COMP2410/6340 Automated Decision Making & Cyber (Physical) Security – Part 2

Hanna Kurniawati

http://users.cecs.anu.edu.au/~hannakur/

Australian National University | RESEARCH SCHOOL OF COMPUTER SCIENCE

This set of videos

✓Part-1: Intro

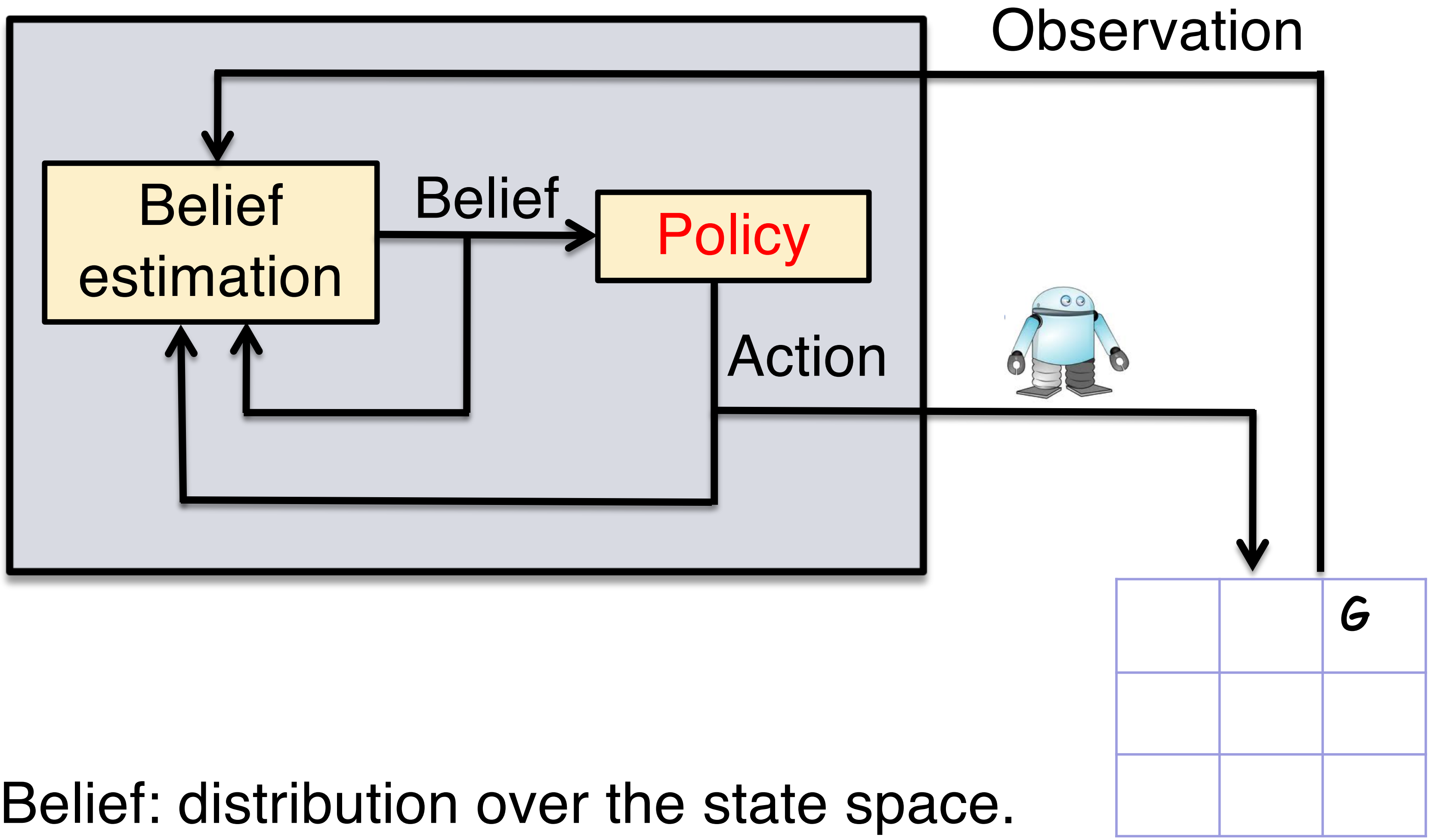
- ✓Automated decision-making

- Part-2: Intro to POMDP
  - Framework for decision-making under uncertainty
  - Solving, aka. generating strategic decisions
- Part-3: Example of POMDP in Cyber security
  - Autonomous pen-testing

