Q3-(a) $V(\pi) = \sum_a \pi(a|s) \sum_{s',r} P(s', r | s, a)\left[r + \gamma V_\pi(s')\right]$

For A: $s' \in \{B, t\}$. ($t \to$ terminal) $V_\pi(A) = \pi(\to, A) \sum_{s',r} P(s', r | A, \to)\left[r + \gamma V_\pi(s')\right]$

$+ \pi(\leftarrow, A) \sum P(s', r | A, \leftarrow)\left[r + \gamma V_\pi(s')\right]$

$\to V_{\pi_A} = 0.9 \times \left[P(B, r | A, \to)(r + \gamma V_\pi(B)) + \cancel{P(t, r | A, \to)(r + \gamma V_\pi(t))}^{0}\right]$

$+0.1\left[\cancel{P(B, r | A, \leftarrow)(r + \gamma V_\pi(B))}^{0} + P(t, r | A, \leftarrow)\cancel{(r + \gamma V_\pi(t))}^{0}\right] = 0.9 \times 1 \times (0 + \gamma V_\pi(B))$

$+0.1 \times 1 \times (100) = 0.9\gamma V_\pi(B) + 10$

$= 0.9 V_\pi(B) + 10$ (1)

For B: $s' \in \{A, C\}$ $V_\pi(B) = \pi(\to, B) \sum_{s',r} P(s', r | B, \to)\left[r + \gamma V_\pi(s')\right]$

$+\pi(\leftarrow, B) \sum_{s',r} P(s', r | B, \leftarrow)\left[r + \gamma V_\pi(s')\right]$

$\to V_{\pi_B} = 0.9\left[P(C, r | B, \to)(r + \gamma V_\pi(C)) + \cancel{P(A, r | B, \to)(r + \gamma V_\pi(A))}^{0}\right]$

$+0.1\left[P(C, r | \cancel{B, \leftarrow})^{0}(r + \gamma V_\pi(C)) + P(A, r | B, \leftarrow)(r + \gamma V_\pi(A))\right] = 0.9 \times 1 \times (0 + \gamma V_\pi(C))$

$+0.1 \times 1 \times (0 + \gamma V_\pi(A))$

$= 0.9 V_\pi(C) + 0.1 V_\pi(A)$ (2)

For C: $s' \in \{B, t\} \Rightarrow V_\pi(C) = \pi(\to, C) \sum_{s',r} P(s', r | C, \to)\left[r + \gamma V_\pi(s')\right]$

$+ \pi(\leftarrow, C) \sum_{s',r} P(s', r | C, \leftarrow)\left[r + \gamma V_\pi(s')\right]$

$\to V_\pi = 0.9\left[P(t, r | C, \to)(r + \gamma V_\pi(t)) + \cancel{P(B, r | C, \to)}^{0}(r + \gamma V_\pi(B))\right]$

$+0.1\left[\cancel{P(t, r | C, \leftarrow)}^{0}(r + \gamma V_\pi(t)) + P(B, r | C, \leftarrow)(r + \gamma V_\pi(B))\right] = 0.9 \times 1 \times (-1 + \gamma \times 0) + 0.1 \times 1 \times (0 + \gamma V_\pi(B))$

$= -0.9 + 0.1 V_\pi(B)$ (3)

(1), (2), (3) $\Rightarrow V_\pi(A) \simeq 10.2085$ $V_\pi(B) \simeq 0.2317$ $V_\pi(C) \simeq -0.8768$

(b) For A: $\max Q(s, a) = 2$ (for left) ε-greedy left : $\frac{1}{4} \times 0.25 + 0.75 = 0.8125$

up : $\frac{1}{4} \times 0.25 = 0.0625$

down : $\frac{1}{4} \times 0.25 = 0.0625$

right : $\frac{1}{4} \times 0.25 = 0.0625$

For B : $\max Q(s, A) = 1.2$ (for right)

ε-greedy left : $\frac{1}{4} \times 0.25 = 0.0625$

up : $\frac{1}{4} \times 0.25 = 0.0625$

down : $\frac{1}{4} \times 0.25 = 0.0625$

right : $\frac{1}{4} \times 0.25 + 0.75 = 0.8125$

$\Rightarrow$

state	up	down	left	right
A	0.0625	0.0625	0.8125	0.0625
B	0.0625	0.0625	0.0625	0.8125

d) $Q(A, up) \leftarrow Q(A, up) + \alpha\left[R_{t+1} + \gamma\left(\sum_a \pi(a|B)\, Q(B, a)\right) - Q(A, up)\right]$

we have: $Q(A, up) = -1.2$, $R_{t+1} = -1$, $\gamma = 0.99$, $\alpha = 0.1$ $\quad a \in \{up, down, left, right\}$

$$\rightarrow Q(A, up) \leftarrow -1.2 + 0.1\left[-1 + 0.99\left(0.0625 \times (-0.2 + 0.1 - 1.1) + 0.8125 \times (1.2)\right) + 1.2\right] = -1.0909$$

Q5: (a) A particular part of a city might have had a bad reputation regarding crime, because of poverty or a culture. But even if these circumstances change over the years, and the next generation became good people, the historical data of their home location predicts a higher chance of being criminal for those people. It somehow even relates to the bias towards black people.

(b) The algorithm relies on input data from human. Some feature that the algorithm learns, are biased, since the dataset was imbalanced. For example, the network might not have observed enough input data from minorities. Or in some cases, the algorithm itself might not do well on some races and skin colors.