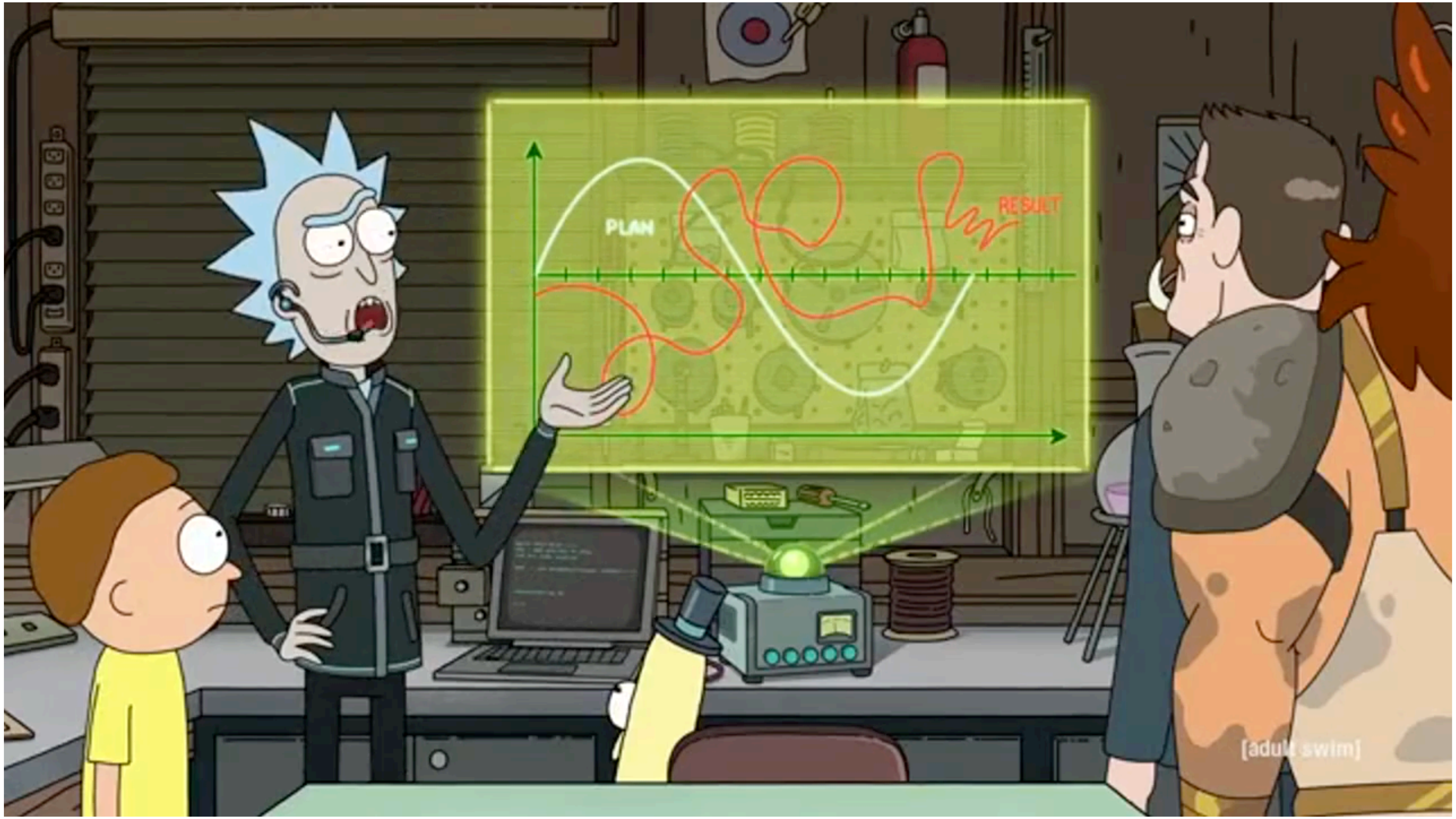
When do you explore & when do you exploit?

