Machine Learning Engineer Nanodegree

Reinforcement Learning

Project: Train a Smartcab to Drive

Welcome to the fourth project of the Machine Learning Engineer Nanodegree! In this notebook, template code has already been provided for you to aid in your analysis of the *Smartcab* and your implemented learning algorithm. You will not need to modify the included code beyond what is requested. There will be questions that you must answer which relate to the project and the visualizations provided in the notebook. Each section where you will answer a question is preceded by a **'Question X'** header. Carefully read each question and provide thorough answers in the following text boxes that begin with **'Answer:'**. Your project submission will be evaluated based on your answers to each of the questions and the implementation you provide in `agent.py`.

> **Note:** Code and Markdown cells can be executed using the **Shift + Enter** keyboard shortcut. In addition, Markdown cells can be edited by typically double-clicking the cell to enter edit mode.

Getting Started

In this project, you will work towards constructing an optimized Q-Learning driving agent that will navigate a *Smartcab* through its environment towards a goal. Since the *Smartcab* is expected to drive passengers from one location to another, the driving agent will be evaluated on two very important metrics: **Safety** and **Reliability**. A driving agent that gets the *Smartcab* to its destination while running red lights or narrowly avoiding accidents would be considered **unsafe**. Similarly, a driving agent that frequently fails to reach the destination in time would be considered **unreliable**. Maximizing the driving agent's **safety** and **reliability** would ensure that *Smartcabs* have a permanent place in the transportation industry.

Safety and **Reliability** are measured using a letter-grade system as follows:

Grade	Safety	Reliability
A+	Agent commits no traffic violations, and always chooses the correct action.	Agent reaches the destination in time for 100% of trips.
A	Agent commits few minor traffic violations, such as failing to move on a green light.	Agent reaches the destination on time for at least 90% of trips.
B	Agent commits frequent minor traffic violations, such as failing to move on a green light.	Agent reaches the destination on time for at least 80% of trips.
C	Agent commits at least one major traffic violation, such as driving through a red light.	Agent reaches the destination on time for at least 70% of trips.
D	Agent causes at least one minor accident, such as turning left on green with oncoming traffic.	Agent reaches the destination on time for at least 60% of trips.
F	Agent causes at least one major accident, such as driving through a red light with cross-traffic.	Agent fails to reach the destination on time for at least 60% of trips.

To assist evaluating these important metrics, you will need to load visualization code that will be used later on in the project. Run the code cell below to import this code which is required for your analysis.

In [1]:
```
# Import the visualization code
import visuals as vs

# Pretty display for notebooks
%matplotlib inline
```

Understand the World

Before starting to work on implementing your driving agent, it's necessary to first understand the world (environment) which the *Smartcab* and driving agent work in. One of the major components to building a self-learning agent is understanding the characteristics about the agent, which includes how the agent operates. To begin, simply run the `agent.py` agent code exactly how it is -- no need to make any additions whatsoever. Let the resulting simulation run for some time to see the various working components. Note that in the visual simulation (if enabled), the **white vehicle** is the *Smartcab*.

Question 1

In a few sentences, describe what you observe during the simulation when running the default `agent.py` agent code. Some things you could consider:

- *Does the Smartcab move at all during the simulation?*
- *What kind of rewards is the driving agent receiving?*
- *How does the light changing color affect the rewards?*

Hint: From the `/smartcab/` top-level directory (where this notebook is located), run the command

```
'python smartcab/agent.py'
```

**Answer: 1. The Smartcab does not move during the simulation. 2. The driving agent receives points for no action taken. 3. If the driving agent is idled at a green light it receives positive points, if it is idled during a green light points are subtracted. **

Understand the Code

In addition to understanding the world, it is also necessary to understand the code itself that governs how the world, simulation, and so on operate. Attempting to create a driving agent would be difficult without having at least explored the *"hidden"* devices that make everything work. In the `/smartcab/` top-level directory, there are two folders: `/logs/` (which will be used later) and `/smartcab/`. Open the `/smartcab/` folder and explore each Python file included, then answer the following question.

Question 2

- *In the* `agent.py` *Python file, choose three flags that can be set and explain how they change the simulation.*
- *In the* `environment.py` *Python file, what Environment class function is called when an agent performs an action?*
- *In the* `simulator.py` *Python file, what is the difference between the* `'render_text()'` *function and the* `'render()'` *function?*
- *In the* `planner.py` *Python file, will the* `'next_waypoint()` *function consider the North-South or East-West direction first?*

**Answer:

1. In the Agent.py three flags that can be set are: Display : The visual simulation is made possible by PyGame GUI if is set to True. The default setting is True.

