WFDR for a1true-a2false-b1false-c1false, reward function 1 0.8 WFDR for 400 experiments 0.6 0.2 ann mlfq bandit-eps bandit-opt bayes-high bayes-low qlearning random tableau