Readings for today

- Wilson, R. C., Bonawitz, E., Costa, V. D., & Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. Current Opinion in Behavioral Sciences, 38, 49-56.

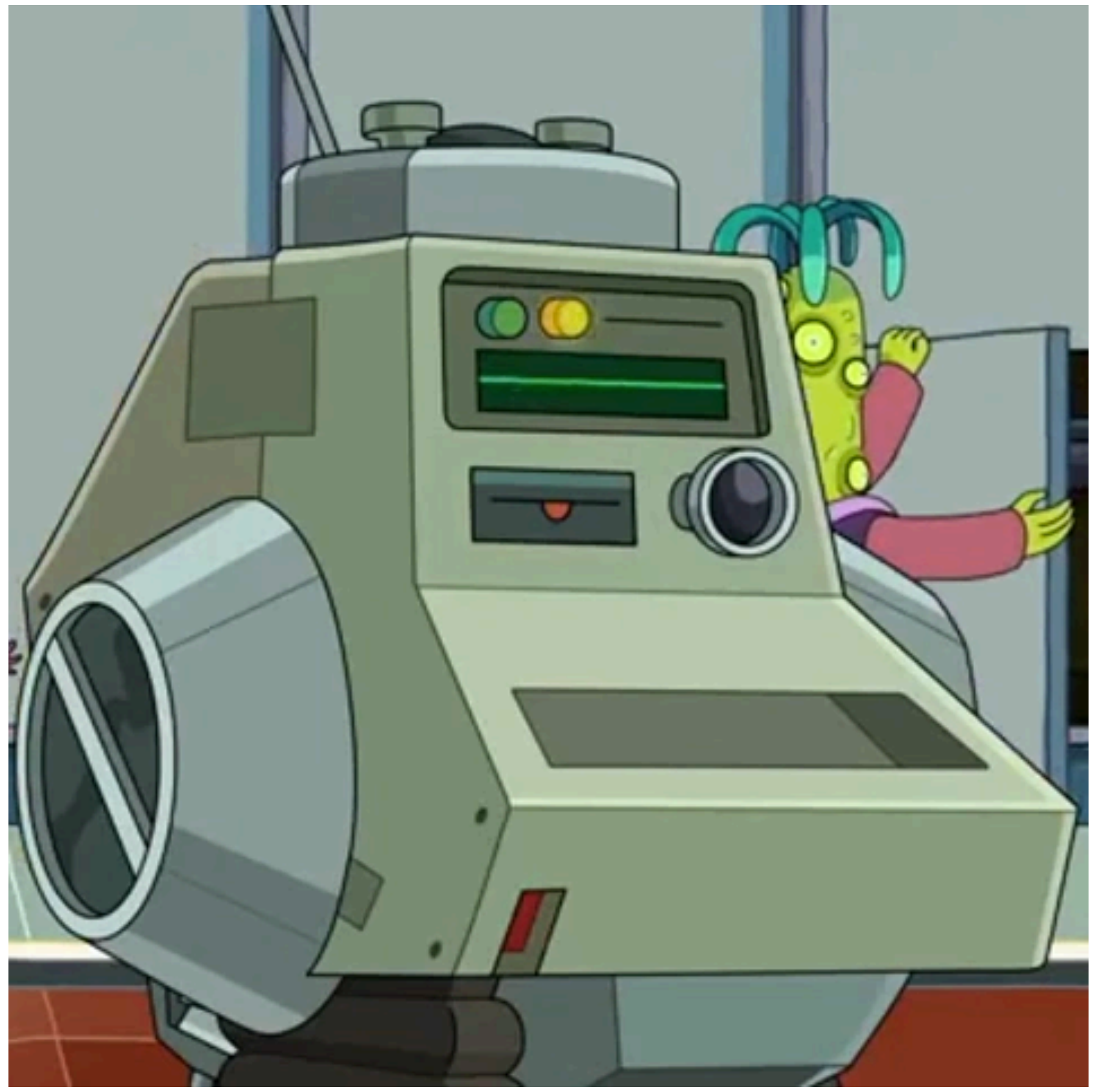
The dilemma



Rick & Morty: One Crew Over the Crewcoo's Morty

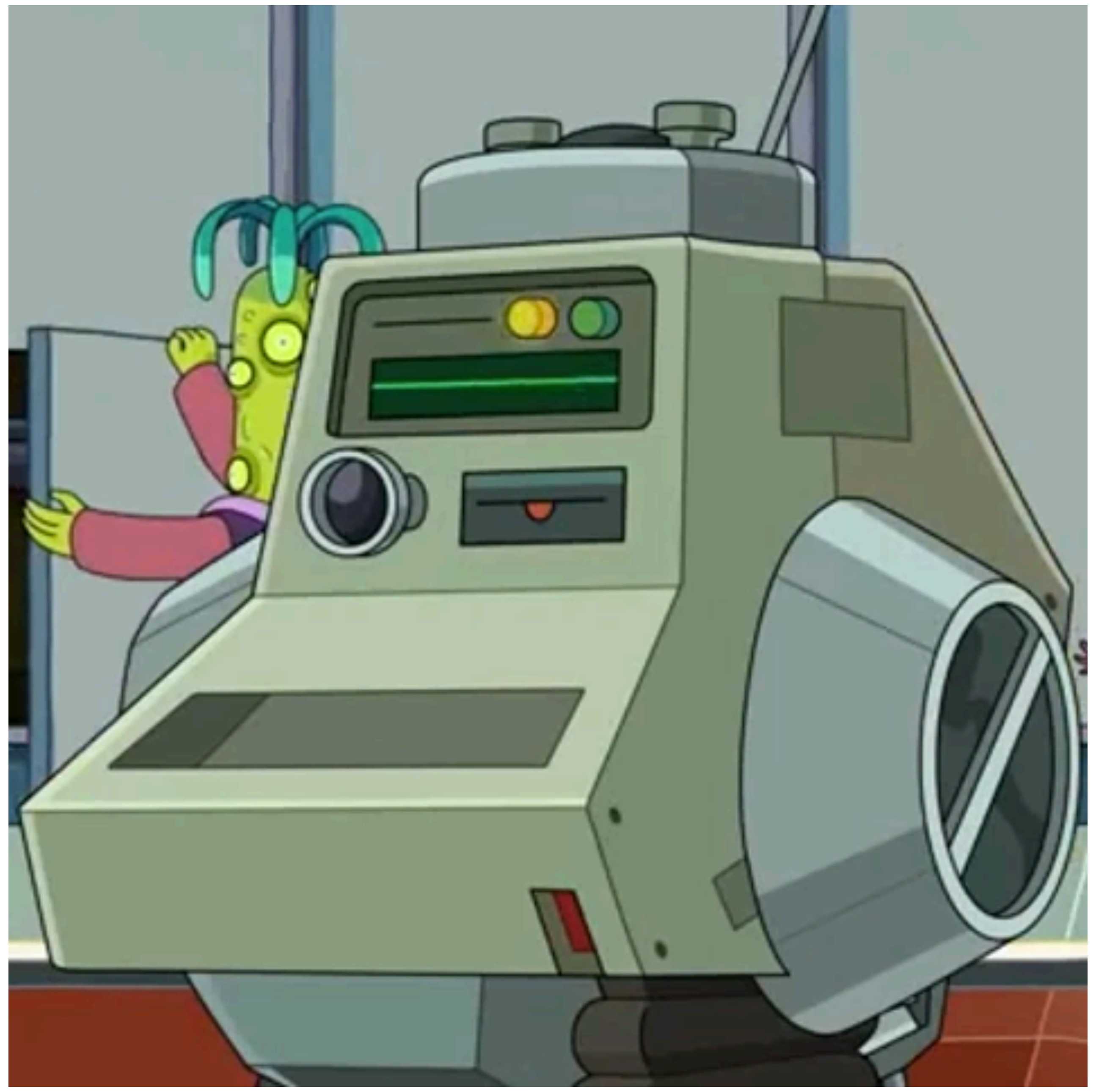
Battle of the bots

Heistotron



- Exploitative
- Strategic
- Resource maximizing

Randotron



- Exploratory
- Random
- Entropy maximizing

Rick & Morty: One Crew Over the Crewcoo's Morty

The exploitation-exploration (e-e) dilemma

Exploitation: Choosing a behavior that is most likely to produce the best outcome.