   Learning: Set to false by default. If set to true, the agent is expected to learn using Q-learning. This means that at the moment, the agent isn't expected to change is behavior during the simulation. Learning is afffected by epsilon (continuous value for the exploration value) and alpha(continuous value for th elearning rate). Optimized: Set to True to change the default log file name.
2. The environment.py

   The Environment class function called when an agent performs an action is act(). Examples: "Environment.act(): Primary agent has reached destination!" and "Environment.act(): Step data: {}".format(self.step_data)
3. In the simulator.py the difference between 'render_text()' function and 'render()' funtion : 'render_text()', This is the non-GUI render display of the simulation. Simulated trial data will be rendered in the terminal/command prompt. 'render()', This is the GUI render display of the simulation. Supplementary trial data can be found from render_text. In summary the render_text() function display the current state in text while the 'render() display in a Graphical User Interface(GUI)
4. In Planner.Py , will the 'next_wayPoint()' function consider the North-South or East-West direction. The East-West is first considered then North-South direction. **

Implement a Basic Driving Agent

The first step to creating an optimized Q-Learning driving agent is getting the agent to actually take valid actions. In this case, a valid action is one of `None`, (do nothing) `'left'` (turn left), `right'` (turn right), or `'forward'` (go forward). For your first implementation, navigate to the `'choose_action()'` agent function and make the driving agent randomly choose one of these actions. Note that you have access to several class variables that will help you write this functionality, such as `'self.learning'` and `'self.valid_actions'`. Once implemented, run the agent file and simulation briefly to confirm that your driving agent is taking a random action each time step.

Basic Agent Simulation Results

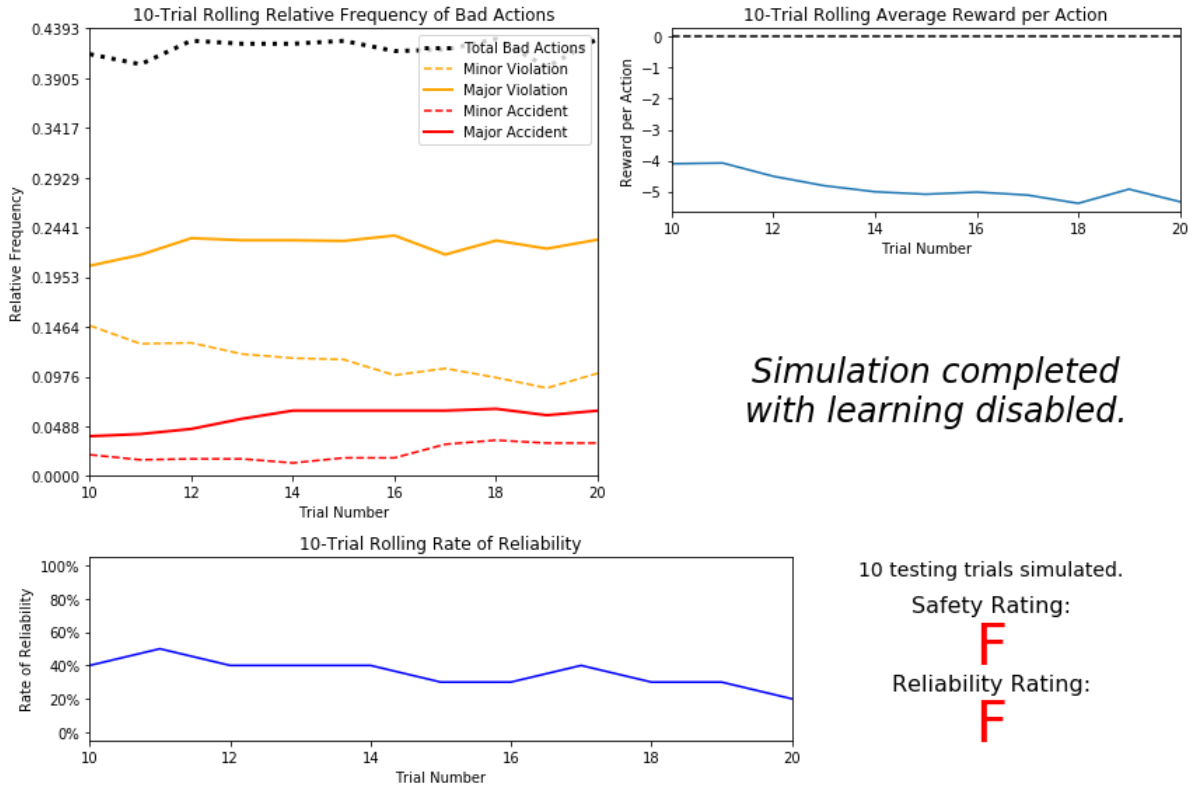
To obtain results from the initial simulation, you will need to adjust following flags:

- `'enforce_deadline'` - Set this to `True` to force the driving agent to capture whether it reaches the destination in time.
- `'update_delay'` - Set this to a small value (such as `0.01`) to reduce the time between steps in each trial.
- `'log_metrics'` - Set this to `True` to log the simluation results as a `.csv` file in `/logs/`.
- `'n_test'` - Set this to `'10'` to perform 10 testing trials.

Optionally, you may disable to the visual simulation (which can make the trials go faster) by setting the `'display'` flag to `False`. Flags that have been set here should be returned to their default setting when debugging. It is important that you understand what each flag does and how it affects the simulation!

