Neural Basis of Reinforcement Learning and Decision Making

Daeyeol Lee et al. 2012
Introduction

★ Two categories of decision making:
  ○ normative theories: assume decision maker is ideal, fully informed, computes with perfect accuracy, rational
    ■ e.g. expected utility theory: probabilities of outcomes are uncertain
  ○ descriptive theories: assume decision makers are behaving on some set of consistent rules
    ■ e.g. prospect theory: probabilities of outcomes are known

★ Normative and descriptive theories are complementary
★ Neither approach accounts for how decision makers acquire their preference for certain choices
Aim of the paper

★ Describe the functions of brain areas which are involved in experience-dependent decision making

○ Which parts of the brain estimate the expected rewards and update them based on the animal’s experience?
○ How are the potential action value signals transformed into the final decision?
○ How is model-based reinforcement learning implemented in the brain?
  ■ model-free: knowledge of the environment is strictly based on experience of reward and penalty
  ■ model-based: gaining knowledge about the environment without any experience of reward or penalty
Economic utilities and value functions

- Economic theories calculate values for actions so the best possible action has the highest value
  - utilities: values for a set of actions
- Reinforcement learning theories describe how value functions change based on the animal’s experience
  - soft max = action chosen probabilistically
Value functions

The goal is to maximise future rewards

Value functions are empirical estimates for future rewards

Two types of value functions

- Action
  - sum of the expected future rewards for a certain action in a certain environment
  - used for evaluating action outcomes
- State
  - sum of the expected future rewards in a certain environment
  - calculated based on action value functions
  - needed for selecting an action
Updating value functions

★ Model-free
  ○ updates value functions based on rewards and penalties
  ○ uses reward prediction error: signed difference between actual rewards and predicted rewards

★ Model-based
  ○ updates value functions based on the animal’s motivational state and knowledge of the environment
  ○ value function adjustment whenever new information is provided
  ○ uses fictive/counterfactual reward prediction error: simulating alternatives to events that have already happened
Animals and food choices with model-free and model-based reinforcement learning
Learning in social decision making

★ Game: action outcomes are determined by multiple players
★ Model-free gradually nears the optimal decisions
★ Model-based is more flexible, predicts the decisions of other players (theory of mind/belief learning)
★ Pure belief learning
  ○ rewards and penalties do not affect the decisions
  ○ not applicable on humans, hybrid models used instead
## Neural signals

### Action value function:
1. posterior parietal cortex
2. dorsolateral prefrontal cortex
3. premotor cortex
4. medial frontal cortex
5. striatum

### Difference calculation:
1. posterior parietal cortex
2. dorsolateral prefrontal cortex
3. premotor cortex
4. dorsal striatum
5. supplementary eye field

### (Post-decision) State value function:
1. posterior parietal cortex
2. dorsolateral prefrontal cortex
3. medial frontal cortex
4. dorsal, ventral striatum
5. anterior cingulate cortex
6. amygdala
7. orbitofrontal cortex
8. supplementary eye field

★ Neural signals for action value functions can greatly vary depending on the properties of the task
Neural mechanisms of action selection

★ During decision making, neural activity related to action value functions get converted to signals related to the chosen action and carried over to motor structures

★ An action is selected when the threshold of neural activity for that action is reached

★ In brain areas related to motor control the neural activity builds up gradually over time before any movement is made -> may be involved in action selection

★ Making decisions based on:
  ○ external sensory stimuli - orbitofrontal cortex
  ○ internal sensory stimuli - medial frontal cortex

★ Neural activity of a movement that is yet to happen appears first in medial frontal cortex than anywhere else
Updating value functions

★ The problem of temporal credit assignment:
  ○ The rewards/penalties from an action are learned later, after other actions have been taken
  ○ Hard to associate the correct action and reward
  ○ Unclear how to update value functions

★ Lesions in the orbitofrontal cortex interfere with reversal learning tasks

★ Solutions to the problem of temporal credit assignment:
  ○ Introduce intermediate states between the action and the reward
  ○ Eligibility traces: use short-term memory signals which are related to the states/actions chosen

★ Neural signals of previous choices have been recorded
★ Records of neurons encoding the conjunction of an action and its outcome
Uncertainty and learning rate

★ Risk/expected uncertainty: probabilities of outcomes are known
  ○ stable environment (e.g. flipping a coin)
  ○ learning rate close to 0
★ Ambiguity/unexpected uncertainty: probabilities of outcomes are unknown
  ○ volatile environment
  ○ high learning rate
★ Humans are good at adjusting their learning rate
★ Anterior cingulate cortex: may be important in adjusting the learning rate
Model-free reinforcement learning

★ Neurons updating value functions use:
  ○ neurons related to the chosen value
  ○ reward prediction error

★ The reward prediction error signals are gradually gathered from different areas in the brain
  ○ dopamine neurons for relaying signals

★ Both signals converge in the orbitofrontal cortex, striatum: role in updating value functions
Possible implementation of model-based learning in the brain:

○ learn separately:
  i. information about the environment
  ii. outcomes of actions in a certain environment

○ combine the information

○ hippocampus: may provide information about the environment for updating value functions
Model-free and model-based

★ Storing value functions of both methods:
  ○ represented in different brain areas
  OR
  ○ combined to reach a new estimate of the outcomes
★ Chosen values and reward prediction errors seem to be represented by the same brain area (ventral striatum)
Hypothetical outcomes

★ After making a decision:
  ○ the decision maker can come up with hypothetical results of taking another action
  ○ based on this, update action value functions for actions not chosen

★ Hypothetical outcomes may be processed where actual outcomes are processed
  ○ prefrontal cortex
  ○ striatum

★ The actual outcome and the hypothetical outcomes of the same action may also be processed in the same brain areas
  ○ dorsolateral prefrontal cortex
  ○ orbitofrontal cortex
  ○ used for updating value functions