- Choosing a “hot” slot machine
- Going to your regular restaurant
- Buying a Honda Civic

Exploration: Choosing a behavior with a less certain outcome on the chance that it will produce more desirable outcome.

- Trying a new slot machine
- Going to a restaurant that has just opened
- Buying a Tesla

The ϵ-greedy method

Action value

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \,|\, A_t = a}{\sum_{i=1}^{t-1} A_t = a}$$

Best action

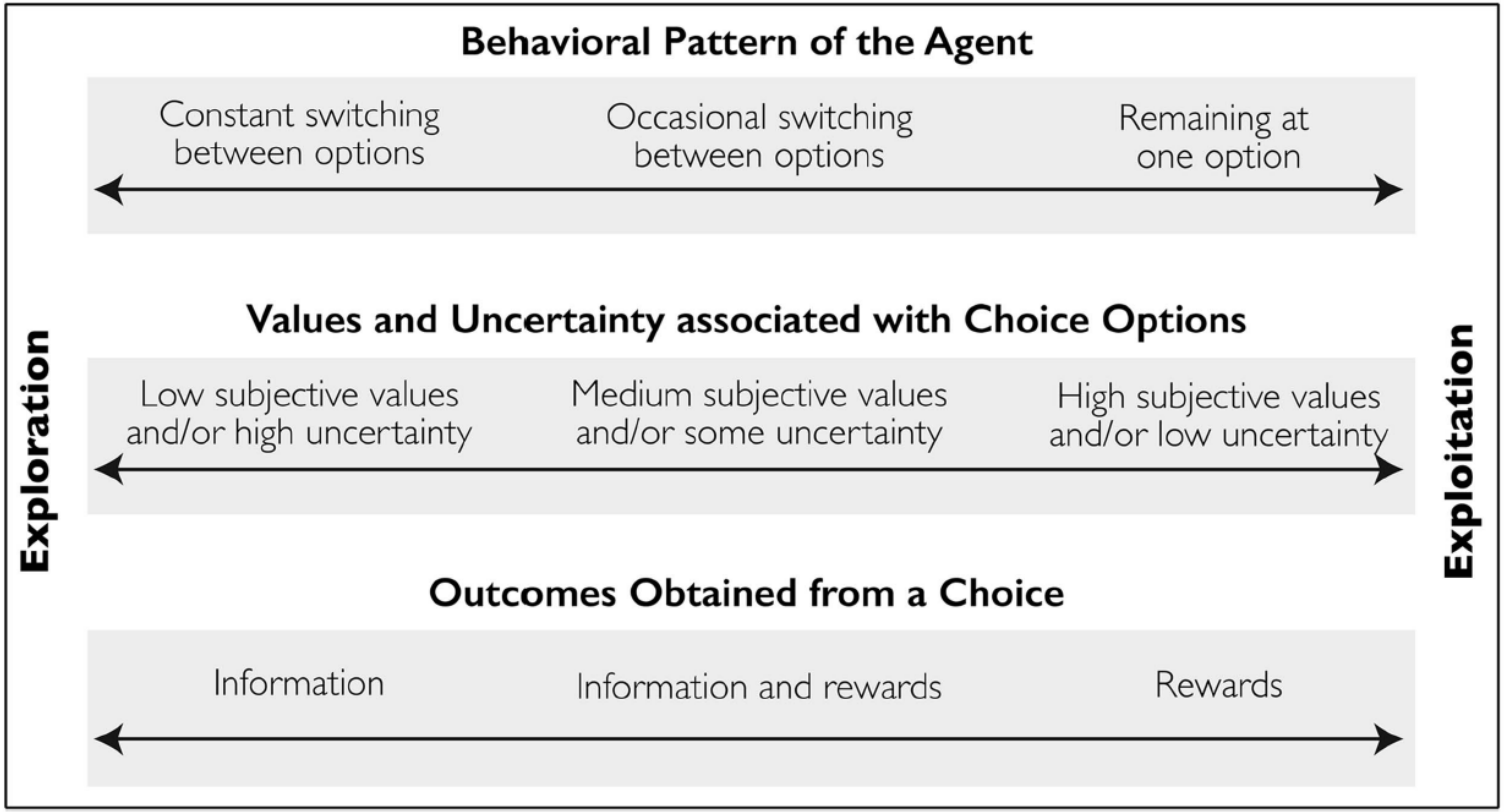
$$A_t = \arg\max_{a} Q_t(a)$$

Decision policy

$$\begin{array}{rr} \max Q_t(a), & \text{with probability } 1-\epsilon \\ \text{any } a, & \text{with probability } \epsilon \end{array}$$

(Sutton & Barto, 2020)

The e-e dilemma



Exploration

Exploitation

Behavioral Pattern of the Agent

Constant switching between options — Occasional switching between options — Remaining at one option

Values and Uncertainty associated with Choice Options

Low subjective values and/or high uncertainty — Medium subjective values and/or some uncertainty — High subjective values and/or low uncertainty

Outcomes Obtained from a Choice

Information — Information and rewards — Rewards

(Melhorn et al. 2018)

Factors that drive the e-e dilemma

Individual Factors

- Cognitive capacity (e.g., memory span)
- Aspiration levels (e.g., greediness)
- Internal latent state (e.g., energy level, drive)
- Prior knowledge (e.g., experience-dependent expectations)
- Morphology (e.g., larger animals more likely to explore)
- Demographics (e.g., delayed discounting changes with age)
- Neurotransmitters (e.g., levels of norepinephrine determine exploration)

(Melhorn et al. 2018)