Once you have successfully completed the initial simulation (there should have been 20 training trials and 10 testing trials), run the code cell below to visualize the results. Note that log files are overwritten when identical simulations are run, so be careful with what log file is being loaded! Run the agent.py file after setting the flags from projects/smartcab folder instead of projects/smartcab/smartcab.

In [2]:
```
# Load the 'sim_no-learning' log file from the initial simulation results
vs.plot_trials('sim_no-learning.csv')
```



Question 3

Using the visualization above that was produced from your initial simulation, provide an analysis and make several observations about the driving agent. Be sure that you are making at least one observation about each panel present in the visualization. Some things you could consider:

- *How frequently is the driving agent making bad decisions? How many of those bad decisions cause accidents?*
- *Given that the agent is driving randomly, does the rate of reliability make sense?*
- *What kind of rewards is the agent receiving for its actions? Do the rewards suggest it has been penalized heavily?*
- *As the number of trials increases, does the outcome of results change significantly?*
- *Would this Smartcab be considered safe and/or reliable for its passengers? Why or why not?*

**Answer:

1. The driving agent made bad decisions around 45% of the time . The relative frequency of major accidents is around 6.5% of the time and minor accidents around 3%; for total accidents of about 9.5% of the time. Therefore these bad decisions resulted in accident approximately 21% of the time.

2.Given the agent is driving randomly 10% reliability is a likely outcome therefore it makes sense for these simulations.

1. The agent started with rewards being a little over negative -4 but decline to -6 as the number of trials approached 20. The results suggest it is getting heavily penalized as the simulations increased.
2. As number of trials increases bad actions increase slightle then tapered off at a steady 45%, however minor violations decrease 16.5% to around 9%. Major violations was steady at around 22% and both major and minor accidents increase slightly. Reliability started at about 10% improved to around 5% and remained constant. Reward per action decline. The results for safet and reliability at F are justified from these results
3. With a safety rating and reliability rating at F it is not safe or reliable for passengers. The frequency of accidents and late arrivals are why it is not considered safe or reliable. **

----- ## Inform the Driving Agent The second step to creating an optimized Q-learning driving agent is defining a set of states that the agent can occupy in the environment. Depending on the input, sensory data, and additional variables available to the driving agent, a set of states can be defined for the agent so that it can eventually *learn* what action it should take when occupying a state. The condition of `'if state then action'` for each state is called a **policy**, and is ultimately what the driving agent is expected to learn. Without defining states, the driving agent would never understand which action is most optimal -- or even what environmental variables and conditions it cares about!

Identify States

Inspecting the `'build_state()'` agent function shows that the driving agent is given the following data from the environment:

- `'waypoint'`, which is the direction the *Smartcab* should drive leading to the destination, relative to the *Smartcab*'s heading.
- `'inputs'`, which is the sensor data from the *Smartcab*. It includes
  - `'light'`, the color of the light.
  - `'left'`, the intended direction of travel for a vehicle to the *Smartcab*'s left. Returns `None` if no vehicle is present.
  - `'right'`, the intended direction of travel for a vehicle to the *Smartcab*'s right. Returns `None` if no vehicle is present.
  - `'oncoming'`, the intended direction of travel for a vehicle across the intersection from the *Smartcab*. Returns `None` if no vehicle is present.
- `'deadline'`, which is the number of actions remaining for the *Smartcab* to reach the destination before running out of time.

Question 4

*Which features available to the agent are most relevant for learning both **safety** and **efficiency**? Why are these features appropriate for modeling the* Smartcab *in the environment? If you did not choose some features, why are those features* not *appropriate? Please note that whatever features you eventually choose for your agent's state, must be argued for here. That is: your code in agent.py should reflect the features chosen in this answer.*

NOTE: You are not allowed to engineer new features for the smartcab.

**Answer:

1. Waypoints: This specify the direction to the destination, this is important for both safety and efficiency. Knowing how to get to a destination will results in optimal outcome.
2. Input: light- The driving agent must be aware of the meaning of the colors of the lights. It will be penalized for sitting at a green light and rewarded for properly responding to a red. In traffic a large percentage of accidents and violations are caused by disobeying traffic lights; therefore obeying them results in learning how to be both safe and efficient.
3. Input: left- Due to the Right-of-Way rule in the USA it is important to know when to yield, waiting for the lights are required for making left turns. Accurate handling of left turn will help the driving agent in learning safety and efficiency.
4. Oncoming :Driving into oncoming traffic can be catastrophic, the driving agent must learn how to deal with vehicles coming into an intersection or coming from the opposite direction to be safe and efficient.

Features not appropriate :The ‘deadline’ feature is not crucial,since waypoint is the only information we need to go to the destination. Also, it is better not to break the traffic rules to keep the deadline. The ‘right’ of the input is not very important, however for NYC it should be included. (Unlike many North American cities, Right Turns on Red (RTOR) are severely restricted in New York City. Within the five boroughs, this movement is permitted only where posted and has been most prevalent in Staten Island, where lower traffic and pedestrian volumes allow for the safe movement of both vehicles and pedestrians[1]. For this implementation however I think it safe to say we could leave out the 'right' and “deadline” features because more features mean state space will be bigger; the result a more efficient learning algorithm.

