Cogmaster

Methods in Computational Neuroscience

February 21st 2019

Exploration-exploitation dilemma

Manuel Beiran
manuel.beiran@ens.fr

Ex. 2: Computational models of behavior

see:

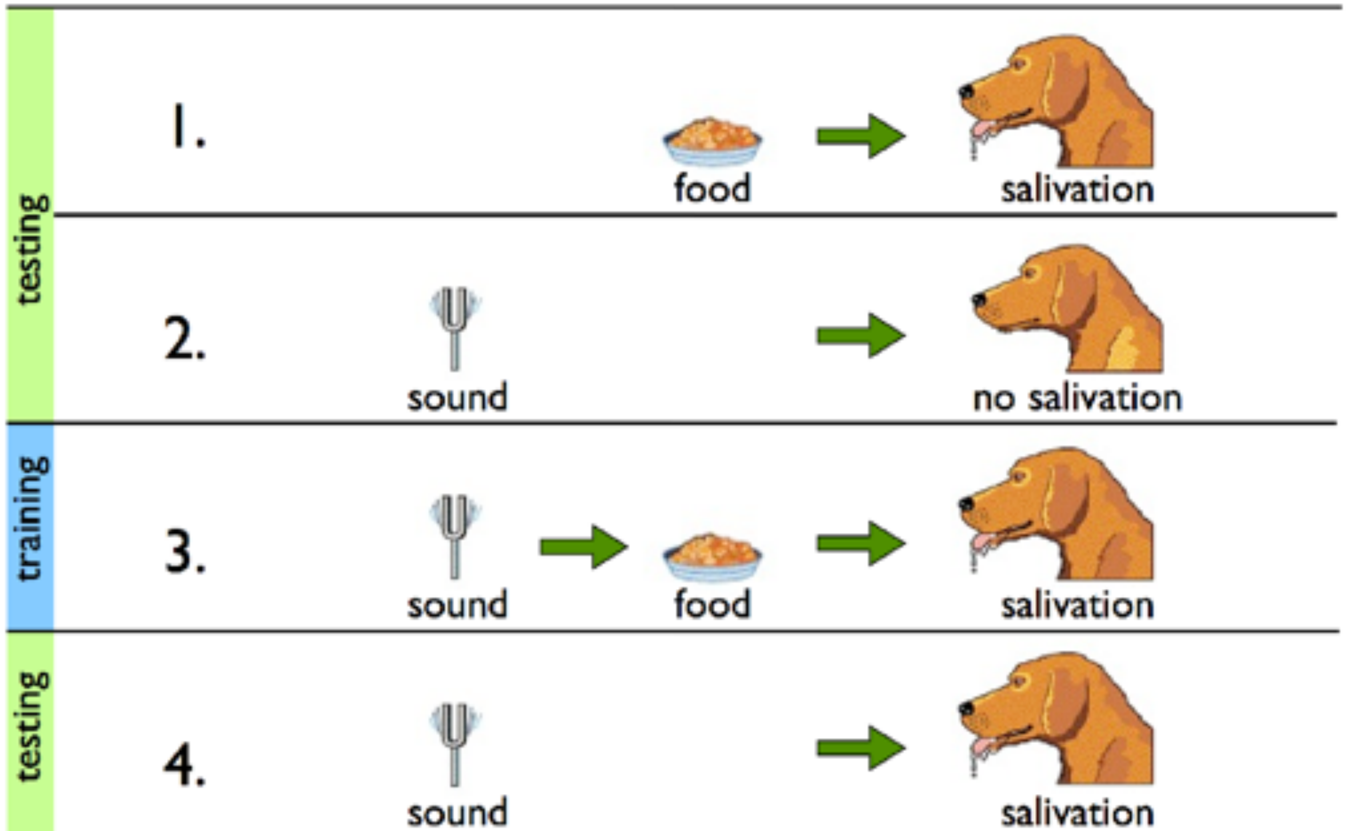
Dayan and Abbott, *Theoretical Neuroscience*, chap. 9
C06 course

Study the ability of animals of taking actions according only to the received **reward** and **punishment:**

REINFORCEMENT LEARNING

Experiments of: - **classical** (Pavlovian) conditioning;
- **instrumental** conditioning