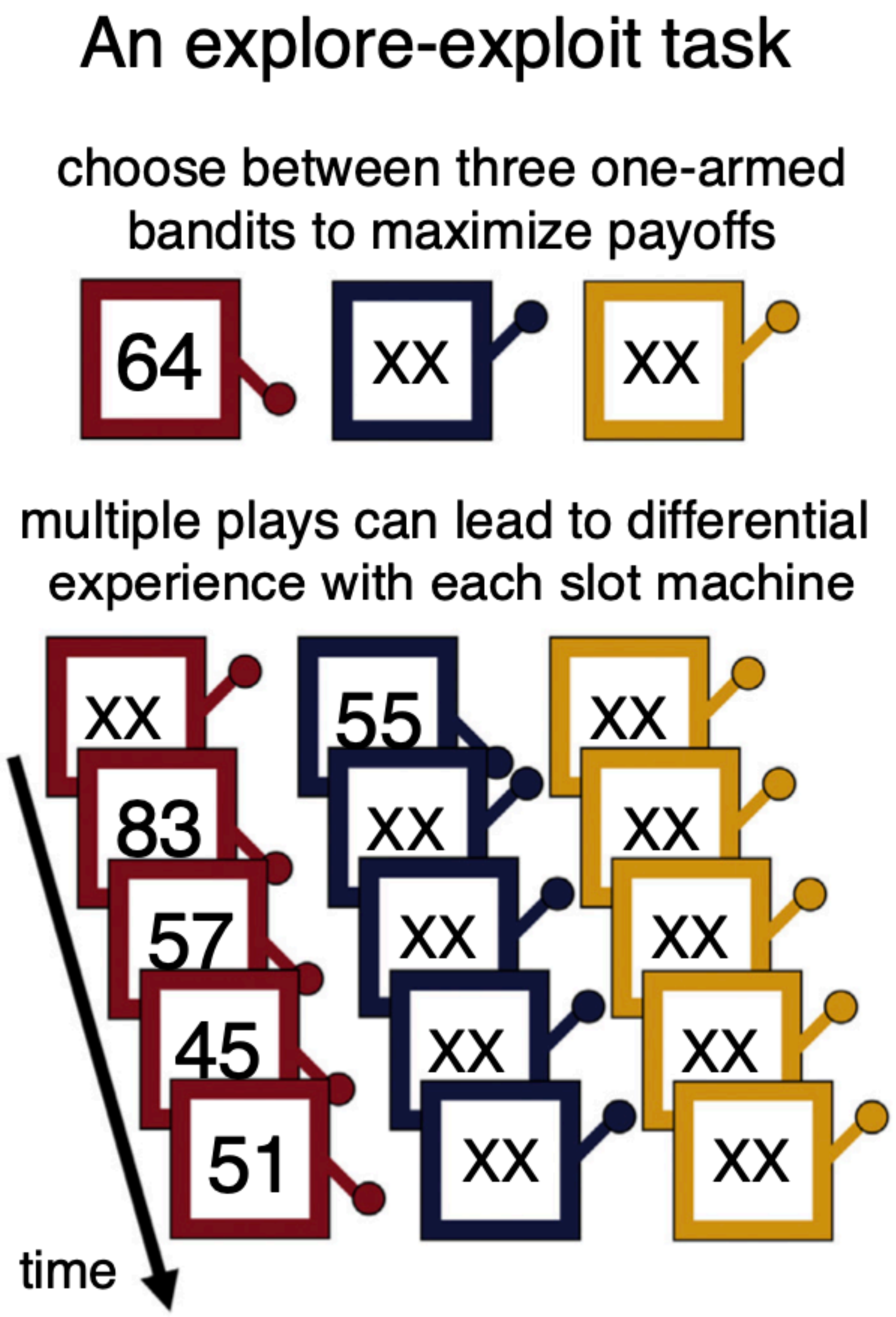
Factors that drive the e-e dilemma

Environmental Factors

- Availability of resources (e.g., depletion of food sources)
- Availability of information about options (e.g., foregone payoff information)
- Cost of information vs. value of reward (e.g., search effort)
- Structure of the environment (e.g., distribution of food sources)
- Probability of gains and losses (e.g., over exploring during “rare disasters”)
- Stability of environmental contingencies (e.g., volatility)
- Shape of reward distributions (e.g., bimodal distributions = more sampling)
- Range of possible actions (i.e., the behavioral “horizon”)

(Melhorn et al. 2018)

