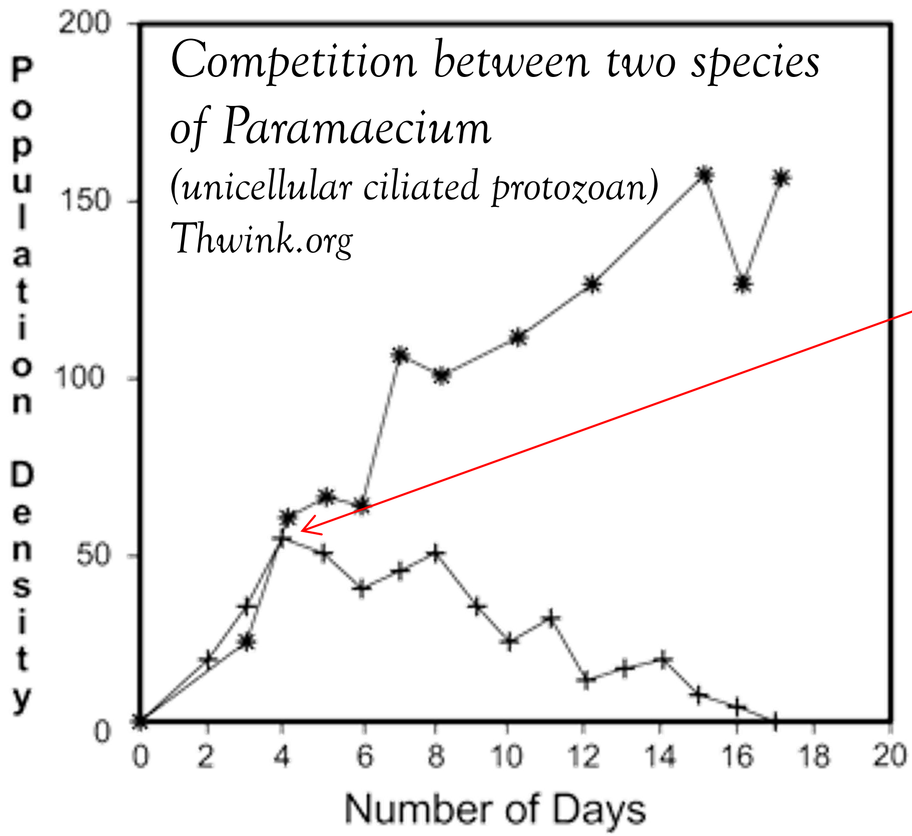
Approach behaviour and economic decisions

Rewards are attractive, worth working for.
Testable using experimental economics,
based on economic decision theory.

Positive emotions & mental states

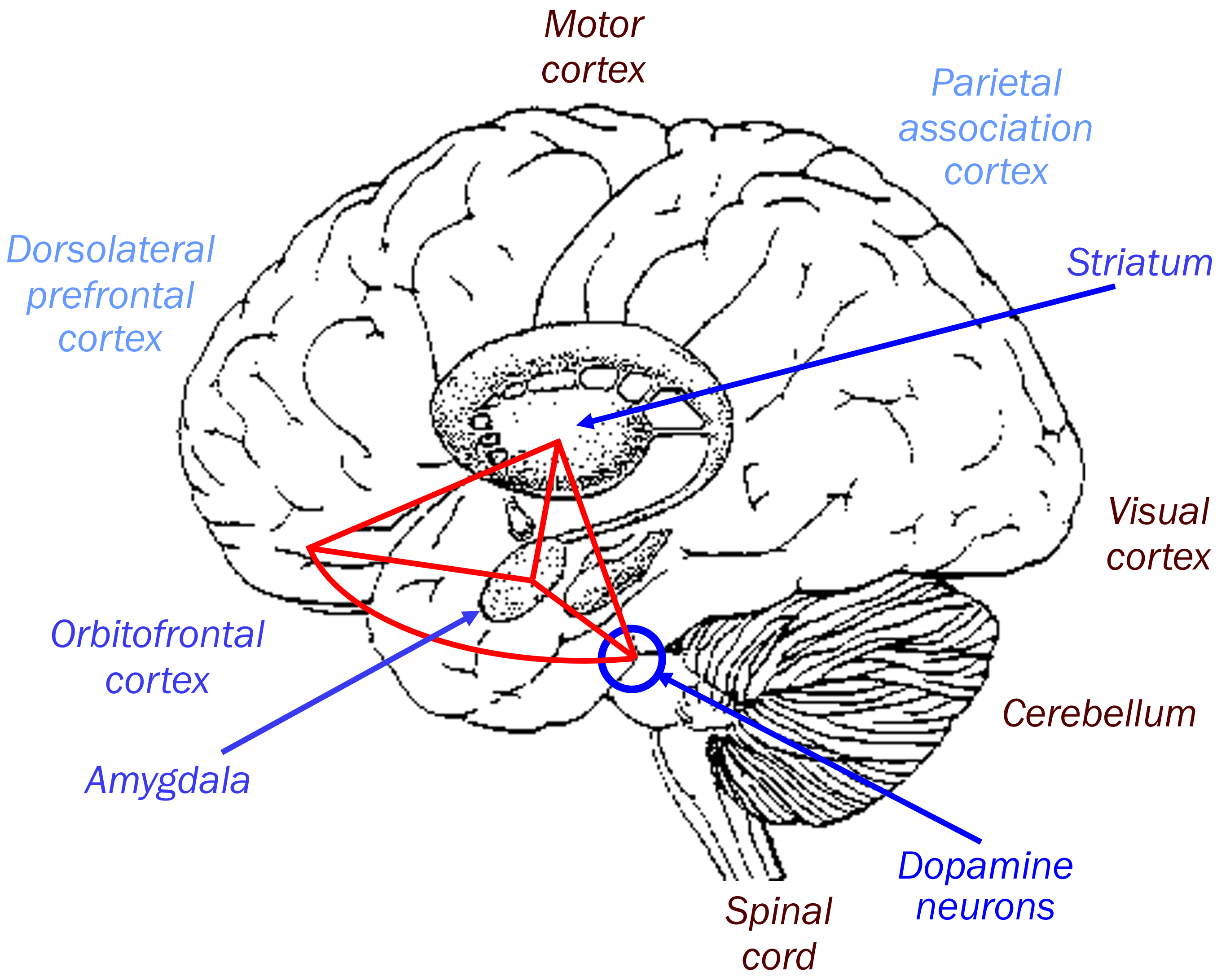
Pleasure (~liking) reaction => state of happiness
Desire (~wanting) => goal, purpose, free will

Utility maximisation as basis for evolutionary fitness: Surviving by getting more reward than others



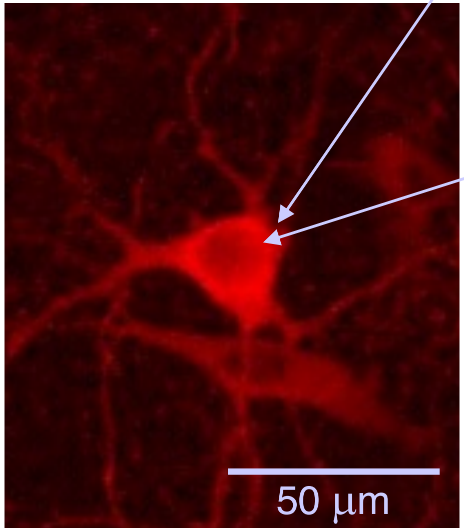
Both did well for four days, then one species disappeared, whereas the other survived.

Principal brain structures for reward

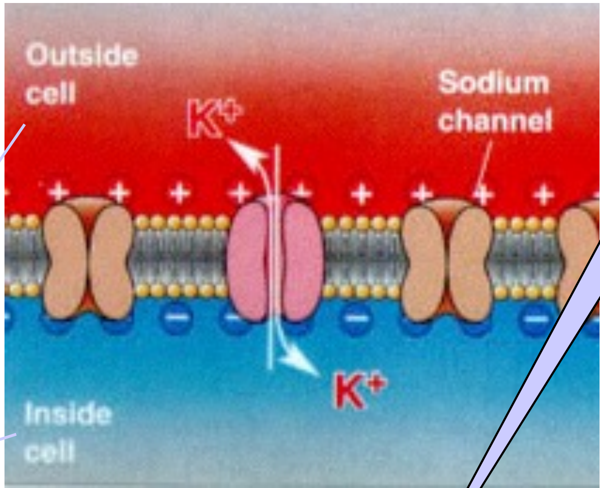


Extracellular recordings from individual dopamine neurons

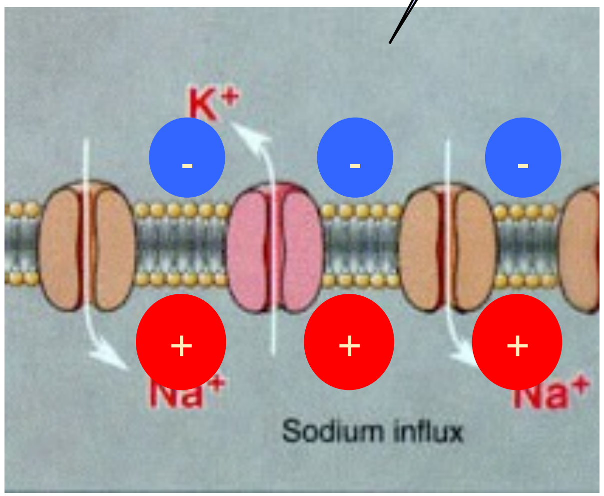
Definition:
A dopamine neuron is a neuron that releases a neurotransmitter called dopamine.

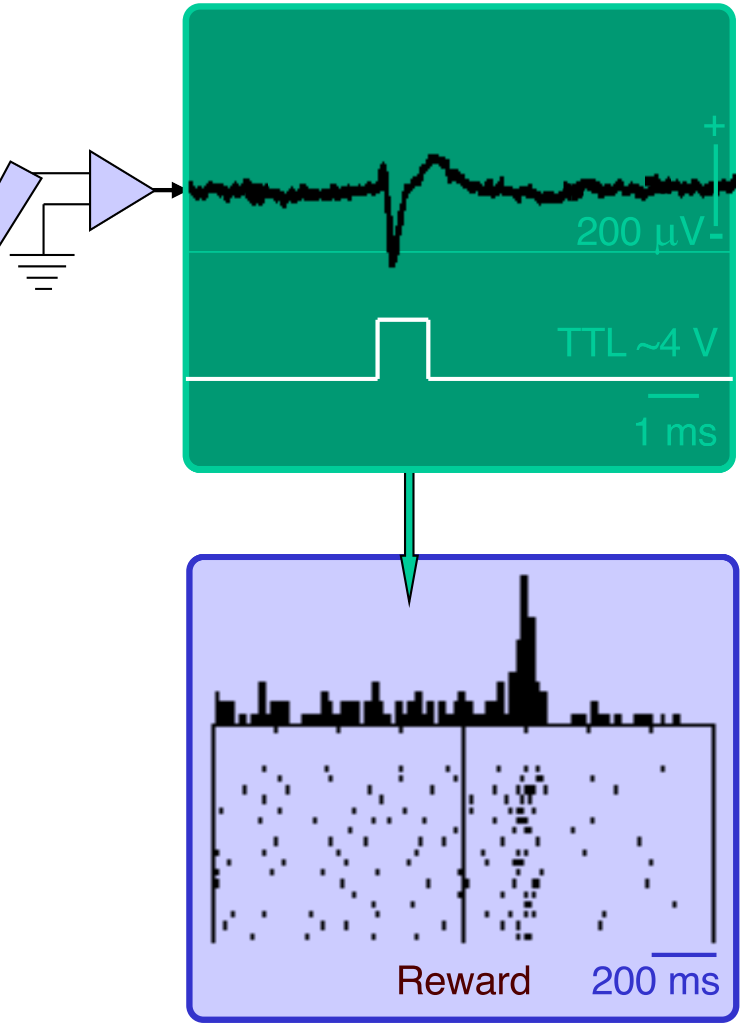


Resting state



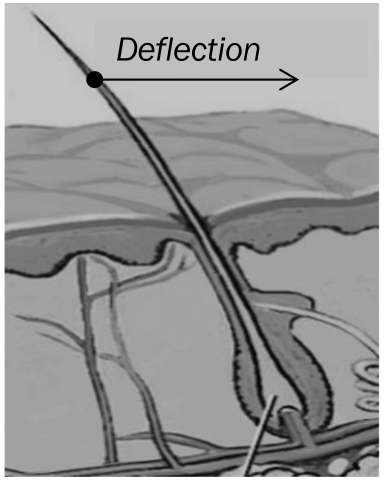
Excited state (action potential)