Ex. 2.1: Classical conditioning



Rescorla-Wagner-rule

$$w \rightarrow w + \epsilon \delta_i u_i$$

"delta-rule"

u_i stimulus ( ) in trial i: $u_i = 0$ or $u_i = 1$

r_i reward ( ) in trial i: $r_i = 0$ or $r_i = 1$

v_i reward that the dog expects ( ) in trial i $v_i = w u_i$

Ex. 2.2: Instrumental conditioning

The value of the reward depends on the action taken by the animal

see:

Dayan and Abbott, *Theoretical Neuroscience*, 9.3
C06 course

Experiments of: - **static action choice:**
reward is delivered immediately after the choice;

- **sequential action choice:**
reward is delivered after a series of actions
(long term planning)

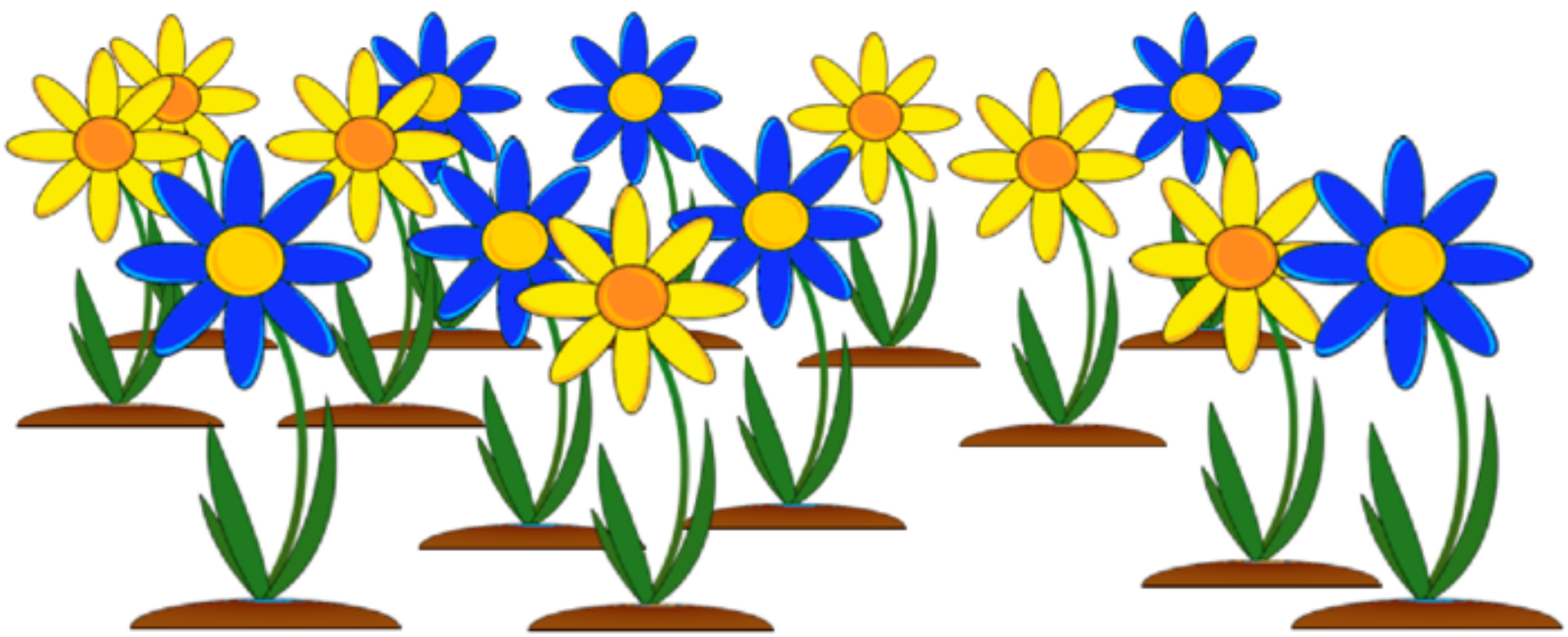
Ex. 2.2: Decision strategies

The animal acts to maximize its expected reward

Ex. 2.2: Decision strategies

The animal acts to maximize its expected reward

Possible choices (actions) of a bee:
land on a blue or yellow flower



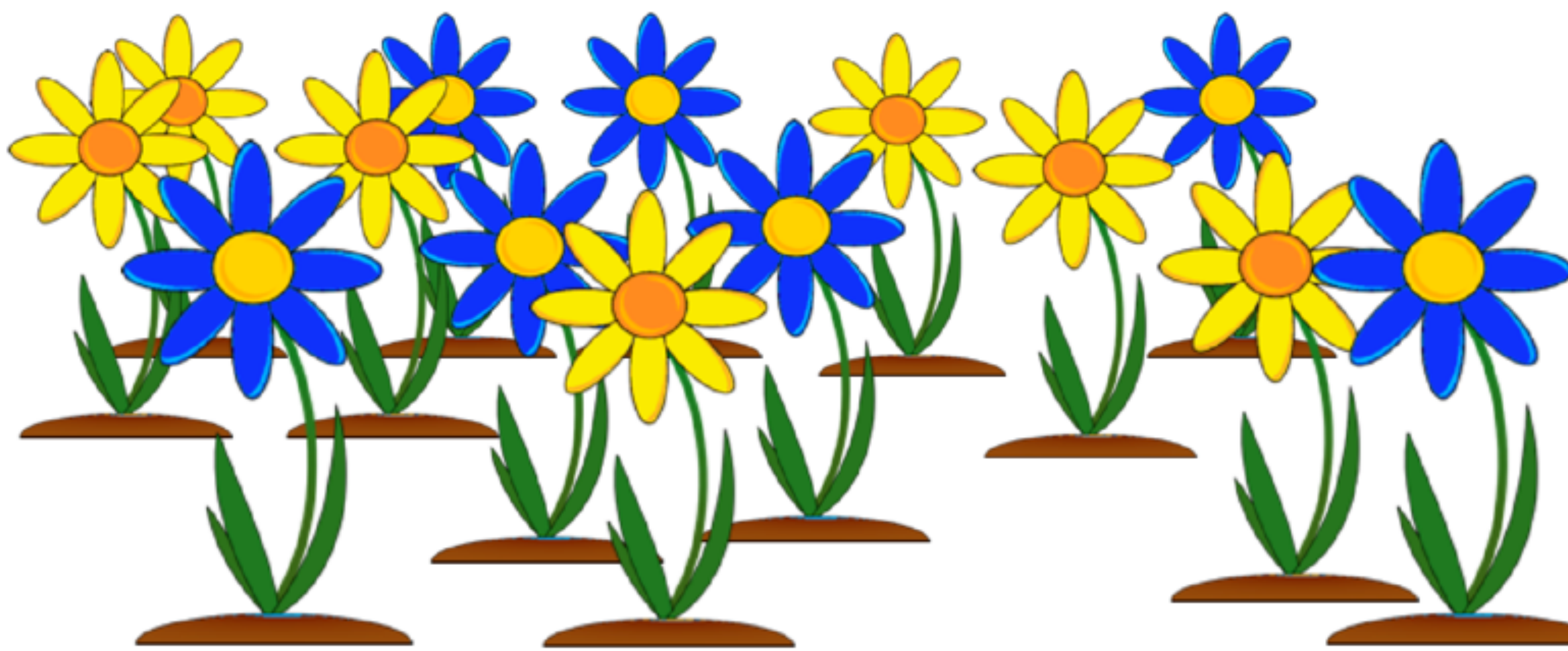
Bee searching for nectar

Possible choices (actions) of the bee:
land on a blue or yellow flower
rewards (in drops of nectar)

 $r_b = 8$

 $r_y = 2$

$$\langle R \rangle = r_y \cdot p\,(a = \text{yellow}) + r_b \cdot p\,(a = \text{blue})$$

