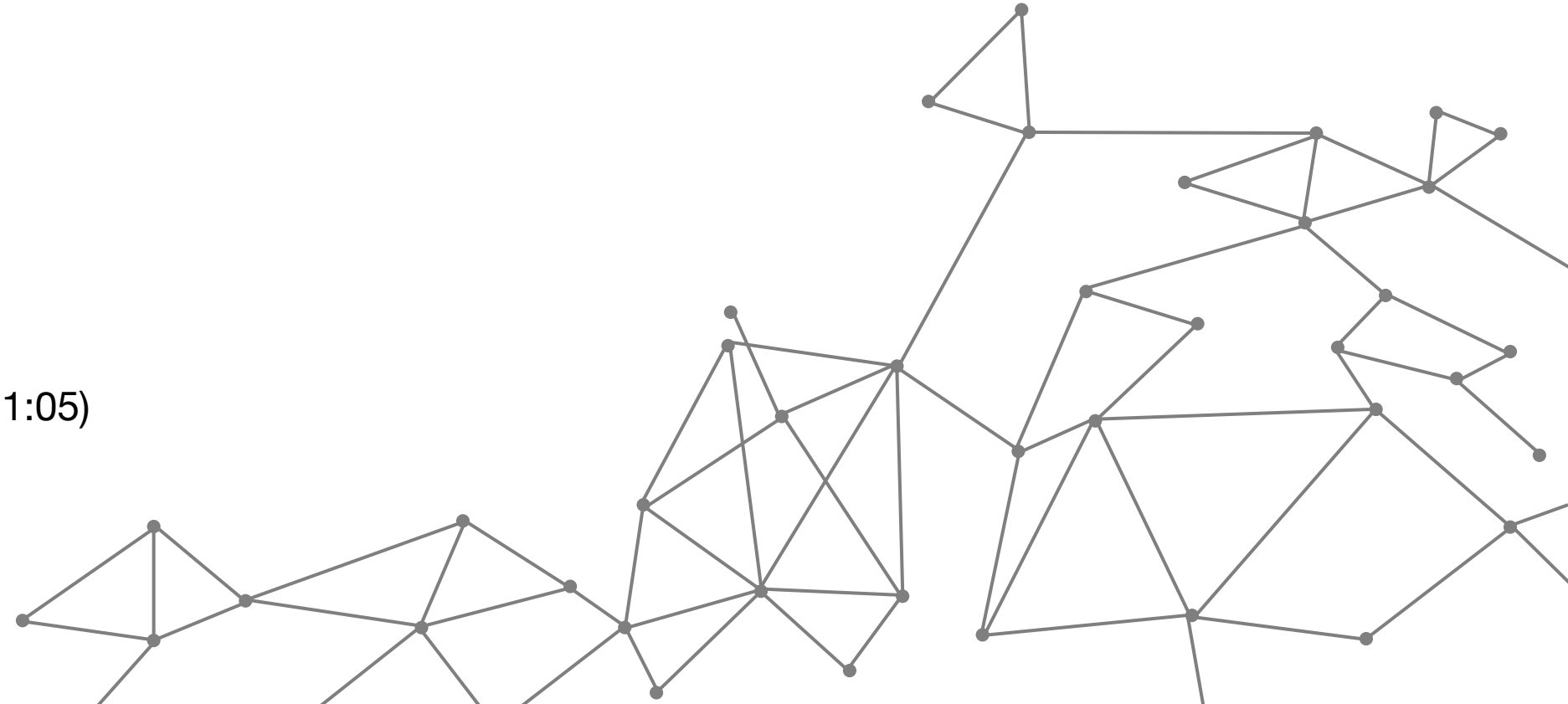
Modeling Markets, Pandemics, and Peace: The Mathematics of Multi-Agent Systems

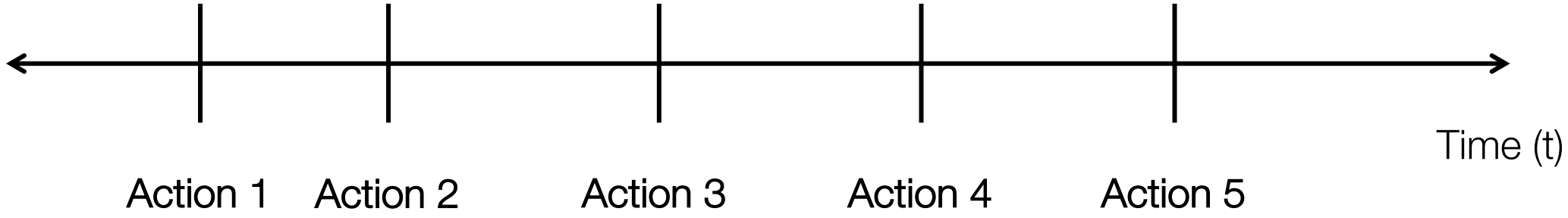
Lecture 3

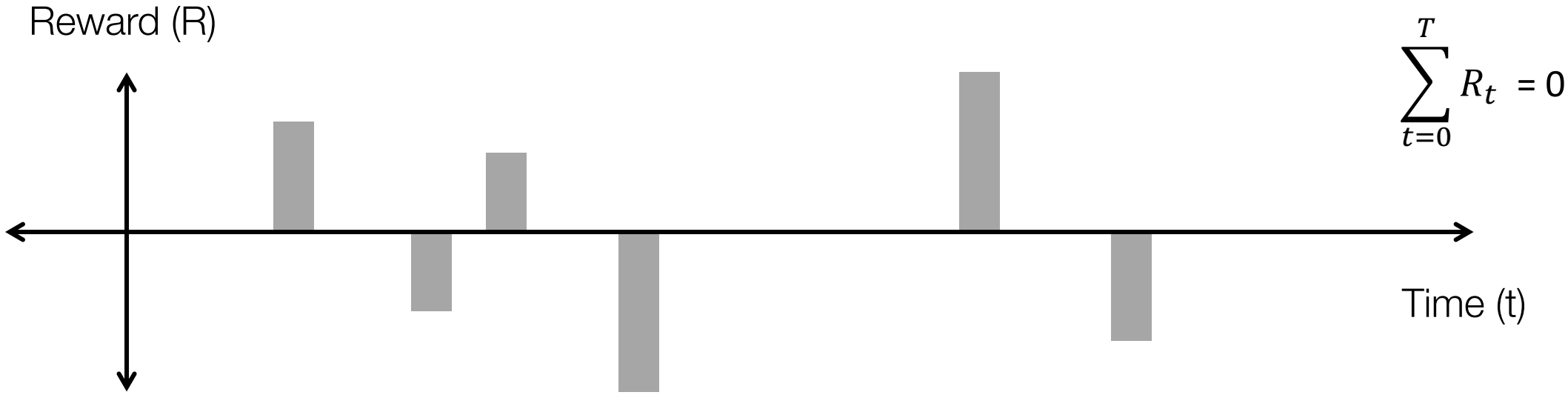
How people learn



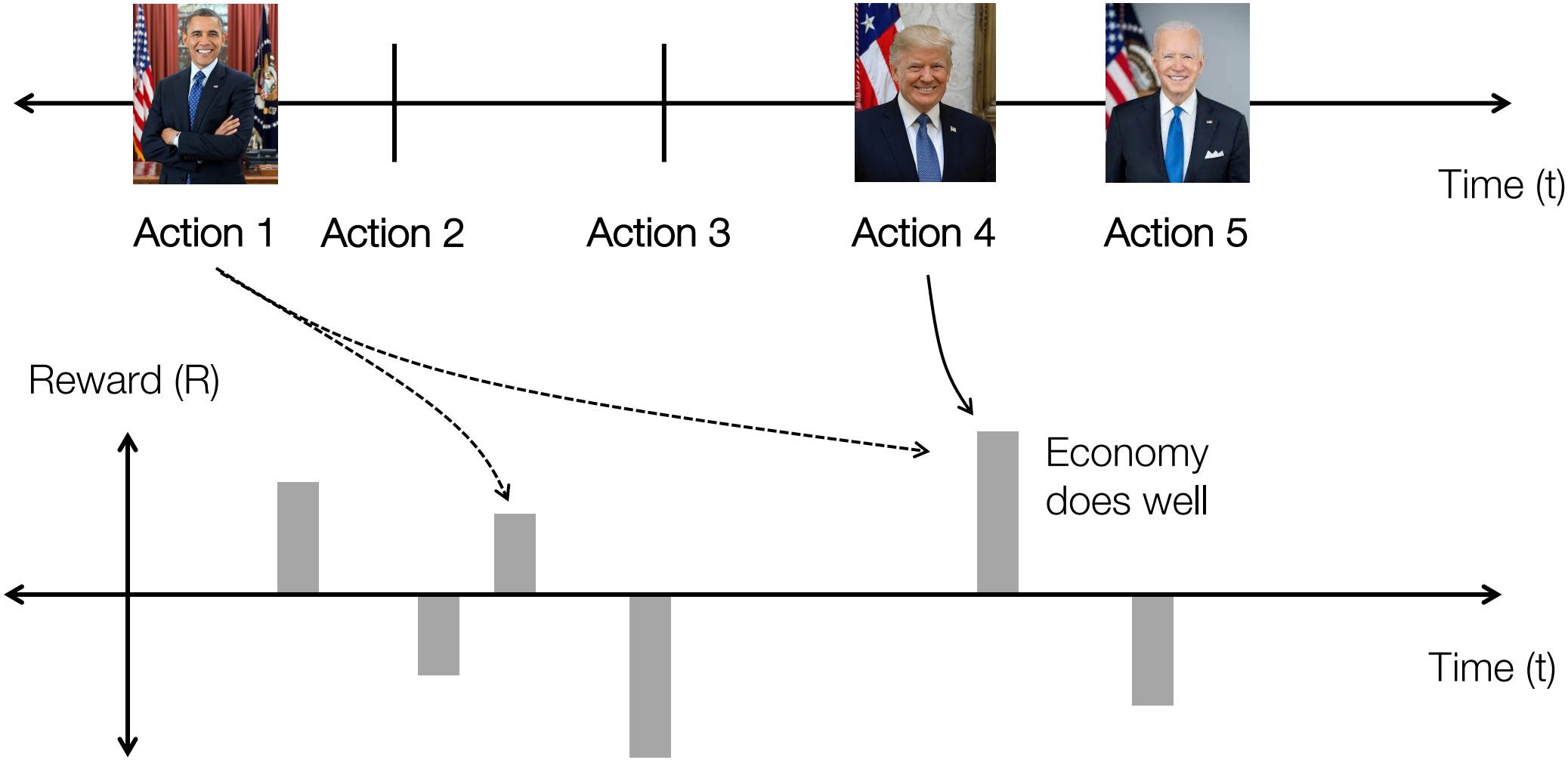
MIT HSSP
July 23rd, 2022 (Starting 1:05)

