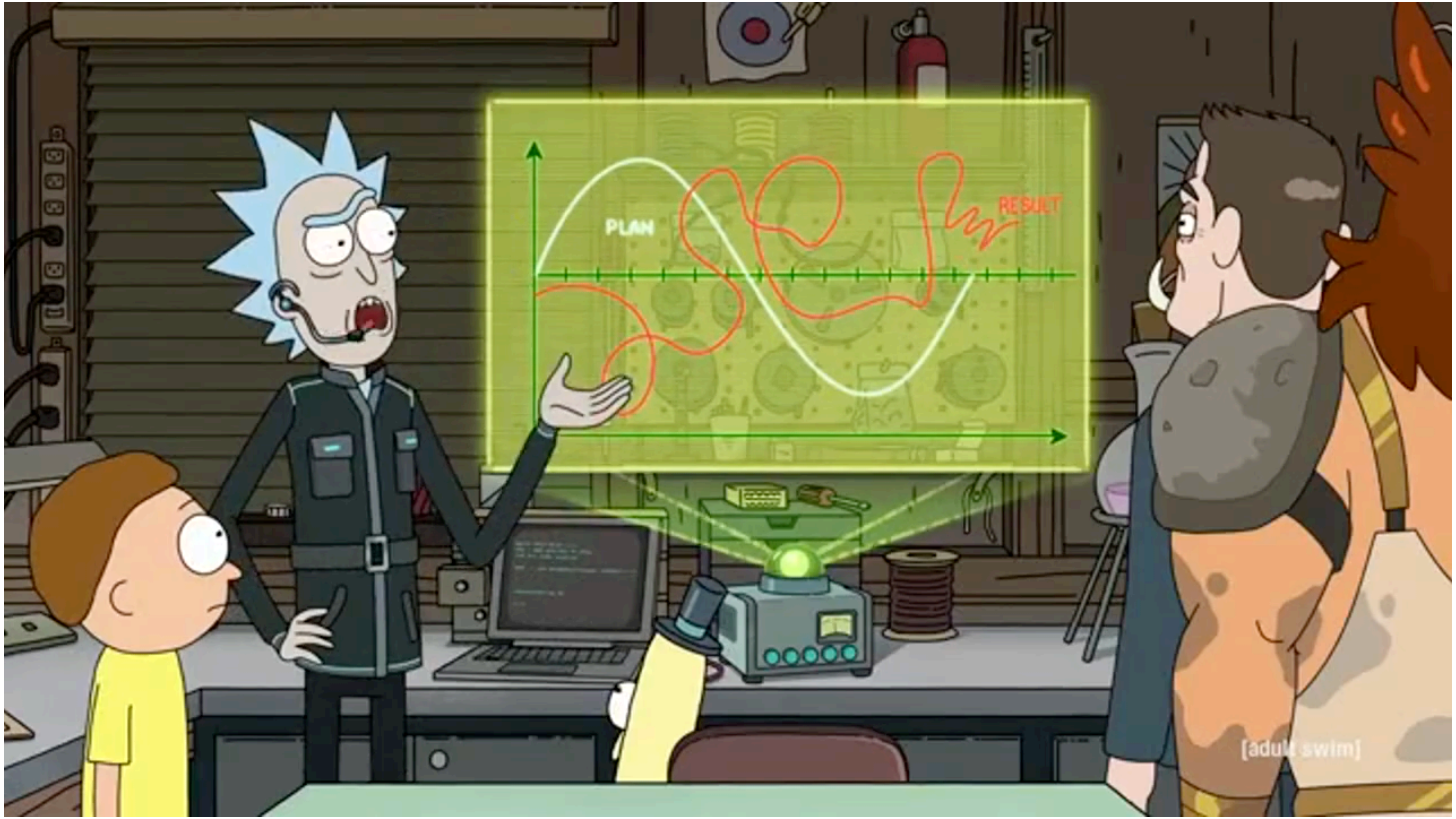
When to explore rather than exploit?

Readings for today

- Peterson, E. J., & Verstynen, T. D. (2022). Embracing curiosity eliminates the exploration-exploitation dilemma. bioRxiv, 671362.