Recall: Our agent is uncertain in …

- Effects of actions, aka. non-deterministic
- Observation it can perceive, aka. partially observable
- The above types of uncertainty occur in many problems, including robotics and cyber-security
  - We will discuss an example of the problems in the next videos
  - For now, we'll discuss a mathematically principled framework for solving this type of decision-making problems: The Partially Observable Markov Decision Process (POMDP)

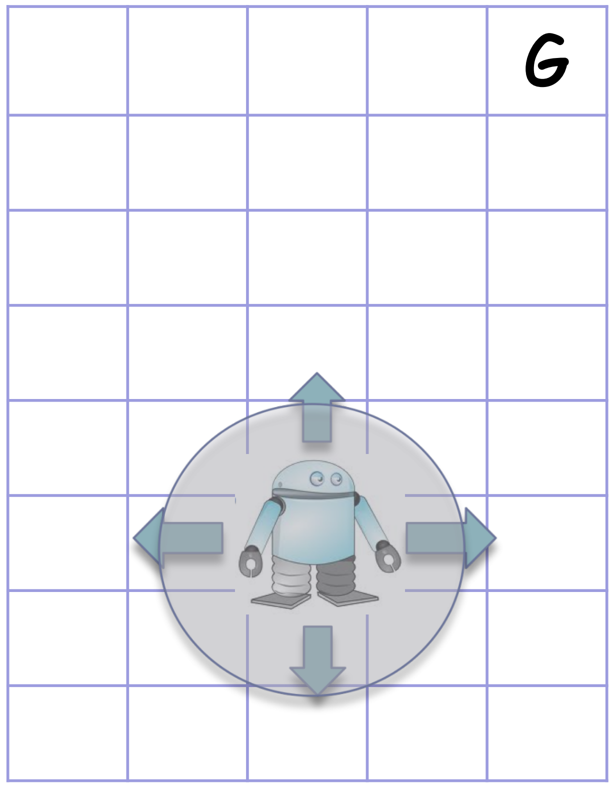
Partially Observable Markov Decision Processes (POMDP)



- Belief: distribution over the state space.
- Strategy/policy: Best mapping from beliefs to actions.

POMDP Model

- A 6-tuples (S, A, O, T, Z, R):
  - State space (S)
  - Action space (A)
  - Observation space (O)



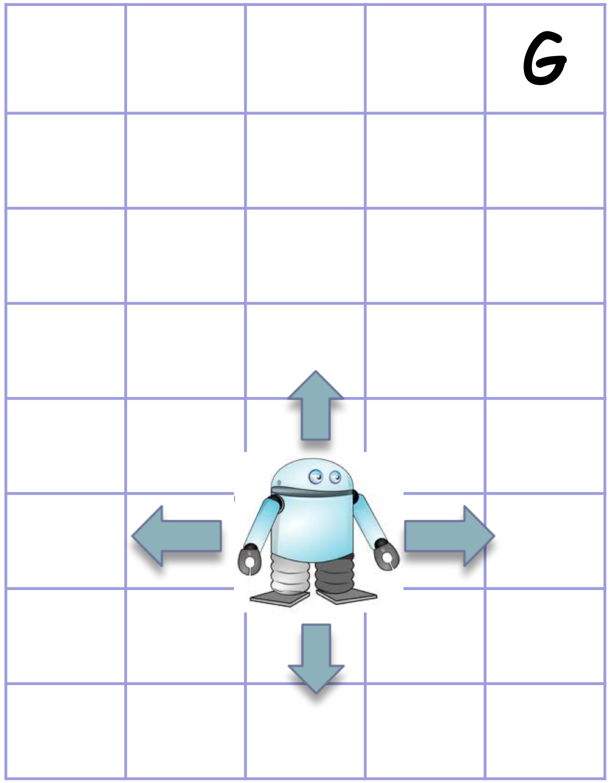
POMDP Model

- A 6-tuples (S, A, O, T, Z, R):
  - State space (S) ← Not known
  - Action space (A)
  - Observation space (O)
  - Transition function (T)

