Austin Dong
CS325 hw5

1.

a. For deterministic, an action leads to a new state. For stochastic, an action leads to a probability distribution of next states.
b. Discount factor is used to decrease reward, which can depend on current state and uncertainty. Using it avoids infinite reward in cyclic processes.
c. Optimal policy is the policy that maximizes sum of rewards.
d. Movement cost is another method of decaying the reward value.
e. V(*s*) is value iteration; Q(*s*, *a*) is Q-learning – an improvement to value iteration. For both, *s* is state and *a* is action. Q(s, a) takes action into account while V(s) does not.

2. $\gamma = 0.9$; R = -0.1

a.

	1	2	3	4	5
a	0.063	-1		0.8	+1
b	0.181	0.312	0.458	0.62	0.8
c	0.06	0.18		-1	+1

Optimal policy:

	1	2	3	4	5
a	↓	-1		→	+1
b	→	→	→	↑	↑
c	↑	↑		-1	+1

b.

	1	2	3	4	5
a	-0.165	-1		0.35	+1
b	-0.146	-0.101	-0.003	0.215	0.35
c	-0.165	-0.146		-1	+1

Optimal policy:

	1	2	3	4	5
a	↓	-1		→	+1
b	→	→	→	↑	←
c	→	↑		-1	+1

b5 after convergence: 0.35

c.

	1	2	3	4	5
a		-1			+1
b	S				
c				-1	+1

3. a. I set up my learning algorithm so that for different mindsets, different types of blocks would have different values. e.g. in the explore mindset, unexplored blocks are most attractive; in the gather mindset, blocks with balls are most attractive; and in the deposit mindset, the block with the deposit bin is most attractive. My agent explores before gathering balls.