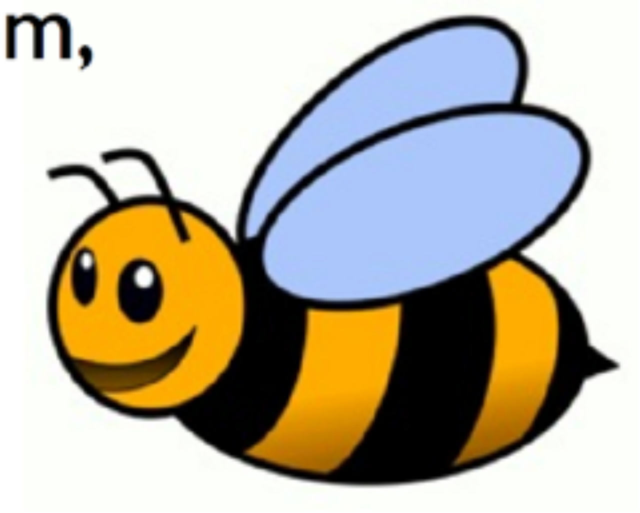


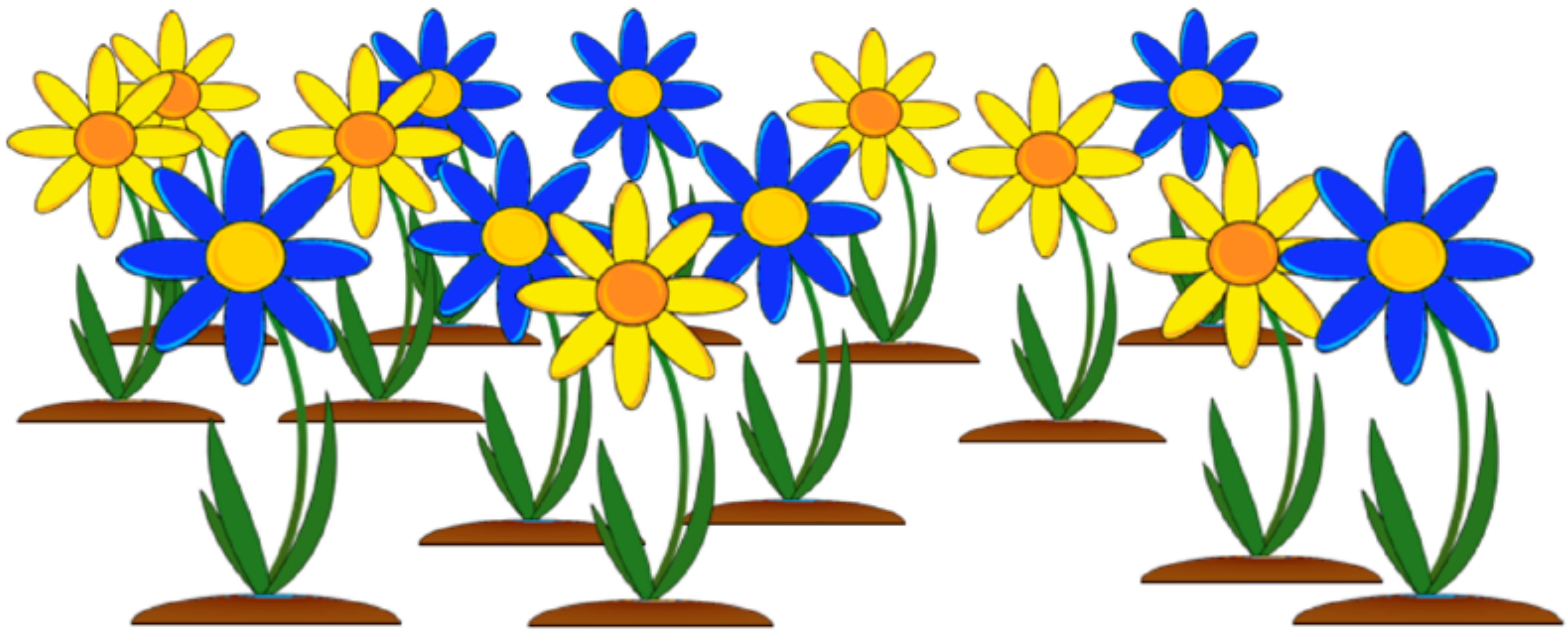
"Policy": Bee's plan of action

Assume: choices or actions a are taken at random, according to a probabilistic "policy":

