Theoretical Neuroscience II
Teacher: Prof. Jochen Braun
Assistant Teacher: Ehsan Kakaei
Name: Viviana Villafañe Barraza & Tasnuba Tazeen & Lee Po Shing
Date: 24th May, 2018

Exercise 6: Policy Learning

Introduction

When considering the learning process under more biological setting, one may probably think of foraging behaviour in animals. Imagining a bee which harvests nectar from two different populations of flowers, for instance the blue and the yellow ones, they have to learn that in which seasons the yield is higher in which flowers. Therefore they have to learn that at a point, they have to change their visiting habits from one group of flowers to another one. A key factor in achieving this behaviour relies on a hypothetical action value, which is updated when the decision of visiting is made and correspondingly affecting the probabilities of visiting the flowers. This circular relationship between action values, probabilities and rewards formed that basis of the model for policy learning. In this model, two other constants can be introduced to alter the variability of the decision and also the speed of learning.

In this assignment, the two different types of actor, the indirect one and the direct one, were modelled and compared to study how can be more successful in achieving a better policy learning. In addition, the variability of the decision and the speed of learning were also varied in order to show their own effects.

Method

In this exercise, a model imitating a foraging bee was considered to study the policy learning with indirect and direct actor schemes. In brief, the bee had complementary probabilities in visiting two different populations of flowers, and they could only choose either one of them and received an amount of reward correspondingly. These would update the action values of the two populations differentially with different learning schemes and consequently the related probabilities. Since the mean rewards for the two populations would change after half of the session, the bee needed to learn to change its policy (i.e. the tendency of visiting a particular population) in order to gain more rewards. The probabilities of it visiting blue flowers and yellow flowers (P_b and P_y, P_b and P_y in the code) were based on the action value of the two populations (m_b and m_y, m_b and m_y in the code) and also a constant β which determines the degree of variability as the following equations:

$$P_b = \frac{1}{1 + e^{\beta(m_y - m_b)}} \qquad P_y = \frac{1}{1 + e^{\beta(m_b - m_y)}}$$

In the beginning of the programme, m_b and m_y were both 0 and progressively changed based on the two different policy learning schemes described as below; and β as 1. The mean reward in the first half of each scheme for blue and yellow flowers were 1 and 3, and the later reversed. The variance for both flowers in each half of each scheme was kept as 1, and so as the standard deviation (sigma in the code). The reward sets with the statistics above were calculated as (r_b and r_y, r_b and r_y in the code) The number of trials in each half was set as 100 (trial in the code). While the sessions continued, the instantaneous rewards and the cumulative rewards for individual population were calculated and plotted together with the action values and probabilities.

Action values were adjusted in different ways for indirect and direct actor schemes. In the scheme of an indirect actor, the action value of the chosen population would change by the value of the reward received minus the current action value weighted by a constant ϵ = 0.1 (epsilon in the code); whereas that of the one not chosen would not change. In the case of a direct actor, the amount of change for the chosen population was the value of the reward received minus the average rewards received in the past 10 trials, weighted again by ϵ and also the probability of the abandoned option; whereas that of the other choice was reduced by the same value. The above can be formularised as followed:

Indirect actor	Direct actor
$\delta = r_i - m_j$	$\delta = r_i - \bar{r}$
$m_i \rightarrow m_i + \epsilon\delta$	$m_i \rightarrow m_i + \epsilon(1 - P_i)\delta$
$m_j \rightarrow m_j$	$m_j \rightarrow m_j - \epsilon P_j \delta$

In order to understand the performance of the schemes, the same was replied with β = 0.1 or 2 or ϵ = 0.01 or 0.9 to assess, in the above model, how probabilistic the decision made and how quickly a learning took place affect the time required the policy to be changed and thus the cumulative rewards received.

