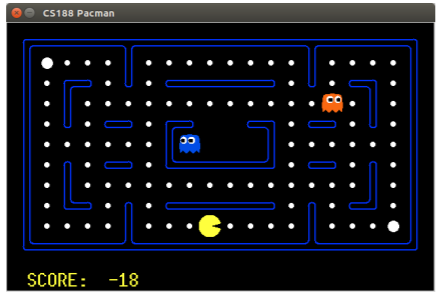
Complex Decision Making

Example 1: Pacman



One decision leads to another.

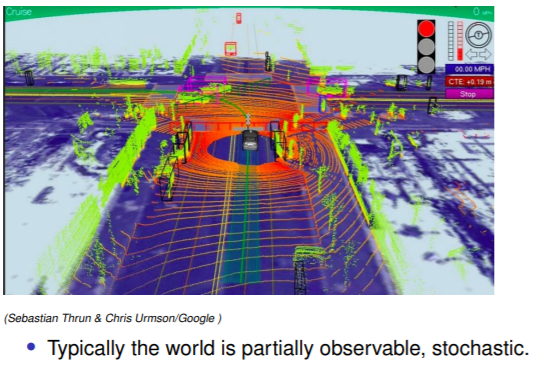
Each decision depends on the ones before and affects the ones after.

* These sequences of decisions are one of the reasons that decision making is difficult.
* It is hard enough to choose a decision when all the relevant information about a given world is available

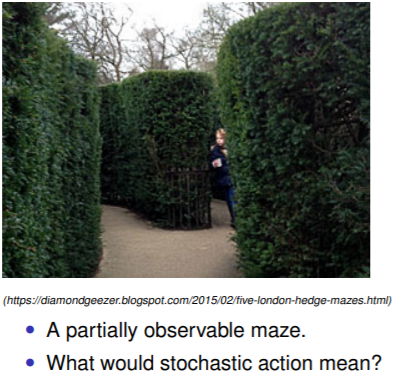
This is more complex in the real world because the world is not a fully observable deterministic process.

Factors that affect the Decision-Making Process

* Fully observable: You have all the information about a system like Pacman.
* Partially Observable: You only have some information and must make use of that information to decide.
* Deterministic World: This means there is only one consequence no matter the action you take.
* Stochastic: There are multiple possible consequences each linked to an action.
* Static World: Nothing changes except what the specific agent we are thinking about, does
* Dynamic World: The world changes with and without interference.
* Are there one or multiple agents in a world



The key point is that the car does not have a complete view of the world due to the range of its sensors, its “area of attention”.



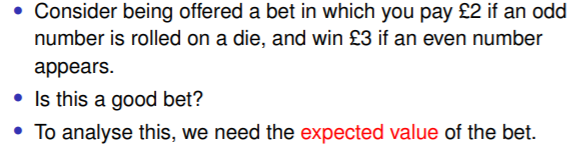
Stochastic action would mean that the robot may chose to go left but still have a probability of going right and vice-versa. This also includes not doing anything.

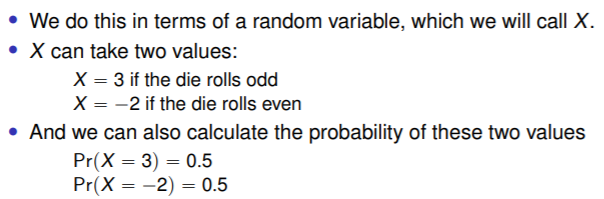
More examples:

If a robot is commanded to go 10 meters forward it may go 11 meters instead or 9 meters. There is a distribution over the result of the command due to a variety of issues such as wheel slippage or avoiding an obstacle causing the robot to deviate from its path etc. Robot may also go forward and to the left or forward and to the right. The same for backwards if the path is not blocked. As if the path is not blocked it is automatically a possible outcome in a stochastic world.

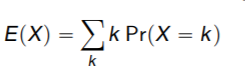
**Deciding what to do**

Start simple and make a single decision.



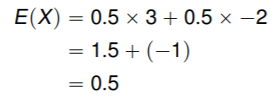


Expected value: The weighted sum of the values where the weights are the probabilities



Where the summation is over all the values of k for which P (X = k) != 0.

Plugging in the values from the previous slide the expected value is:



The expected value of X is £0.5 and we take this to be the value of the bet.

Is this a good bet?

Expected value of not taking the bet is £0 therefore it is.

