Exercise 1

Task 1 - Multi-armed Bandits

a) The probability that the greedy action is selected is $1 - \epsilon = 1 - 0.5 = 0.5$

b) A random action definitely occurred at time step 2 and 5, and could possibly have occurred at time step 1 and 3.
Calculate $Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$ for each action and time step → if the action with the largest value was taken, a greedy action occurred, otherwise it would be random.

			$\underset{a}{\arg\max} Q_t(a)$	A_t	R_t
$Q_1(1) = 0$	$Q_1(2) = 0$	$Q_1(3) = 0$	1, 2, 3	1	1
$Q_2(1) = 1$	$Q_2(2) = 0$	$Q_2(3) = 0$	1	2	1
$Q_3(1) = 1$	$Q_3(2) = 1$	$Q_3(3) = 0$	1, 2	2	2
$Q_4(1) = 1$	$Q_4(2) = 3$	$Q_4(3) = 0$	2	2	2
$Q_5(1) = 1$	$Q_5(2) = 5$	$Q_5(3) = 0$	2	3	0

Task 2 - Action Selection Strategies

c)

epsilon greedy search performs better, as it also includes exploration and not only exploitation.

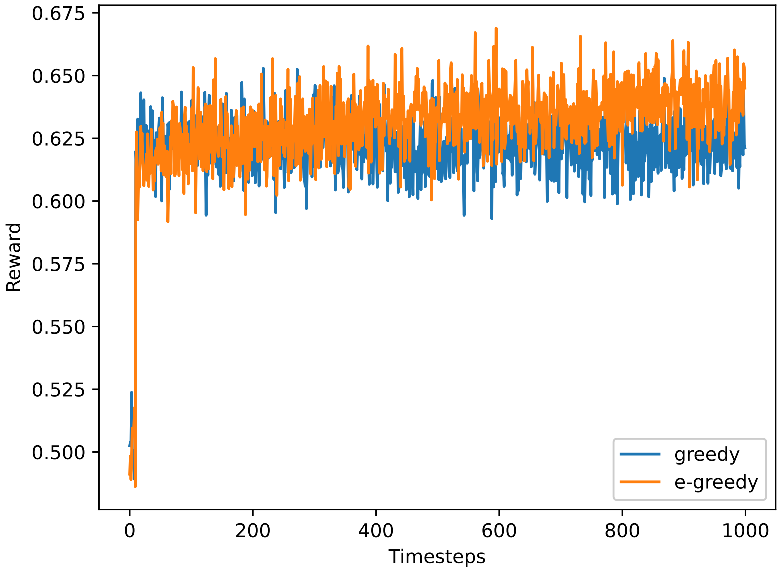

Figure 1: Output bandit_strategies.eps

d)

decay epsilon over time.