At a high-level, Reinforcement Learning is learning through interaction. The four necessary elements defined in Reinforcement Learning include the agent, environment action and reward. The goal of Reinforcement Learning is to have the agent interact with its environment where the consequences of the agent’s actions will lead to a reward. Based on the reward, the agent will learn whether an action taken is positive or negative which enables the agent to learn over time how to choose most optimal action to take for any given situation. At first glance, the context of Reinforcement Learning may appear to be alike other machine learning approaches where an agent might be exposed to the correct action to take during training scenarios. However, Reinforcement Learning differs in the aspect that the agent will strictly only learn through trial and error.

An agent is defined to be autonomous, usually being a program that utilizes a machine learning algorithm to interact with an environment. Everything that an agent interacts with is defined as the environment. The goal of the agent is to be able to maximize its predictability capabilities to predict anomalies within the environment. This prediction is known as an action where once an action is taken, the agent will transition into a new state. The agent will transition to a new state whenever it takes an action as once an action is taken, it will be exposed to a new set of uncertainties, requiring the agent to take another action. Since the objective is to maximize an agent’s predictability capabilities through rewards, the agent aims to develop a sequence of actions that will enable it to achieve this goal. In order to develop the optimal sequence of actions, the agent utilizes a Markov Decision Process so that the agent can take the best action in each state, known as an optimal policy.

The Markov Decision Process (MDP) is defined by two key characteristics: The Markov Property and the Markov Chain. The Markov Property states that the conditional property of states to occur in the future will not depend on the current sequence of states, but instead only the current state. In other words, the next state will only be dependent on the current state. The Markov Chain utilizes the Markov Property in a random process where the Markov Chain contains a set of states where each state contains probabilities describing the likelihood of transitioning to each of the next possible states. The MDP is built upon a Markov Chain but also includes a set of possible actions for the agent, a reward function that determines the immediate reward of an action, and a discounted reward function that determines the cumulative rewards amassed by the agent as additional variables to consider. The reward function is one of the most critical components of the MDP. Its goal is to evaluate the reward of the action taken by the agent while also having to consider balancing the importance of immediate and long-term rewards. The balancing of the importance of immediate and long-term rewards is done by the disc0unt function. Since each environment contains numerous uncertainties, the agent will continuously improve its estimation of achieving the optimal policy as the progression of transitioning through states occurs. The expected result is for an agent to be able to maximize the accumulation of rewards once it has been trained a sufficient amount of times.

In hindsight, the key strengths of Reinforcement Learning is that it can solve complex problems autonomously without the reliance on humans, it can create an appropriate model for a specific problem, and that it is self-learning where it will correct its own mistakes. However, it holds the assumption that a system is Markov which is not true as in reality, most events are the result of multiple events preceding it. Reinforcement Learning can also be problematic when models are overtrained (have too many states) as it may make the results less credible or inaccurate.