

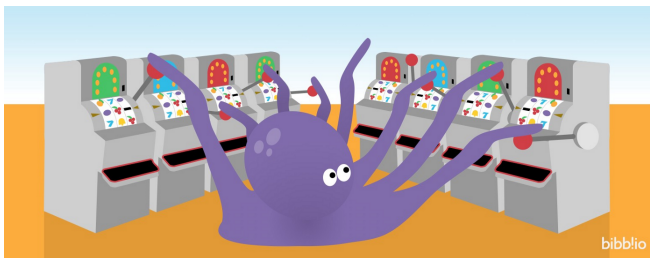
# Multi-Arm Bandit

Corrado Possieri

Machine and Reinforcement Learning in Control Applications

# Problem

- Choose repetitively which arm to pull from those available
- Each arm returns a reward
- The objective is to maximize the expected total reward



## ① **Deterministic and stationary**

- Rewards are equal to  $q_*(a)$
- $q_*(a)$ 's don't change over time

## ② **Stochastic and stationary**

- Normally distributed rewards with mean  $q_*(a)$
- $q_*(a)$ 's don't change over time

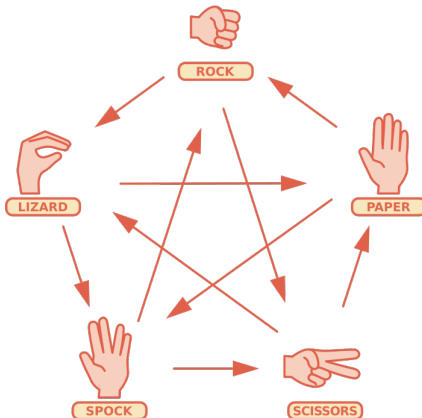
## ③ **Stochastic and non-stationary**

- Normally distributed rewards with mean  $q_*(a)$
- $q_*(a)$ 's change over time

- The arm to pull is chosen according to the following policies:
  - $\epsilon$ -greedy sample-average
  - Upper confidence bound
  - Preference updates
- Comparison of the results obtained with the various policies and parameters

# Assignment #1

- Model the **Rock, Paper, Scissors, Lizard, Spock** game



# Assignment #1

- Model the **Rock, Paper, Scissors, Lizard, Spock** game
  - See the episode 5x17 of "The Big Bang Theory" for rules
  - Choose repetitively which action to play between: Rock, Paper, Scissors, Lizard, Spock
  - The opponent plays randomly
  - Reward  $+1$  for winning and  $-1$  for losing
  - The objective is to maximize the rewards
  - Analyze the trends of the expected rewards for each action