Results & Discussion

Indirect actor schemes allowed policy learning

With an indirect actor scheme, the effect of the degree of exploratory/exploitive behaviour, as seen by the variability of the probabilities as defined by β, and that of the speed of learning as defined by ϵ could be observed on the results obtained. The basic pattern with β = 1 and ϵ = 0.1 was shown in *Fig. 1* and those with β changed to 0.1 or 2 or ϵ to 0.01 or 0.9 were shown in *Fig. 2*. In the beginning of the session with basic conditions, the bee determined the yellow flowers would give a higher return of rewards, and therefore after a while, with an increase in the probability of visiting it brought by an increase of the action value, it visited them more often and gained rewards mostly from them, as seen in *Fig. 1(b)*. Consequently the cumulative rewards by the yellow flowers increased steadily. However, there was changes in the mean rewards for the two

populations at trial 100, and the bee realised soon after that and gradually switched to the blue flowers which gave the more rewards. However, it took roughly 50 trials for the policy to completely reversed, which was relatively slow (*Fig 1(a)*). Correspondingly, the difference between action values followed the same trend (*Fig. 1(c)*). Consequently, the odd of choosing blue flowers and gaining rewards from them dominated in the latter half of the session (*Fig. 1(b)*), causing the rise in the cumulative reward of blue flowers in that phase and simultaneously a plateau for the yellow flowers (*Fig. 1(d)*).