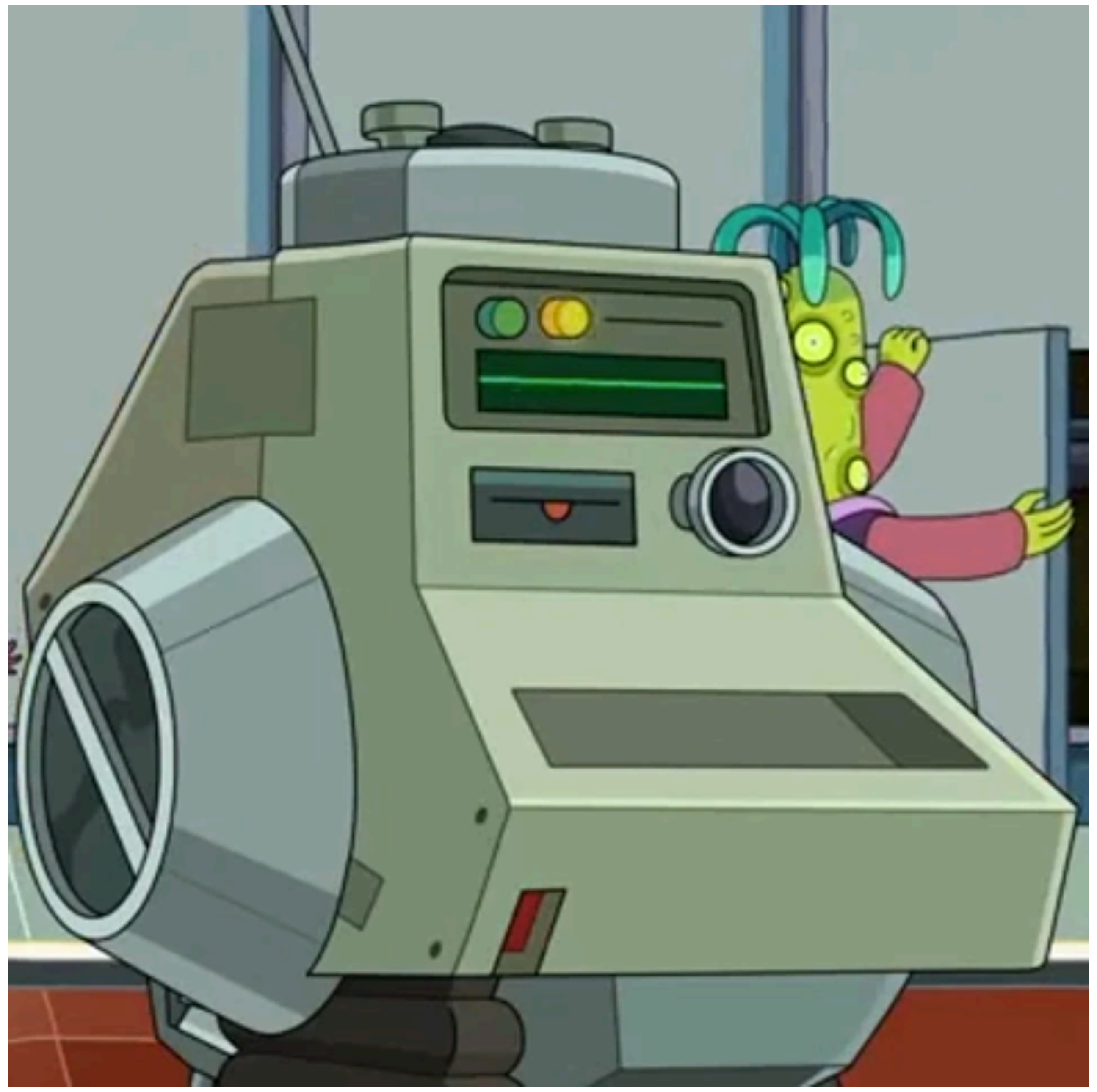
The dilemma



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

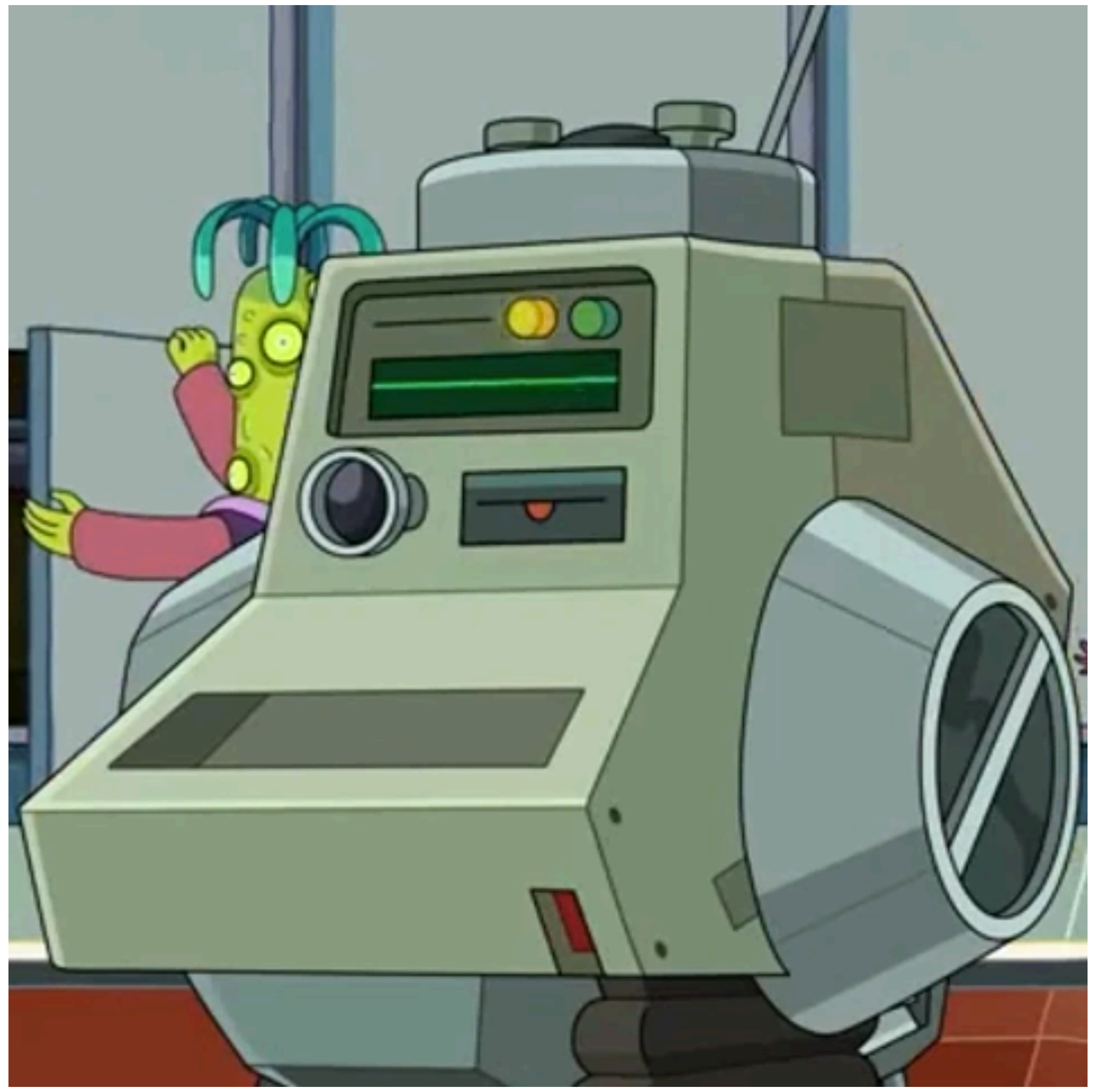
Battle of the bots

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

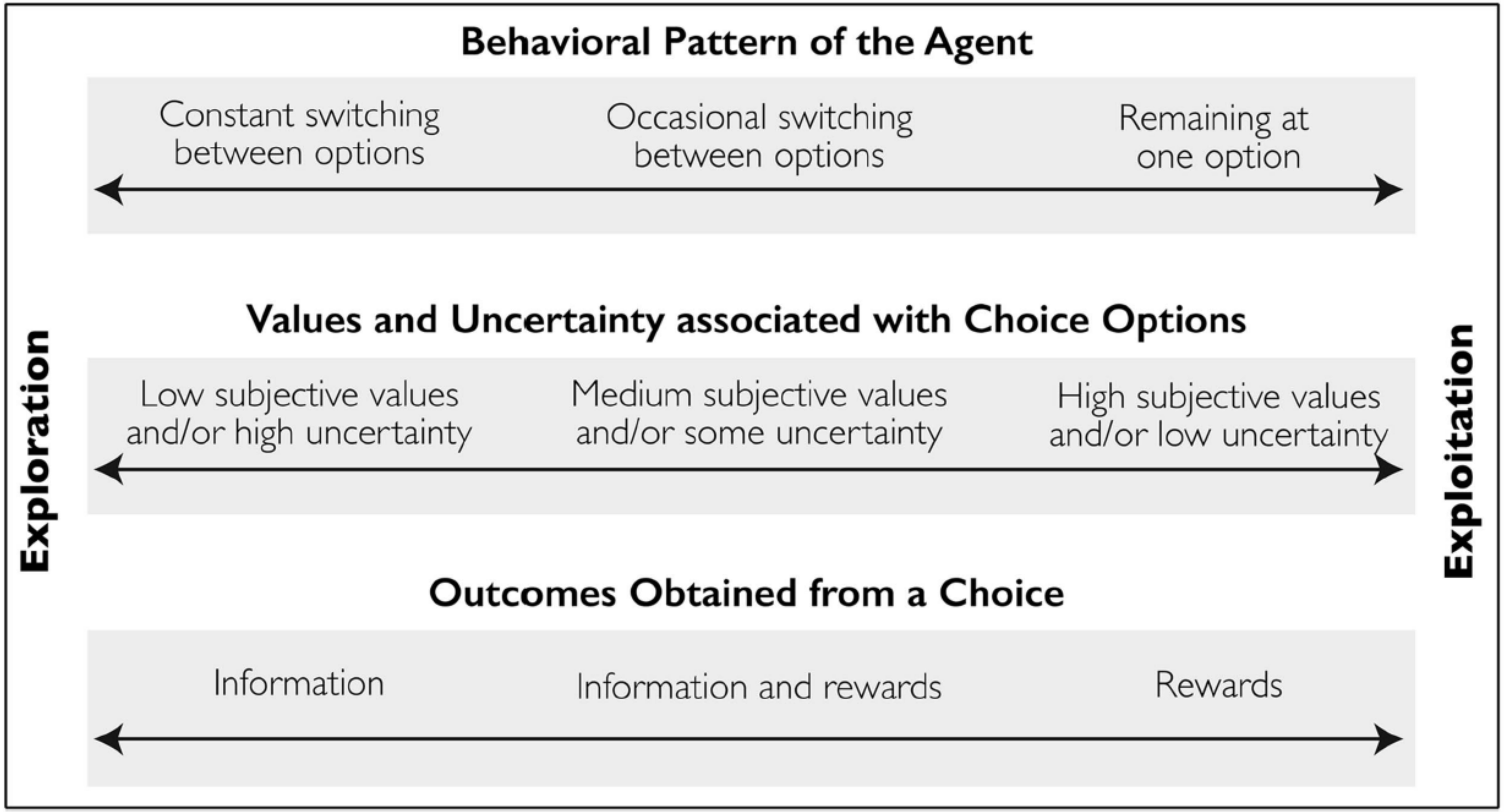
$$A_t = \arg\max_a Q_t(a)$$

Decision policy