$$p(a = \text{yellow})$$

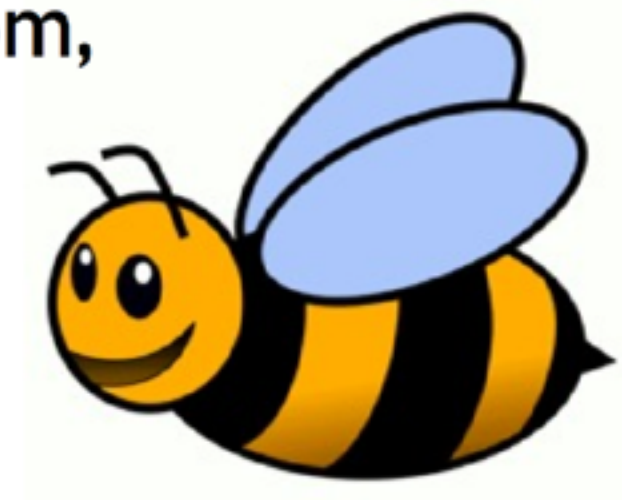
$$p(a = \text{blue})$$





"Policy": Bee's plan of action

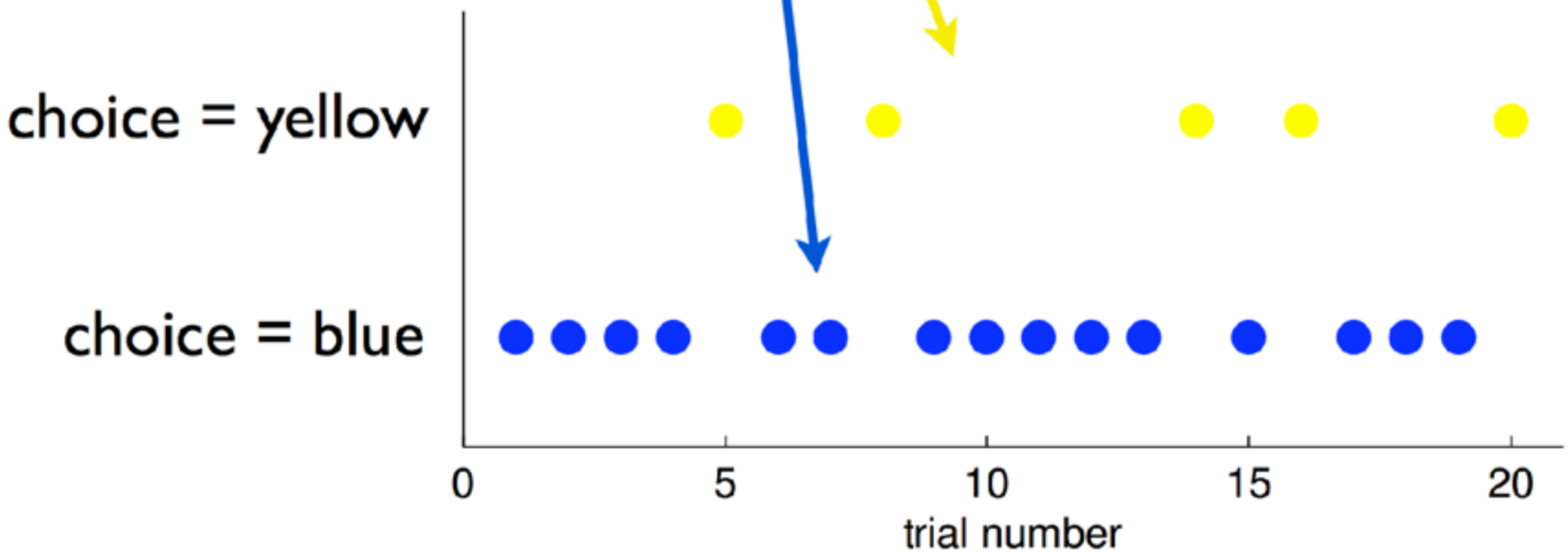
Assume: choices or actions a are taken at random, according to a probabilistic "policy":



$$p(a = \text{yellow}) = 0.2$$

$$p(a = \text{blue}) = 0.8$$

$$p(a = \text{blue}) + p(a = \text{yellow}) = 1$$



"Optimal Policy": The greedy bee

Optimal policy:

$$p(a = \text{blue}) = 1$$
$$p(a = \text{yellow}) = 0$$

Rewards:

 $r_b = 8$

 $r_y = 2$

BUT: What happens if the environment changes?

