How to proceed:

1.Explore the world to gather experience

2.Use experience to guide future decisions

**Trajectory** is the set of elements that are produced when agent moves from one state to another e.g. (S0, A0, R1, S1, A1, R2, S2).

**Episode** is a trajectory that starts from the initial state and ends in a terminal (final) state.

**Reward** is the Immediate result that our actions produce.

**Return** is the sum of rewards from a certain point in time until the task is completed.

The goal is to maximize the episodes’ return.

**Discount factor (γ):** reduces the reward as time goes by. 0.99 chosen here.

**Policy** is mapping from past experience to actions. It is a function that decides what action to take in a particular state, it takes the state as an input and outputs an action.

Here, we chose an ε-greedy policy to have room for exploration. The ε value starts at 1 for maximum exploration and is reduced to 0.1 over the first 1 million frames linearly.

**Replay memory** is the memory that stores the state transitions that the agent experiences. The memory has a predefined capacity and when it fills up new transitions will replace the oldest ones. Here, we chose 50000 as we had limitations with the memory size.

To update the neural network, we randomly choose a **batch** of transitions from memory each time. We chose the batch size of 32 transitions here.