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

(Sutton & Barto, 2020)

The e-e dilemma



(Melhorn et al. 2018)

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

"temperature" parameter

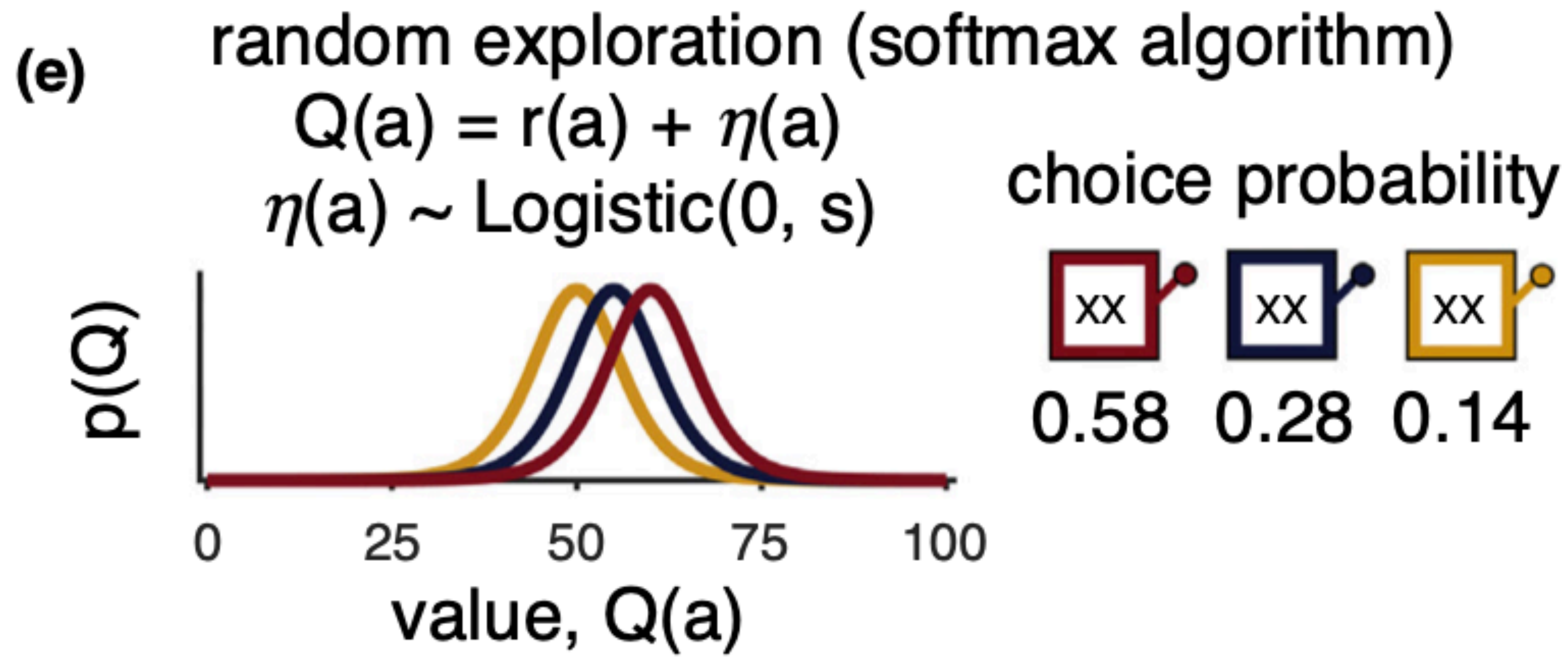
$$p(a) = \frac{e^{Q(a)/\tau}}{\sum_{i=1}^{A} e^{Q(i)/\tau}}$$