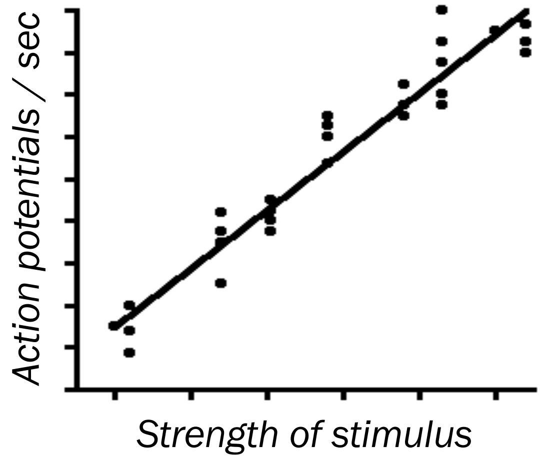




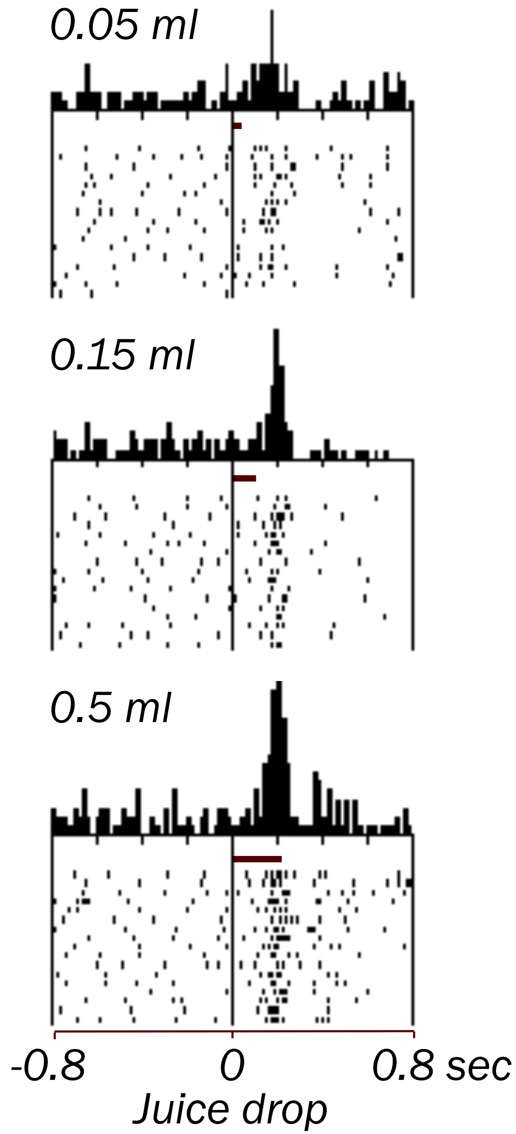
The intuitive metric of neuronal information is a rate code: number of action potentials/second.

The neuronal rate code originating from the opening of Na–channels in sensory receptors serves as a neuronal metric for stimulus strength (Adrian & Zotterman 1926).





The neuronal reward signal: action potentials provide a rate code for reward.