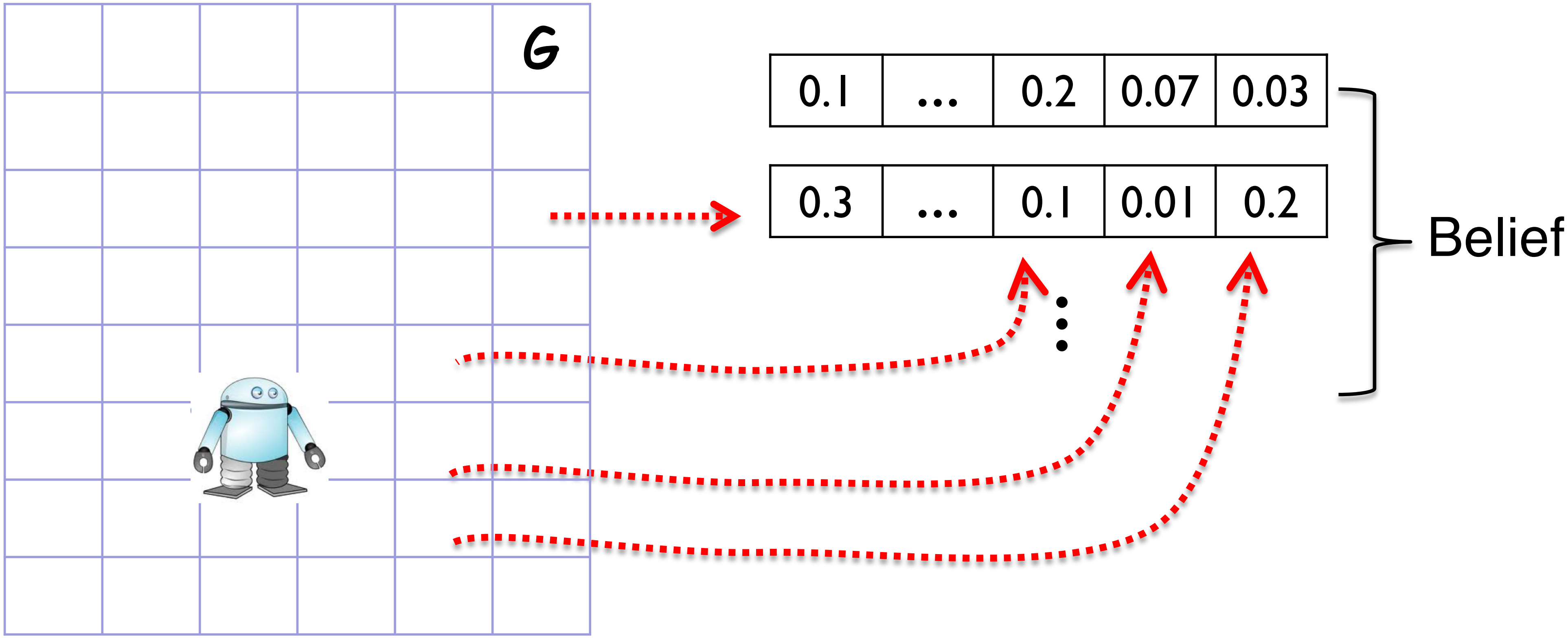
    $T(s, a, s') = P(S_{t+1} = s' \mid S_t = s, A_t = a)$
  - Observation function (Z)

    $Z(s, a, o) = P(O_{t+1} = o \mid S_{t+1} = s, A_t = a)$
  - Reward function (R)

    $R(s, a)$



POMDP Model



- Belief: distribution over the state space.
- Strategy/policy: Best mapping from beliefs to actions.

"Best" policy

- Maps each belief to an action that satisfies the following objective function

$$V^*(b) = \max_{a \in A}\left(\underbrace{\sum_{s \epsilon S} R(s,a)b(s)}_{\text{Expected immediate reward}} + \gamma \underbrace{\sum_{o \epsilon O} P(o|b,a)V^*(b')}_{\text{Expected total future reward}}\right)$$

$P(o|b,a)$: The probability of perceiving observation o after the system at belief b performs action a

b': next belief after the system at belief b performs action a and observes o

γ: discount factor, (0,1)

Just for completeness…

$$b'(s') = \frac{Z(s', a, o) \sum_{s \epsilon S} T(s, a, s') b(s)}{P(o|a, b)}$$

$P(o|a,b)$: can be computed as a normalizing factor

Derivation is not in this class, but I'll talk about them next semester in Advanced AI class

POMDP Solution

- The policy that maximizes the value of all beliefs
- Computing such a policy is PSPACE-hard
  [Papadimitriou & Tsikilis'87, Madani, et.al.'99]
- In practice,
  - Approximate the value function
  - Policy that maximizes the approximated value of the initial beliefs b_0
- Many practical methods for solving:

  <adMode = on>
  - Not in this class, but I'll talk about some of them next semester in Advanced AI class
  - Software, e.g.: http://rdl.cecs.anu.edu.au/software

  <adMode = off> ☺

This set of videos

- ✓ Part-1: Intro
  - ✓ Automated decision-making
- ✓ Part-2: Intro to POMDP
  - ✓ Framework for decision-making under uncertainty
  - ✓ Solving, aka. generating strategic decisions
- Part-3: Example of POMDP in Cyber security
  - Autonomous pen-testing