Credit assignment problem



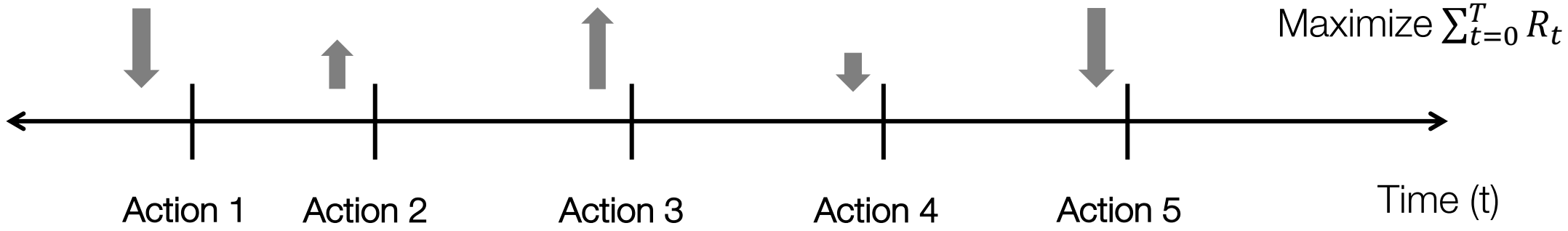


Credit assignment problem



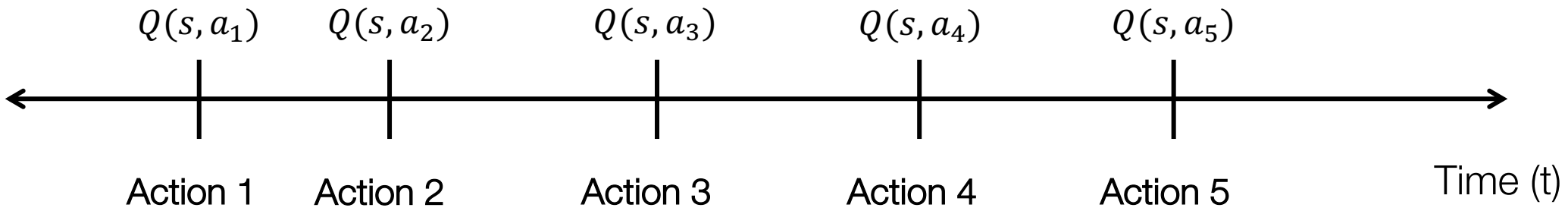
Two ways to address credit assignment

1. Policy gradient (the REINFORCE algorithm)



Link the total reward to all the actions before it and nudge our policy in different ways to see what happens.

2. Q-learning (associating values to state-action pairs)

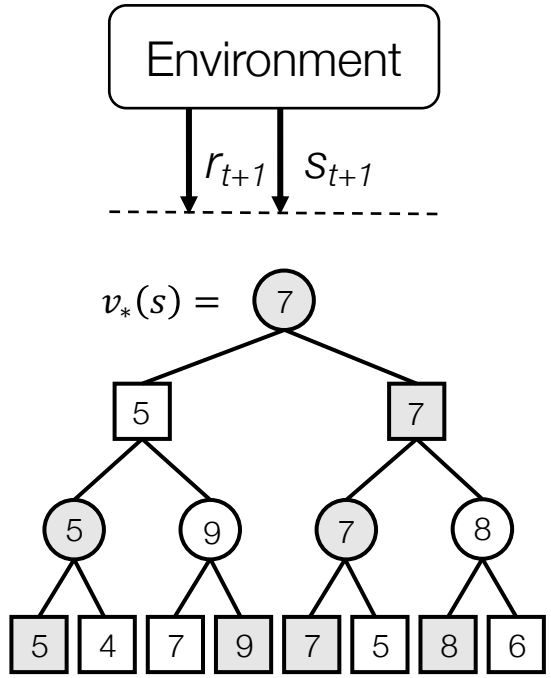


Learn a Q-value via temporal difference methods: if the action is good, the Q value goes up.

Statistical vs. causal vs. mechanistic model

Mechanistic model

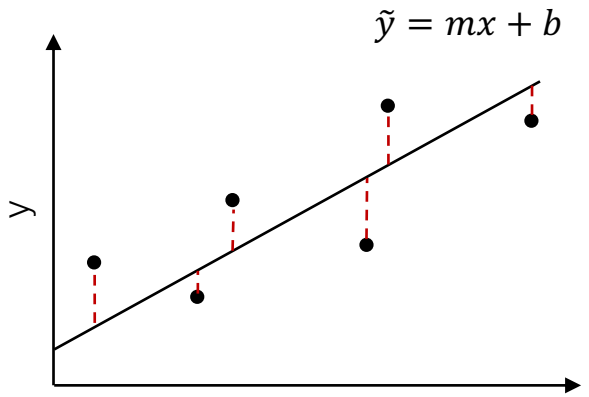
Full, precise description of what generates the data.



Statistical model

Correlation only: every time I sit and raise my paw, my owner gives me a treat.

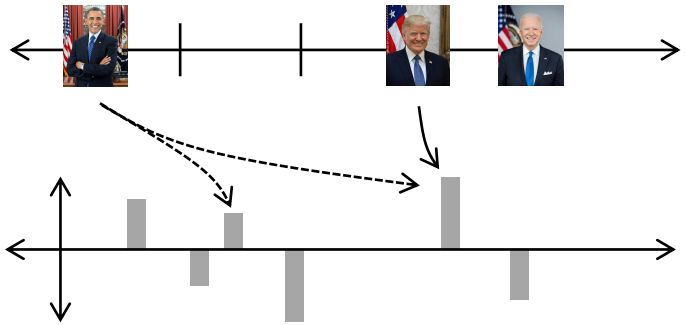
Reinforce the action of sitting and raising my paw to obtain more treats.



Causal model

Correlation is not sufficient: rain is correlated to people bringing umbrellas. But people not bringing umbrellas will not stop rain.

Counterfactual cases are required.

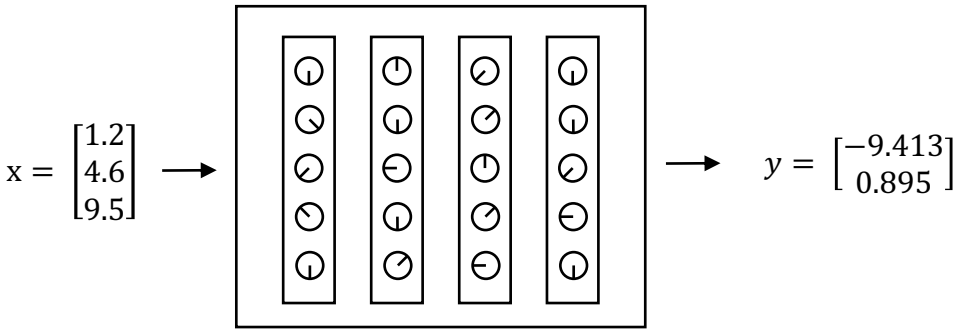


For extra reading, see the paper Towards Causal Representation Learning.

Credit assignment as a general framework*

*Abram Demski provides an excellent exposition of this idea.

Backpropagation assigns loss to each knob and tells us how to tweak them



Natural selection assigns credit to aspects of an organism that maximizes the chance of reproduction

Economies (ideally) assign credit to incentivize good behavior and actors who create social value

Social norms (ideally) help with the proper assignment of fault/credit and enforce good aggregate behavior.

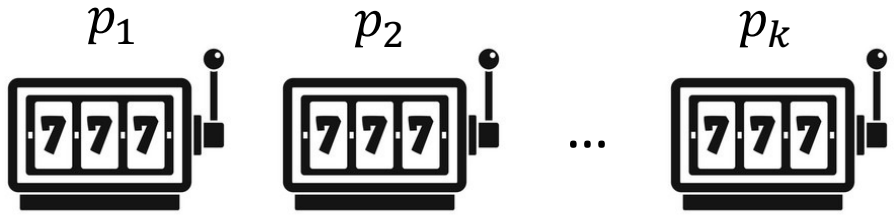
Mapping reinforcement learning to human and social behavior

Human-machine correspondence

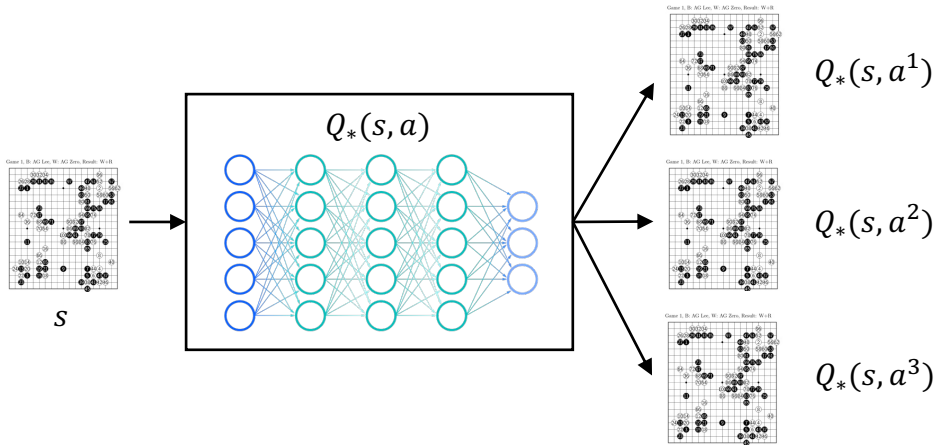
Behavior reinforcement (actor)

$$w_{n+1} = w_n + \gamma \sum_{t=0}^{T-1} \nabla_w \log \pi_w (s_t, a_t) \sum_{t'=t+1}^{T} R_{t'}$$