Reference

[1] http://www.nyc.gov/html/dot/downloads/pdf/ssi09_rightonred.pdf (http://www.nyc.gov/html/dot/downloads/pdf/ssi09_rightonred.pdf) **

Define a State Space

When defining a set of states that the agent can occupy, it is necessary to consider the *size* of the state space. That is to say, if you expect the driving agent to learn a **policy** for each state, you would need to have an optimal action for *every* state the agent can occupy. If the number of all possible states is very large, it might be the case that the driving agent never learns what to do in some states, which can lead to uninformed decisions. For example, consider a case where the following features are used to define the state of the *Smartcab*:

`('is_raining', 'is_foggy', 'is_red_light', 'turn_left', 'no_traffic', 'previous_turn_left', 'time_of_day')`.

How frequently would the agent occupy a state like `(False, True, True, True, False, False, '3AM')`? Without a near-infinite amount of time for training, it's doubtful the agent would ever learn the proper action!

Question 5

*If a state is defined using the features you've selected from **Question 4**, what would be the size of the state space? Given what you know about the environment and how it is simulated, do you think the driving agent could learn a policy for each possible state within a reasonable number of training trials?*
Hint: Consider the *combinations* of features to calculate the total number of states!

**Answer:

```
Features  ---------     States                Total


Waypoint  ---------    forward, left, right      3
```

Inputs:

```
light             red, green                   2
left              None, forward, left,right    4
oncoming          None, forward, left, right   4
```

Size of the states space: 3 x 2 x 4 x 4=96.

96 is not a large number; given what I have learned about the environment the driving agent could learn a policy for each possible state with a couple hundred or reasonable number of training trials.

**

Update the Driving Agent State

For your second implementation, navigate to the `'build_state()'` agent function. With the justification you've provided in **Question 4**, you will now set the `'state'` variable to a tuple of all the features necessary for Q-Learning. Confirm your driving agent is updating its state by running the agent file and simulation briefly and note whether the state is displaying. If the visual simulation is used, confirm that the updated state corresponds with what is seen in the simulation.

Note: Remember to reset simulation flags to their default setting when making this observation!

Implement a Q-Learning Driving Agent

The third step to creating an optimized Q-Learning agent is to begin implementing the functionality of Q-Learning itself. The concept of Q-Learning is fairly straightforward: For every state the agent visits, create an entry in the Q-table for all state-action pairs available. Then, when the agent encounters a state and performs an action, update the Q-value associated with that state-action pair based on the reward received and the iterative update rule implemented. Of course, additional benefits come from Q-Learning, such that we can have the agent choose the *best* action for each state based on the Q-values of each state-action pair possible. For this project, you will be implementing a *decaying,* ϵ*-greedy* Q-learning algorithm with *no* discount factor. Follow the implementation instructions under each **TODO** in the agent functions.

Note that the agent attribute `self.Q` is a dictionary: This is how the Q-table will be formed. Each state will be a key of the `self.Q` dictionary, and each value will then be another dictionary that holds the *action* and *Q-value*. Here is an example:

```
{ 'state-1': {
    'action-1' : Qvalue-1,
    'action-2' : Qvalue-2,
     ...
   },
  'state-2': {
    'action-1' : Qvalue-1,
     ...
   },
   ...
}
```

Furthermore, note that you are expected to use a *decaying* ϵ *(exploration) factor*. Hence, as the number of trials increases, ϵ should decrease towards 0. This is because the agent is expected to learn from its behavior and begin acting on its learned behavior. Additionally, The agent will be tested on what it has learned after ϵ has passed a certain threshold (the default threshold is 0.05). For the initial Q-Learning implementation, you will be implementing a linear decaying function for ϵ.

Q-Learning Simulation Results

To obtain results from the initial Q-Learning implementation, you will need to adjust the following flags and setup:

- `'enforce_deadline'` - Set this to `True` to force the driving agent to capture whether it reaches the destination in time.
- `'update_delay'` - Set this to a small value (such as `0.01`) to reduce the time between steps in each trial.
- `'log_metrics'` - Set this to `True` to log the simluation results as a `.csv` file and the Q-table as a `.txt` file in `/logs/`.
- `'n_test'` - Set this to `'10'` to perform 10 testing trials.
- `'learning'` - Set this to `'True'` to tell the driving agent to use your Q-Learning implementation.