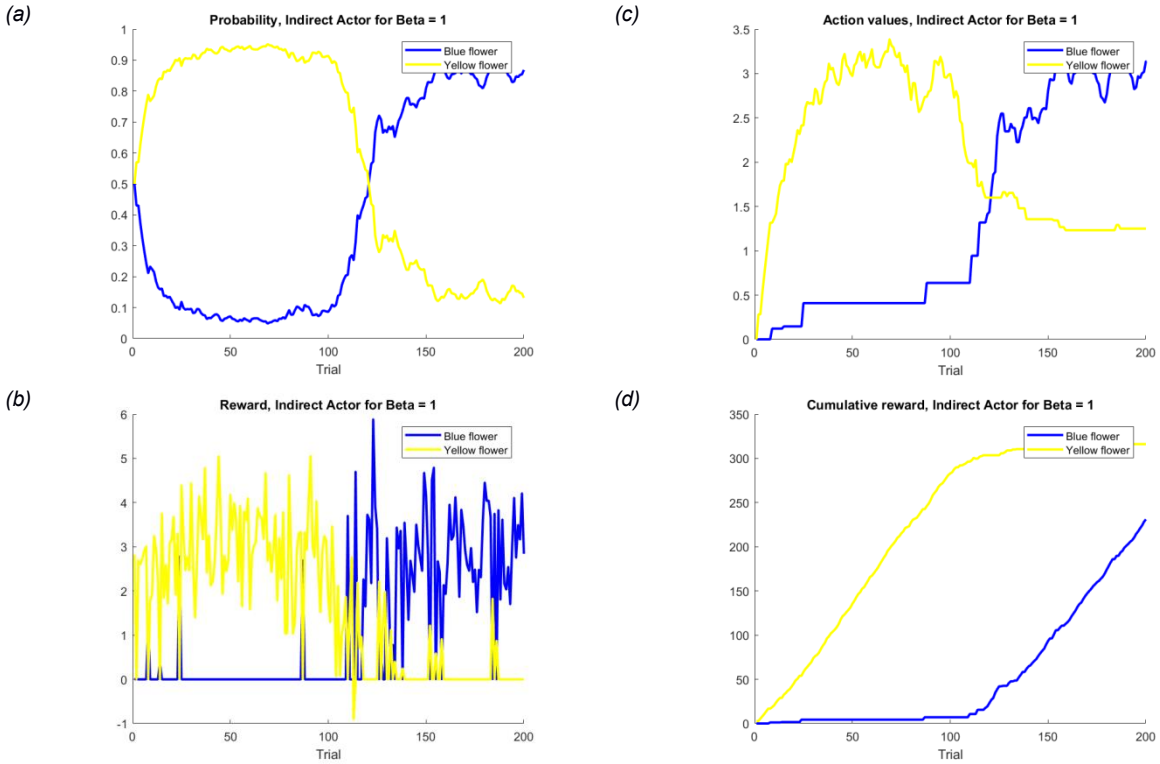


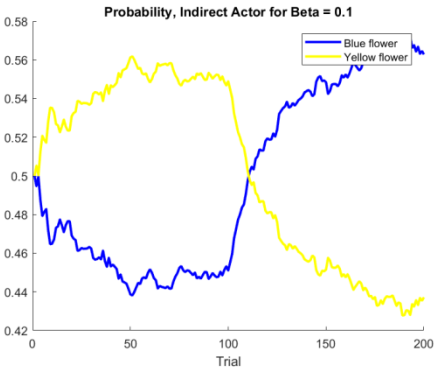
Fig. 1 Policy learning with an indirect actor scheme of β = 1 and ϵ = 0.1. With an indirect actor scheme, only the action value of the chosen option was updated, and thus its probability. With a moderate degree of variability and learning pace, the bee noticed the change in environment and progressively changed its policy *(a)* and action values *(b)*. With a change in policy, the time of visiting the blue flowers increased *(c)* and thus the cumulative reward from them also increased *(d)*.

Effect of β and ϵ on policy learning with indirect actor scheme

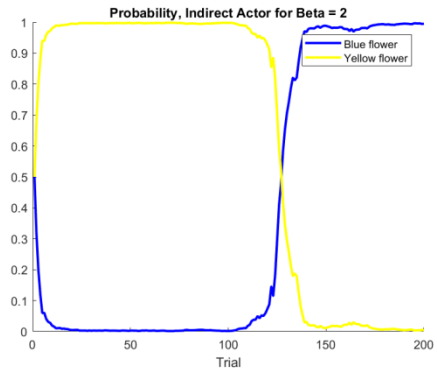
To study the effect of the degree of variability of choice and the speed of learning, the value of either *β* or *ϵ* was changed (*Fig. 2*). With a change in *β*, the smaller it was the higher variability in decision could be observed (*(a)* & *(b)*), i.e. a curve with more fluctuations as in *(a)*. Since the probabilities developed quicker with higher *β*, therefore the chance of the bee to choose the one with less probability, i.e. blue flowers in the first half phase, was less. This is the reason why the action value of them was almost zero as in *(f)*. In another case, with lower *β*, the bee was less deterministic and hence the learning process for the action values was observable in both populations *(e)*. Notably, with a higher *β*, the intercept of the two probabilities after half of the session shifted to a later trial, which means the time it took to realise the switch between the mean rewards was longer. This is due to the fact that the bee kept on exploiting the first choice once it found it with a higher reward, as seen in *(j)*. However, this doesn't imply a higher reward for the bee due to the probabilistic behaviour (*Table 1*). In this experiment, a higher *β* gave a higher cumulative reward; yet it is believed that when it reaches a certain threshold, this exploitative policy was not beneficial, as it lacks the flexibility to learn a change in the environment. Lastly, the intercept also found to correspond to the time when a change in the slopes in the cumulative rewards.

With a change in *ϵ*, the constant determining the learning rate, the performance of policy learning was also affected. A higher *ϵ* means the bee rapidly updates the action value according the reward it received. This is, however, not optimal when it is too high, as the bee will constantly adjust its behaviour after every decision. In *Fig. 2(c)* and *(d)*, a lower learning rate decreased the rate of adjustment on the behaviour. As a result, the policy did not change even at the end of the session. On the other hand, a high *ϵ* led to a strong fluctuation in the policy but in general a switch in policy corresponded to the change in environment could be observed. Similarly, the action values in the case of lower *ϵ* showed a smoother development; whereas in the case of higher *ϵ* strong oscillations with in a general switch in pattern were observed. For the rewards obtained, an increase in *ϵ* caused the bee to response better to the change in the environment and to choose the flowers with a higher reward more often (*(k)* and *(l)*). Combining all these effects, the cumulative reward was shown proportional to the learning rate (*Table 1*). However, it is recommended to use an intermediate learning rate that gives a less fluctuating behaviour.