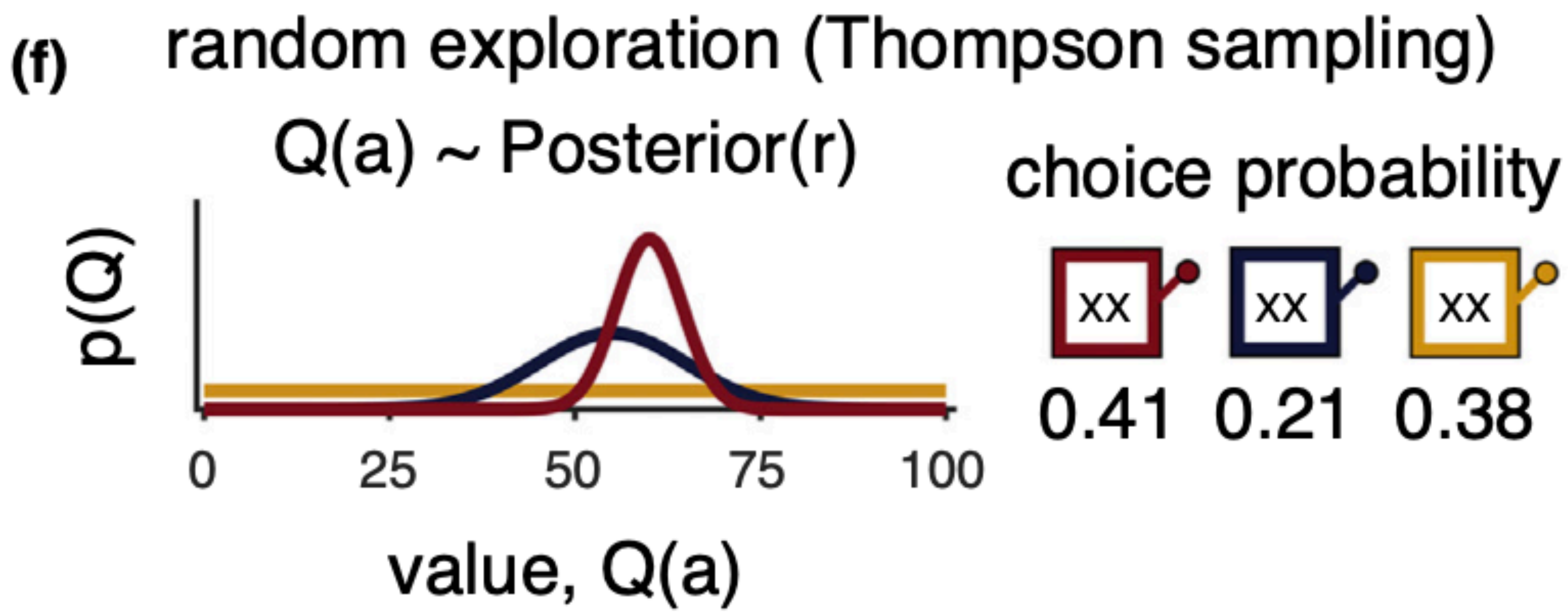
larger τ = more random

(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

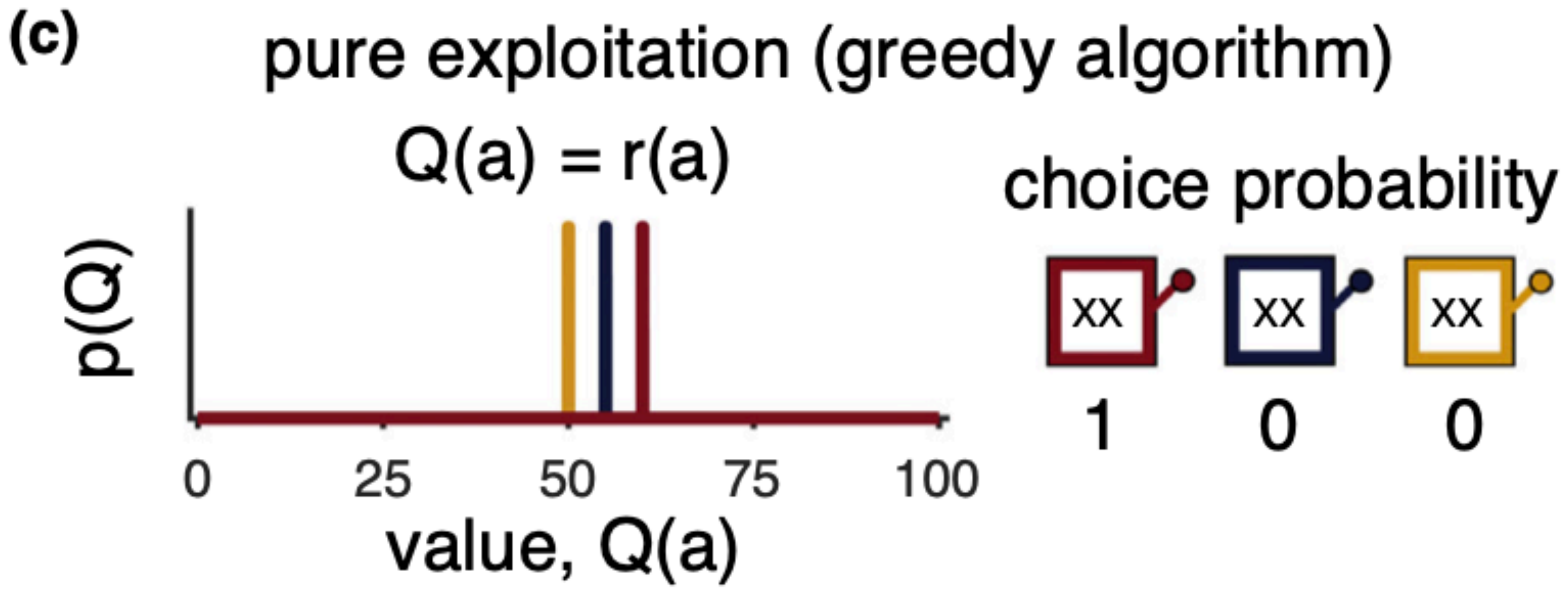
variance of the posterior distribution

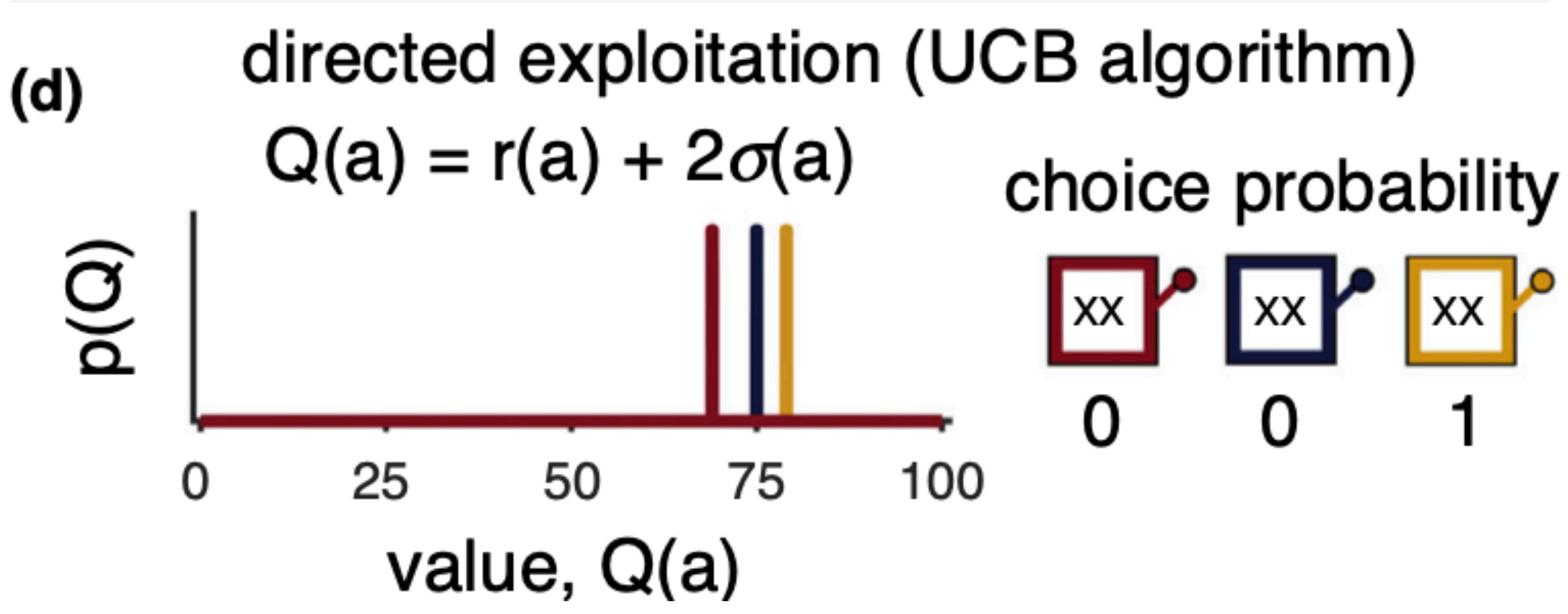
$$p(a) = Q(a) + 2\sigma(a)$$

(Wilson et al. 2020)