0.5 is not the value you will get. For each bet you either win £3 or lose £2, £0.5 is simply the average you would earn if you took the bet multiple times

Another bet example:

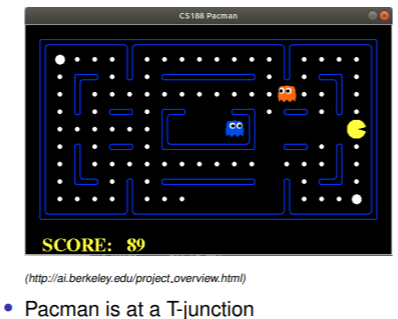
You get £1 if a 2 or 4 is rolled, £5 is a 6 is rolled and pay £3 otherwise

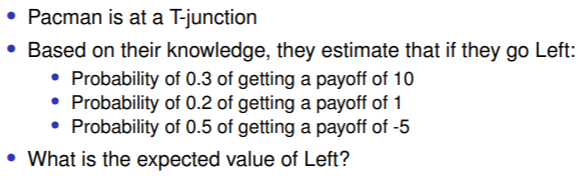
Expected value:

E(X) = 1 \* 1/3 + 5 \* 1/6 \* - 3

= -0.33

Example 4: Pacman at a T-Junction





0.3 \* 10 + 0.2 \* 1 + 0.5 \* -5 = 0.7

How agents decide what to do

Consider an agent with a set possible actions “A”.

Each “a” is one of many, sums up to a total “A” and has a set of possible outcomes sa. (s stands for states).Which action should the agent pick?

The action a\* which a rational agent should choose is that which maximises the agent’s utility. This means we assume that the agent looks at all the actions that it might do and all the outcomes of those, how good those different outcomes are, and then pick the action that maximises the utility.

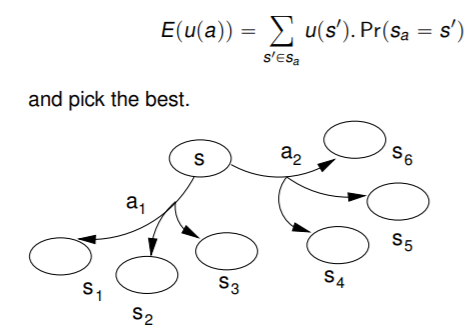
In other words, the agent should pick:



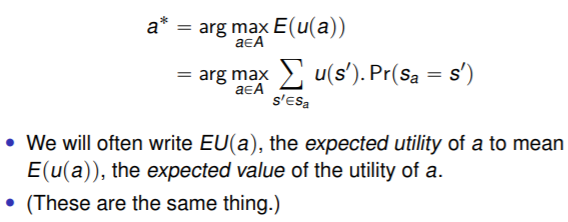
The problem is that in any realistic situation, we don’t know which sa will result from a given a, so we don’t know the actual utility of a given action.

Instead, we have to calculate the expected utility of each action and make the choice on that basis.

This means, for each action a with a set of outcomes sa, the agent should calculate:



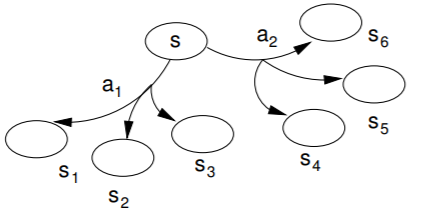
The action that has the greatest expected utility is picked:



The “rational” thing for an agent to do is the action that maximises the expected utility (which we can calculate) as we cannot pick the action based on actual utility yet since this is apparent after the action has occurred.

Non-deterministic

Note that we are dealing with non-deterministic actions here.



A given action has several possible outcomes

We have no way of knowing in advance which outcome will occur.

The lesser of evils approach

There are other criteria for decision making than maximising expected utility. One method of picking an action is to look at the action with the least terrible outcome.

This maximin criterion can be formalised in the same framework as MEU, making the rational (in this sense) action:



Its effect is to ignore the probability of outcomes and concentrate on optimising the worst-case outcome.

The opposite attitude, that of optimistic risk-seeker, is captured by the Maximax criterion: 

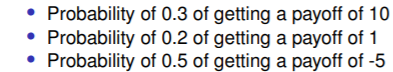
This will ignore possible bad outcomes and just focus on the best outcome of each action.