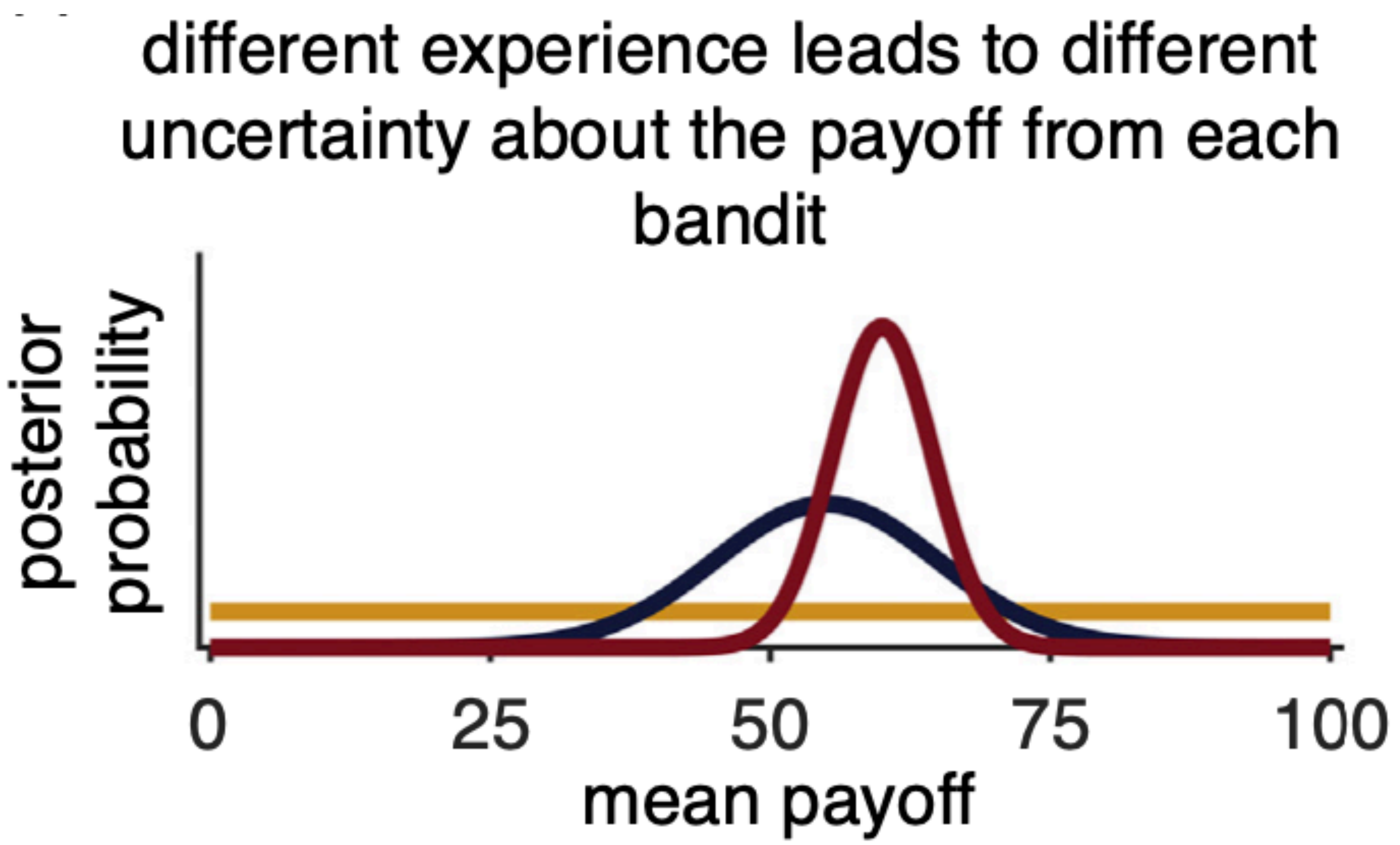
The bandit task



An explore-exploit task

choose between three one-armed bandits to maximize payoffs

multiple plays can lead to differential experience with each slot machine



different experience leads to different uncertainty about the payoff from each bandit

(Wilson et al. 2020)