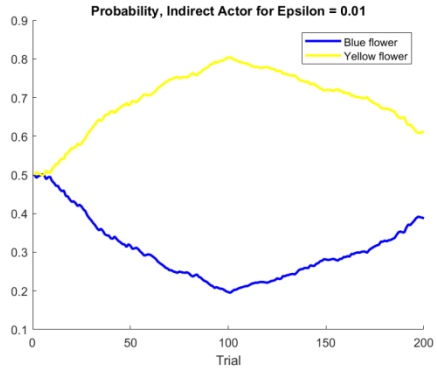
(a)


Probability, Indirect Actor for Beta = 0.1
Blue flower
Yellow flower
Trial

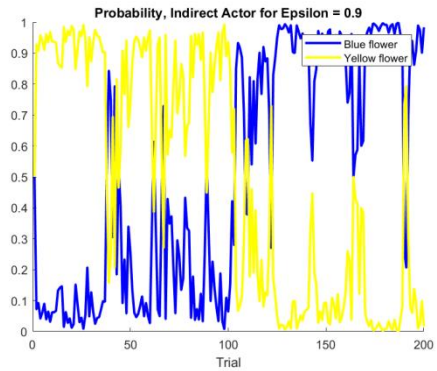
(b)


Probability, Indirect Actor for Beta = 2
Blue flower
Yellow flower
Trial

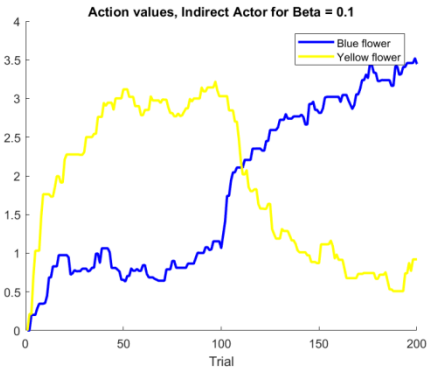
(c)


Probability, Indirect Actor for Epsilon = 0.01
Blue flower
Yellow flower
Trial

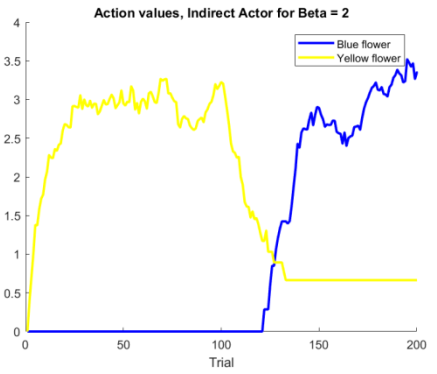
(d)


Probability, Indirect Actor for Epsilon = 0.9
Blue flower
Yellow flower
Trial

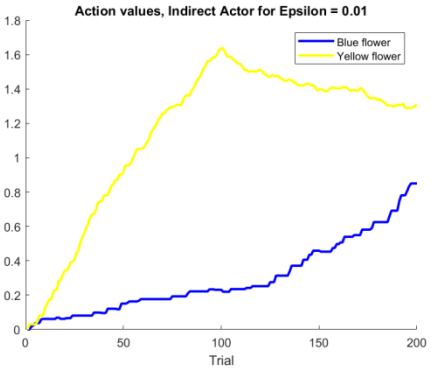
(e)


Action values, Indirect Actor for Beta = 0.1
Blue flower
Yellow flower
Trial

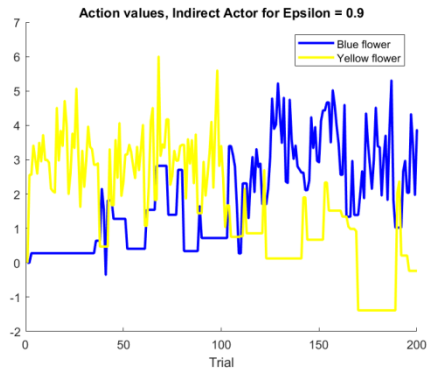
(f)