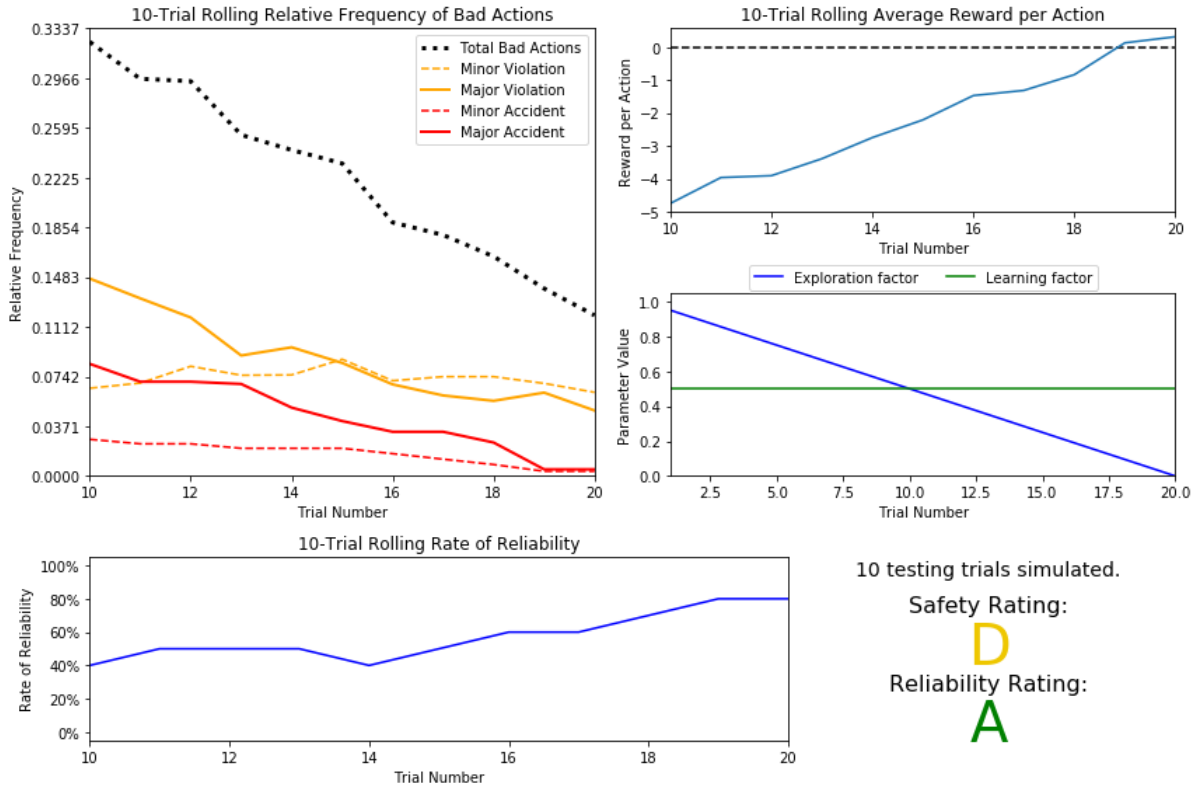
In addition, use the following decay function for ϵ:

$$\epsilon_{t+1} = \epsilon_t - 0.05, \quad \text{for trial number } t$$

If you have difficulty getting your implementation to work, try setting the `'verbose'` flag to `True` to help debug. Flags that have been set here should be returned to their default setting when debugging. It is important that you understand what each flag does and how it affects the simulation!

Once you have successfully completed the initial Q-Learning simulation, run the code cell below to visualize the results. Note that log files are overwritten when identical simulations are run, so be careful with what log file is being loaded!

In [3]:
```
# Load the 'sim_default-learning' file from the default Q-Learning simulation
vs.plot_trials('sim_default-learning.csv')
```



Question 6

Using the visualization above that was produced from your default Q-Learning simulation, provide an analysis and make observations about the driving agent like in **Question 3**. Note that the simulation should have also produced the Q-table in a text file which can help you make observations about the agent's learning. Some additional things you could consider:

- *Are there any observations that are similar between the basic driving agent and the default Q-Learning agent?*
- *Approximately how many training trials did the driving agent require before testing? Does that number make sense given the epsilon-tolerance?*
- *Is the decaying function you implemented for ϵ (the exploration factor) accurately represented in the parameters panel?*
- *As the number of training trials increased, did the number of bad actions decrease? Did the average reward increase?*
- *How does the safety and reliability rating compare to the initial driving agent?*

**Answer:

1. The main observations of similarity is that the Safety rating is still very bad; the driving agent from Question 3 scored F and the default Q-learning.
2. The tolerance of 0.05 which was the default value was what I used to decrement linearly. My initial ϵ was 1, and decreasing 0.05 each time resulted in the number of training trials simply becoming [(1-0.05)/0.05] 19. The first trial at $\epsilon = 1$, resulted in a total of 20 trials. This number of trials made sense given the decremented value of 0.05
3. Yes, the decaying function was properly represented. The constant ϵ is a straight line with a negative slope of 0.05. The α is constant at 0.5 as represented in the graph.
4. As the number of training trials increased the number of bad action actually decreased, and the average reward increased.More trials could result in higher number of rewards.
5. The 'no-learning' ratings were both Fs. The default-learning driving agent had a reliability rating of an “A” which is much better than the 'no-learning' simulation. The Safty Rating is a 'D' , which means that the driving agent had some accidents and had some driving violations.