Tobler, Fiorillo & Schultz 2005

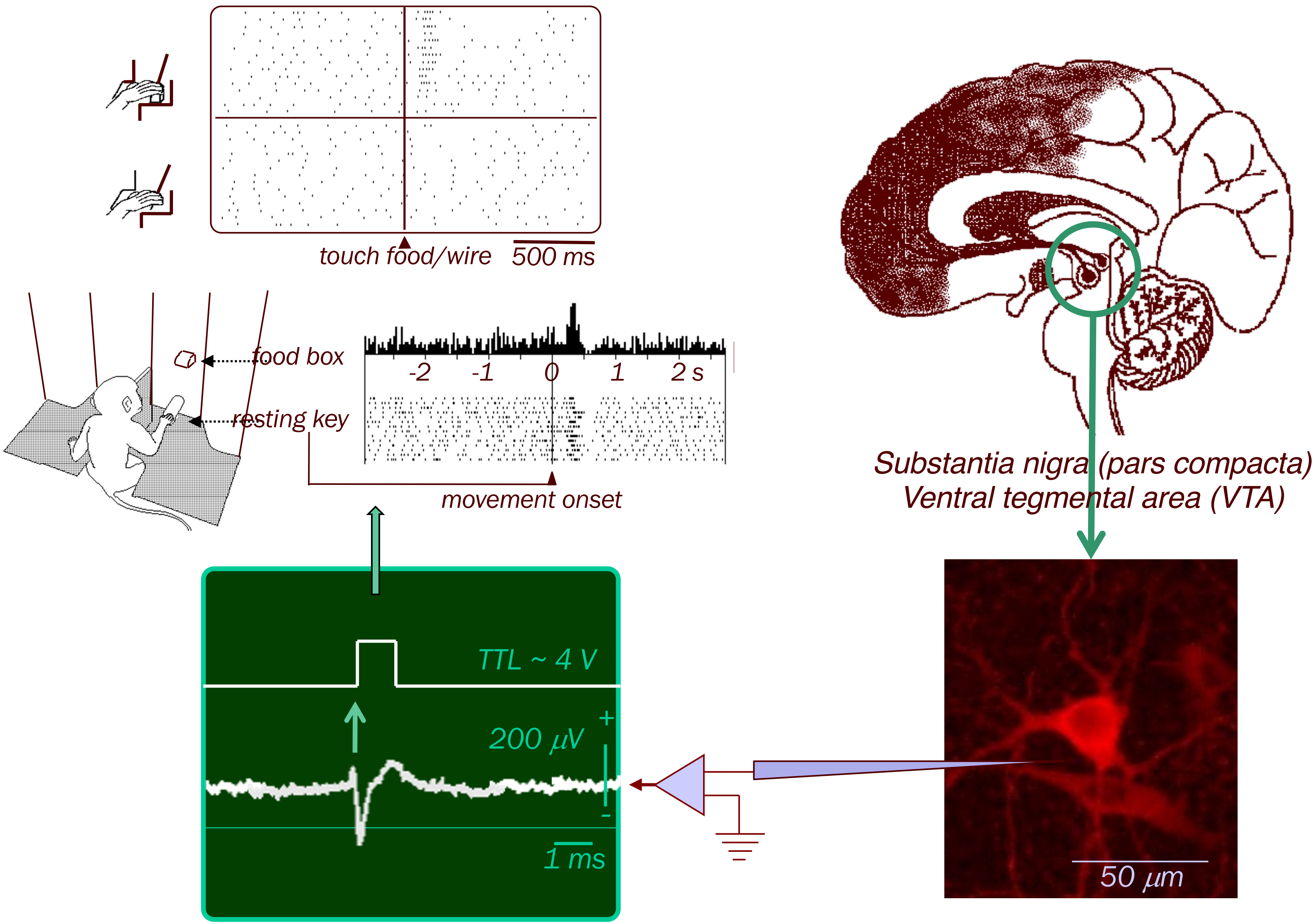
Behavioural reward functions

Learning

Approach & choice

Positive emotions

The dopamine reward signal

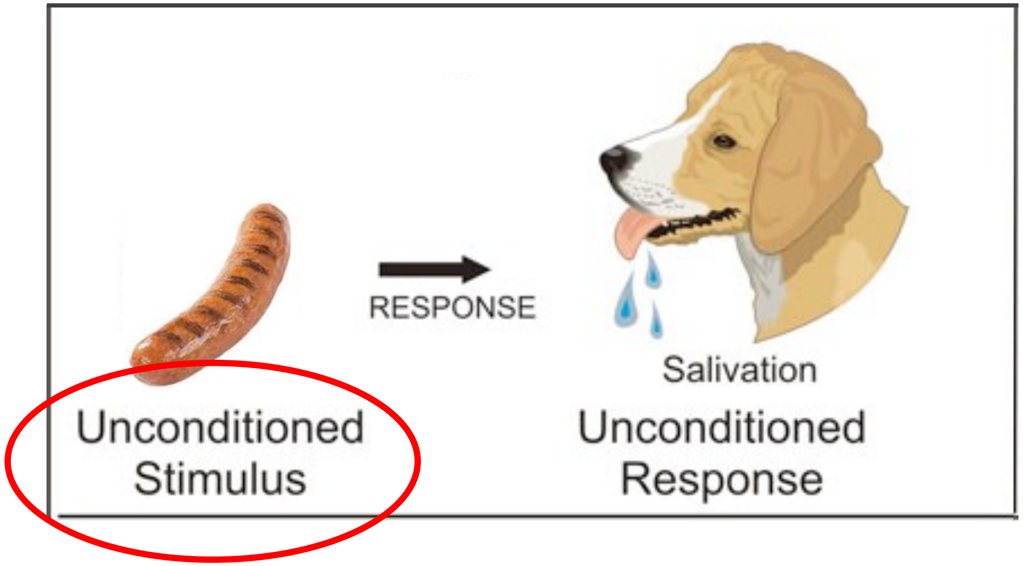


Romo & Schultz J Neurophysiol 1990

Pavlovian conditioning

Making a stimulus predictive

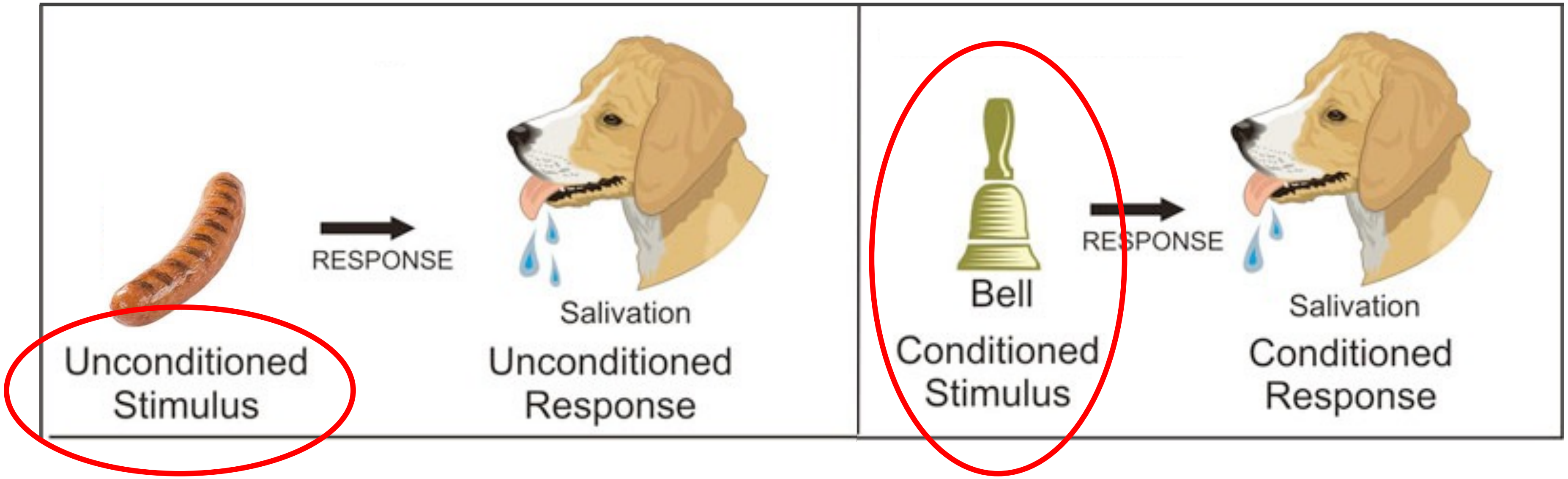
Before



Pavlovian conditioning

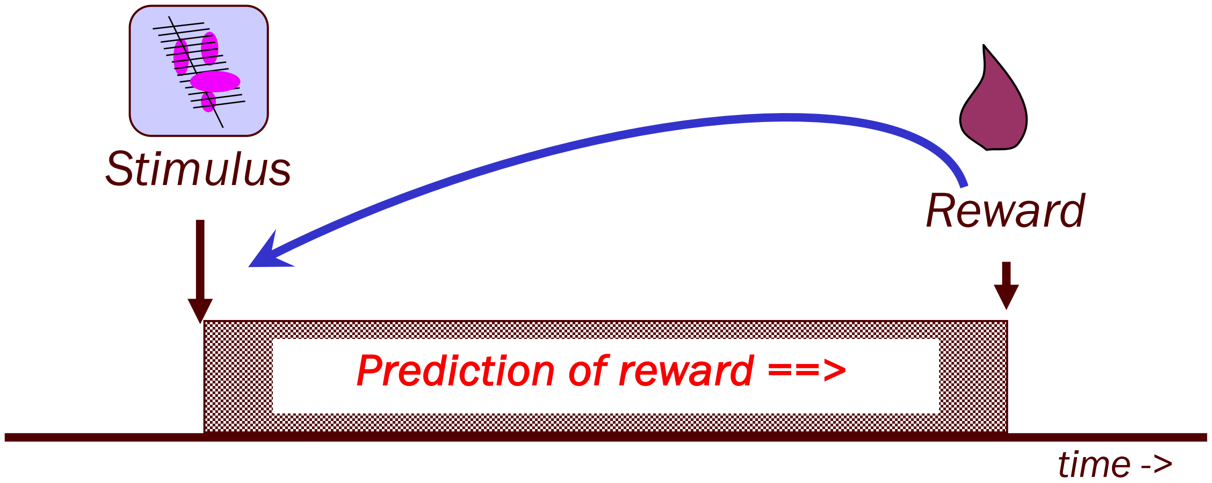
Making a stimulus predictive

Before *After*

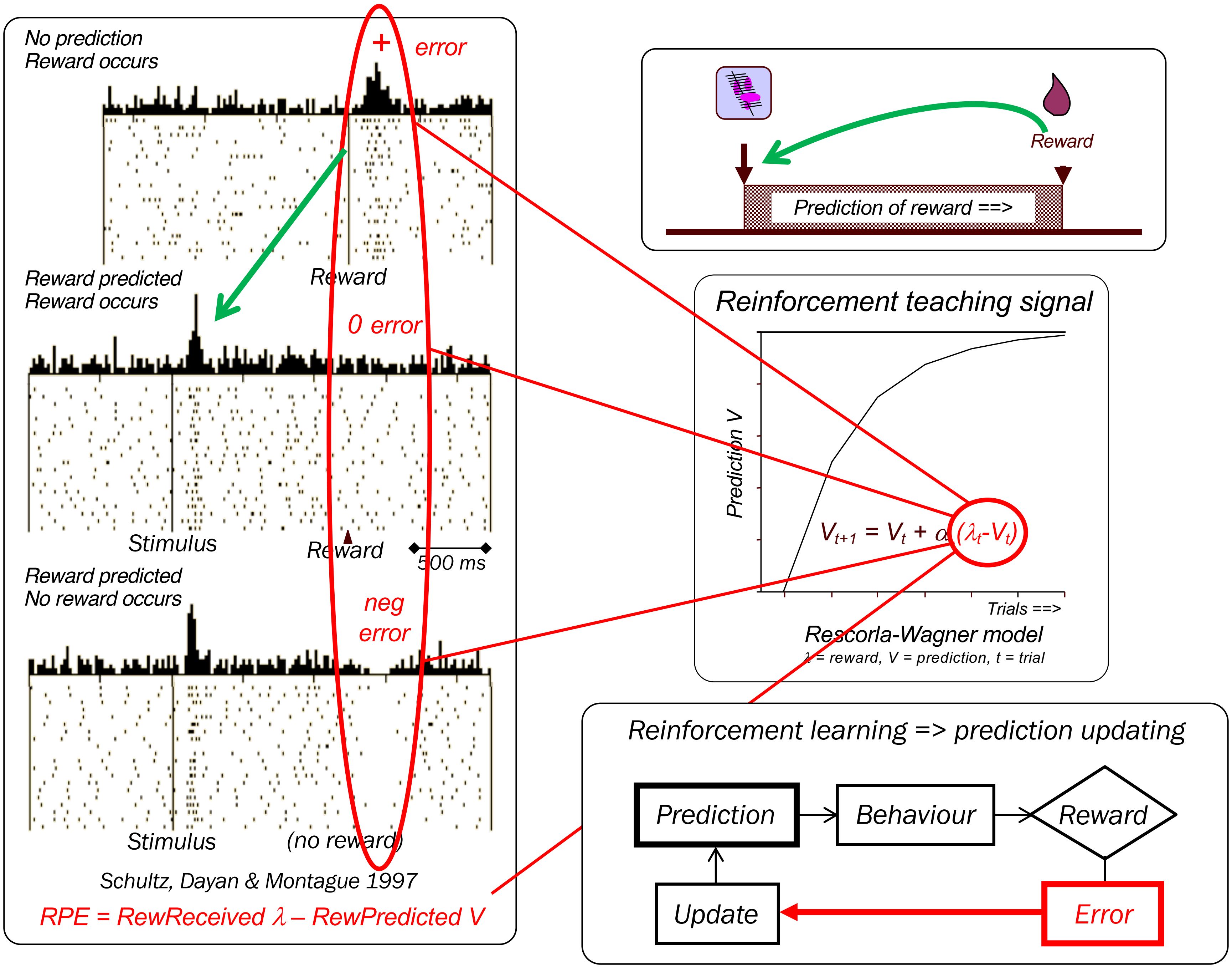


Pavlovian conditioning

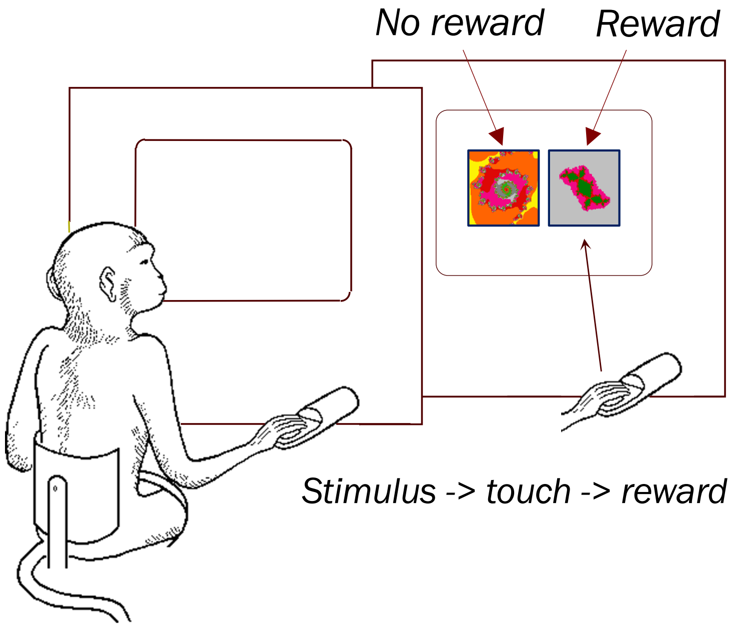
Making a stimulus predictive

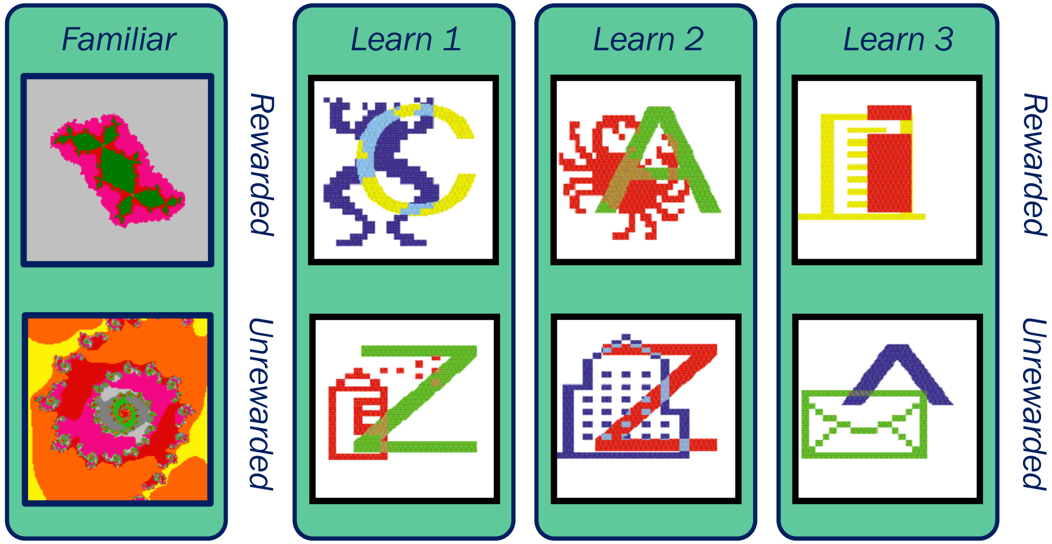


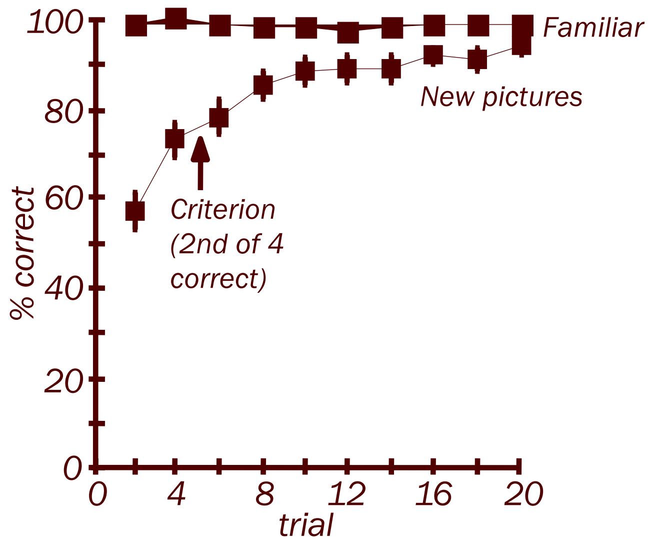
Dopamine neurons report reward prediction errors (RPE).

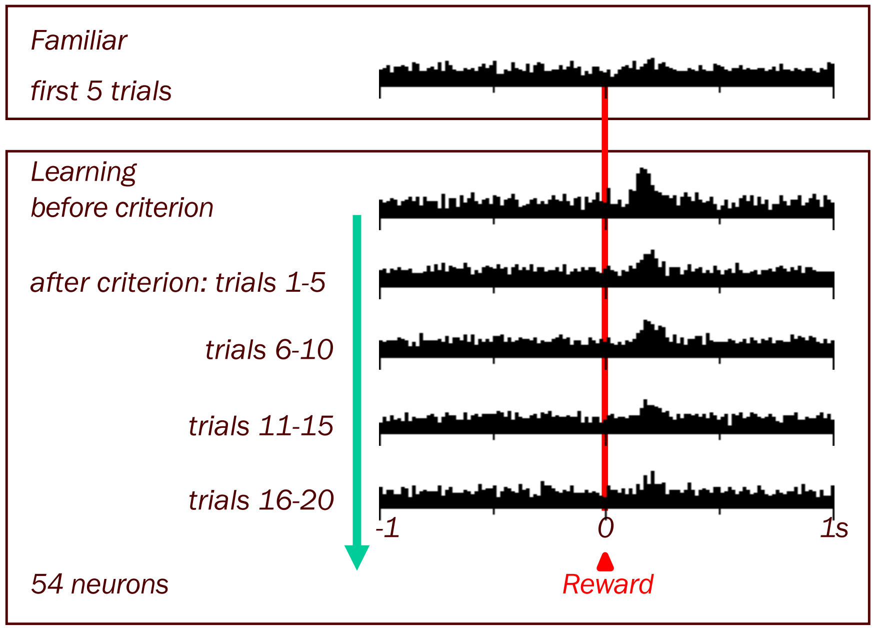


Positive dopamine prediction error signal during learning



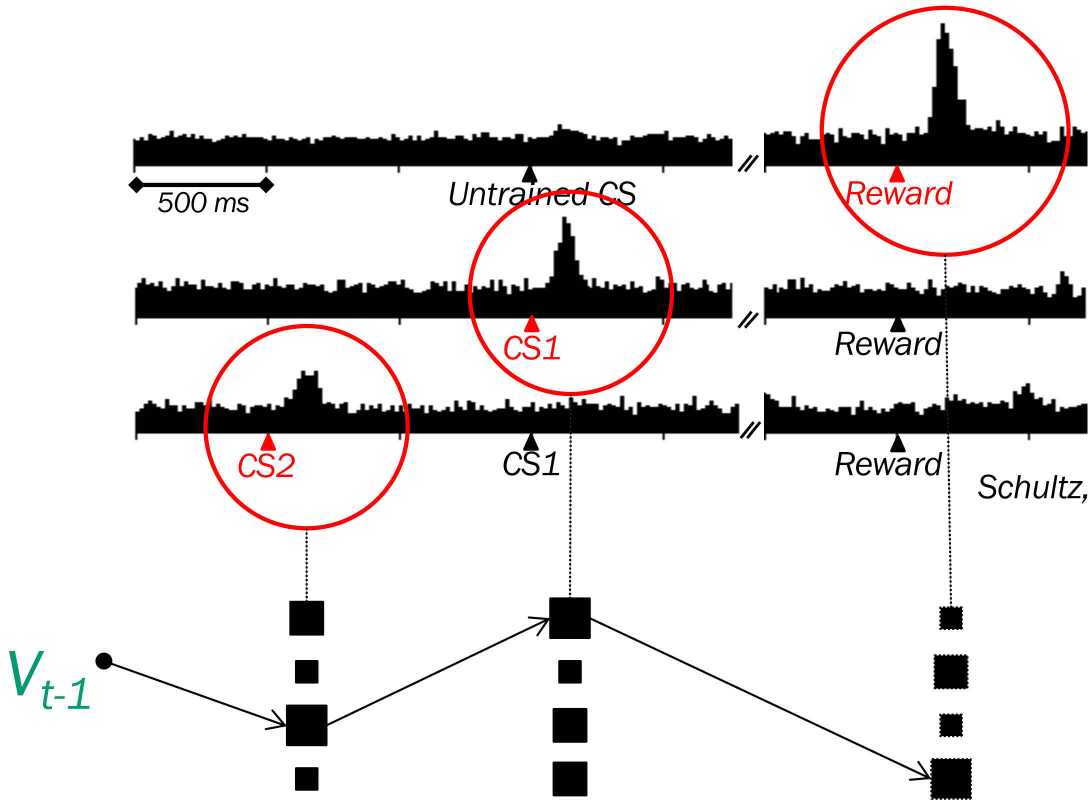






Hollerman & Schultz Nature Neuroscience 1998

Dopamine neurons report RPEs for higher-order rewards, complying with Temporal Difference (TD) learning.