Action values, Indirect Actor for Beta = 2
Blue flower
Yellow flower
Trial

(g)


Action values, Indirect Actor for Epsilon = 0.01
Blue flower
Yellow flower
Trial

(h)


Action values, Indirect Actor for Epsilon = 0.9
Blue flower
Yellow flower
Trial

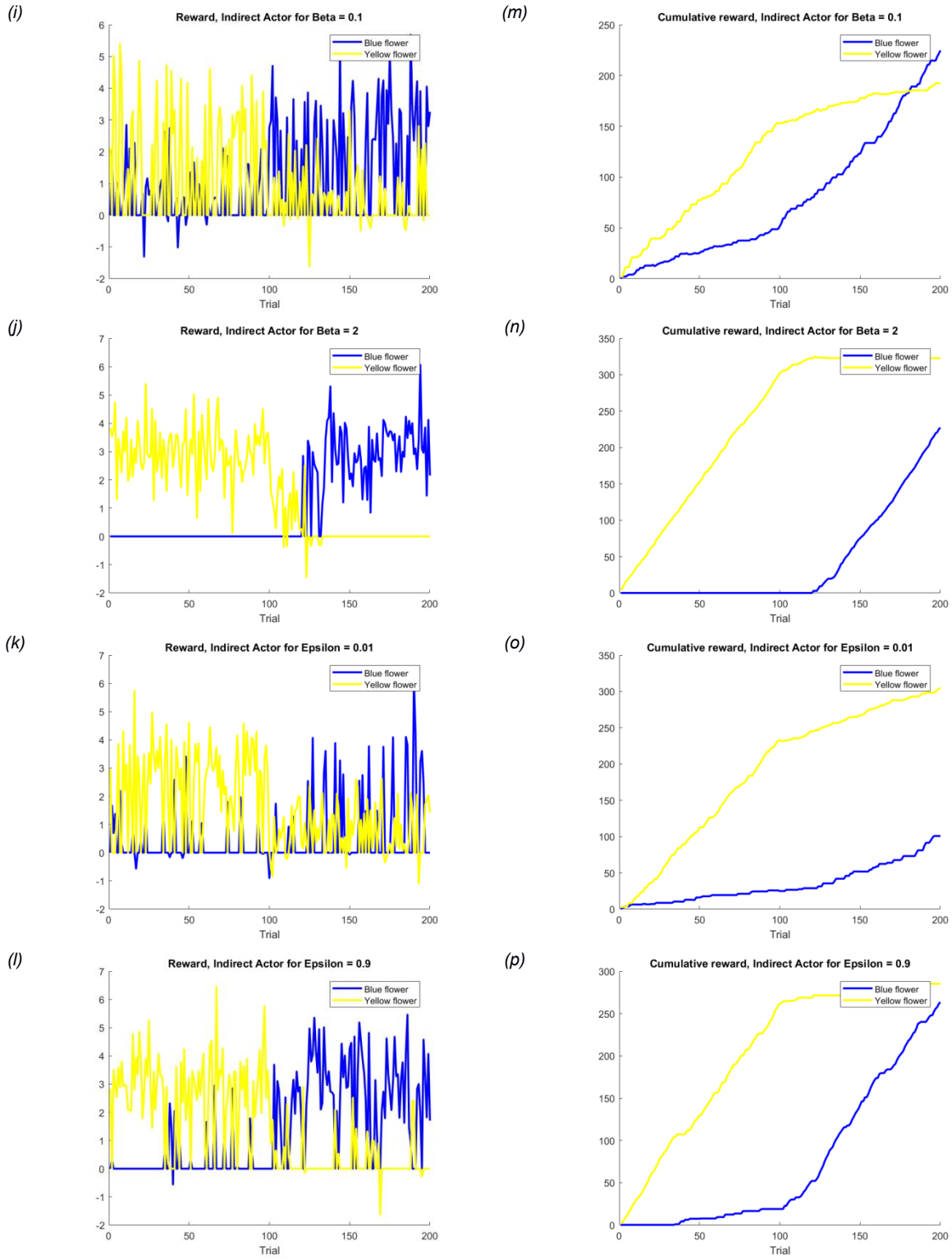


Fig. 2 Effect of β and ϵ on policy learning with indirect actor scheme. Using different value of *β* or *ϵ*, the same scheme was used to study their effects on different parameters, namely the probabilities *(a)-(d)*, the action values *(e)-(h)*, the instantaneous rewards *(i)-(l)* and the cumulative rewards *(m)-(p)* for the blue and yellow flowers (the blue and yellow lines). The values of variables changed were as shown in the title.