Two types of ways to explore

Random exploration

$$Q(a) = r(a) + \eta(a)$$

How good we expect a to be

Random noise

Example: softmax selection policy

$$p(a) = \frac{e^{Q(a)/\tau}}{\sum_{i=1}^{A} e^{Q(i)/\tau}}$$

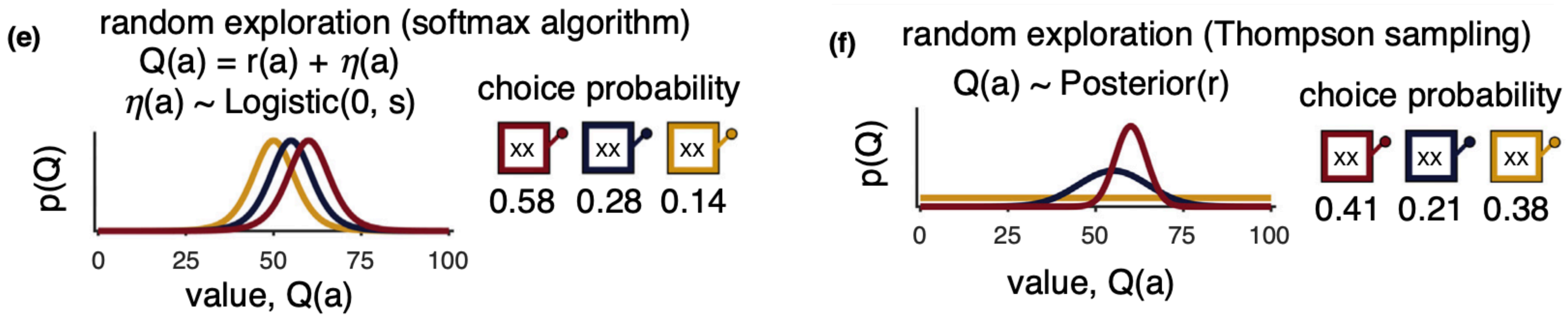
“temperature” parameter

larger τ = more random

(Wilson et al. 2020)