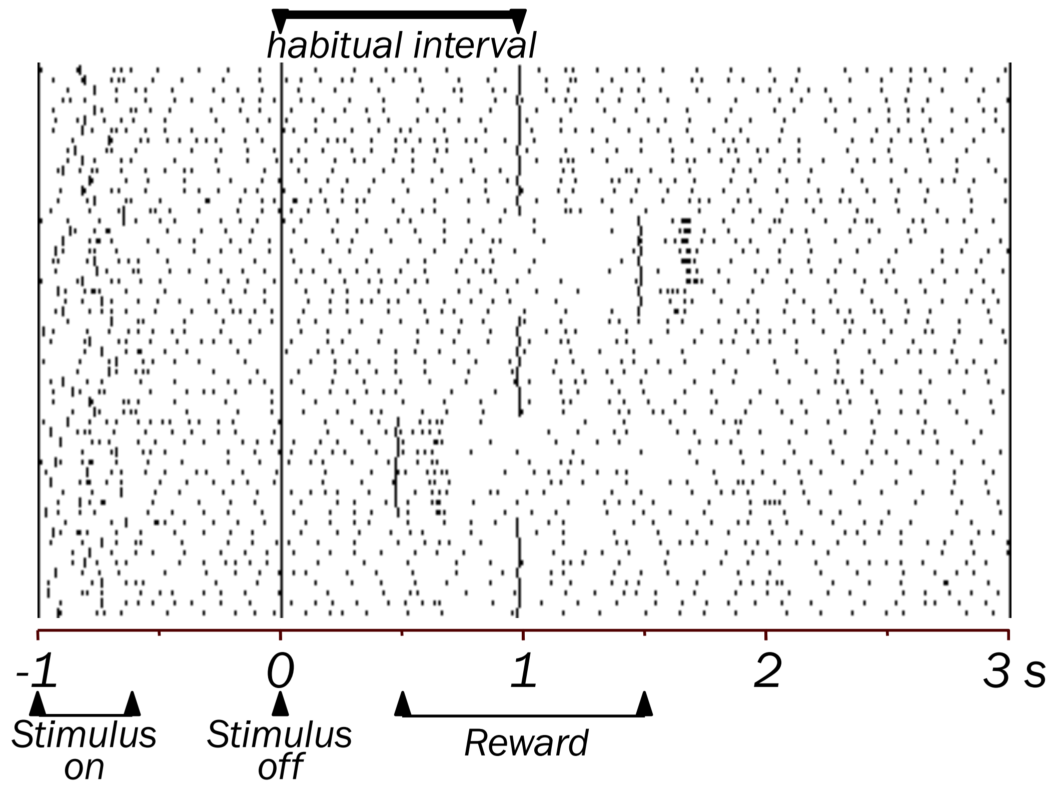


Schultz, Apicella & Ljungberg JN 1993
Schultz JNP 1998

TD

$$\text{Prediction Error} = (V_t + \gamma V_{t+1} + \gamma^2 \lambda_{t+2}) - V_{t-1}$$

Time sensitivity of dopamine signal: excitation with unpredicted reward, and inhibition with reward omission at time of expected reward



The dopamine reward signal reflects RPE not just across trials (λ – V; Rescorla-Wagner RL) but RPE across time steps (Δv/Δt; Temporal Difference RL)

Hollerman & Schultz NN 1998

Maximising reward via Machine Learning

Current reward

Discounted sum of all future rewards

Bellman Equation and Dynamic Programming define optimal value function V_t (1956).

$$V_t = max\, (\lambda_t + \gamma^k \sum_{k=1}^{\infty} V_{t+k}\,)$$

Temporal Difference Learning (TD) achieves optimal value function (Sutton & Barto 1981).

Richard Sutton

Andrew Barto

$$\text{TD Prediction Error} = (\lambda_t + \gamma^k \sum_{k=1}^{\infty} V_{t+k}\,) - V_t$$

(how far away from V_t)

$$\text{TD learning} \quad V_{t+1} = V_t + \alpha\, \{\text{TD-PE}\}$$

TD Learning derives from Rescorla-Wagner learning rule (1972).

Robert Rescorla

Allan R Wagner

$$V_{t+1} = V_t + \alpha\, \underbrace{\{\, \lambda_t - V_t \,\}}_{\text{RW error}}$$

The Start: Pavlov

V : value function, value, prediction, associative strength

λ : reward

α : learning coefficient

γ : temporal discounting coefficient

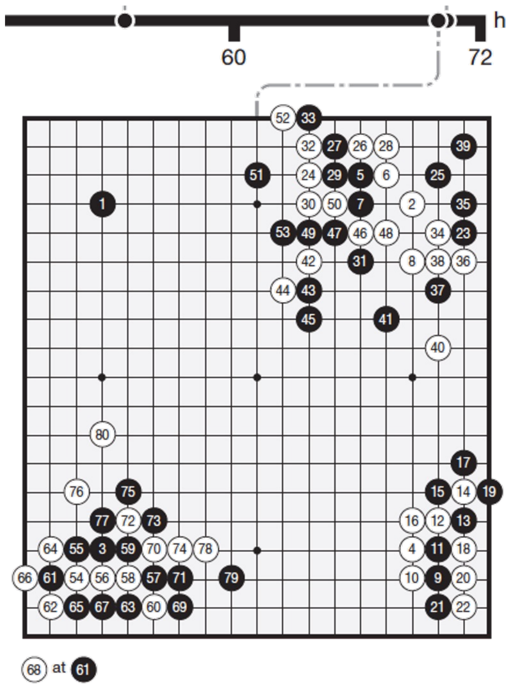
Machine Learning becomes biologically plausible due to the neuronal (dopamine) implementation of prediction error.

Now, Reinforcement Learning outsmarts human intelligence.

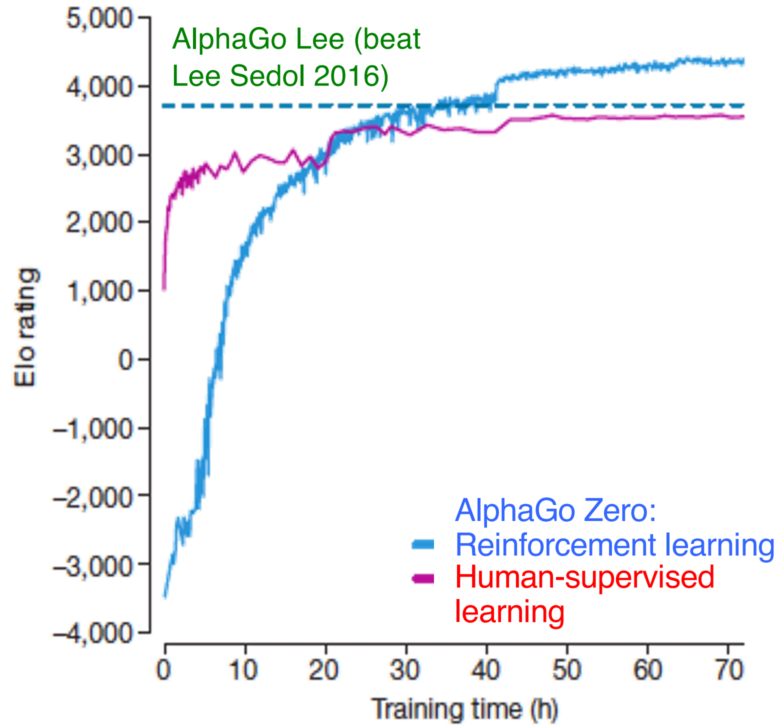
Learning to play world-class Go in 3 days to Elo rating > 4,000

Mastering the game of Go without human knowledge

David Silver[1*], Julian Schrittwieser[1*], Karen Simonyan[1*], Ioannis Antonoglou[1], Aja Huang[1], Arthur Guez[1], Thomas Hubert[1], Lucas Baker[1], Matthew Lai[1], Adrian Bolton[1], Yutian Chen[1], Timothy Lillicrap[1], Fan Hui[1], Laurent Sifre[1], George van den Driessche[1], Thore Graepel[1] & Demis Hassabis[1]

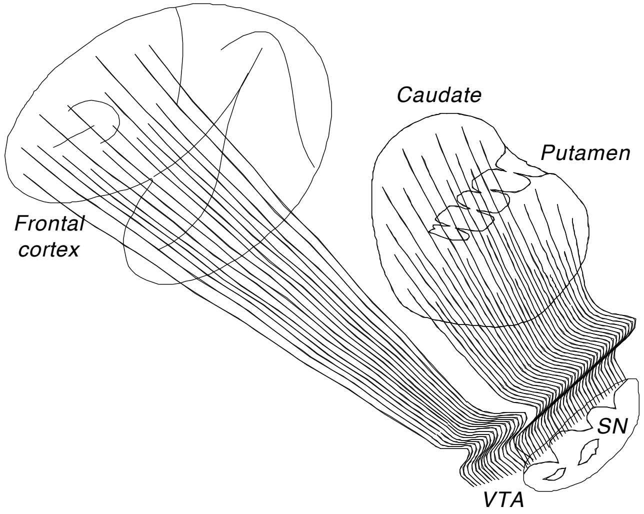


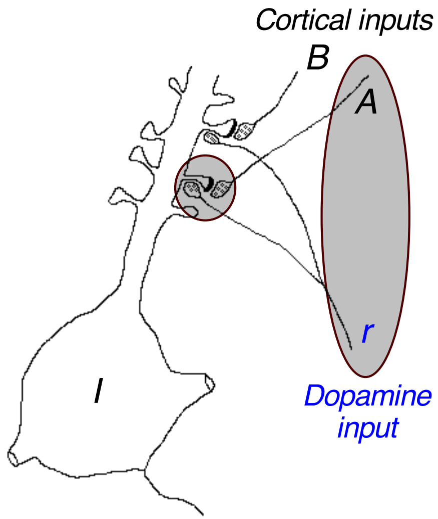
Better learning on its own (blue) than supervised by humans (red)



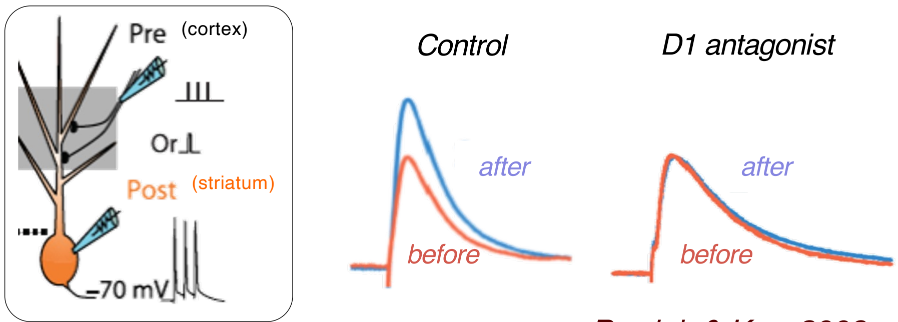
AlphaGo Zero: Silver et al (Hassabis) Nature 2017

Postsynaptic effects of phasic dopamine signal





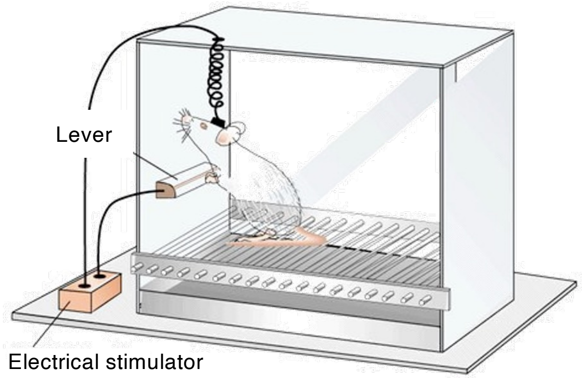
Long-term potentiation in striatum



Pawlak & Kerr 2008

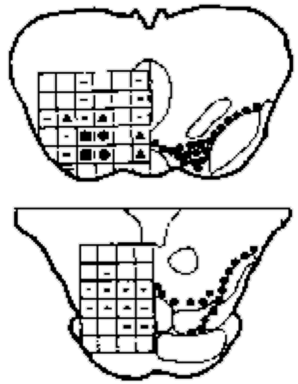
Excitation and inhibition of dopamine neurons induces behavioral learning and unlearning (positive and negative reinforcement).

