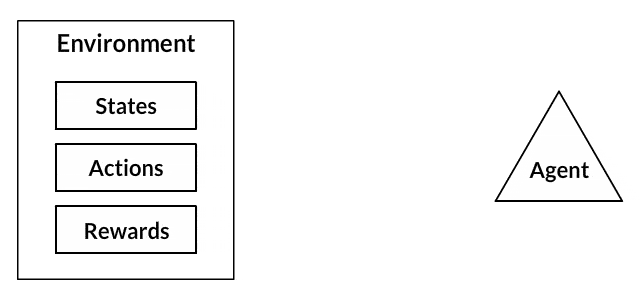
**LEARN HOW TO BUILD AN AI** – *Self Driving Car Simulation*

* Reinforcement Learning
* Bellman Equation
* The Plan
* Markov Decision Process (MDP)
* Policy Vs Plan
* Adding a “living penalty”
* Q-Learning Intuition
* Temporal Difference
* Q – Learning Visualization

*Reinforcement Learning* –

It involves an [agent](https://en.wikipedia.org/wiki/Intelligent_agent), a set of states [S] and a set of actions [A] per state. By performing an action, the agent moves from one state to other state. After performing an action in a specific state provides the agent gets a reward (a numerical score). The goal of the agent is to maximize its total (future) reward.



Action

State/ Reward

*Bellman Equation* –

It is the basic block of solving reinforcement learning. It helps us to solve MDP. To solve means finding the optimal policy and value functions.

V(s) = max (R(s, a) + γ V(s’))

Where R = reward in state‘s’ by performing action ‘a’

γ = discount factor

*The Plan* –

Well the plan is simply like a treasure map for artificial intelligence instead of looking at the values we got from bellman equation, we just replace them with arrows which indicate in which direction the agent should go. It knows the value of being in each state and therefore we can come up with this map.

*Markov Decision Process (MDP)* –

1. If the conditional probability distribution of future states of the process (conditional and both past and present state) depends only upon the present state not on the sequence of events that preceded it. A process with this property is called a Markov process.
2. It provides a mathematical framework for modeling decision making in situations where the outcome is random and sometimes it is under the control of the decision maker.
3. As we have randomness in our process, the bellman equation modified as below;

V(s) = max [R(s, a) + γ where s’ = possible number of states that agent can get into as it is a non-deterministic search involves randomness.

*Living Penalty –*

So far we are getting rewards at the end i.e. if we finish at the end line we get +1 reward or if the agent ends in some other wrong place like in a fire pit we are getting -1 reward. But from now, a reward i.e. we consider minus 0.04 throughout the journey except at the final tiles that is continuously given to the agent throughout the game not just at the end. That’s what called a living penalty because no matter where he goes he will always get this negative reward except for these final tiles because that's the end of the game.

*Q – Learning Intuition* -

The 'Q' in Q-learning stands for quality. Quality in this case represents how useful a given action is in gaining some future reward.

V(s) = max [R(s, a) + γ

Q(s, a) = R(s, a) + γ

Q(s, a) = R(s, a) + γ

*Temporal difference* –

Before taking an action, the value of a state where the agent is - Q(s, a)

After taking an action, the value of a state – R(s, a) + γ max Q (s’, a’)

So, here we calculate the temporal difference by;

TD(s, a) = R(s, a) + γ max Q (s’, a’) – Q (s, a)

We are re-writing this with the learning rate ‘α’ -> Q (s, a) = R(s, a) + α TD(s, a)