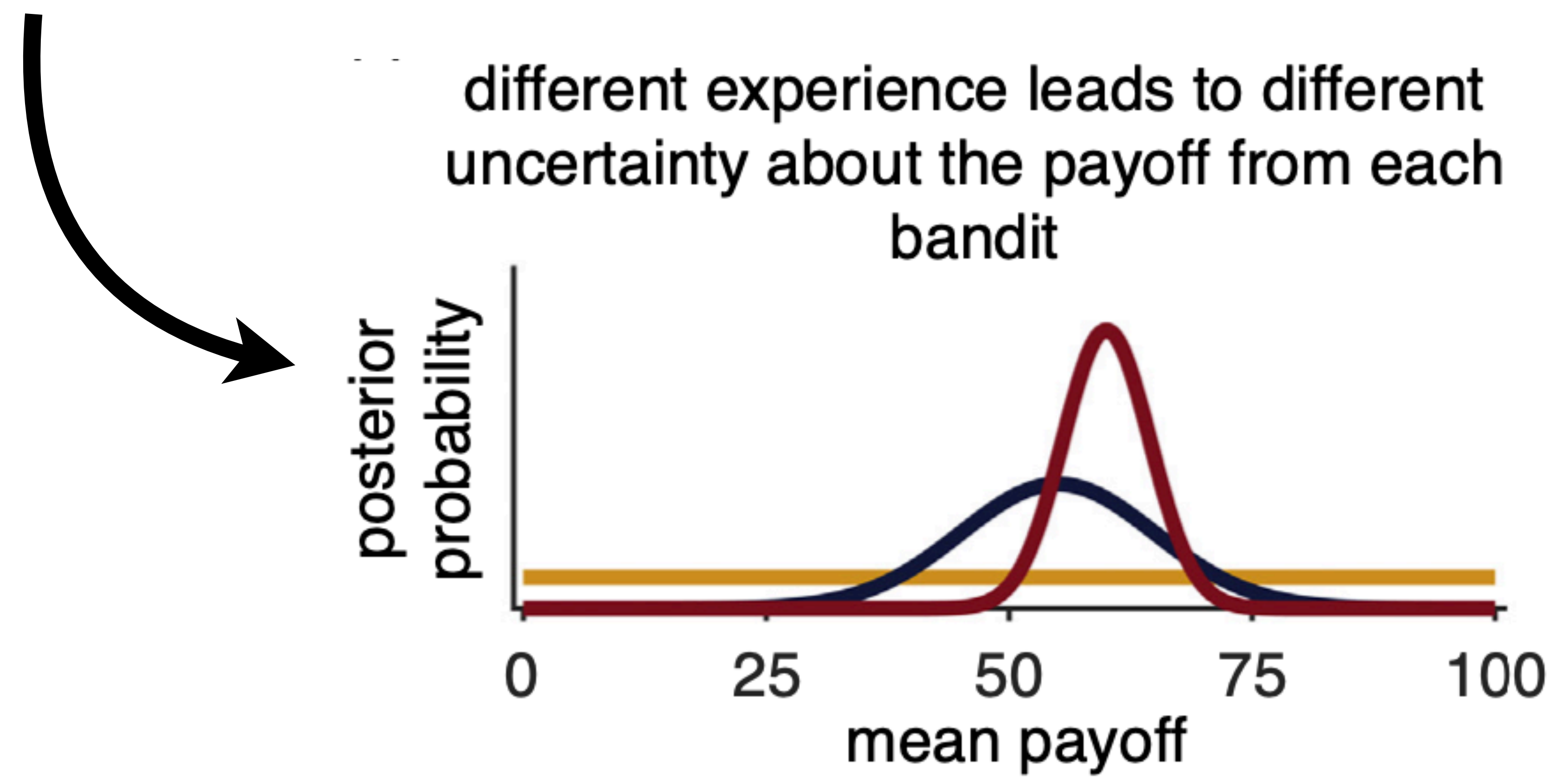
Two types of ways to explore

Directed exploration



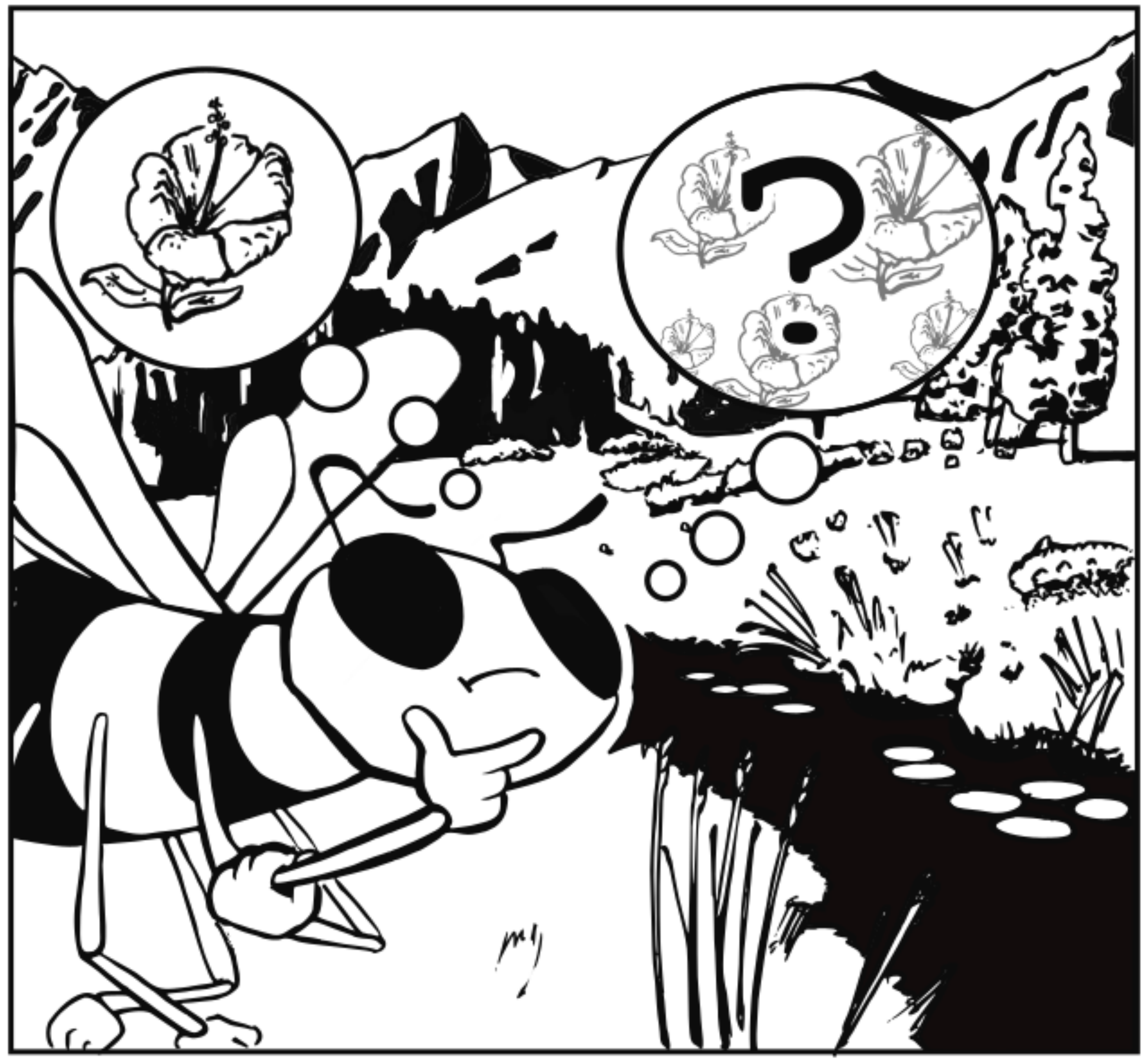


Curiosity!



(Wilson et al. 2020)

Rethinking the dilemma



- Value-based decision making is dominated by explore-exploit policy, π.