Exploration-exploitation tradeoff



Action evaluation (critic)



Experience replay

Replay memory

$$\underbrace{e_t, e_{t-1}, \ldots, e_{t-N}}$$

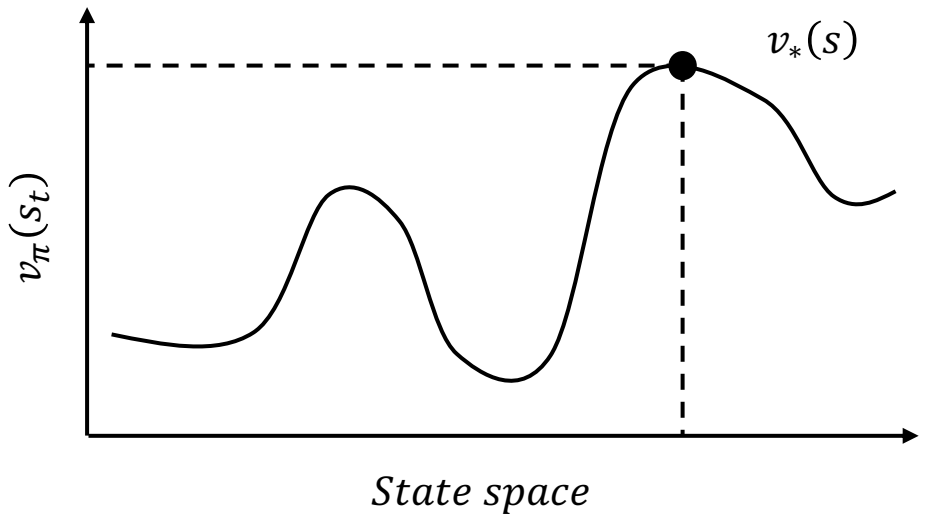
Store last N experiences

Expected utility hypothesis (reinforcement learning)

We can program our reinforcement learning agent by telling it to maximize some utility,

$$v_\pi(s_t) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} R_{t+k+1} | S = s_t \right] \qquad v_*(s) = \text{maximize}_\pi(v_\pi(s))$$

where we define the rewards (e.g., # apples eaten, money earned…) to shape its behavior.



Models in social science

Economics: “In a free market, the forces of supply and demand will cause the economy to settle into its most efficient state.”

Political science: “Country A and country B will never go to war if their leaders want to maximize their own political survival.”

Sociology: “People go to protest if their direct benefits is sufficient to overweigh the costs.”

Design the incentive structure ⟶ Agent behavior

Reverse-engineer the incentive structure ⟶ Agent behavior

The ubiquity of utility theory

General equilibrium theory, Nobel 1972

The separation theorem for convex sets asserts that if two convex sets are disjoint, there is a hyperplane which separates them, so that one set is on one side and the other set on the other. In symbols, if C_1 and C_2 are disjoint convex sets in n-dimensional space, there exists numbers $p_i (i = 1, \ldots, n)$, not all zero, c, such that $\sum_{i=1}^{n} p_i x_i \geqq c$ for all $x = (x_1, \ldots, x_n)$ in C_1, $\sum_{i=1}^{n} p_i x_i \leqq c$ for all x in C_2. Let us apply this theorem to the present case. The non-positive orthant is obviously a convex set; let us assume for the moment that Z is convex. Then we can find numbers pi $(i = 1, \ldots, n)$, not all zero, c such that,

$$\sum_{i=1}^{n} p_i z_i \geqq c \text{ for } z = (z_1, \ldots, z_n) \text{ in } Z,$$

Modern portfolio theory, Nobel 1990

PORTFOLIO SELECTION*

HARRY MARKOWITZ
The Rand Corporation

THE PROCESS OF SELECTING a portfolio may be divided into two stages. The first stage starts with observation and experience and ends with beliefs about the future performances of available securities. The second stage starts with the relevant beliefs about future performances and ends with the choice of portfolio. This paper is concerned with the second stage. We first consider the rule that the investor does (or should) maximize discounted expected, or anticipated, returns. This rule is rejected both as a hypothesis to explain, and as a maximum to guide investment behavior. We next consider the rule that the investor does (or

Game theory, Nobel 1994

Equilibrium Point:
An n-tuple $\mathfrak{s}$ is an *equilibrium point* if and only if for every i

$$p_i(\mathfrak{s}) = \max_{\text{all } r_i\text{'s}} [p_i(\mathfrak{s}; r_i)] . \tag{1}$$

Thus an equilibrium point is an n-tuple $\mathfrak{s}$ such that each player's mixed strategy maximizes his payoff if the strategies of the others are held fixed. Thus each player's strategy is optimal against those of the others. We shall occasionally abbreviate equilibrium point by eq. pt.

We say that a mixed strategy s_i *uses* a pure strategy $\pi_{i\alpha}$ if $s_i = \sum_\beta c_{i\beta}\pi_{i\beta}$ and $c_{i\alpha} > 0$. If $\mathfrak{s} = (s_1, s_2, \cdots, s_n)$ and s_i uses $\pi_{i\alpha}$ we also say that $\mathfrak{s}$ uses $\pi_{i\alpha}$.

From the linearity of $p_i(s_1, \cdots, s_n)$ in s_i,

$$\max_{\text{all } r_i\text{'s}} [p_i(\mathfrak{s}; r_i)] = \max_{\alpha} [p_i(\mathfrak{s}; \pi_{i\alpha})]. \tag{2}$$

Behavioral economics, Nobel 2002

Judgment under Uncertainty: Heuristics and Biases

Biases in judgments reveal some heuristics of thinking under uncertainty.

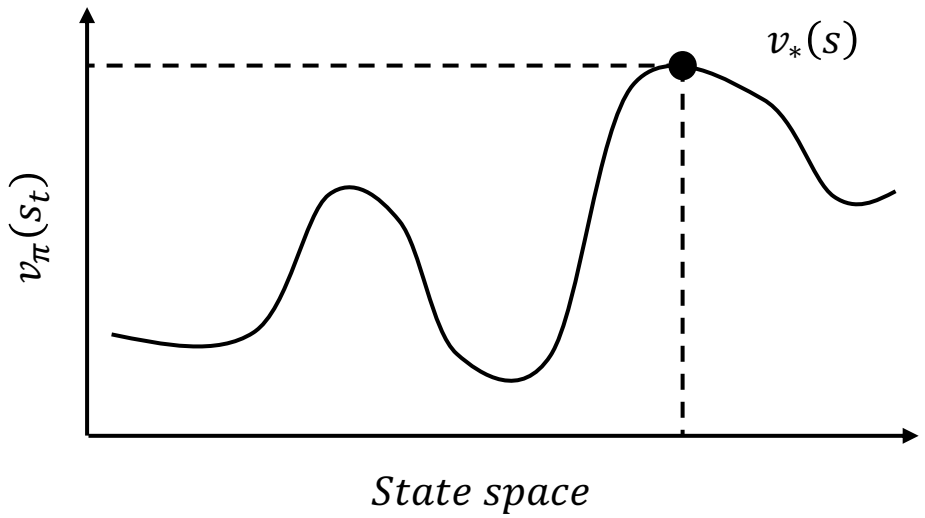
Amos Tversky and Daniel Kahneman

Expected utility hypothesis (social science)

To determine how actors behave, we can always define them to be utility maximizers

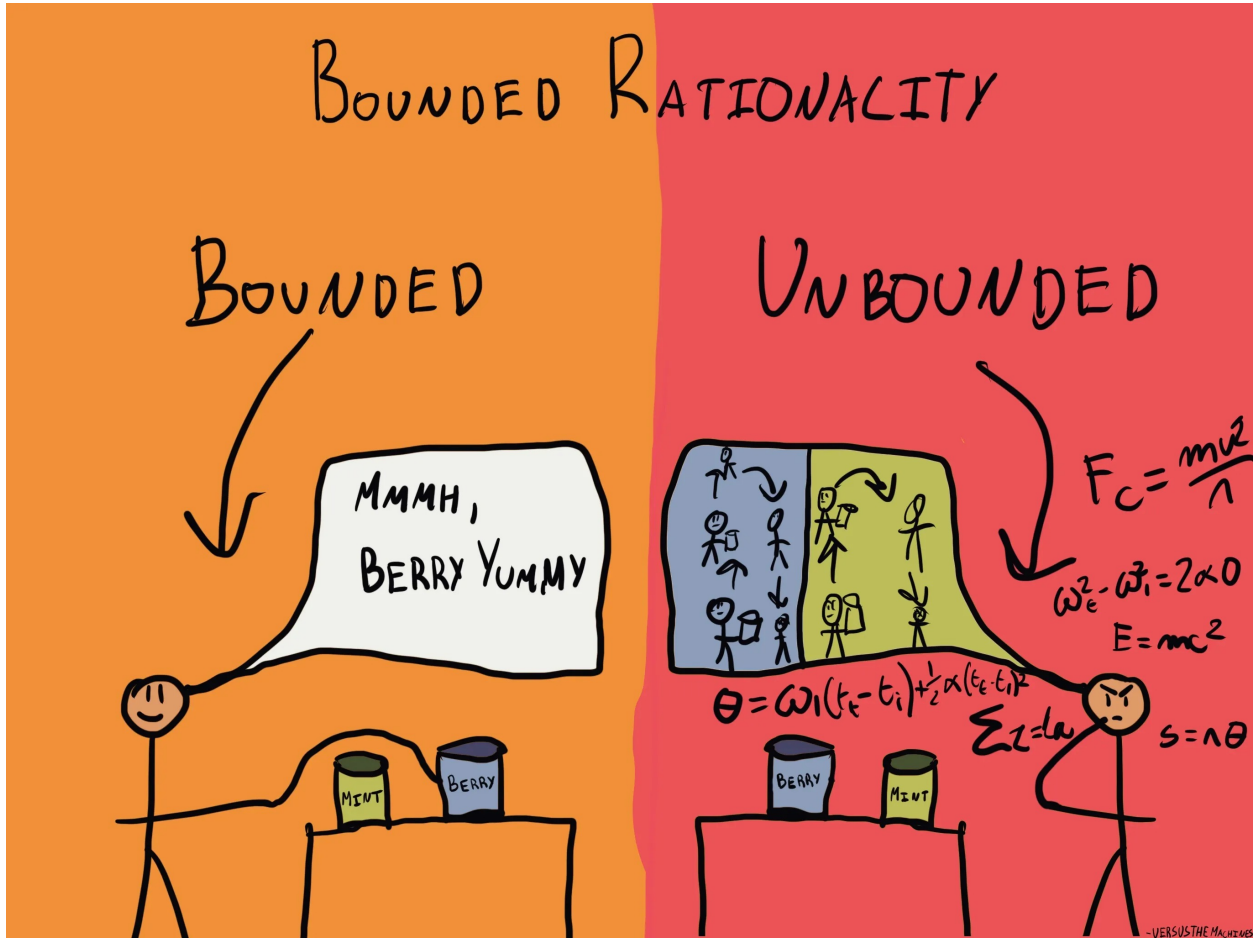
$$v_\pi(s_t) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} u_{t+k+1} | S = s_t\right] \qquad v_*(s) = \text{maximize}_\pi(v_\pi(s))$$

for some definition of $v_*(s)$ – as happiness, monetary value, or other rewards.