Electrical self stimulation



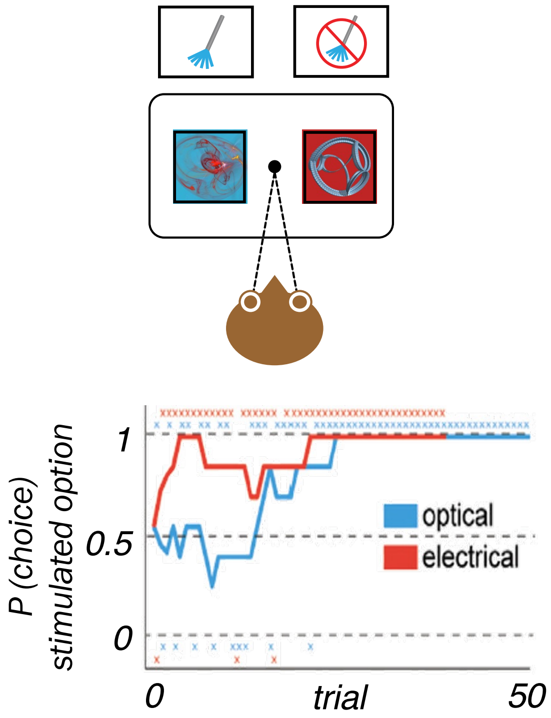
Olds & Milner 1954

Electrical self stimulation involves dopamine neurons



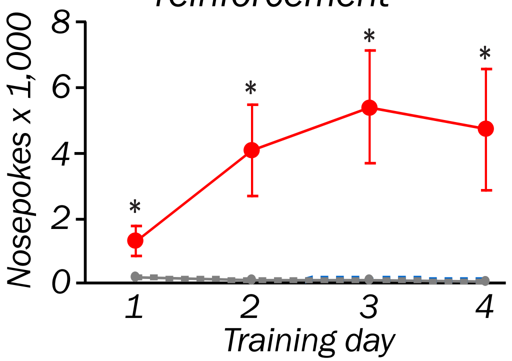
Corbett & Wise 1980

Monkey optogenetic dopamine excitation: positive reinforcement



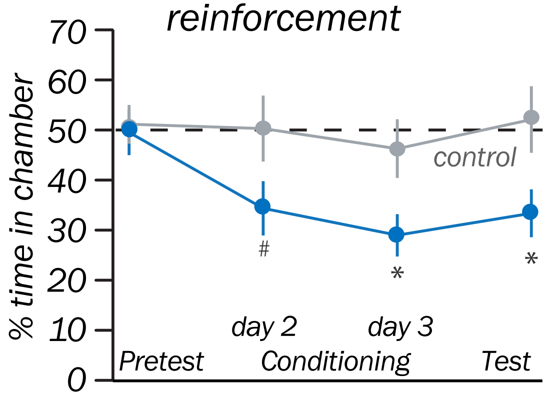
Stauffer, Lak ... Schultz Cell 2016

Rodent optogenetic dopamine excitation: positive reinforcement



Witten et al 2011

Rodent optogenetic dopamine inhibition: negative reinforcement



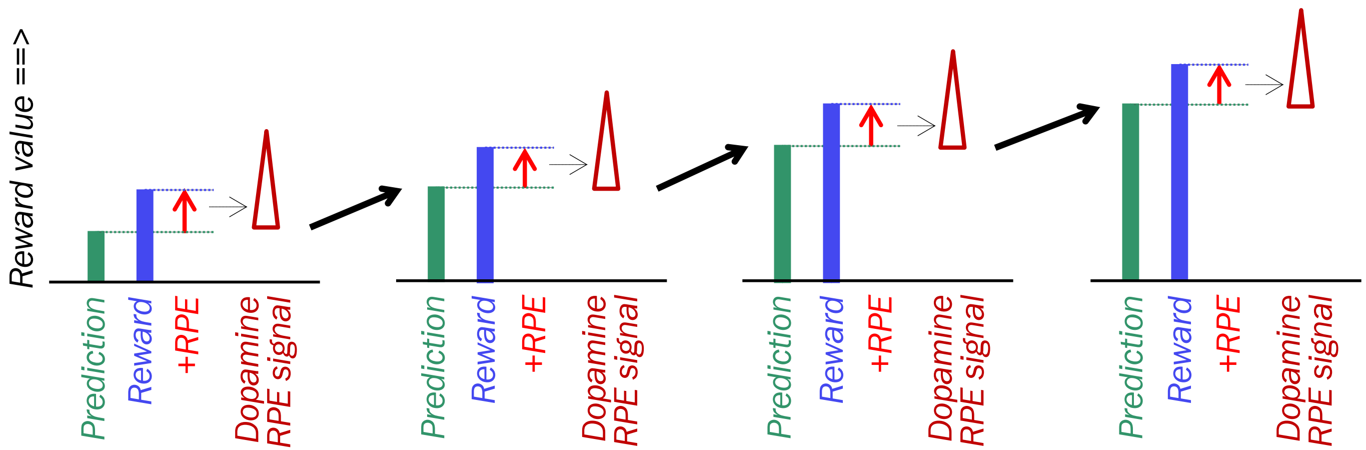
Tan et al 2012

Dopamine signals drive agents

- *towards more reward*
- *away from less reward*

Reward maximization by recursive dopamine RPE coding:

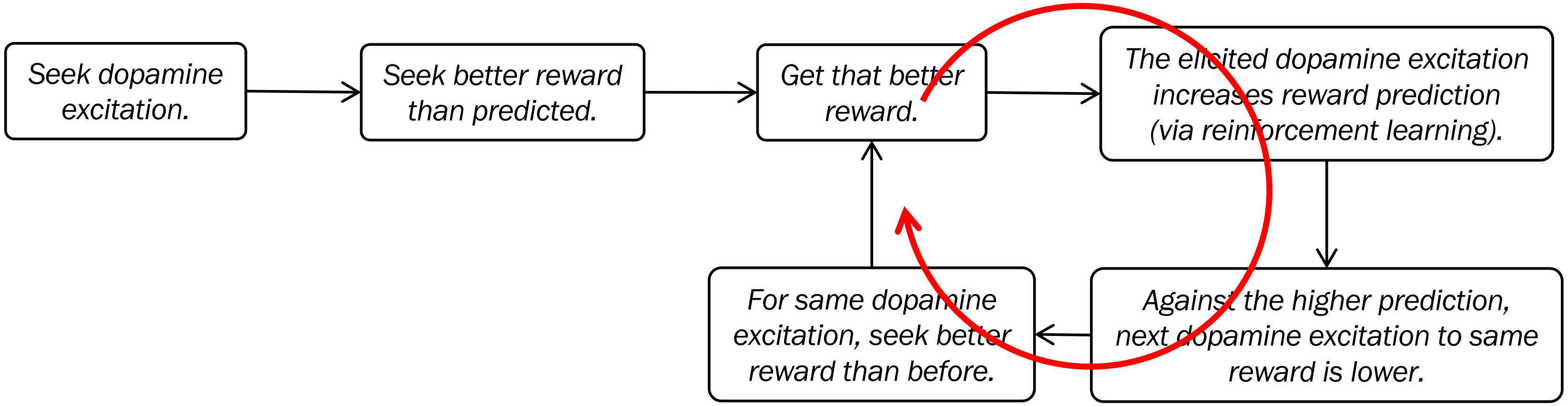
positive dopamine RPE signal drives agents to more reward
in order to get positive RPE signals again.



Schultz PNAS 2024

A dopamine mechanism for reward maximization:

Iteration of dopamine reward prediction error signal and reinforcement leads to continuous reward seeking

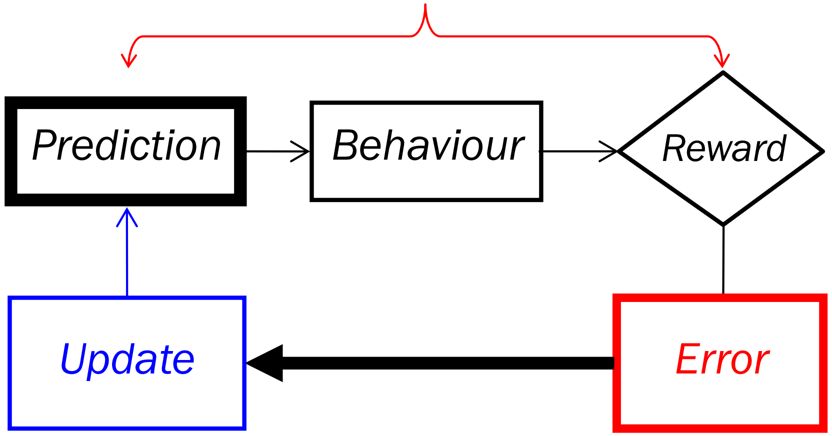


Iteration leads to ever more reward seeking.

Error-driven mechanisms

Reward learning

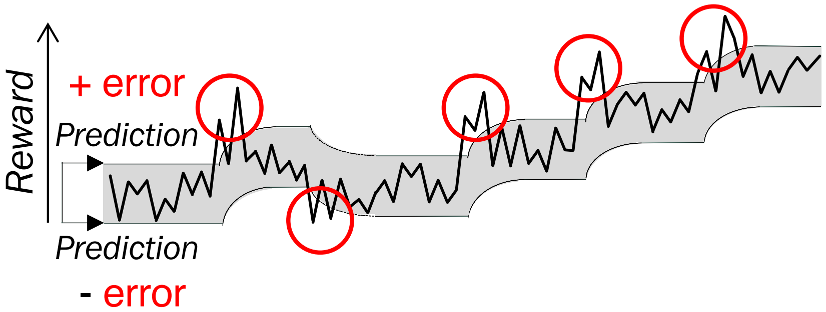
Error = reward - prediction



⇒ Change prediction

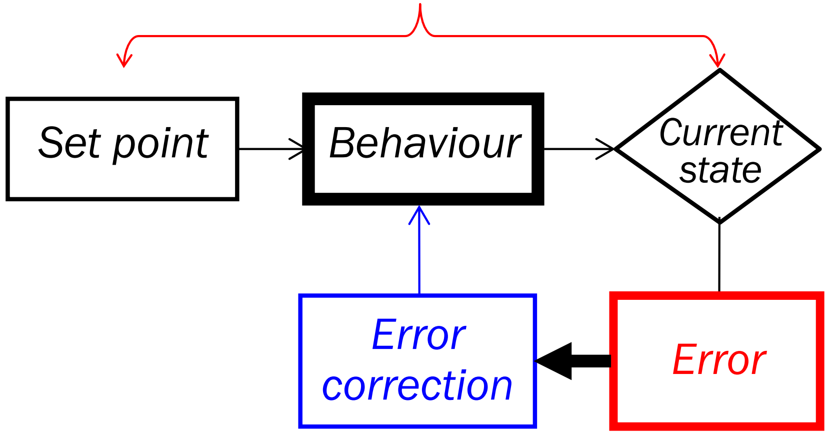
Seek positive error - Avoid negative error

=> INCREASE OVERALL REWARD



Error correction

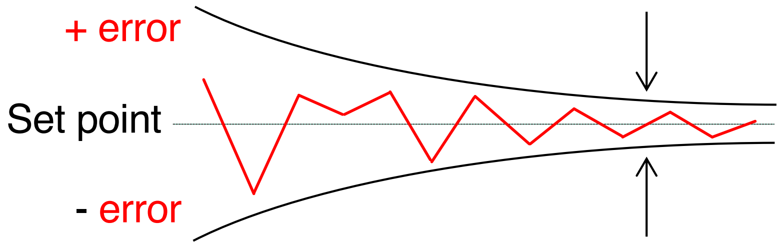
Error = current state – set point



⇒ Change behaviour

Avoid positive and negative error

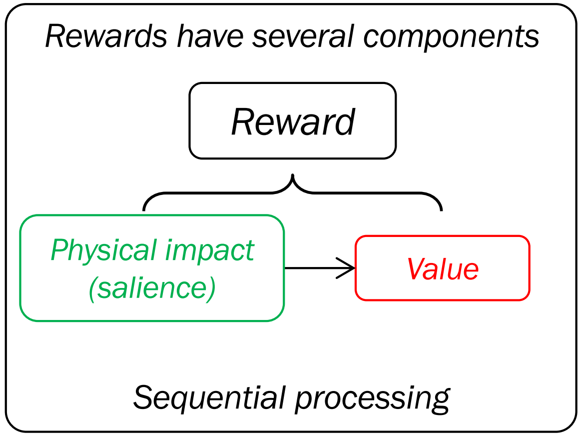
=> MINIMISE ERROR
(STAY WITHIN BOUNDS)

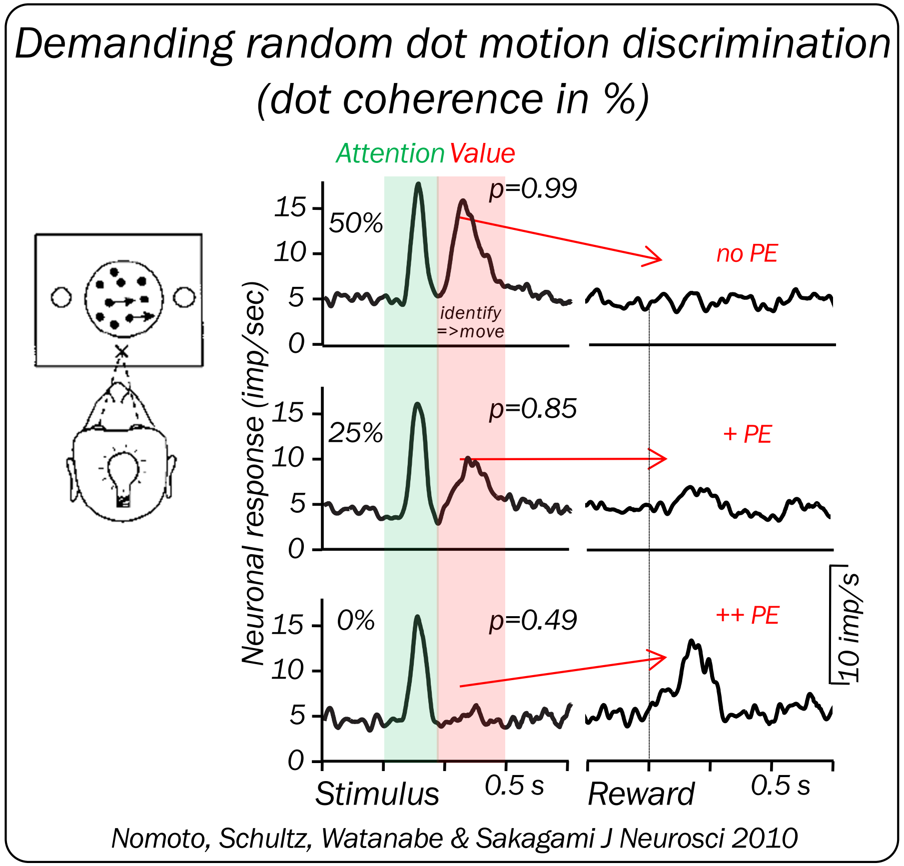


Why do we go to this pub?
We seek excitation of our dopamine neurons.
Actually, we seek rewards just to get dopamine excitation.