- Mathematically optimal solution is intractable for reward collection.

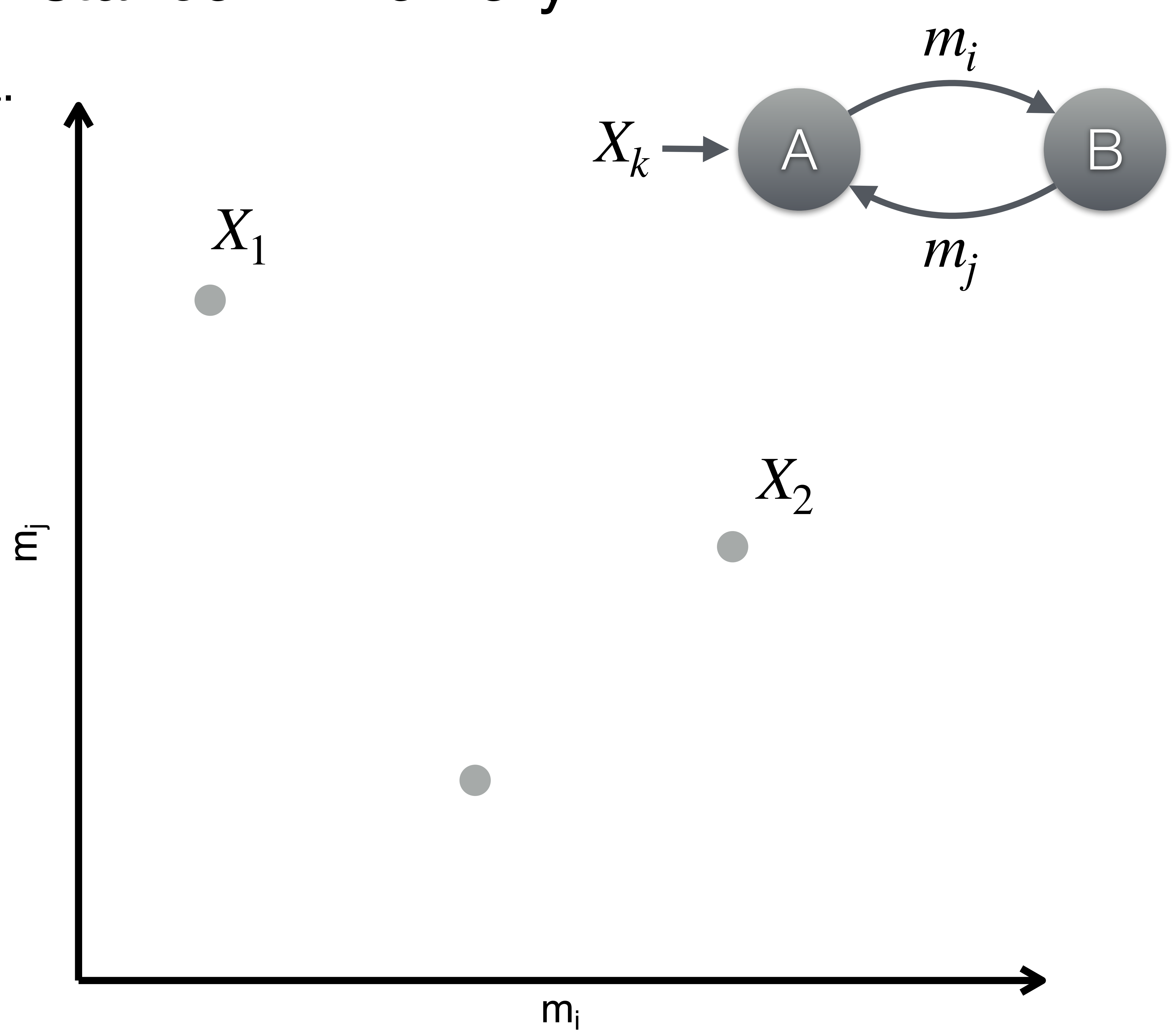
$$\max \sum_{s \in S,T} \gamma \mathbf{R}$$

(Peterson & Verstynen, 2022)

Information value: E

Distance in memory M

a.