Theoretically, this is a sound assumption. We can always define a function that peaks where we want the behavior to be. But…

Bounded rationality*



Picture credits: [The Decision Lab]

*Herbert Simon, Nobel Prize in Economics (1978)

- We do not have a perfect model of the world
- We are not supercomputers, and even if we are, there is a “cost” to thinking too much
- There may be multiple value functions that are hard to reconcile (e.g., justice, fairness)

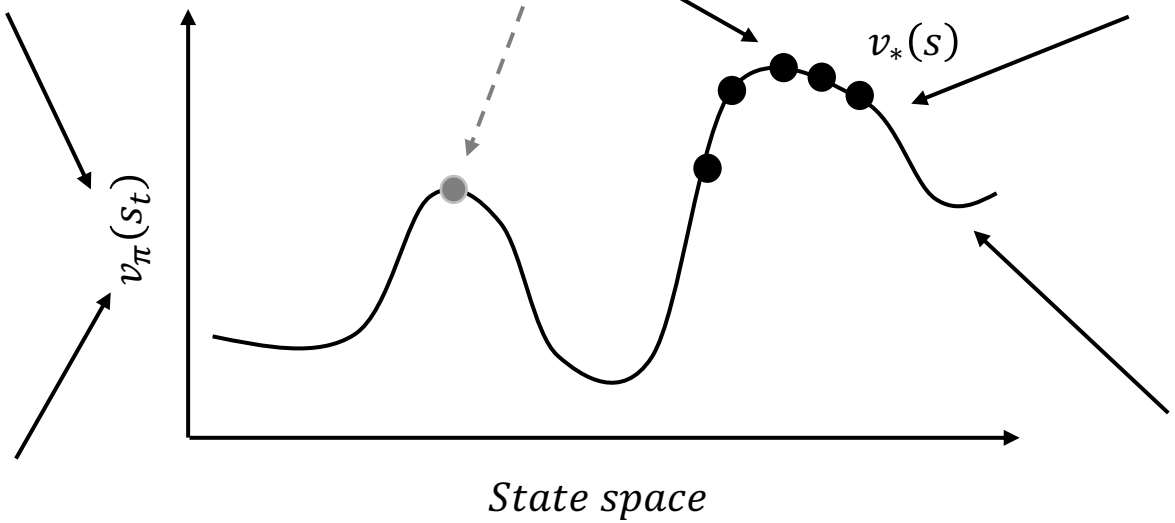
Therefore, we employ heuristics (e.g., our Q-network) to solve complex problems.

Issues with expected utility

Can actors always find this maximum point?

Is the state truly Markovian? (i.e., are there hidden variables?)

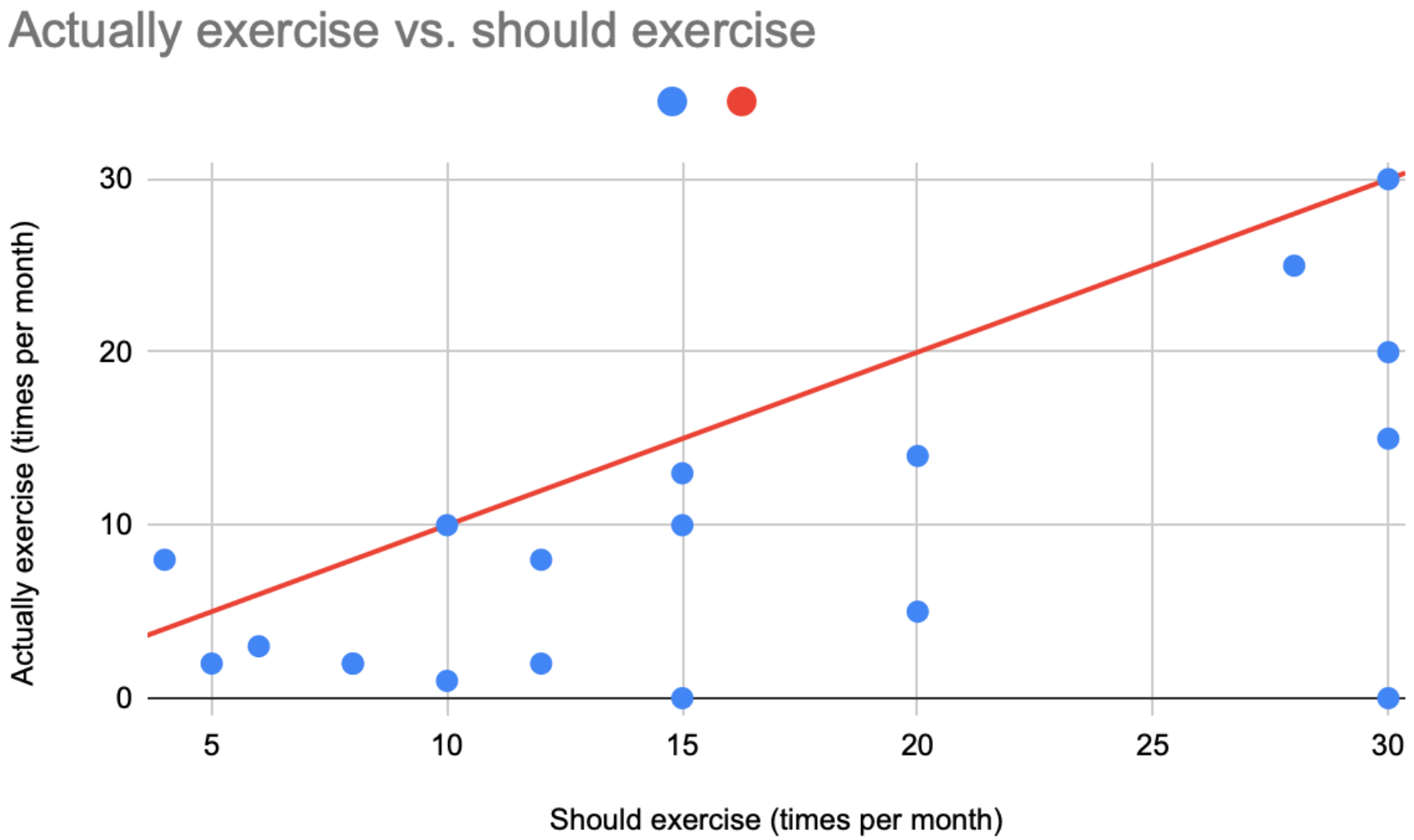
Do people always play rationally?



Does everyone in your model share this value function?

Does the agent have perfect knowledge of the system/have unlimited computational resources?

Do you exercise as much as you think you should?



Choices over several time periods

Present-bias: Putting more value on immediate reward than long-term reward

How do we model this?

Exponential discounting

Instead of maximizing total utility,

$$u_0 + u_1 + u_2 + u_3 + \cdots = \sum_{t=0}^{\infty} u_t,$$

the agent maximizes **exponentially discounted utility**,

$$u_0 + \delta u_1 + \delta^2 u_2 + \delta^3 u_3 \ldots = \sum_{t=0}^{\infty} \delta^t u_t,$$

where δ is the **discount factor** (usually ≤ 1).

At time t, utility discounted by δ^t

Bellman's equation with discounting

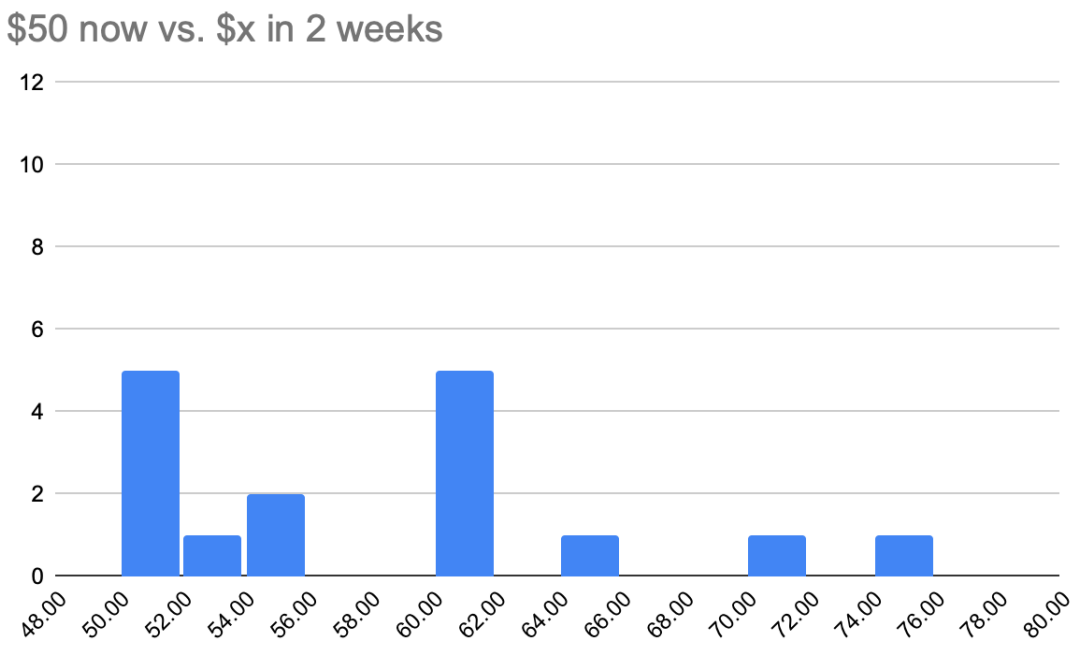
$$v_\pi(s_t) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \delta^k R_{t+k+1} | S = s_t \right] = r(s_t, a_t) + \delta \mathbb{E}[v_\pi(s_{t+1})]$$