Day	1	2	3	...
r_b	8	2	3	...
r_y	2	8	5	...

Bee needs to explore and exploit

"greedy" policy

$$p(a = \text{blue}) = 1$$
$$p(a = \text{yellow}) = 0$$

BAD IF THINGS CHANGE!

Bee needs to explore and exploit

"greedy" policy

$$p(a = \text{blue}) = 1$$
$$p(a = \text{yellow}) = 0$$

BAD IF THINGS CHANGE!

"ε–greedy" policy ($\epsilon \ll 1$)

$$p(a = \text{blue}) = 1 - \epsilon$$
$$p(a = \text{yellow}) = \epsilon$$

Bee needs to explore and exploit

"greedy" policy

BAD IF THINGS CHANGE!

$$p(a = \text{blue}) = 1$$
$$p(a = \text{yellow}) = 0$$

"ε–greedy" policy ($\epsilon \ll 1$)

$$p(a = \text{blue}) = 1 - \epsilon$$
$$p(a = \text{yellow}) = \epsilon$$

softmax Gibbs-policy (depends on rewards!)

$$p(a = \text{blue}) = \exp(\beta r_b) / \big(\exp(\beta r_b) + \exp(\beta r_y)\big)$$
$$p(a = \text{yellow}) = \exp(\beta r_y) / \big(\exp(\beta r_b) + \exp(\beta r_y)\big)$$

Softmax-Gibbs Policy

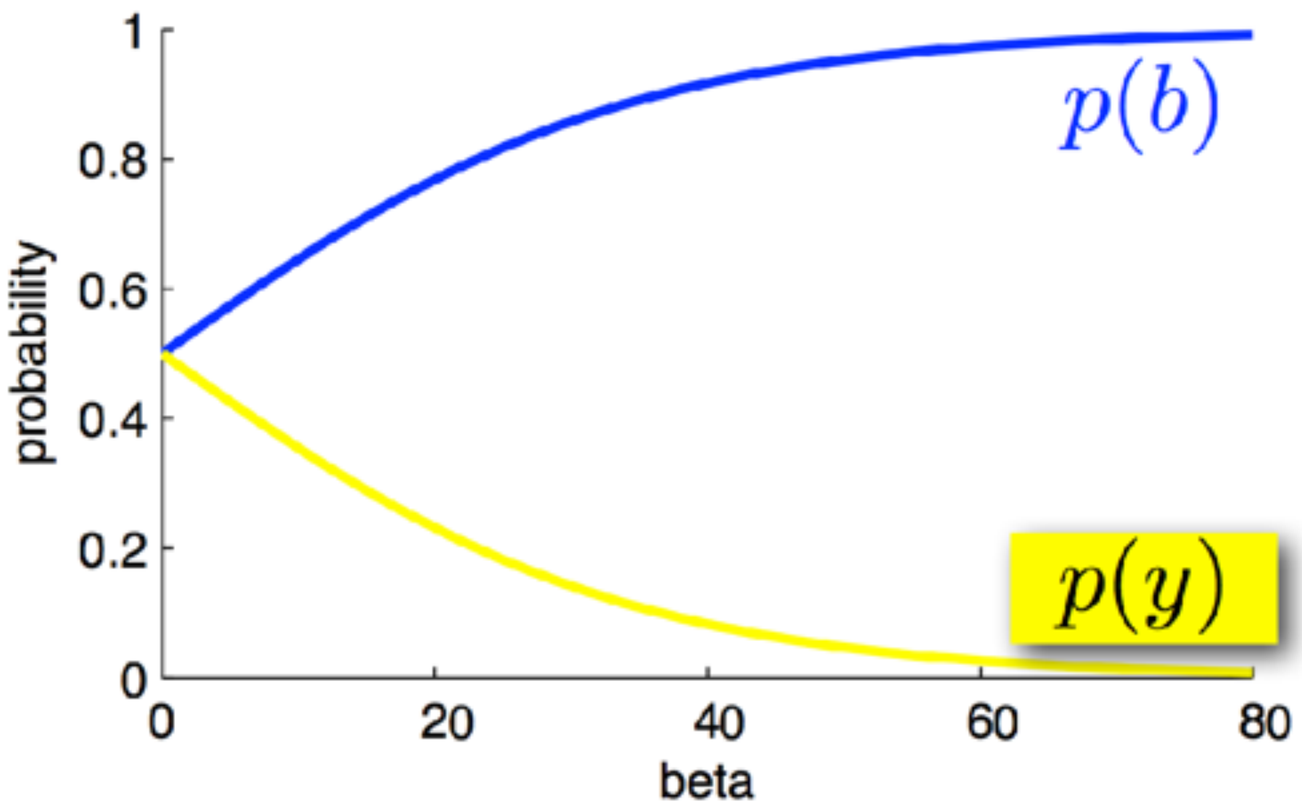
$$p(b) = \frac{\exp(\beta r_b)}{\exp(\beta r_b) + \exp(\beta r_y)}$$

