# DRL Homework 01

Group 10

April 2022

## 1 Task 01

Within the game of chess the **agent** is the player and the **environment** is the chessboard and the opponents pieces.
The **set of states** can be thought of as all possible configurations of pieces on the field.This could either be represented as an 8x8 grid with the respective pieces in each field or as a 2x16 state representation where each piece is denoted with its current position. While the former might be more readable the second might be more computationally efficient. The **set of actions** is the set of all allowed moves (by all respective pieces), this changes depending on the state of the board. Each time **time step** consists of the action performed by the agent and the responding action of the opponent.
Since the end goal of chess is to checkmate the opponents king, states depicting this should result in a positive **reward** (e.g. 1). If the opposite is true (the agent's king is in checkmate) a penalty should be given (a negative reward like -1). Since winning the game as soon as possible is preferred a slight action penalty could be introduced (a small negative reward in each state that does not lead to ending the game e.g. 0.1). Maybe there could also be a small reward for decimating the opponents pieces, thus encouraging attacks, but a small penalty for losing a piece, therefore also encouraging defense.
The **policy** for an action in a current state that would result in a checkmate state would be high.

## 2 Task 02

The same principles from before also hold for the Lunar Lander as and MDP. The Agent would be the Lander (or the entitiy steering it). The Environment is represented in the planet the agent is landing on.
The first part of the reward is the negative distance to the landing pat, thus the objective can be thought of as minimizing the penalty (rather then the regular reward maximization). The second part is the actual landing of the Lander. If it lands successfully it receives a positive (100 points) reward while a crash would result in a negative one (-100). In addition to this, the agent receives a positive reward of 10 for each leg ground contact as well as a slight penalty of

0.3 for firing the engine. Solving i.e. landing successfully in the landing pad grants a reward of 200.
The state is defined by the position of the landing pad and the position of the lander itself.
The action space is 4. Namely, the agent can either fire the main engine, do nothing or fire the right/left engine.
Thus, the policy would be the probability of either one of those for action given the current position of the lander and the landing pad.

# 3 Task 03

1. The reward function defines the probability of reaching a specific reward $R_{t+1}$, given a previous state $s$ and a performed action $a$: $p(R_{t+1}|s, a)$.

   For example the expected reward $E(R_{t+1}|s, a)$ of performing the action $a$ "right engine" in a state $s$ where the landing pad is to the right of the lander is relatively high, since the lander is now closer to the goal state of landing.

   The transition function models the probability of encountering a state $s_{t+1}$ after performing an action $a$ in a previous state $s_t$.

   In an uncertain environment with lots of different factors that are unaccounted for, like for example a Lander in the Lunar environment, there is no certainty that performing a specific action will in fact result in a specific state.

2. The policy iteration makes use of these environment dynamics. Thus, the environment dynamics have to be modeled beforehand. Oftentimes for this only the goal/end state has to be known. Thus if these and the possible actions are given the problem can be solved with RL. Nonetheless, in some cases the state transition probabilities might be unknown, meaning it is hard to model the exact physics and properties of the environment. This can be a problem in real world applications like autonomous driving, as we might not be able to model all contingencies. Hence, in some cases model-free learning might be more beneficial since the environment physics don't have to be known.