Direct actor schemes gained less than indirect actor in policy learning test

Another bee performed with direct actor rule which only learnt from the rewards gained from the previous ten trials was considered. The environment changed in the same manner. It is interesting to notice the drop in action value of the blue flowers in the first half of the session, and that in the second half of the session for the yellow flowers (*Fig. 3(c)*). This is due to the fact that a direct actor modulated both action values with an opposite manner. Correspondingly, a drop in the probability was observed for blue flowers when the difference between the two action values got larger (*(a)*). When comparing the two different schemes, there was never a policy learning (a reversal of probabilities) in the indirect actor. Therefore the total reward gained in the latter half for the blue flowers was relatively low, and thus the total amount gained (*Table 1*). In practical, this mode relied much on the knowledge or understanding of the world by simultaneously monitoring the action values of both populations.

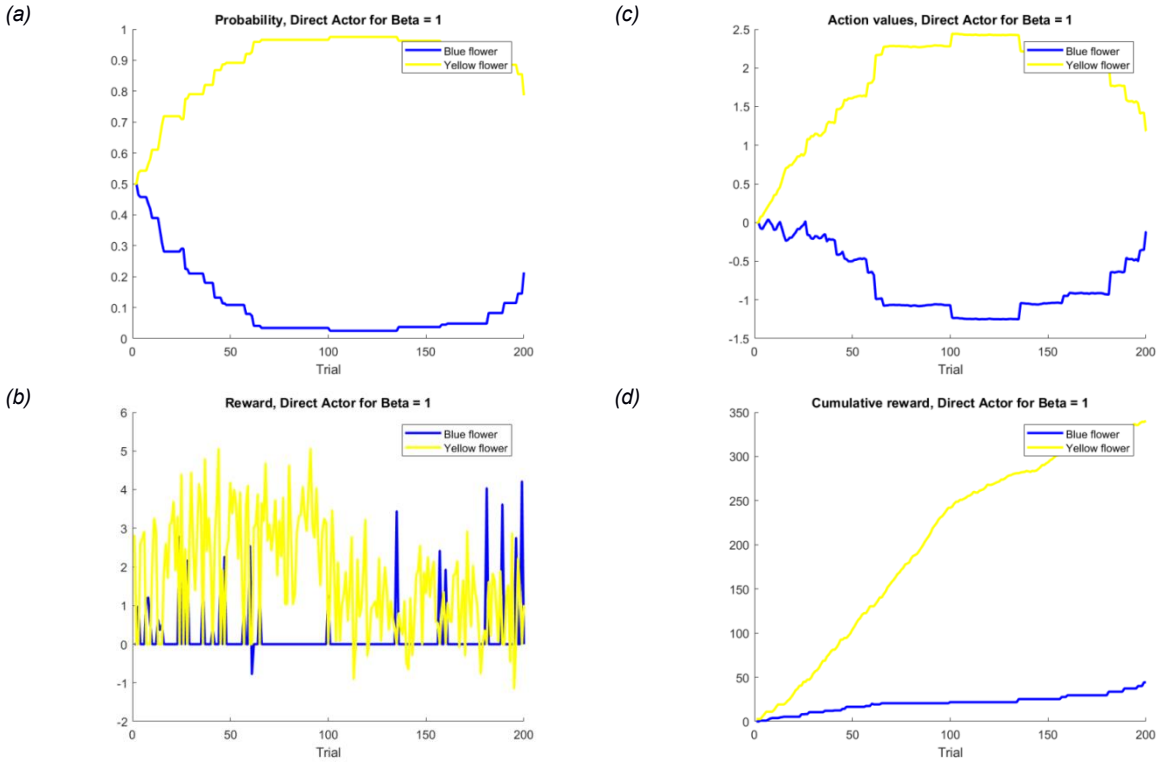


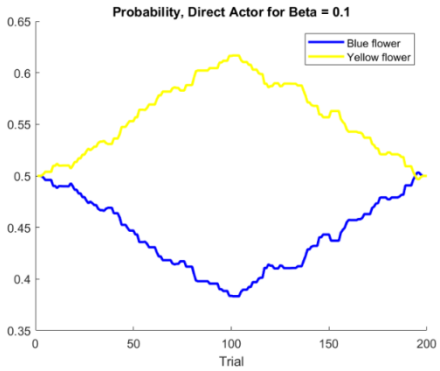
Fig. 3 Policy learning with a direct actor scheme of β = 1 and ϵ = 0.1. With a direct actor scheme, only the action value of the chosen option was updated, and thus its probability. With a moderate degree of variability and learning pace, the bee noticed the change in environment and progressively changed its policy *(a)* and action values *(b)*. With a change in policy, the time of visiting the blue flowers increased *(c)* and thus the cumulative reward from them also increased *(d)*.

Effect of β and ϵ on policy learning with indirect actor scheme

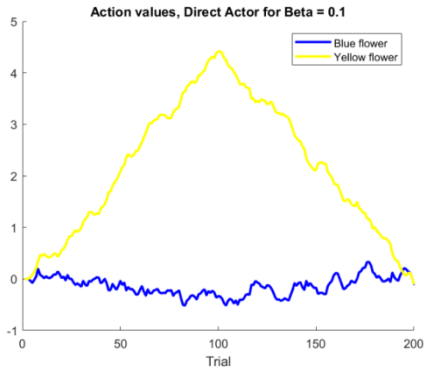