**

Improve the Q-Learning Driving Agent

The third step to creating an optimized Q-Learning agent is to perform the optimization! Now that the Q-Learning algorithm is implemented and the driving agent is successfully learning, it's necessary to tune settings and adjust learning paramaters so the driving agent learns both **safety** and **efficiency**. Typically this step will require a lot of trial and error, as some settings will invariably make the learning worse. One thing to keep in mind is the act of learning itself and the time that this takes: In theory, we could allow the agent to learn for an incredibly long amount of time; however, another goal of Q-Learning is to *transition from experimenting with unlearned behavior to acting on learned behavior*. For example, always allowing the agent to perform a random action during training (if $\epsilon = 1$ and never decays) will certainly make it *learn*, but never let it *act*. When improving on your Q-Learning implementation, consider the implications it creates and whether it is logistically sensible to make a particular adjustment.

Improved Q-Learning Simulation Results

To obtain results from the initial Q-Learning implementation, you will need to adjust the following flags and setup:

- `'enforce_deadline'` - Set this to `True` to force the driving agent to capture whether it reaches the destination in time.
- `'update_delay'` - Set this to a small value (such as `0.01`) to reduce the time between steps in each trial.
- `'log_metrics'` - Set this to `True` to log the simluation results as a `.csv` file and the Q-table as a `.txt` file in `/logs/`.
- `'learning'` - Set this to `'True'` to tell the driving agent to use your Q-Learning implementation.
- `'optimized'` - Set this to `'True'` to tell the driving agent you are performing an optimized version of the Q-Learning implementation.

Additional flags that can be adjusted as part of optimizing the Q-Learning agent:

- `'n_test'` - Set this to some positive number (previously 10) to perform that many testing trials.
- `'alpha'` - Set this to a real number between 0 - 1 to adjust the learning rate of the Q-Learning algorithm.
- `'epsilon'` - Set this to a real number between 0 - 1 to adjust the starting exploration factor of the Q-Learning algorithm.
- `'tolerance'` - set this to some small value larger than 0 (default was 0.05) to set the epsilon threshold for testing.

Furthermore, use a decaying function of your choice for ϵ (the exploration factor). Note that whichever function you use, it **must decay to `'tolerance'` at a reasonable rate**. The Q-Learning agent will not begin testing until this occurs. Some example decaying functions (for t, the number of trials):

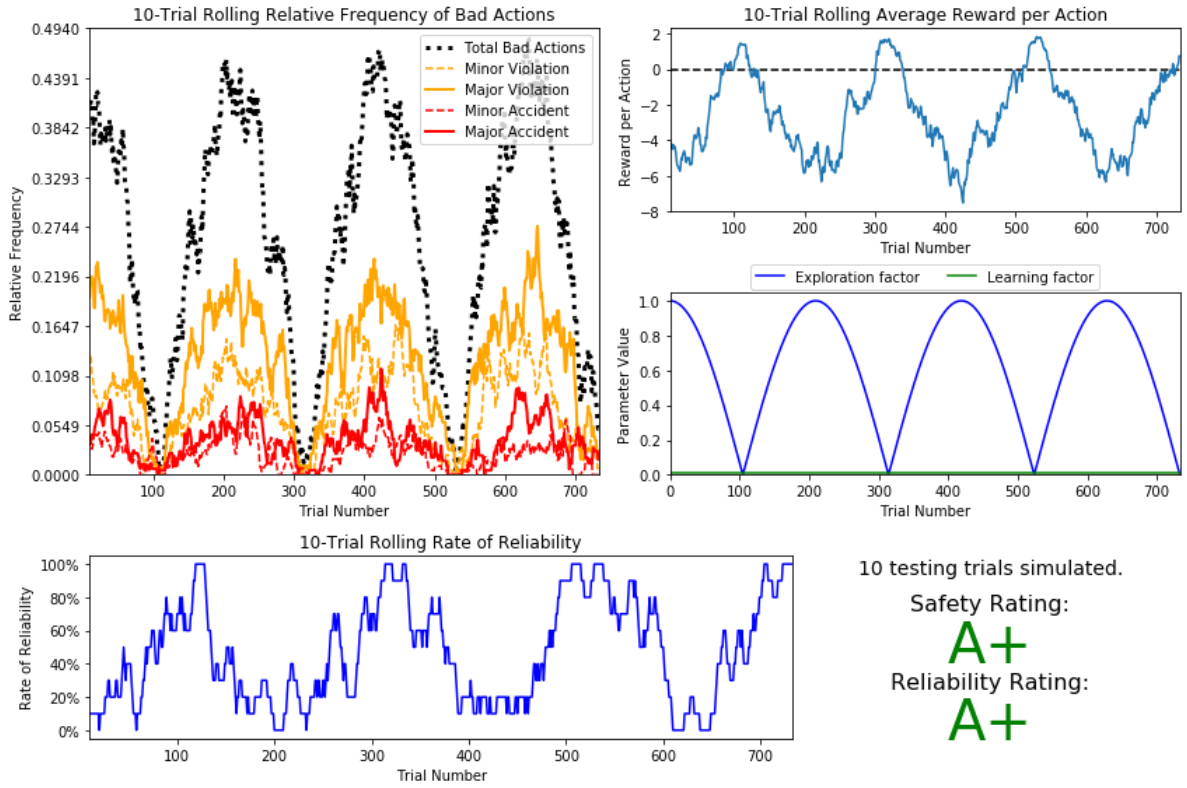
$$\epsilon = a^t, \text{for } 0 < a < 1 \qquad \epsilon = \frac{1}{t^2} \qquad \epsilon = e^{-at}, \text{for } 0 < a < 1 \qquad \epsilon = \cos(at), \text{for } 0 < a < 1$$