One brain system = one function?

An attentional dopamine response component preceding the dopamine reward prediction error (RPE) signal

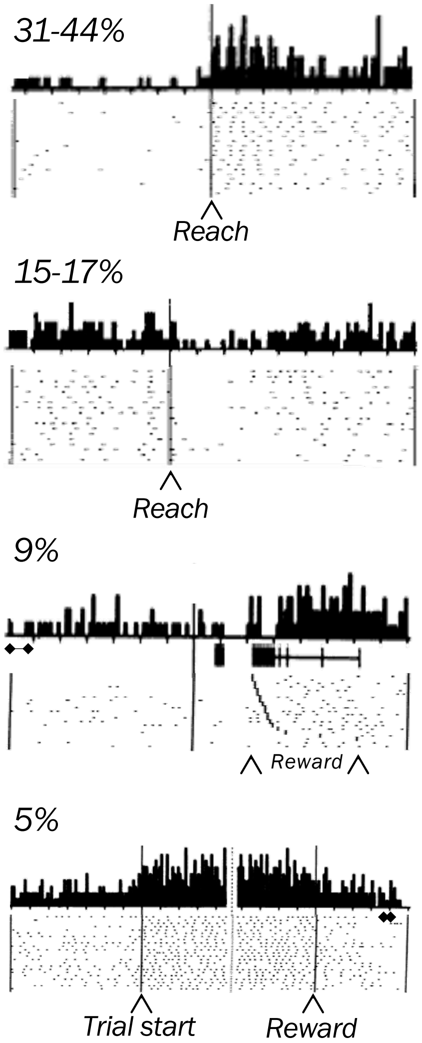




Distinct phasic dopamine signals

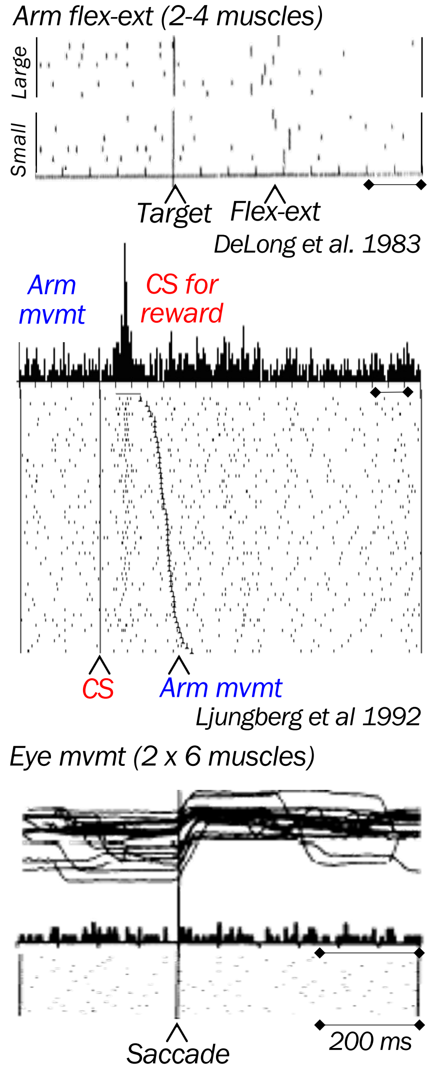
Reward prediction error vs. behavioural activation (including movement)

Dopamine changes with large movements in monkeys (> 35 muscles active)



Schultz et al. 1983; Schultz 1986

No dopamine change with well controlled movements in monkeys

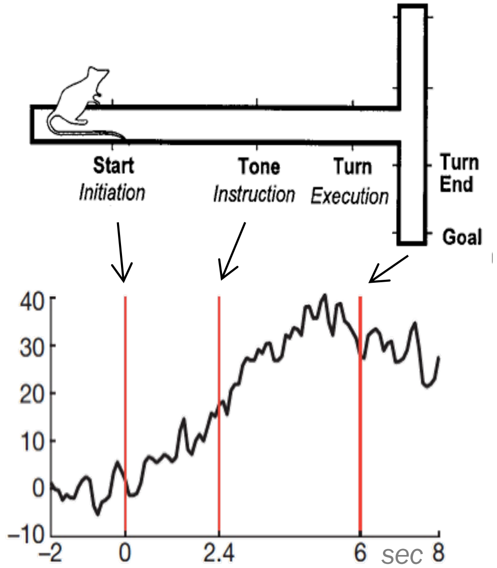


Schultz & Romo 1990

Optogenetics kindled interest in rodents: again dopamine changes with movements

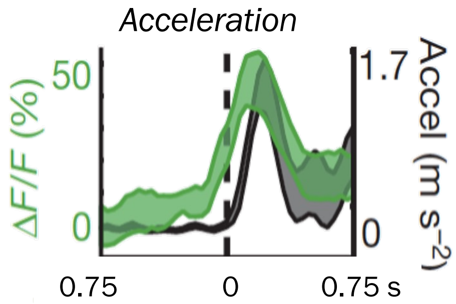
(hundreds of muscles, sensory receptors, cognition)

T-maze

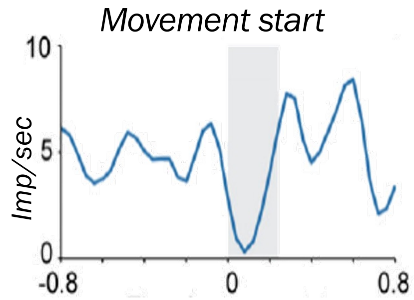


Howe,..., Phillips, Graybiel Nature 2013

Running wheel or track ball

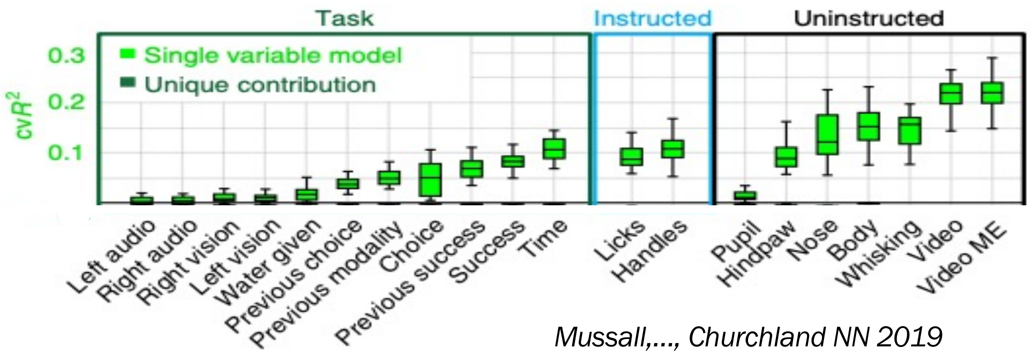


Howe & Dombeck Nature 2016



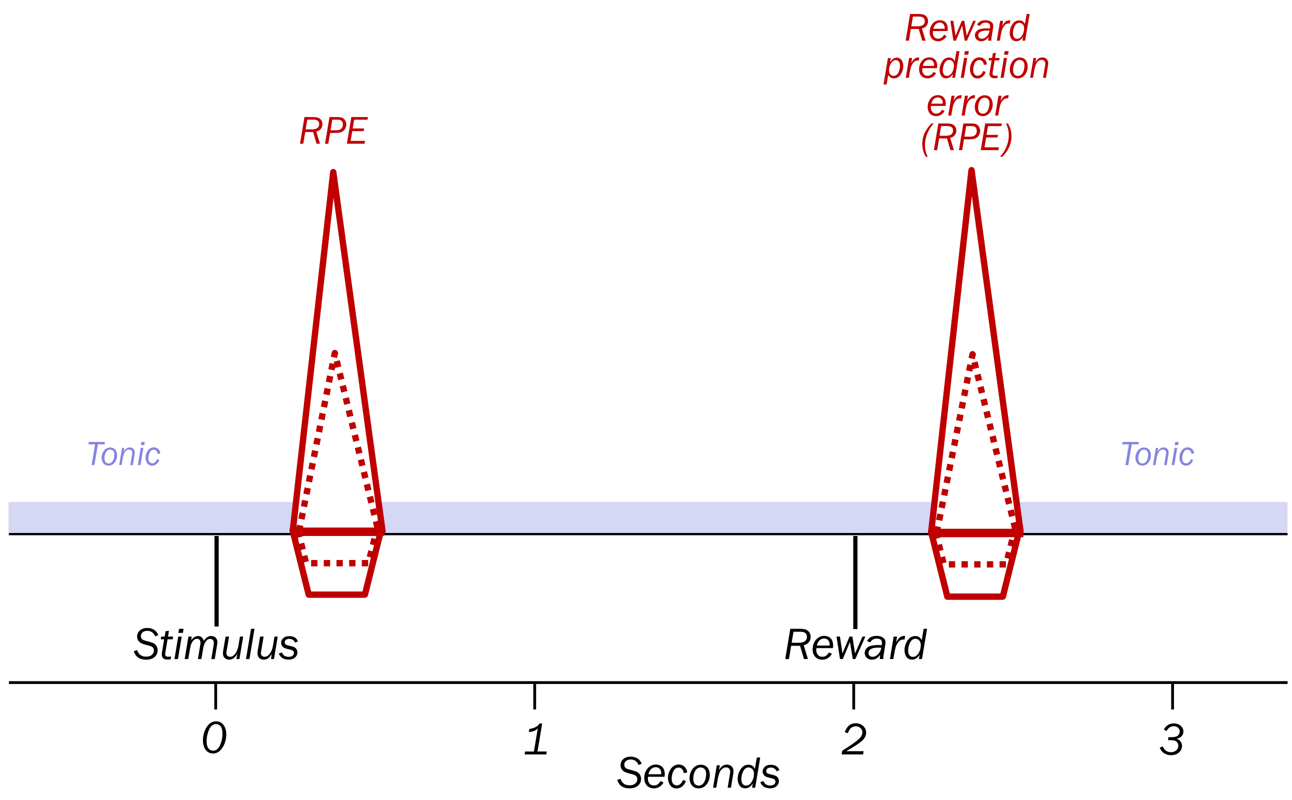
Dodson et al. (Magill) PNAS 2016

Explanation: rodent tasks involve plenty of movements (evidenced here in cortical activity)