$$p(y) = \frac{\exp(\beta r_y)}{\exp(\beta r_b) + \exp(\beta r_y)}$$

Rewards:

$r_b = 8$

$r_y = 2$

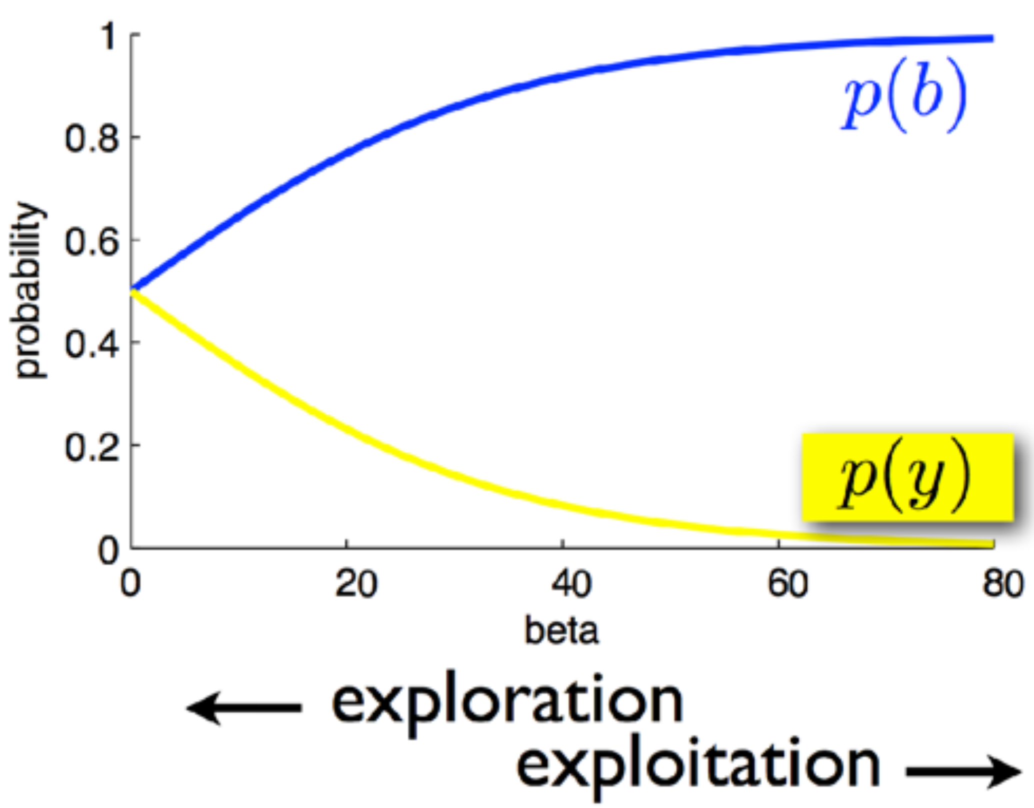


← exploration

exploitation →

$$p(b) = \frac{\exp(\beta r_b)}{\exp(\beta r_b) + \exp(\beta r_y)}$$

$$p(y) = \frac{\exp(\beta r_y)}{\exp(\beta r_b) + \exp(\beta r_y)}$$



Some properties:

- *p(b)+p(y)=1* (normalization)
- *p(b)* is a **sigmoid** of *(r_b-r_y)*
- beta encodes the 'exploration-exploitation' balance (temperature parameter)
- Very good for avoiding strong punishment (exponential negative decay)

Changing the policy online

The animal does not know the reward,
it can only estimate the reward.

