Exercise

In groups of UP TO FOUR:

1. Implement a MAB:
   - Let each arm give rewards from a Gaussian of variance 1, and means drawn from a Gaussian of mean 0, variance 3 when they are created.
   - You should be able to "pull" an arm (select an action) and receive a random reward.
2. Implement the ϵ-greedy, greedy with optimistic initialisation, and UCB algorithms.
3. Run the three algorithms with different parameter settings on a 10-arm bandit.

By next week's lecture, submit on Moodle:

1. A plot of reward over time (averaged over 100 runs each) on the same axes, for ϵ-greedy with $\epsilon = 0.1$, greedy with $Q_1 = 5$, and UCB with $c = 2$
2. A summary comparison plot of rewards over first 1000 steps for the three algorithms with different values of the hyperparameters
3. Your code