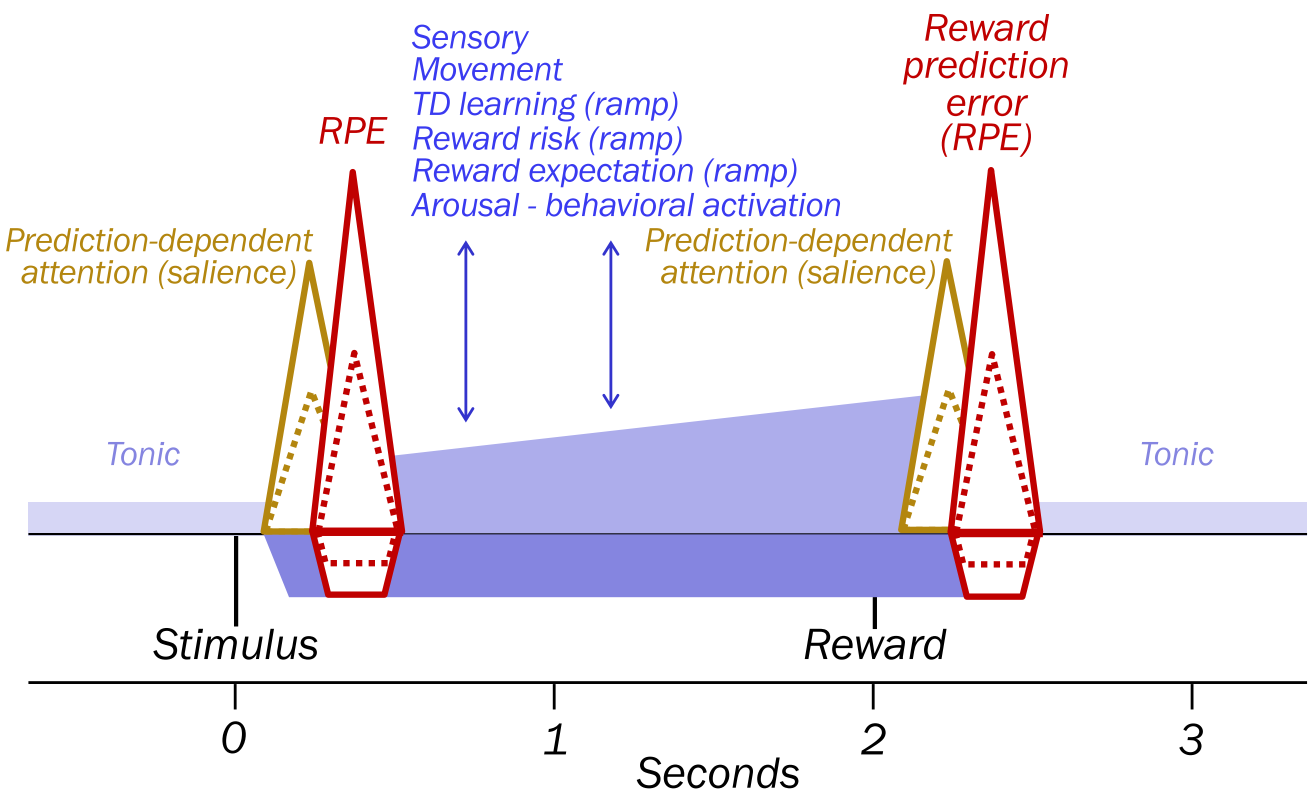


Mussall,..., Churchland NN 2019

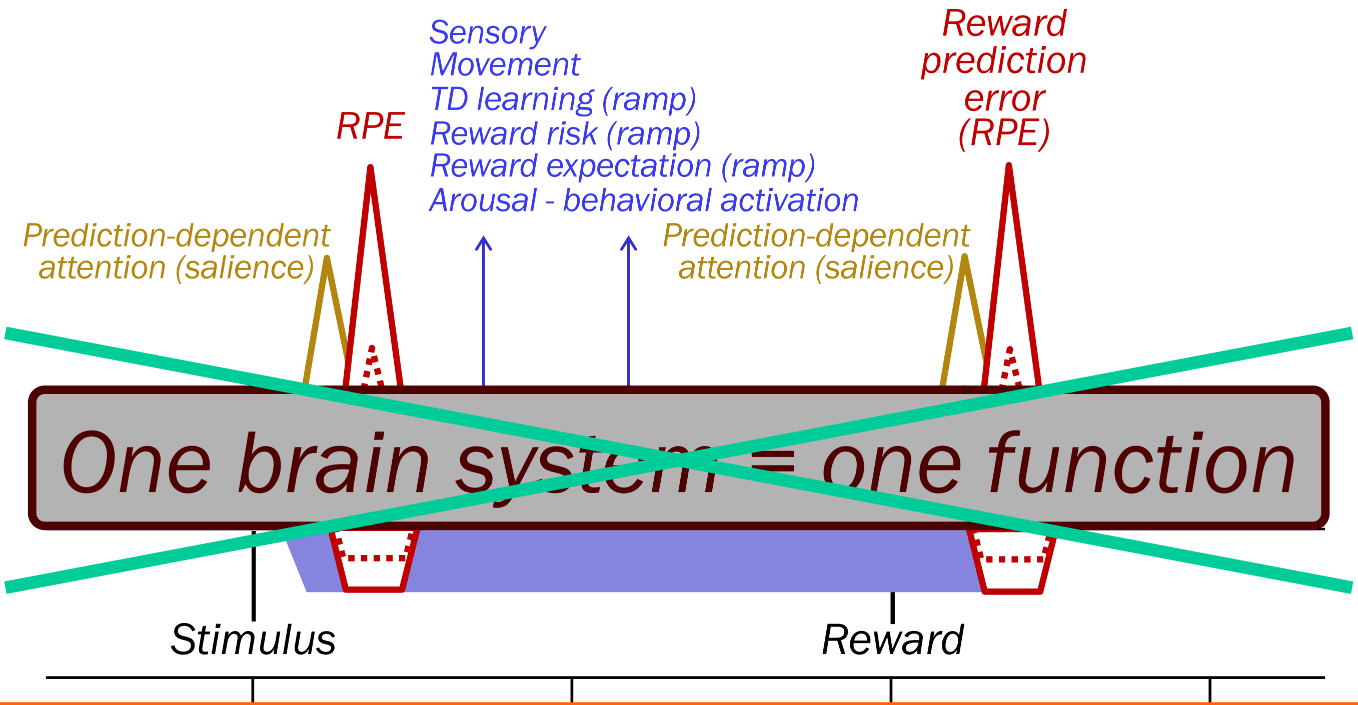
What does 'dopamine' do?



What does 'dopamine' do?



What does 'dopamine' do?



The multifunctionality of dopamine neurons seems appropriate for an evolutionary ancient brain system that remains efficient in the face of changing environmental demands.

Behavioural reward functions

Learning

Approach & choice

Positive emotions

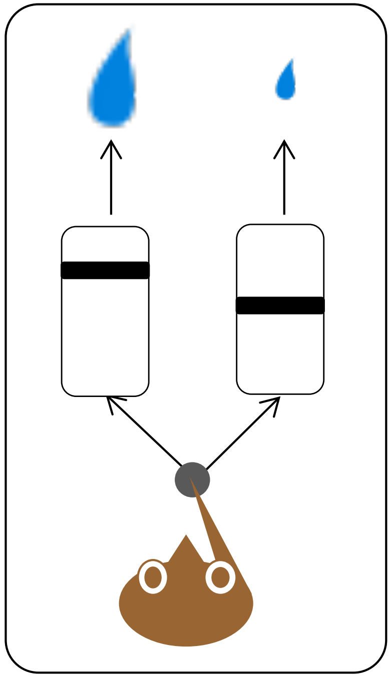
Biological organisms are not silicon machines: Reward value is subjective

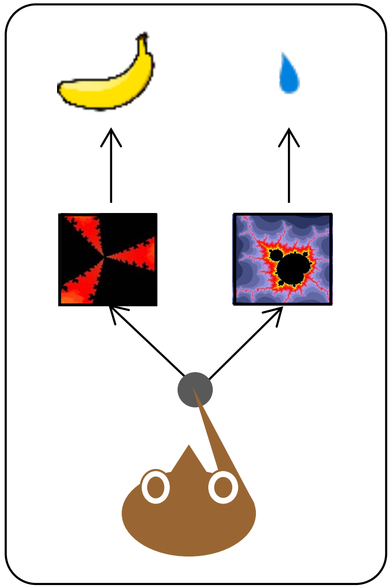
You eat steak # 1

... steak # 2

ENOUGH at steak # 3 !

Subjective steak value decreases with satiety
(while objective steak value stays constant).





Inferring subjective reward value from observable choice

Discrete choice among 2 options

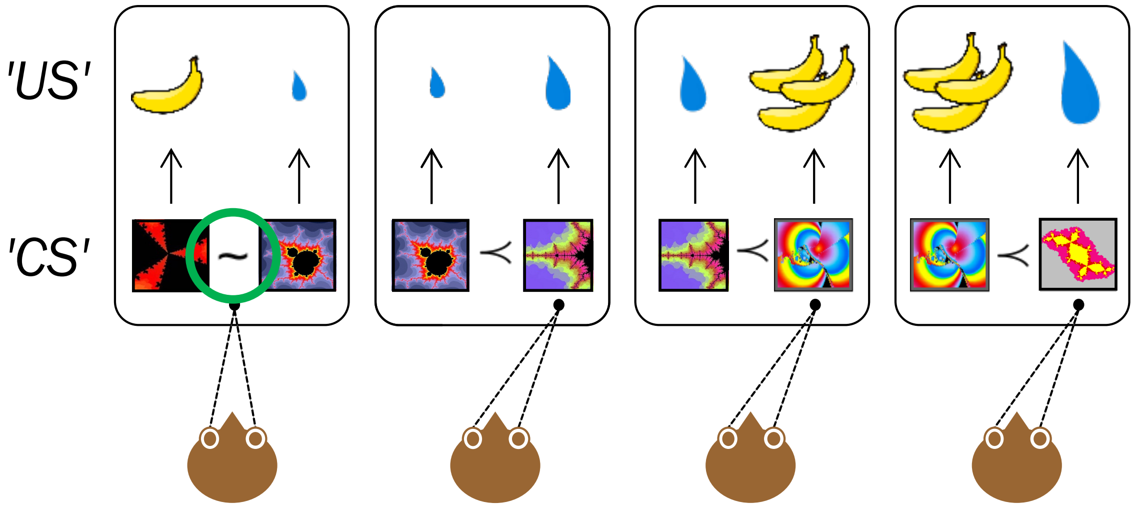
- *option set includes all options (collectively exhaustive)*
- *options are mutually exclusive (choose only one)*
- *options are distinct and well-separated*
- *options alternate pseudorandomly*
- *options appear simultaneously*
- *options cost is constant*

=> everything well-controlled, action distinct from reward

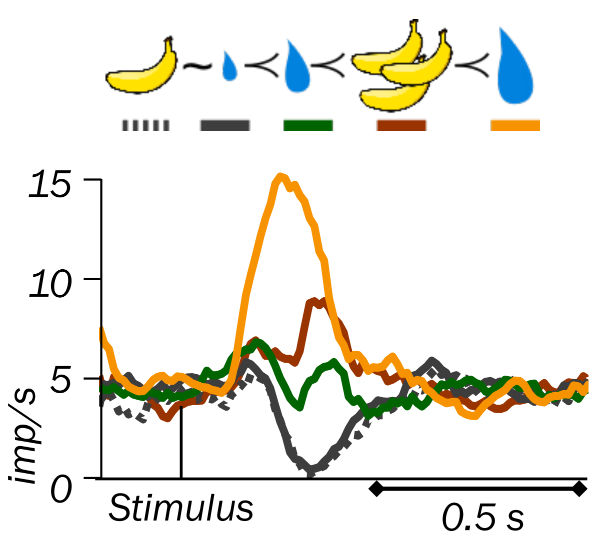
Now we can estimate subjective value
=> same value with equal choice
('choice indifference': immune from slope of choice function)
(repeated testing: stochastic choice)

The dopamine RPE signal reflects subjective reward value.

Subjective value inferred from choice:
more frequent choice => higher value



Dopamine signal follows subjective value



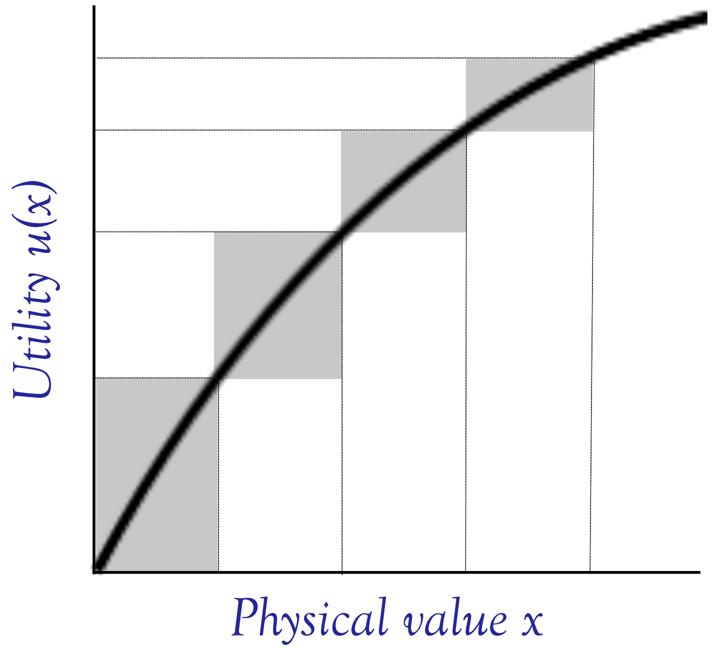
Lak, Stauffer & Schultz PNAS 2014

Economic utility defines subjective reward value

Daniel Bernoulli
1738

Von Neumann &
Morgenstern 1944

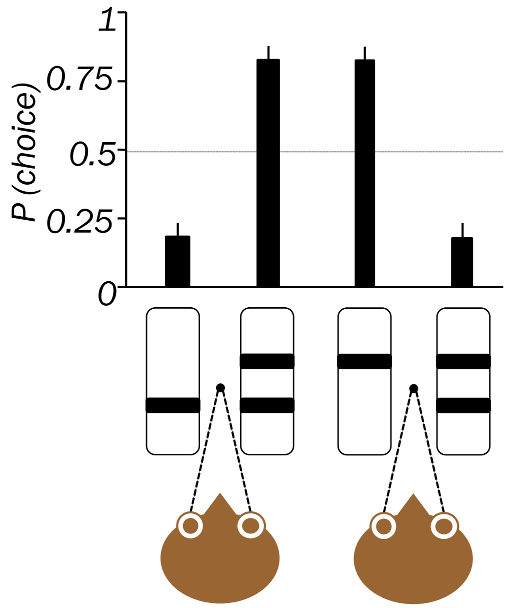
A mathematical function for
subjective reward value



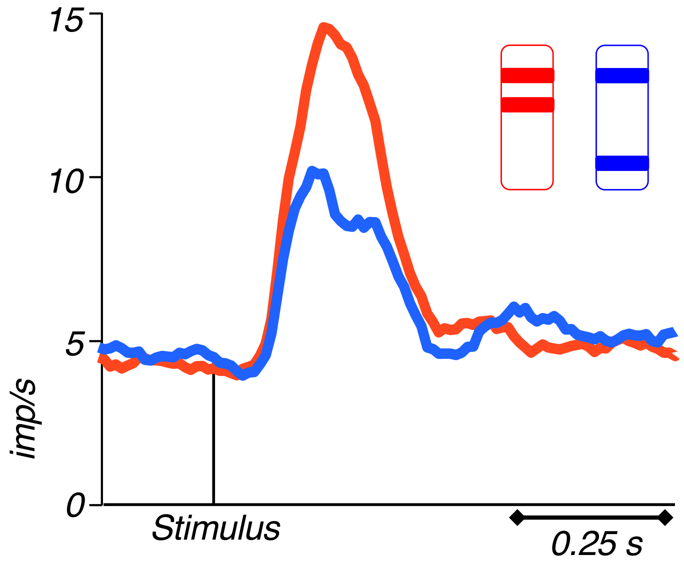
Choice: do monkeys and their reward neurons know what they are doing?