And, what happens if the rewards vary?

Changing the policy online

$$p(b) = \frac{\exp(\beta m_b)}{\exp(\beta m_b) + \exp(\beta m_y)}$$

$$p(y) = \frac{\exp(\beta m_y)}{\exp(\beta m_b + \exp(\beta m_y)}$$

	actual reward	internal estimate
	r_b	m_b
	r_y	m_y

How can the bee learn the rewards?

"greedy" update:

$$m_b = r_{b,i} \qquad\qquad m_y = r_{y,i}$$

"batch" update:

$$m_b = \frac{1}{N}\sum_{i=1}^{N} r_{b,i} \qquad\qquad m_y = \frac{1}{N}\sum_{i=1}^{N} r_{y,i}$$

average reward on last N visits to a blue flower

average reward on last N visits to a yellow flower

How can the bee learn the rewards?

"greedy" update:

$$m_b = r_{b,i} \qquad\qquad m_y = r_{y,i}$$

"batch" update:

$$m_b = \frac{1}{N}\sum_{i=1}^{N} r_{b,i} \qquad\qquad m_y = \frac{1}{N}\sum_{i=1}^{N} r_{y,i}$$

"online" update: **'INDIRECT ACTOR'**

$$m_b \rightarrow m_b + \epsilon(r_{b,i} - m_b)$$

"delta"- rule $m_b \rightarrow m_b + \underbrace{\epsilon}_{\text{learning rate}}\delta$ with $\delta = \underbrace{r_{b,i} - m_b}_{\text{prediction error}}$

Remarks on Ex. 1: content

- An exponential curve doesn’t grow faster in the end.
- Logscale vs linear scale

<u>Remarks on Ex. 1: style</u>

- The report should be a **scientific report** written in English.

- Include an **introduction** and a **conclusion**/summary for each exercise.

- Use questions as a guide to write a coherent explanation of the model.

- Each sentence should be rigorous: ~~“looks like an exponential”, “the model is not very realistic”, “I try to…”, “As a conclusion we can say…”~~

- Don’t start a text sentence with a variable.

Remarks on Ex. 1: style (figures and equations)

- Axis labels: “ magnitude (units) ”.

- Figures should support text:

  “The population growth depends linearly on the initial number of individuals. If we double the initial population, we will obtain a final population twice as large (see **Fig. 3**, blue line vs orange line).

  ~~“In figure 3 we vary the parameter of the initial population and the behavior changes considerably.”~~

- Write equations and symbol in mathematical notation

  The value of ~~alpha~~ α determines…

  Applying the equation ~~“pn+1 = pn + 0.1*pn”~~

$$p_{n+1} = p_n + 0.1 \cdot p_n$$

Remarks on Ex. 1: common mistakes in scientific English

- Sensitivity to initial conditions
- Resources
- Computational
- Literature
- Modeling / Modelling

Remarks on Ex. 1: grading criteria

- **19 - 20:** excellent report, well structured, deep insight into the studied phenomena, correct style. The models are well explained, in a technical way.

- **16 - 18:** Very good reports. Well structured, all pointed issues are clearly explained and no problems with the figures. No evident mistakes.

- **13 - 15:** Good reports. I see the student has well understood the exercise. There are some problems with the structure and/or figures, some explanations are not clear

- **10 - 12:** Important effort. More or less complete report, codes work. Explanations are often very unprecise or even wrong. Results missing. Important problems in the structure.

Programming: some more tricks

- Python functions: masks, find

Introduction to LaTeX

- A good way to start is to use a TeX editor (i.e. TexStudio)
- Useful when the text is combined with mathematical equations
- (Arguably) useful for displaying figures and captions
- Tip: find a template you like and use it.