You may also use a decaying function for α (the learning rate) if you so choose, however this is typically less common. If you do so, be sure that it adheres to the inequality $0 \leq \alpha \leq 1$.

If you have difficulty getting your implementation to work, try setting the `'verbose'` flag to `True` to help debug. Flags that have been set here should be returned to their default setting when debugging. It is important that you understand what each flag does and how it affects the simulation!

Once you have successfully completed the improved Q-Learning simulation, run the code cell below to visualize the results. Note that log files are overwritten when identical simulations are run, so be careful with what log file is being loaded!

In [4]:
```
# Load the 'sim_improved-learning' file from the improved Q-Learning simulation
vs.plot_trials('sim_improved-learning.csv')
```



Question 7

Using the visualization above that was produced from your improved Q-Learning simulation, provide a final analysis and make observations about the improved driving agent like in **Question 6**. Questions you should answer:

- *What decaying function was used for epsilon (the exploration factor)?*
- *Approximately how many training trials were needed for your agent before begining testing?*
- *What epsilon-tolerance and alpha (learning rate) did you use? Why did you use them?*
- *How much improvement was made with this Q-Learner when compared to the default Q-Learner from the previous section?*
- *Would you say that the Q-Learner results show that your driving agent successfully learned an appropriate policy?*
- *Are you satisfied with the safety and reliability ratings of the* Smartcab*?*

**Answer:

1. I tried different decaying functions for epsilon before settling with ABS(COS(at)).
2. The number of training trials needed before testing was 733 using exploration factor ; ABS(COS(at)).
3. I settled on epsilon-tolerance of 0.001 and alpha of 0.015. I tried other variations with the decaying function ABS(COS(at)) however while I was able to get safety rating of A+ the highest rating I got for reliability was an A. I adjusted the values for alpha and epsilon tolerance whilst also trying different decaying functions this to get the best performance.-For ϵ, we need a large number of trials to make sure that the majority of the states has its Q-values filled and thus decrease the probability of facing an illegal new state. The exponential decay starts with a high epsilon which allows exploring new states and not getting stuck in a local minimum. The epsilon then decreases exponentially to minimize the random bahvior and thus allow for learning and correctly updating Q-table.-For the α, at the first time, the Q-table has not been learned well, so the new results that the driving agent is learning will be worthwhile, therefore I assign more weight to it.
4. This Q-Learner had bad actions at very at the begining of the trials but falls precipitously as the trial proceed, as expected the rewards behavior followed suit. The rewards quicky increased from approximatley -4.8 to close to 2 as the trials proceed. The default Q-learner has “D” Safety Rating and “A” Reliability Rating. Now the optimized Q-learner has “A+” rating for both Safty and Reliability.

5.With the safety and Reliability ratings both at an A+ rating we could say the tuning employed of the various factors may have cause the driving agent to learn an appropriate policy. For a real world setting however much more testing would be necessary. The Q-table still could be improve to for better optimization.

1. I am satisfied with both safety and Reliabilty ratings of the Smartcab as the final results are both A+. **

Define an Optimal Policy

Sometimes, the answer to the important question *"what am I trying to get my agent to learn?"* only has a theoretical answer and cannot be concretely described. Here, however, you can concretely define what it is the agent is trying to learn, and that is the U.S. right-of-way traffic laws. Since these laws are known information, you can further define, for each state the *Smartcab* is occupying, the optimal action for the driving agent based on these laws. In that case, we call the set of optimal state-action pairs an **optimal policy**. Hence, unlike some theoretical answers, it is clear whether the agent is acting "incorrectly" not only by the reward (penalty) it receives, but also by pure observation. If the agent drives through a red light, we both see it receive a negative reward but also know that it is not the correct behavior. This can be used to your advantage for verifying whether the **policy** your driving agent has learned is the correct one, or if it is a **suboptimal policy**.