The problem of learning:

Memory

$$M' \leftarrow f(X, M)$$

Updated memory

Observation

Learning function

The problem of forgetting:

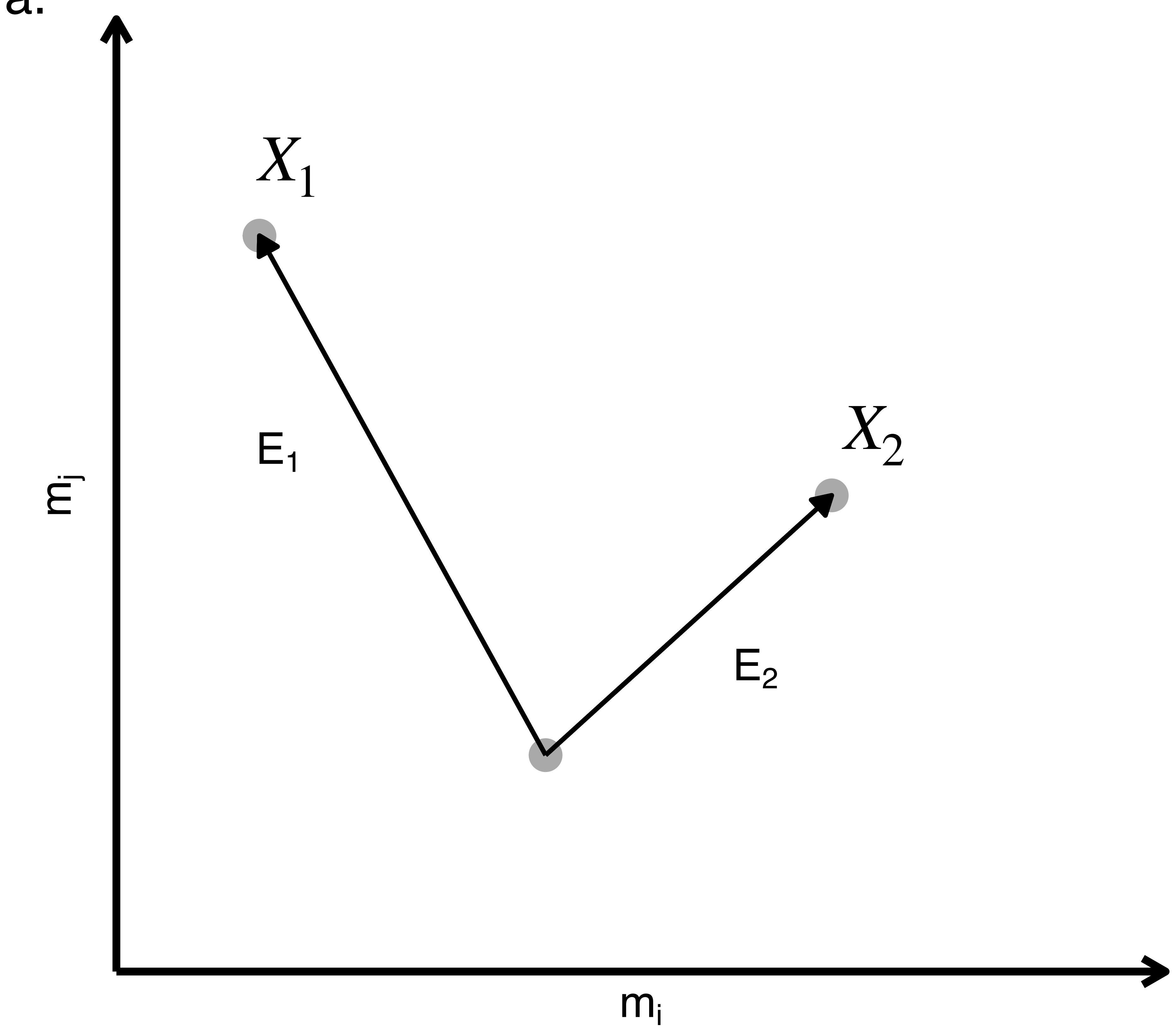
$$f^{-1}(X, M') \rightarrow M$$

(Peterson & Verstynen, 2022)

Information value: E

Distance in memory M

a.



Axiom of Memory:
E depends only on the difference $\Delta\mathbf{M}$ between $\mathbf{M}$ and $\mathbf{M}'$

Axiom of Specificity:
If all $\Delta\mathbf{M}$ are equal, then $E = 0$

Axiom of Scholarship:
$E \geq 0$

Axiom of Equilibrium:
For the same observation E should approach 0 in finite time.

(Peterson & Verstynen, 2022)

A scheduling problem

An alternative view:

- Turn the dilemma into a *two objective problem*
- Mathematically tractable

$$\max \sum_{s \in S,T} R$$

$$\max \sum_{s \in S,T} E$$

Optimal E learning:

- $\hat{E}$ has optimal substructure
- So the optimal learning policy is the Bellman eq.

