**Summary**

[The agent-environment interaction in reinforcement learning. (Source: Sutton and Barto, 2017)](https://classroom.udacity.com/nanodegrees/nd009t/parts/ac12e0fe-e54e-40d5-b0f8-136dbdd1987b/modules/f87db1ea-a332-4007-9f37-5e641d80c92a/lessons/86acfc34-0551-4cc6-8de4-a1ab2e66b5af/concepts/ee28399b-f809-4e2b-936b-5a88d7297899)

**The Setting, Revisited**

* The reinforcement learning (RL) framework is characterized by an **agent** learning to interact with its **environment**.
* At each time step, the agent receives the environment's **state** (*the environment presents a situation to the agent)*, and the agent must choose an appropriate **action** in response. One time step later, the agent receives a **reward** (*the environment indicates whether the agent has responded appropriately to the state*) and a new **state**.
* All agents have the goal to maximize expected **cumulative reward**, or the expected sum of rewards attained over all time steps.

**Episodic vs. Continuing Tasks**

* A **task** is an instance of the reinforcement learning (RL) problem.
* **Continuing tasks** are tasks that continue forever, without end.
* **Episodic tasks** are tasks with a well-defined starting and ending point.
  + In this case, we refer to a complete sequence of interaction, from start to finish, as an **episode**.
  + Episodic tasks come to an end whenever the agent reaches a **terminal state**.

**The Reward Hypothesis**

* **Reward Hypothesis**: All goals can be framed as the maximization of (expected) cumulative reward.

**Goals and Rewards**

* (Please see **Part 1** and **Part 2** to review an example of how to specify the reward signal in a real-world problem.)

**Cumulative Reward**

* The **return at time step**t*t* is G\_t := R\_{t+1} + R\_{t+2} + R\_{t+3} + \ldots*Gt*​:=*Rt*+1​+*Rt*+2​+*Rt*+3​+…
* The agent selects actions with the goal of maximizing expected (discounted) return. (*Note: discounting is covered in the next concept.*)

**Discounted Return**

* The **discounted return at time step**t*t* is G\_t := R\_{t+1} + \gamma R\_{t+2} + \gamma^2 R\_{t+3} + \ldots*Gt*​:=*Rt*+1​+*γRt*+2​+*γ*2*Rt*+3​+….
* The **discount rate**\gamma*γ* is something that you set, to refine the goal that you have the agent.
  + It must satisfy 0 \leq \gamma \leq 10≤*γ*≤1.
  + If \gamma=0*γ*=0, the agent only cares about the most immediate reward.
  + If \gamma=1*γ*=1, the return is not discounted.
  + For larger values of \gamma*γ*, the agent cares more about the distant future. Smaller values of \gamma*γ*result in more extreme discounting, where - in the most extreme case - agent only cares about the most immediate reward.

**MDPs and One-Step Dynamics**

* The **state space**\mathcal{S}S is the set of all (*nonterminal*) states.
* In episodic tasks, we use \mathcal{S}^+S+ to refer to the set of all states, including terminal states.
* The **action space**\mathcal{A}A is the set of possible actions. (Alternatively, \mathcal{A}(s)A(*s*) refers to the set of possible actions available in state s \in \mathcal{S}*s*∈S.)
* (Please see **Part 2** to review how to specify the reward signal in the recycling robot example.)
* The **one-step dynamics** of the environment determine how the environment decides the state and reward at every time step. The dynamics can be defined by specifying p(s',r|s,a) \doteq \mathbb{P}(S\_{t+1}=s', R\_{t+1}=r|S\_{t} = s, A\_{t}=a)*p*(*s*′,*r*∣*s*,*a*)≐P(*St*+1​=*s*′,*Rt*+1​=*r*∣*St*​=*s*,*At*​=*a*) for each possible s', r, s, \text{and } a*s*′,*r*,*s*,and *a*.
* A **(finite) Markov Decision Process (MDP)** is defined by:
  + a (finite) set of states \mathcal{S}S (or \mathcal{S}^+S+, in the case of an episodic task)
  + a (finite) set of actions \mathcal{A}A
  + a set of rewards \mathcal{R}R
  + the one-step dynamics of the environment
  + the discount rate \gamma \in [0,1]*γ*∈[0,1]