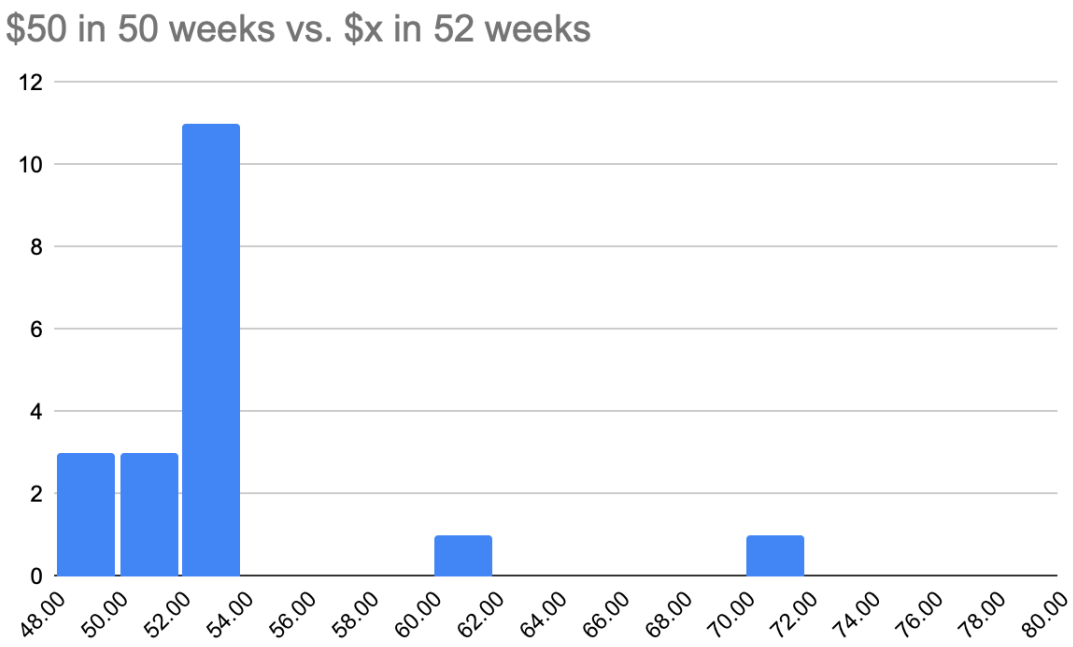
Need discounting in RL because:

1) Agent might not have perfect model of the environment
2) Environment might be stochastic (can't predict next state)
3) Computational trick to guarantee gradient descent convergence

i.e. if $R_t = 1$ for all t, and $\delta = \frac{1}{2}$, then $\sum_{t=0}^{\infty} \delta^t R_t = 1 + \frac{1}{2} + \left(\frac{1}{2}\right)^2 + \cdots = 1 + \underbrace{\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots}_{=1} = 2$

Are discount rates constant over time?





Impatience for 2-week delay now > impatience for 2-week delay in a year

Short-term vs. long-term discounting

Exponential discounting assumes same level of impatience (δ) over time

$$u(r_0) + \delta u(r_1) + \delta^2 u(r_2) + \delta^3 u(r_3) \ldots = \sum_{t=0}^{\infty} \delta^t u(r_t)$$

But evidence shows people are more impatient in the short-term and more patient in long-term

Quasi-hyperbolic discounting

Instead of maximizing exponentially discounted utility,

$$u(r_0) + \delta u(r_1) + \delta^2 u(r_2) + \delta^3 u(r_3) \ldots = \sum_{t=0}^{\infty} \delta^t u(r_t)$$

the agent maximizes **quasi-hyperbolic discounted utility**,

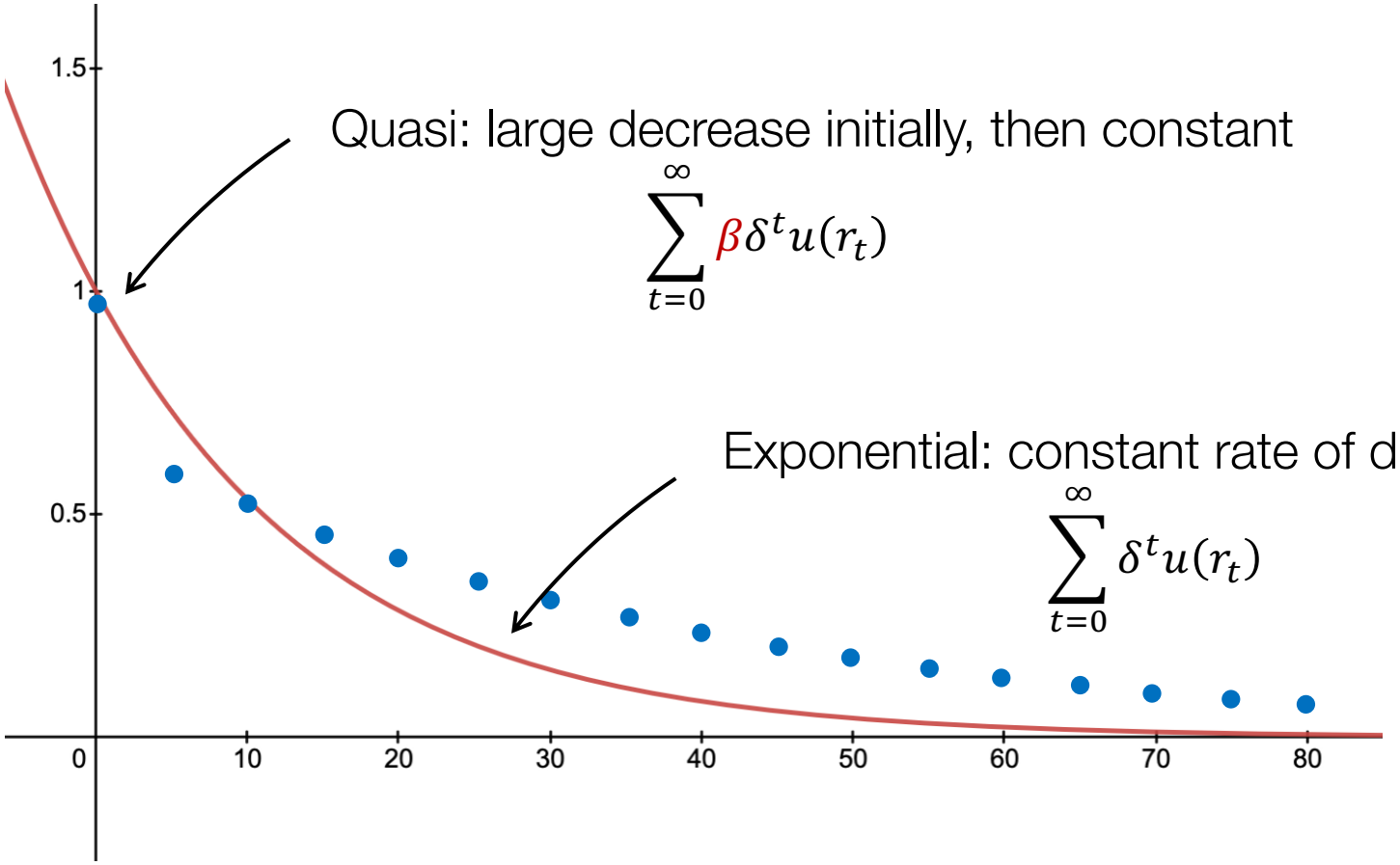
$$u(r_0) + \beta\delta u(r_1) + \beta\delta^2 u(r_2) + \beta\delta^3 u(r_3) \ldots = \sum_{t=0}^{\infty} \beta\delta^t u(r_t),$$

where $\delta \leq 1$is the **long-term discount factor** (usually ≤ 1),

and $0 \leq \beta \leq 1$ is the **short-term discount factor**

At time t, utility discounted by $\beta\delta^t$

Exponential vs. quasi-hyperbolic discounting



Exponential discounting

Quasi-hyperbolic discounting

Example: exponential discounting

Sihao has 3 days to complete slides for his HSSP class: $t = 0,1,2$. The instant cost of completing the slides increases each day as follows:

- Cost at $t = 0$: -18 utils
- Cost at $t = 1$: -24 utils
- Cost at $t = 2$: -30 utils

If he hasn't completed the slides during days $t = 0$ or $t = 1$, he must complete them on day $t = 2$.

Suppose Sihao is an exponential discounter with $\delta = \frac{1}{2}$, maximizing $u_0 + \frac{1}{2}u_1 + \left(\frac{1}{2}\right)^2 u_2$.

When would he complete the slides?

t=0

$$u_{\text{complete at t=0}} = -18 + \frac{1}{2}(0) + \left(\frac{1}{2}\right)^2 (0) = -18$$

$$u_{\text{complete at t=1}} = 0 + \frac{1}{2}(-24) + \left(\frac{1}{2}\right)^2 (0) = -12$$

$$u_{\text{complete at t=2}} = 0 + \frac{3}{4}(0) + \left(\frac{1}{2}\right)^2 (-30) = \boxed{-7.5}$$

t=1

$$u_{\text{complete at t=1}} = -24 + \frac{1}{2}(0) = -24$$

$$u_{\text{complete at t=2}} = 0 + \frac{1}{2}(-30) = \boxed{-15}$$

t=2

$$u_{\text{complete at t=2}} = \boxed{-30}$$

Example: exponential discounting

Sihao has 3 days to complete slides for his HSSP class: $t = 0,1,2$. The instant cost of completing the slides increases each day as follows:

-
-
-

If he

Exponential discounters have **time-consistent** preferences.

They always follow through with their plans

Suppose Sihao is an exponential discounter with $\delta = \frac{1}{2}$, maximizing $u_0 + \frac{1}{2}u_1 + \left(\frac{1}{2}\right)^2 u_2$.