Rational choice requires choice of subjectively best reward:
more is better: first-order stochastic dominance

Monkeys maximise reward

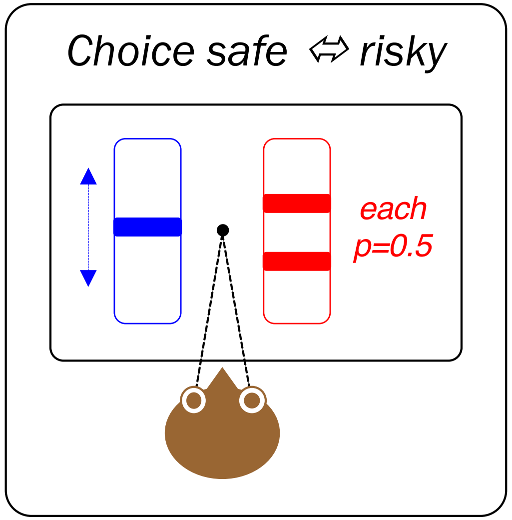


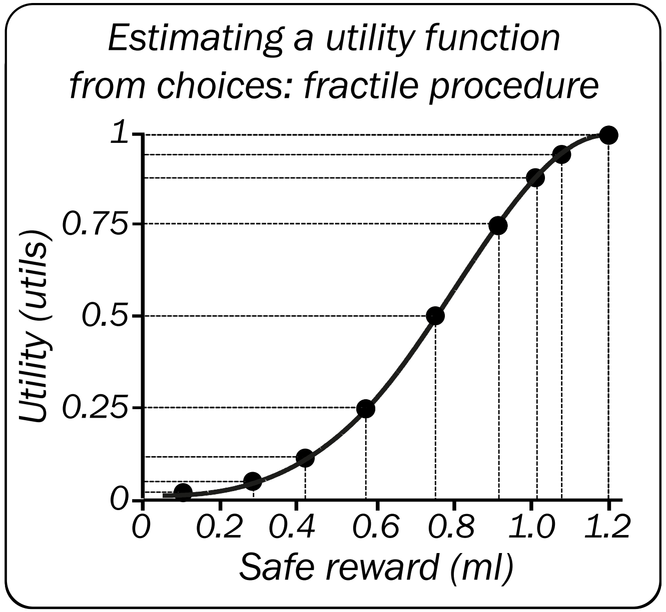
Dopamine neurons signal the best reward



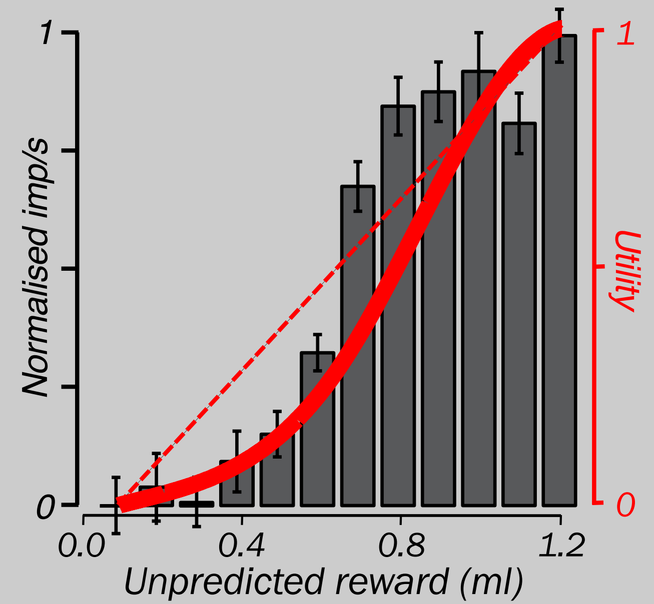
Stauffer, Lak & Schultz CurrBiol 2014

The dopamine utility prediction error signal





Dopamine signal



$$\mathrm{UPE} = u - \mathrm{EU}$$

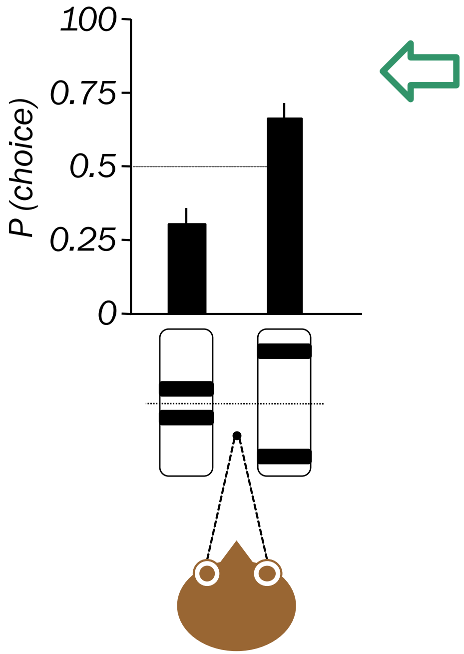
Expected utility $\mathrm{EU} = \Sigma_i \, u(x_i) \bullet p_i(x_i)$

Stauffer, Lak & Schultz Curr Biol 2014

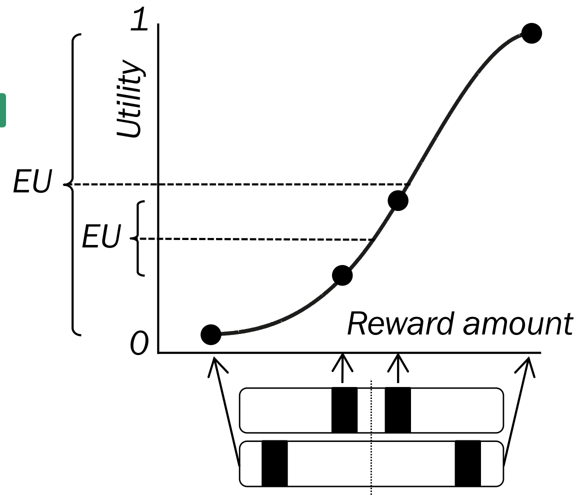
Choice: do monkeys and their reward neurons know what they are doing?

Rational choice means choice of best reward:
choose according to subjective value (not objective value):
mean-preserving spread

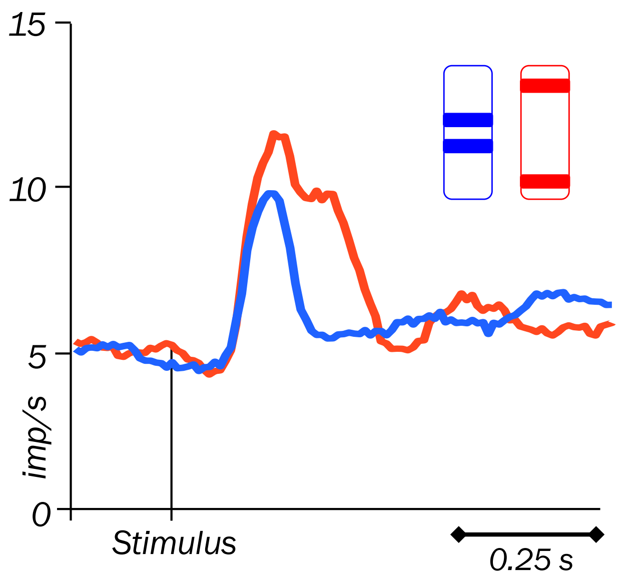
Monkeys maximise subjective value



Distinguish subjective from objective value



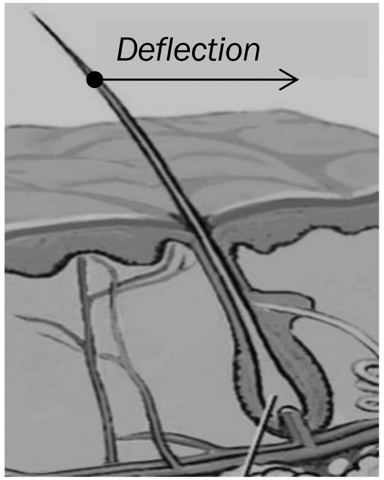
Dopamine neurons signal the subjectively best reward

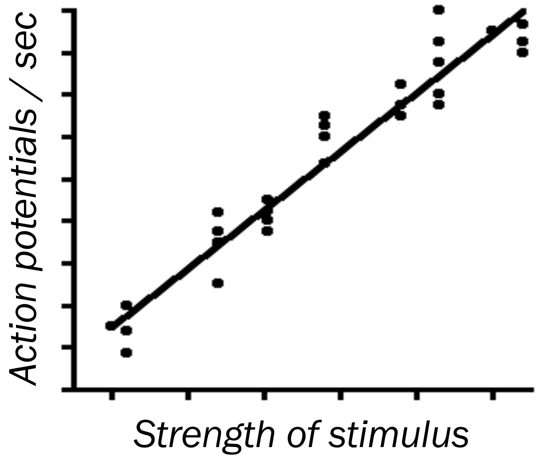


Stauffer, Lak & Schultz CurrBiol 2014

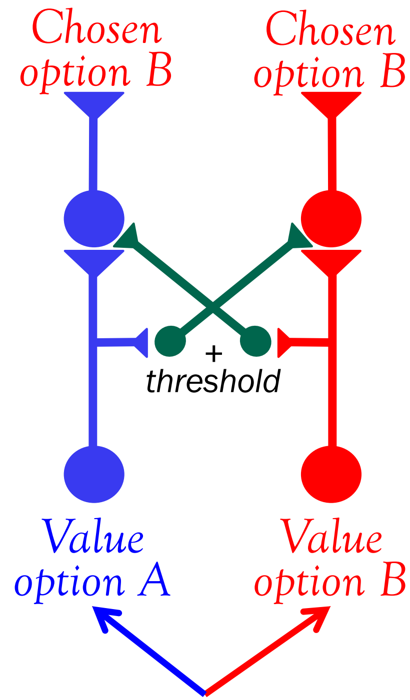
The intuitive metric of neuronal information is a rate code: number of action potentials/second.

The neuronal rate code originating from the opening of Na–channels in sensory receptors serves as a neuronal metric for stimulus strength (Adrian & Zotterman 1926).





Value metric as basis of choice:
Winner-Take-All choice mechanism and its value inputs



Value of each option A and B is composed of:

Objective (physical) value:
Amount, probability, reward type, effort

+ Subjective modifiers:
Utility, weighted probability, weighted effort, reference, risk, delay, satiety

+ Environmental influences:
Personal history, convention, compassion, cooperation, coordination, social norms, moral, ethics, tradition, culture, strategy, heuristics, idiosyncrasy, prejudice, superstition, parochialism, nationalism

Robust activation of human (dopamine-receiving) striatum by rewards

Money

Thut (Leenders, Schultz) et al. 1997

Reward prediction error

McClure (Montague) et al. 2003

Beautiful faces

Aharon (Breiter) et al. 2001

Sports cars

Erk (Spitzer) et al. 2002

Pleasant music

Blood & Zatorre 2001

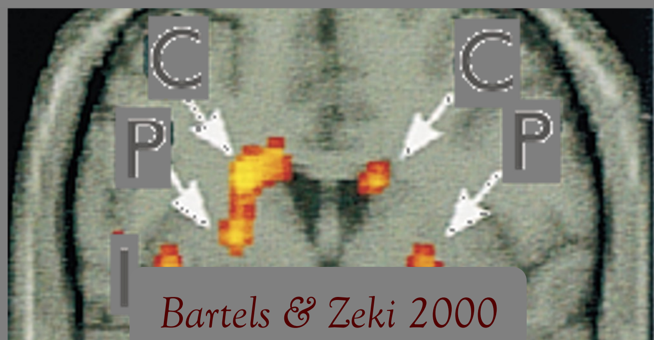
Humor

Mobbs et al. 2003

Placebo

Petrovic et al. 2005

Romantic love



Bartels & Zeki 2000

Altruistic punishment

De Quervain (Fehr) et al. 2005