Example - PACMAN at a T junction



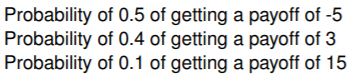
Based on the agent’s knowledge, it estimates that

if it goes left:



* 0.3 \*10 + 0.2 \* 1 + 0.5 \* -5 = E
* Expected utility = 0.7

If it goes right:



* 0.5 \* -5 + 0.4 \* 3 + 0.1 \* 15 = E
* Expected utility = 0.2

Example question 1:

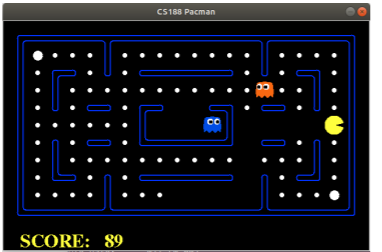
If we use the Maximin criterion the agent may go either left or right because Maximin optimises the worst-case outcome but the worst case outcome in both left and right is -2.5 therefore they are equally as bad.

Example question 2:

If we use the Maximax criterion then the agent would go right because out of all the actions, if we ignore probability, right has the highest payoff of 15.

Left’s highest payoff is only 10 in contrast.

What is being modelled?

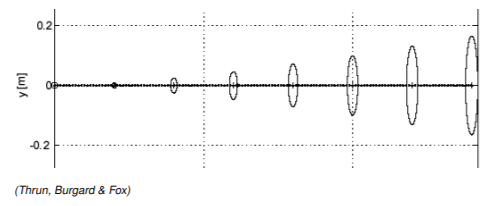


PACMAN in “real-life” is fully observable and deterministic

So, what are we modelling with the probabilities?

* Partial observability
* Stochastic/non-deterministic actions
* Both

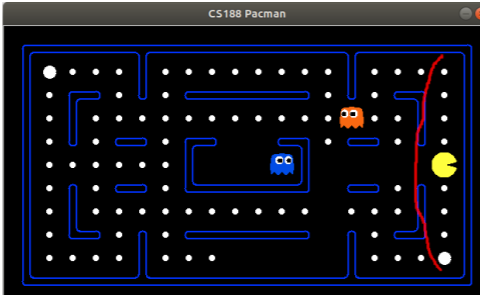
Non-Deterministic actions



This demonstrates the movement of a robot in an experiment as it tries to follow instructions. The ovals demonstrate how the radius of the error increases as the robot gets further from its starting point. This error is created by things such as the size of the wheel, slippage etc.

* Intended movement is not necessarily the actual movement.
* Common occurrence in robotics.

Partial observability



The agent doesn’t know everything, its view is limited by its area of influence. The areas within the red line in this example.

Sequential decision problems

These approaches provide a range of techniques that agents can apply to possible decisions in each stage. However, the techniques aren’t really sufficient.

Agents aren’t usually in the business of taking single decisions.

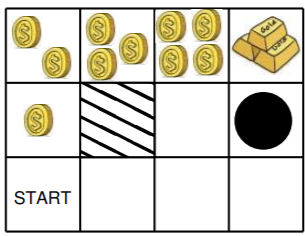
* Life is a continuous series of decisions.

The current best option may not be the best long-term option.

Picking the current best option is know as the greedy approach.

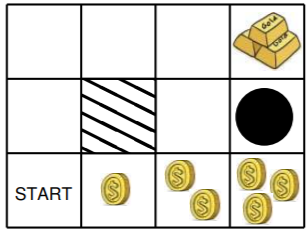
The greedy approach is “myopic” (short-sighted). It only looks at the next step (immediate result of the action). And therefore, does not consider the longer view which could hinder the agent’s performance.

Myopic, greedy approach example 1: “Greedy is GOOD”



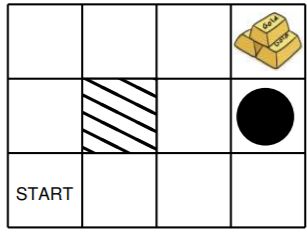
In this case We tell the agent that the greedy approach leads to the big prize

Myopic, greedy approach example 2: “Greedy is BAD”



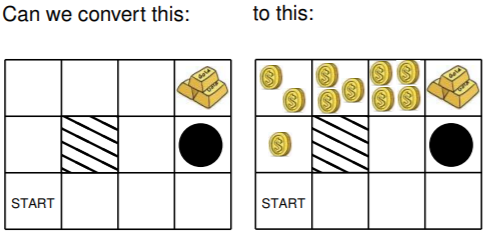
In this case we tell the agent that the greedy approach leads us to a bad place.