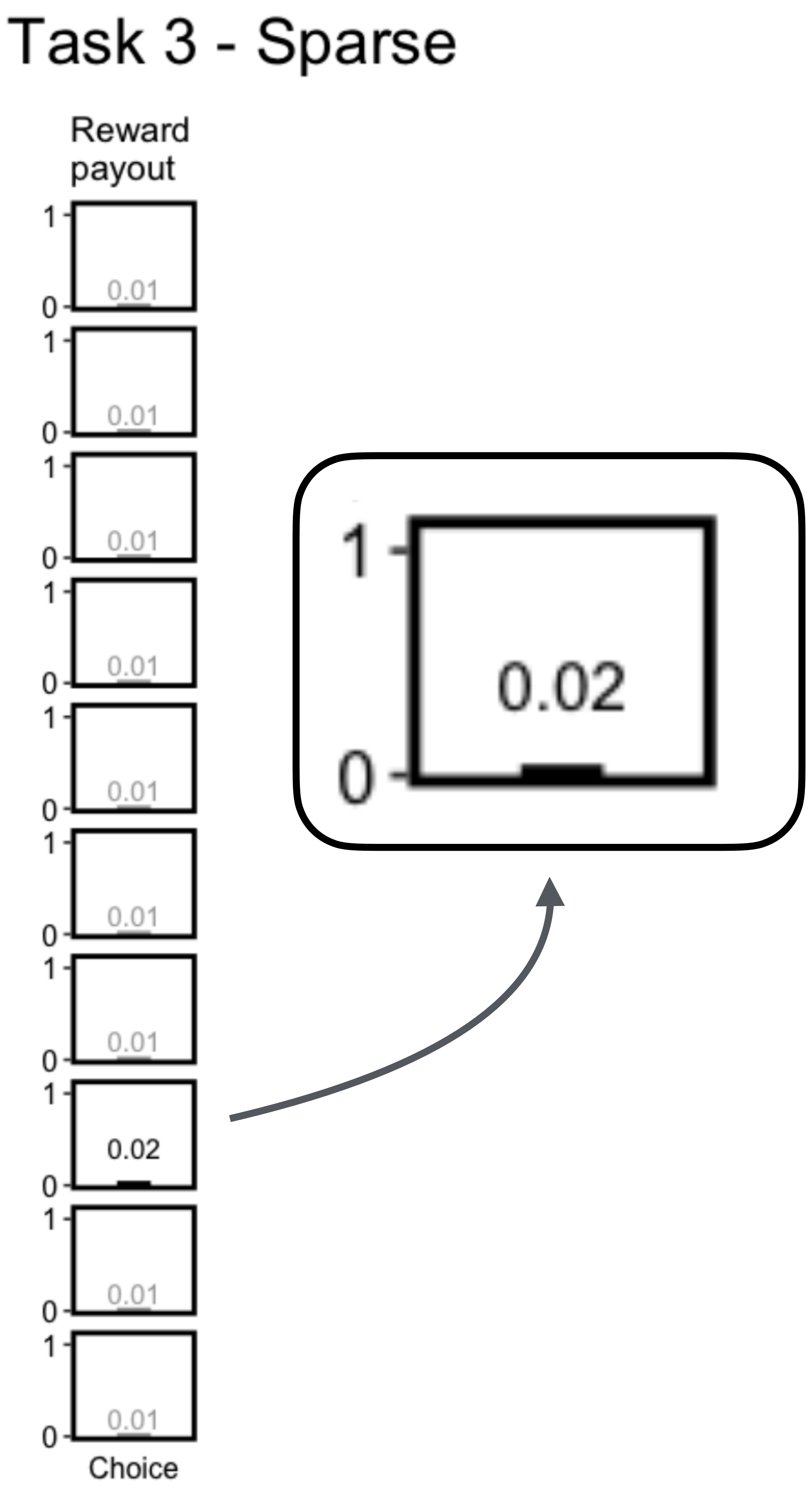
$$\underbrace{V^*_{\hat{E}}(\mathbf{S})}_{\text{State}} = \operatorname*{argmax}_{\mathbf{A} \in \mathbb{A}} \left[\hat{E}_t + V^*_{\hat{E}}(\Lambda(\mathbf{S}, \underbrace{\mathbf{A}}_{\text{Action}})) \right]$$

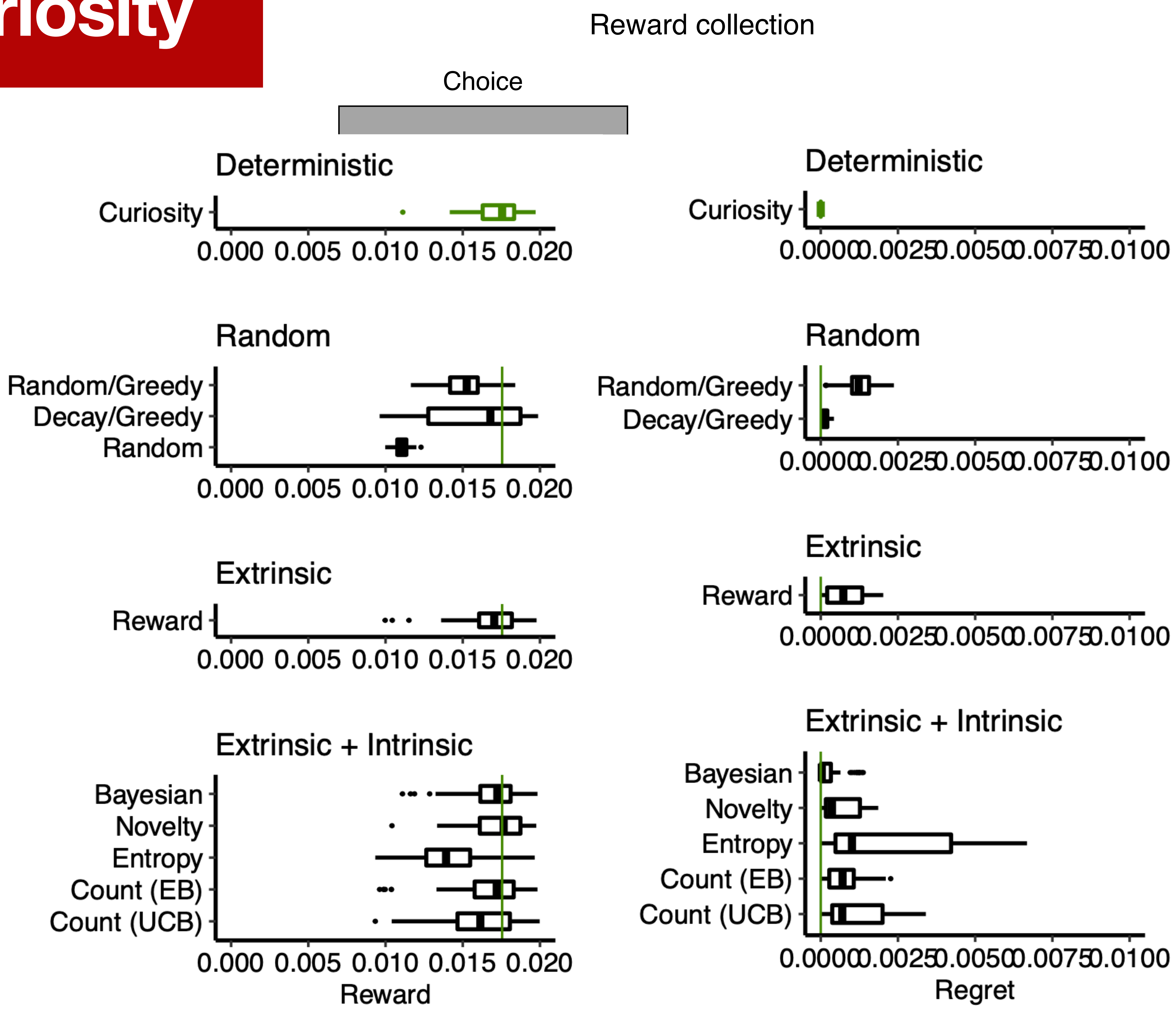
Optimal meta-greedy policy:

$$\Pi_\pi = \begin{cases} \pi^*_{\hat{E}} : E > R \\ \pi_R : E < R \end{cases}$$

(Peterson & Verstynen, 2022)

Evaluating optimal curiosity





(Peterson & Verstynen, 2022)

Take home message

- Exploration can be random or directed (curiosity), with the latter being information seeking.
- If you treat maximizing rewards versus maximizing information as separate objectives, the exploration-exploitation dilemma disappears.