When would he complete the slides?

t=0

$$u_{\text{complete at t=0}} = -18 + \frac{1}{2}(0) + \left(\frac{1}{2}\right)^2 (0) = -18$$

$$u_{\text{complete at t=1}} = 0 + \frac{1}{2}(-24) + \left(\frac{1}{2}\right)^2 (0) = -12$$

$$u_{\text{complete at t=2}} = 0 + \frac{3}{4}(0) + \left(\frac{1}{2}\right)^2 (-30) = \boxed{-7.5}$$

t=1

$$u_{\text{complete at t=1}} = -24 + \frac{1}{2}(0) = -24$$

$$u_{\text{complete at t=2}} = 0 + \frac{1}{2}(-30) = \boxed{-15}$$

t=2

$$u_{\text{complete at t=2}} = \boxed{-30}$$

Example: quasi-hyperbolic discounting

Sihao has 3 days to complete slides for his HSSP class: $t = 0,1,2$. The instant cost of completing the slides increases each day as follows:

- Cost at $t = 0$: -18 utils
- Cost at $t = 1$: -24 utils
- Cost at $t = 2$: -30 utils

If he hasn't completed the slides during days $t = 0$ or $t = 1$, he must complete them on day $t = 2$.

Suppose Sihao is a quasi-hyperbolic discounter with $\delta = 1$, $\beta = \frac{1}{2}$, maximizing $u_0 + \frac{1}{2}u_1 + \frac{1}{2}u_2$.

When would he complete the slides?

t=0

$$u_{\text{complete at t=0}} = -18 + \frac{1}{2}(0) + \frac{1}{2}(0) = -18$$

$$u_{\text{complete at t=1}} = 0 + \frac{1}{2}(-24) + \frac{1}{2}(0) = \boxed{-12}$$

$$u_{\text{complete at t=2}} = 0 + \frac{1}{2}(0) + \frac{1}{2}(-30) = -15$$

t=1

$$u_{\text{complete at t=1}} = -24 + \frac{1}{2}(0) = -24$$

$$u_{\text{complete at t=2}} = 0 + + \frac{1}{2}(-30) = \boxed{-15}$$

t=2

$$u_{\text{complete at t=2}} = \boxed{-30}$$

Example: quasi-hyperbolic discounting

Sihao has 3 days to complete slides for his HSSP class: $t = 0,1,2$. The instant cost of completing the slides increases each day as follows:

-
-
-

If he

Quasi-hyperbolic discounters have **time-inconsistent** preferences.

Their optimal plan may change once they are in the future state.

Suppose Sihao is a quasi-hyperbolic discounter with $\delta = 1$, $\beta = \frac{1}{2}$, maximizing $u_0 + \frac{1}{2}u_1 + \frac{1}{2}u_2$.

When would he complete the slides?

t=0

$$u_{\text{complete at t=0}} = -18 + \frac{1}{2}(0) + \frac{1}{2}(0) = -18$$

$$u_{\text{complete at t=1}} = 0 + \frac{1}{2}(-24) + \frac{1}{2}(0) = \boxed{-12}$$

$$u_{\text{complete at t=2}} = 0 + \frac{1}{2}(0) + \frac{1}{2}(-30) = -15$$

t=1

$$u_{\text{complete at t=1}} = -24 + \frac{1}{2}(0) = -24$$

$$u_{\text{complete at t=2}} = 0 + + \frac{1}{2}(-30) = \boxed{-15}$$

t=2

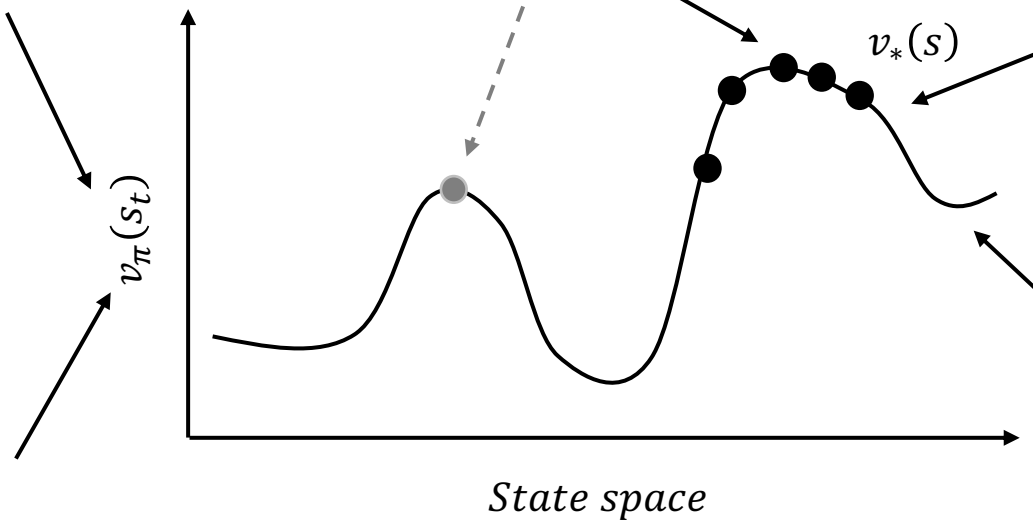
$$u_{\text{complete at t=2}} = \boxed{-30}$$

Issues with expected utility

Can actors always find this maximum point?

Is the state truly Markovian? (i.e., are there hidden variables?)

Do people always play rationally?

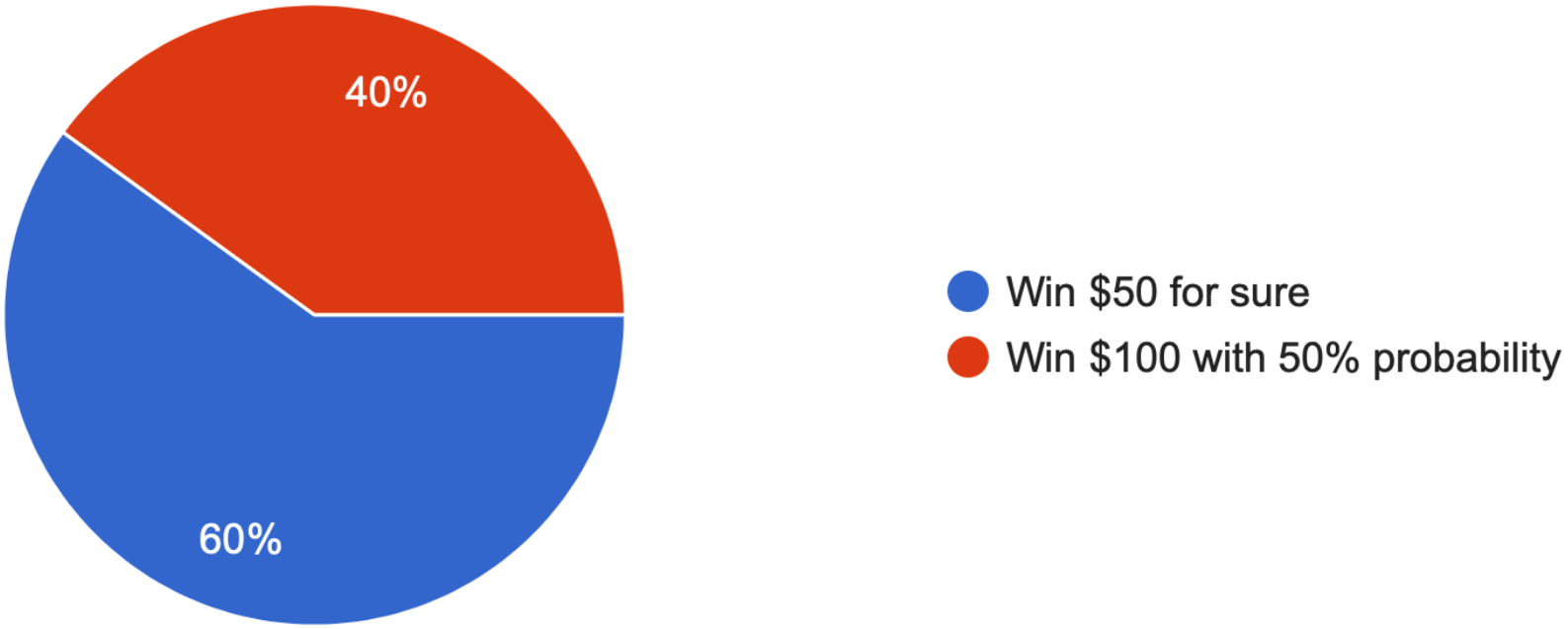


Does everyone in your model share this value function?

Does the agent have perfect knowledge of the system/have unlimited computational resources?

Would you take the gamble?

Would you choose to win $50 for sure OR win $100 with 50% probability (and win $0 otherwise)?

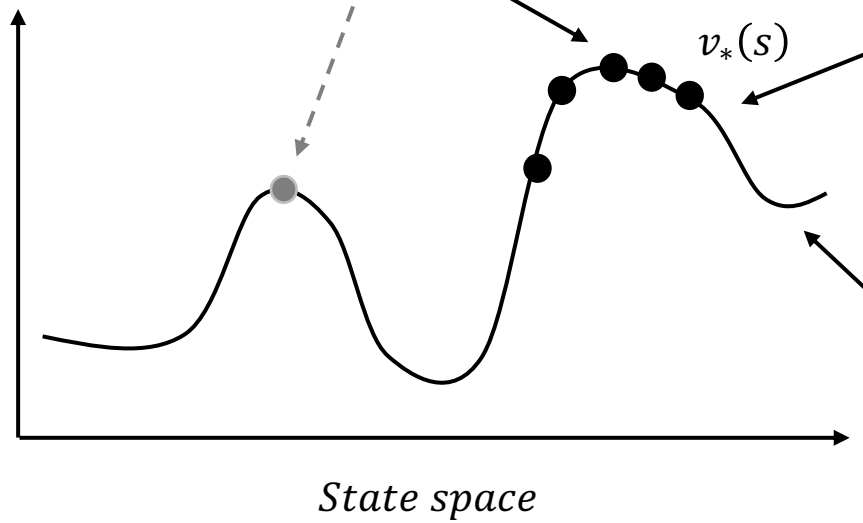


Issues with expected utility

Can actors always find this maximum point?

Is the state truly Markovian? (i.e., are there hidden variables?)

Do people always play rationally?



Does everyone in your model share this value function?