Question 8

1. Please summarize what the optimal policy is for the smartcab in the given environment. What would be the best set of instructions possible given what we know about the environment? *You can explain with words or a table, but you should thoroughly discuss the optimal policy.*
2. Next, investigate the `'sim_improved-learning.txt'` text file to see the results of your improved Q-Learning algorithm. *For each state that has been recorded from the simulation, is the **policy** (the action with the highest value) correct for the given state? Are there any states where the policy is different than what would be expected from an optimal policy?*
3. Provide a few examples from your recorded Q-table which demonstrate that your smartcab learned the optimal policy. Explain why these entries demonstrate the optimal policy.
4. Try to find at least one entry where the smartcab did *not* learn the optimal policy. Discuss why your cab may have not learned the correct policy for the given state.

Be sure to document your `state` dictionary below, it should be easy for the reader to understand what each state represents.

**Answer:

1. A summary of an optimal policy : I will select Waypoint to frame my explanation.

   Waypoint intended is forward:

   ```
                    -> Light is green , go forward

                    -> Wait at light if other than green.
   ```

   Waypoint intended is right:

   ```
                    -> If oncoming vehicle is from the left, driving agent should
       wait

                    -> Otherwise, okay to go right.
   ```

   Waypoint intended is Left:

   ```
                    -> If light is green, and oncoming traffic is moving left or
       no oncoming traffic, go left.
                    -> If light is red, wait.
                    -> Otherwise go forward.
   ```

   2 and 3. Investigating the sim_improved-learning.txt revealed that the policy (the action with the highest value) was correct for the given state in most instances. Here are some examples:

Entry 137 from the latest simulation

left_red_left_left -- forward : -1.79 -- right : 0.11 -- None : 0.79 -- left : -1.56

In the entry 137 we see where the reward for doing the correct thing None(stay put at the red light since intention is to go left) is 0.79 and the reward for going forward on the red light is -1.79 . The Merit for the doing the right thing and the demerit for doing the wrong thing was correct. This is an example of the Optimal policy(action with the highest value).

Another example:

Entry 197

right_green_left_left -- forward : 0.27 -- right : 0.21 -- None : -0.41 -- left : 0.07

In this example the smartcab approaches a green light from the right, here the reward for stopping at the green is -0.41 while doing any of the three other legal functions are rewarded proportionately with merit given the hierarchy of the correct action.

From these examples presented I would say they demonstrate that the smartcab learned the Optimal policy from the number of simulations that were conducted.

1. Upon meticulous examining of the sim_improved-learning.txt Here are two examples:

Entry 275 in latest simulation where the light is green and so turning left should not carry such a high penalty (-1.75), the early trials did have some bad actions so the efficiency is still not 100%. Not withstanding the effort it took to find negative result might be proof that the smartcab is learning the optimal policy as we increased the number of trials.

forward_green_left_right -- forward : 0.37 -- right : 0.02 -- None : -0.44 -- left : -1.75

In the example below "right" corresponds to the action with the highest q value 0.14 over "forward" which only is 0.06; this is incorrect as the forward in this instance should have the biggest reward.

forward_green_forward_right -- forward : 0.06 -- right : 0.14 -- None : -0.34 -- left : -1.47

Not withstanding these instances prove we may need more trials to have the Optimal population for the Q-table since there are possible 96 states.

**

Optional: Future Rewards - Discount Factor, `'gamma'`

Curiously, as part of the Q-Learning algorithm, you were asked to **not** use the discount factor, `'gamma'` in the implementation. Including future rewards in the algorithm is used to aid in propagating positive rewards backwards from a future state to the current state. Essentially, if the driving agent is given the option to make several actions to arrive at different states, including future rewards will bias the agent towards states that could provide even more rewards. An example of this would be the driving agent moving towards a goal: With all actions and rewards equal, moving towards the goal would theoretically yield better rewards if there is an additional reward for reaching the goal. However, even though in this project, the driving agent is trying to reach a destination in the allotted time, including future rewards will not benefit the agent. In fact, if the agent were given many trials to learn, it could negatively affect Q-values!

Optional Question 9

There are two characteristics about the project that invalidate the use of future rewards in the Q-Learning algorithm. One characteristic has to do with the Smartcab *itself, and the other has to do with the environment. Can you figure out what they are and why future rewards won't work for this project?*

**Answer:

1. For the smartcab it is not able to differentiate and make a comparison between two paths to a given destination; its objective is to obey traffic rules and get to the destination. The smartcab will stick to its inputs and focus on the immediate reward; hence future rewards will not improve the process.

1. In the case for the environment; each new goal is randomly selected after the previous training and testing period. Keeping the results of previous trials to create an influence on subsequent simulations may result in having too many trials which could negatively affect the Q-values. Also, during the attempt to get to each destination on time the state at any given time is largely independent of the previous traffic signal except for waypoint in the event of turning right at a red light if there is no on-coming traffic. In conclusion; relating to the real world making a good decision at all the traffic signal to a destination then making one bad decision at the final can be fatal, which I think translate to having each reward for each state independent in the smartcab learning in these simulations.

**

> **Note**: Once you have completed all of the code implementations and successfully answered each question above, you may finalize your work by exporting the iPython Notebook as an HTML document. You can do this by using the menu above and navigating to
> **File -> Download as -> HTML (.html)**. Include the finished document along with this notebook as your submission.