Myopic, greedy approach example 3: “Greedy is USELESS”



In MOST cases the greedy approach doesn’t help at all

The best technique in this scenario would be a random walk as this will be unbiased due to a lack of information.

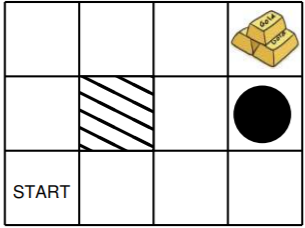


It is possible to turn a situation where we have no information about the word regarding what approach is best, into a situation where we can apply a specific approach. This can be done by using the utility.

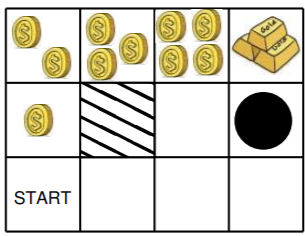
Reward vs. Utility

The key element is to distinguish reward from utility.

In this case, reward is what we get from being in a particular location:



The rewards don’t have to be positive, in this example, they are collecting the gold or falling into the pit. Those are the only states that have rewards in this “grid world”

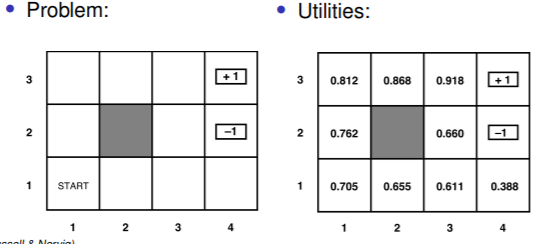


The Utility then illustrates how much being in a location is worth to the agent.

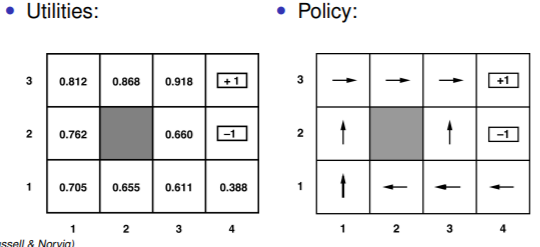
For instance, the agent can know, from the utility that being in a spot next to the goal spot is of benefit and so on and so forth even from the agent’s starting location.

The Utility can also factor in potential future rewards.

A more usual example:



On the left is the reward structure which tells the agent the locations of rewards.

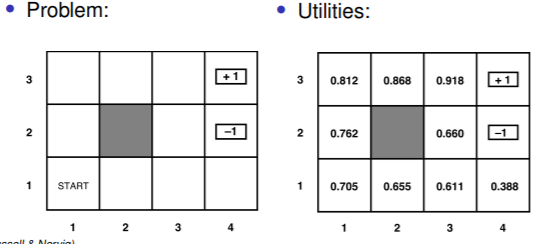
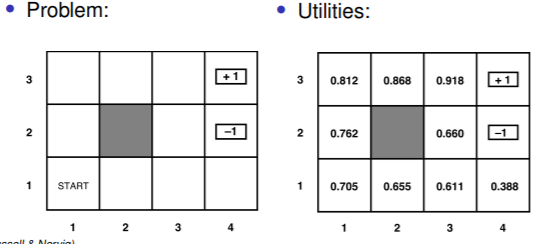
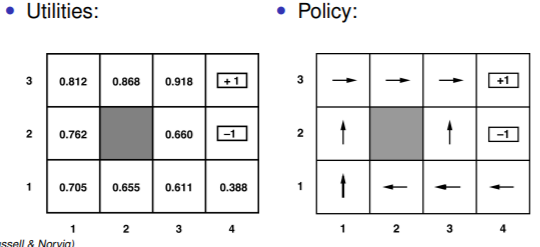


With the information from the rewards, we can then derive utilities (values) for individual states.

With the utility values we can then figure out what the right thing to do is, what the right A\* Action is by looking at the utilities that have been assigned to the various states and applying a “greedy approach”.

If you add all those utilities together you get an expected utility which if you maximise, you get a policy(arrows) based on the expected utility and a particular model of action. The policy represents a choice of action for every state and the choices of actions for each state is indicated by the arrows.

The general approach to solving a problem is to frame it as a Markov decision process and then take the problem which looks like so:

And then the policy

and then turn it into the utilities