Does the agent have perfect knowledge of the system/have unlimited computational resources?

How do we model an individual's "riskiness"?

Expected payoff vs. expected utility

Consider a gamble with payoff x. For example, $x = \begin{cases} \$100 \text{ with prob. } 1/2 \\ \$0 \text{ with prob. } 1/2 \end{cases}$

The expected payoff of this gamble is $\mathbb{E}[x] = \frac{1}{2}(\$100) + \frac{1}{2}(\$0) = \$50$

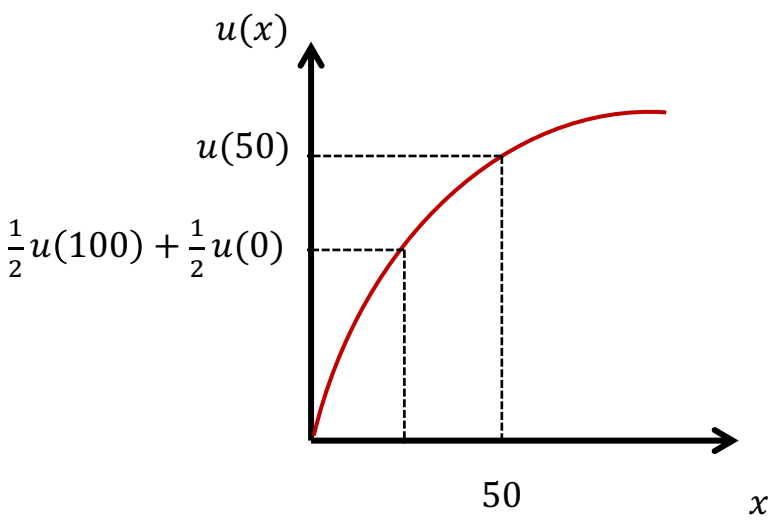
The expected utility of this gamble is $\mathbb{E}[u(x)] = \frac{1}{2}u(\$100) + \frac{1}{2}u(\$0)$

An agent is:

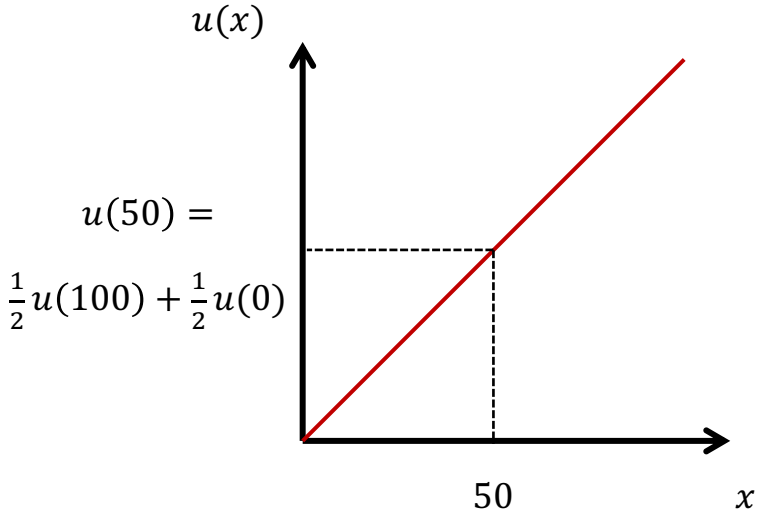
- **Risk-averse** if prefer to guarantee payoff $\mathbb{E}[x]$ over taking the gamble (i.e. $u(\mathbb{E}[x]) > \mathbb{E}[u(x)]$)
- **Risk-neutral** if indifferent between getting $\mathbb{E}[x]$ and taking the gamble (i.e. $u(\mathbb{E}[x]) = \mathbb{E}[u(x)]$)
- **Risk-loving** if they prefer to gamble over guaranteeing payoff $\mathbb{E}[x]$ (i.e. $u(\mathbb{E}[x]) > \mathbb{E}[u(x)]$)

Risk aversion (graphically)

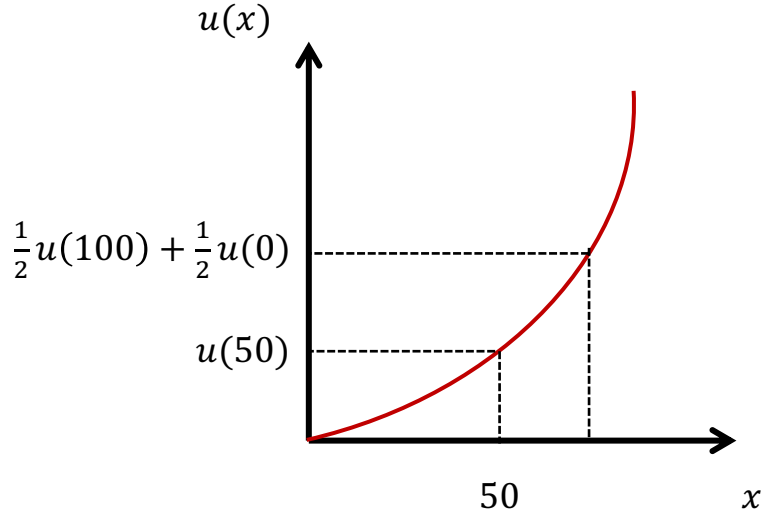
- Risk-averse $\Leftrightarrow u(\mathbb{E}[x]) > \mathbb{E}[u(x)] \Leftrightarrow u(R)$ is concave
- Risk-neutral $\Leftrightarrow u(\mathbb{E}[x]) = \mathbb{E}[u(x)] \Leftrightarrow u(R)$ is linear
- Risk-loving $\Leftrightarrow u(\mathbb{E}[x]) < \mathbb{E}[u(x)] \Leftrightarrow u(R)$ is convex



Risk-averse



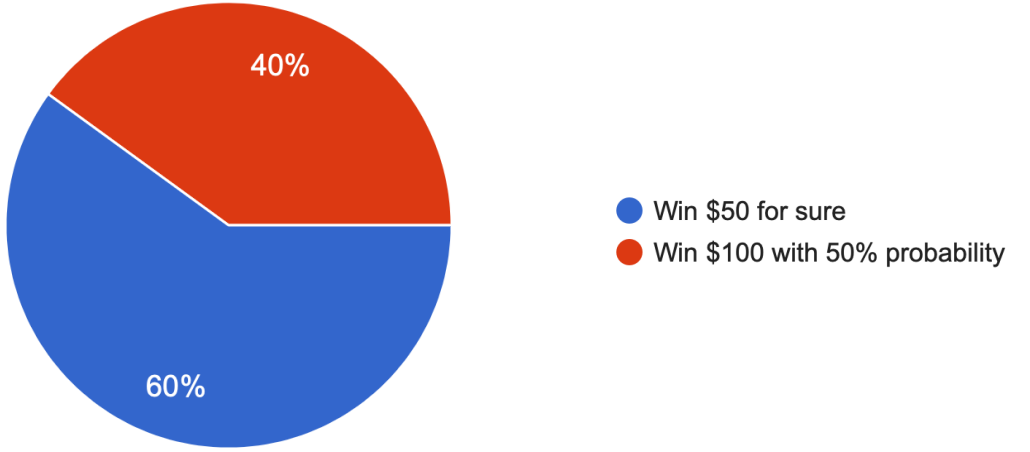
Risk-neutral



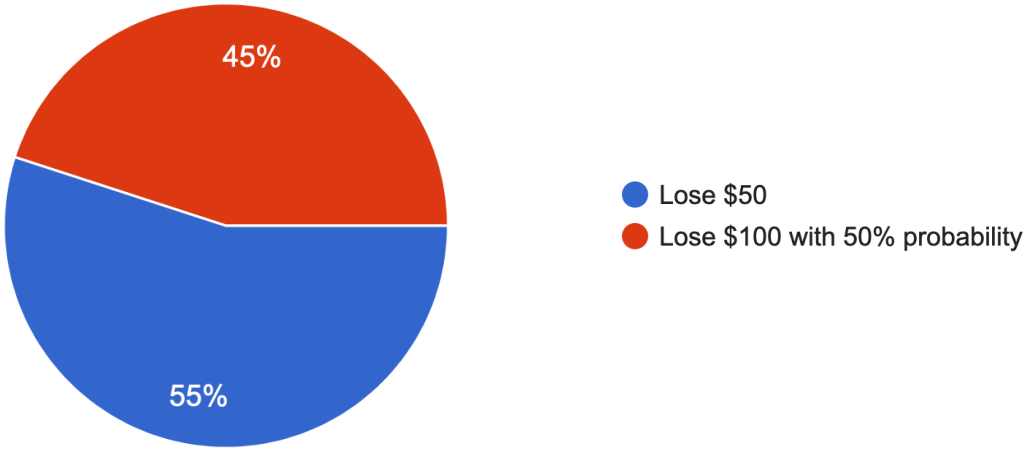
Risk-loving

Winning vs. losing

1) Would you choose to win $50 for sure OR win $100 with 50% probability (and win $0 otherwise)?



2) Suppose you were given $100, but then had to choose one of two options. Would you rather lose $50 for sure, OR lose $100 with 50% probability (and lose $0 otherwise)?



But these gambles are equivalent!

