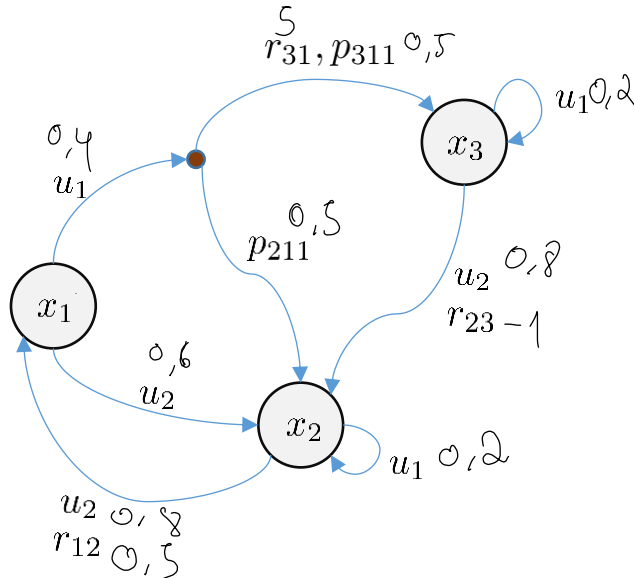


# Artificial Intelligence in Automotive Technology

## Reinforcement Learning

### Exercise 1

An MDP is given in the figure below, with  $x$  describing the states,  $u$  the actions,  $r$  the reward and  $p$  additional transition probabilities. The discount factor to be considered is  $\gamma = 0.5$ . For transitions where no reward is defined, assume  $r=0$ . Round your results to 3 decimal places.



$$r_{31} = 5$$

$$r_{23} = -1$$

$$r_{12} = 0.5$$

$$p_{311} = p(x_3|x_1, u_1) = 0.5$$

$$p_{211} = p(x_2|x_1, u_1) = 0.5$$

The policy  $\pi$  defines the action probabilities as follows:

$$\pi(u_1|x_1) = 0.4$$

$$\pi(u_2|x_1) = 0.6$$

$$\pi(u_1|x_2) = 0.2$$

$$\pi(u_2|x_2) = 0.8$$

$$\pi(u_1|x_3) = 0.2$$

$$\pi(u_2|x_3) = 0.8$$

- Calculate the value function of all the states.
- Calculate the values of the action value function  $Q(x_2, u_1)$ ,  $Q(x_2, u_2)$  and  $Q(x_1, u_1)$ .
- Based on the results in b), would a greedy policy pick action  $u_1$  or  $u_2$  in state  $x_2$ ? For which reason?
- If you derived an  $\epsilon$ -greedy policy based on the Q function from b), with  $\epsilon = 0.2$ , what would be the probability to pick  $u_2$  in state  $x_2$ ?

### Exercise 2

Open the provided Jupyter notebook and complete the exercise on Q-learning.