Two types of ways to explore

Random exploration



(Wilson et al. 2020)

Two types of ways to explore

Directed exploration

$$Q(a) = r(a) + IB(a)$$

How good we expect a to be

Information bonus

Example: Upper confidence bound

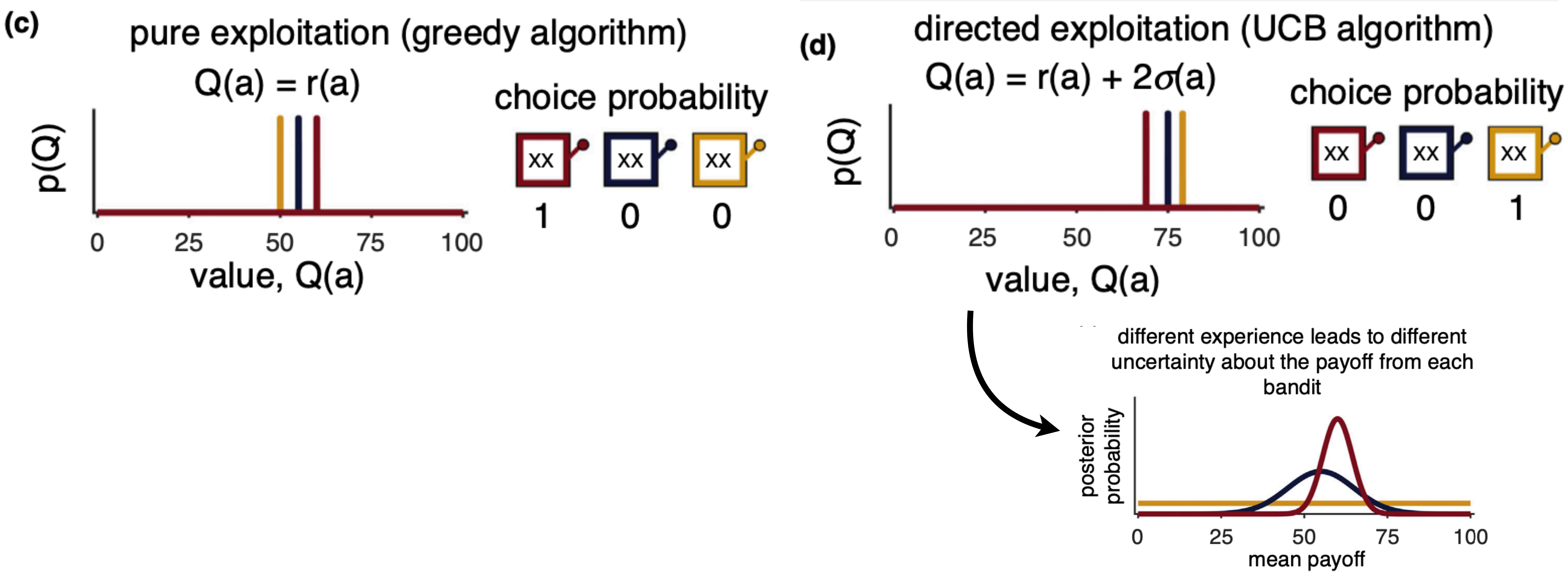
variance of the posterior distribution

$$p(a) = Q(a) + 2\sigma(a)$$

(Wilson et al. 2020)

Two types of ways to explore

Directed exploration



(Wilson et al. 2020)

Food for thought

Small group (2-3 people) exercise:

Come up with two “real world” examples of situations where someone would have to make an exploratory decisions:

1. Where directed exploration would be the best decision.
2. Where random exploration would be the best decision

Justify why each decision policy would be the most appropriate.