Loss aversion

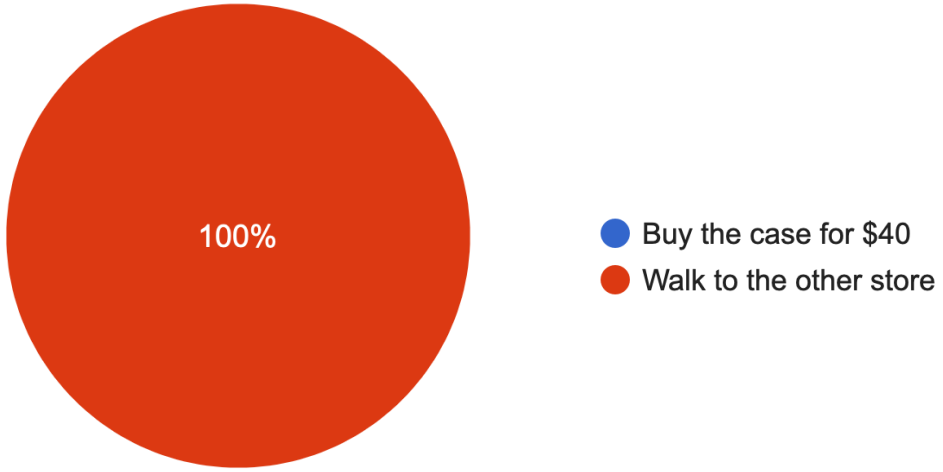
Most people are **loss-averse** (losses hurt more than equivalent gains help)

Ex: Investors sell stocks when they're up, and hold when they're down

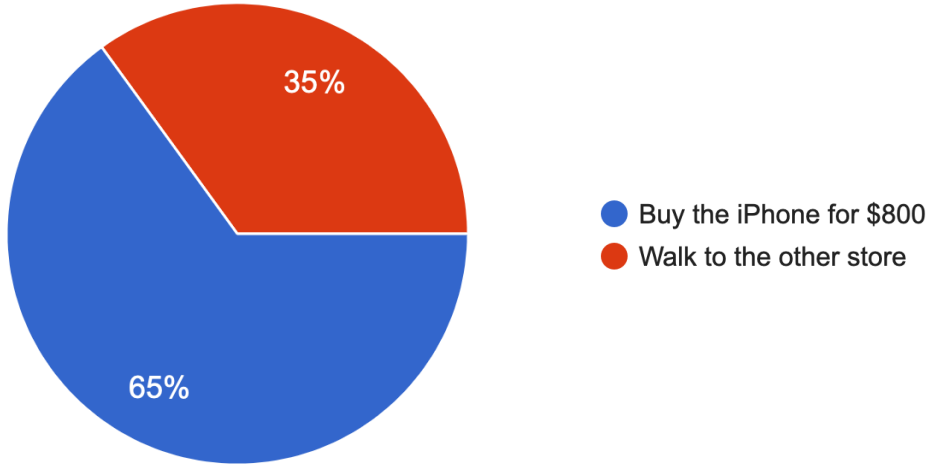
Ex: People are more likely to buy a car after test driving it

$40 reference point vs. $800

1) You want to purchase an iPhone case for $40. The salesperson tells you that you can get the exact same case in a nearby store for $20 off. You would need to walk for 30 minutes in total. Would you go to the other store?



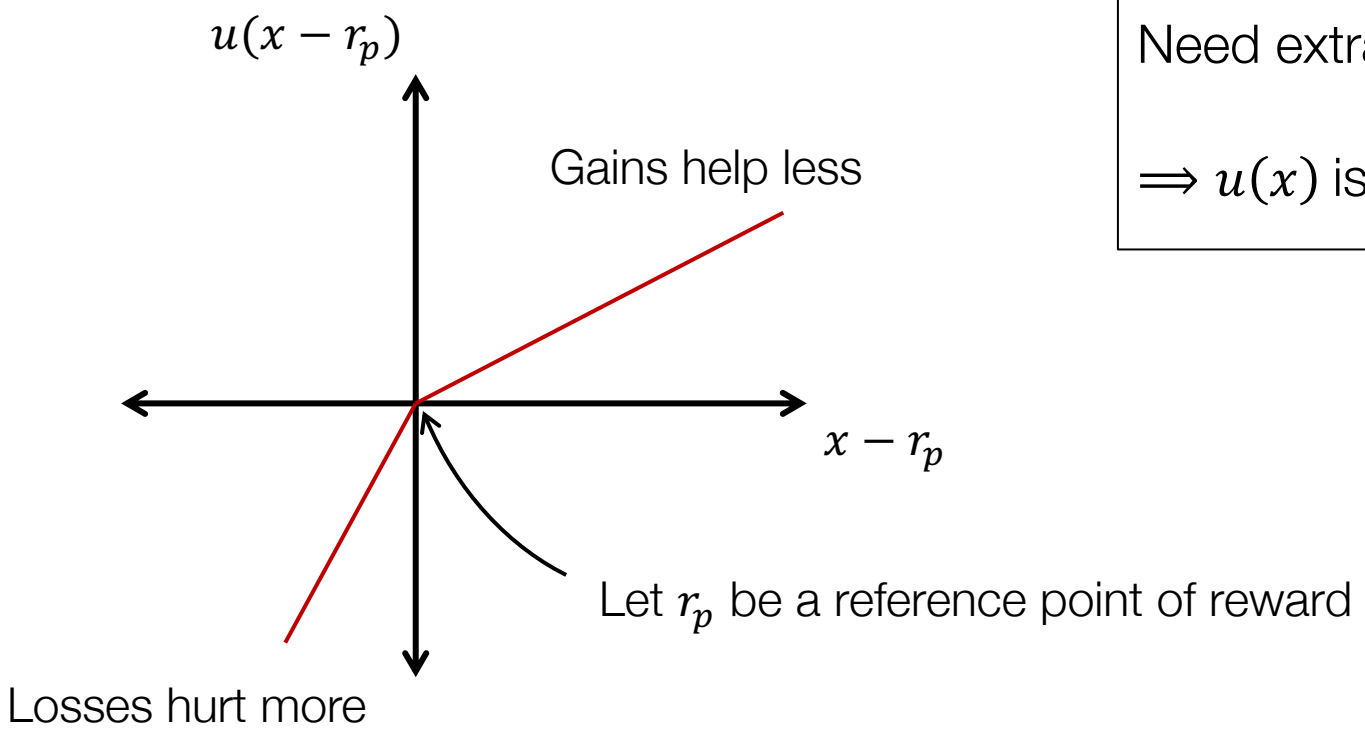
2) You want to purchase an iPhone for $800. The salesperson tells you that you can get the exact same case in a nearby store for $20 off. You would need to walk for 30 minutes in total. Would you go to the other store?



It's all relative

Reference-dependent utility: outcomes are evaluated relative to a reference point

Most people are **loss-averse:**



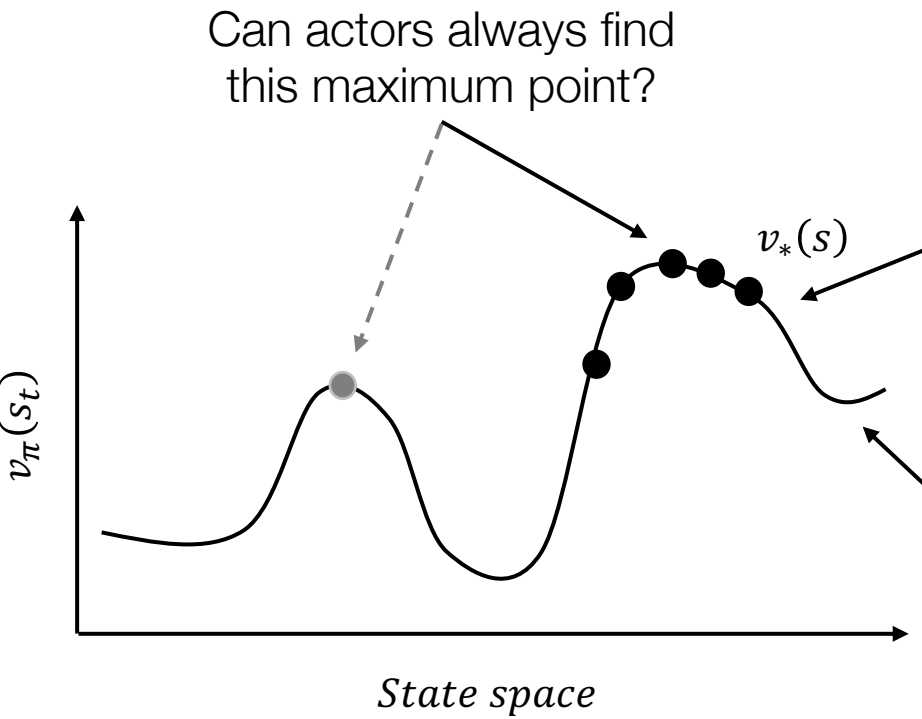
Need extra information R_p to evaluate utility

$\Longrightarrow u(x)$ is not Markovian

Issues with expected utility

Is the state truly Markovian? (i.e., are there hidden variables?)

Does everyone in your model share this value function?

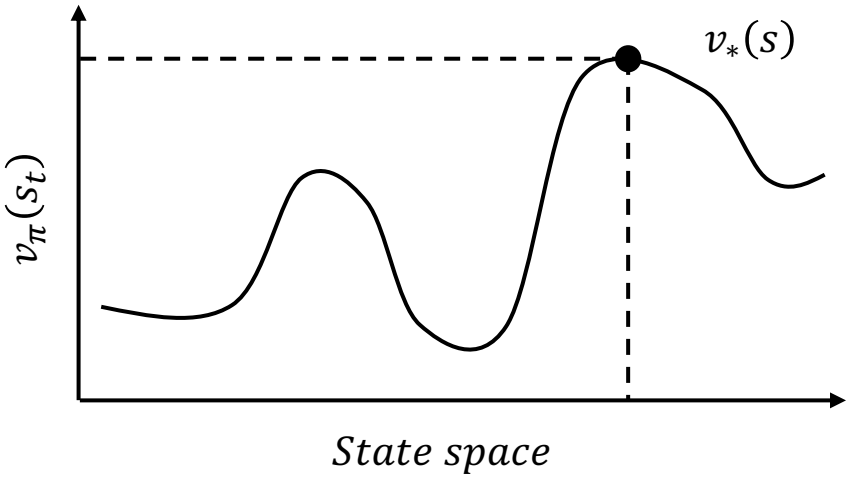


Do people always play rationally?

Does the agent have perfect knowledge of the system/have unlimited computational resources?

Recap

- Utility theory in reinforcement learning and social science
- Bounded rationality: decision-making under uncertainty and limited resources
- Discounting (exponential and quasi-hyperbolic)
- Risk aversion and reference-dependent utility



Lecture 1	Lecture 2	Lecture 3	Lecture 4	Lecture 5	Lecture 6
Introduction and the RL problem	How computers learn	How people learn	Multi-agent systems	Interactions on graphs	Complex systems science

References and additional resources

- Towards Causal Representation Learning by Bernhard Schölkopf et al.
- MIT 14.13 – Psychology and Economics lecture notes and videos
- “Thinking, Fast and Slow” by Daniel Kahneman
- Hyperbolic Discounting and Learning Over Multiple Horizons by Fedus et al.