Similar effect of the degree of variability of choice and the speed of learning on the direct actor scheme could be seen as in the indirect actor one (*Fig. 4*). The development of probabilities was slower in the case of lower *β* (*(a)* & *(b)*). Interestingly, in both cases policy learning was successful. This may because in the case of exploitative behaviour with *β* = 0.1, the highest probability of the yellow flower did not develop into a high value, and thus the unlearning process was possible. In the case of the more exploratory behaviour, since the probabilities heavily dependent on the difference in the action values, therefore once a less probable choice was chosen, due to the fact that direct actor monitored both action values together in a contradictory manner, the difference greatly dropped and thus accelerated the development of the probabilities. As a result, the learning process sped up. In fact, the one with the highest *β* was the fastest one with successful policy learning and gained the most rewards in case of direct actor scheme (*Table 1*).

In case of the learning rate *ϵ*, a higher value means a greater response in the change of action values. As seen in *(h)*, the range of action values was much higher, meaning the probabilities would also become more dispersed *(d)*. As a result, the change in environment was not even noticed by the direct actor and thus kept exploiting the previous preferred option, i.e. the yellow flowers *(j)*. Thus it didn't even gain much rewards from the blue flowers as represented by the cumulative values *(p)*. Direct actor worked better with a smaller learning rate and a more explorative behaviour (*Table 1*).

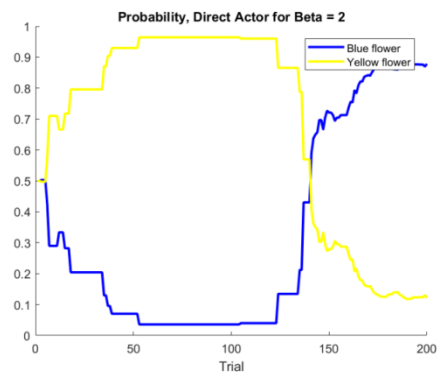
(a)


Probability, Direct Actor for Beta = 0.1
Blue flower
Yellow flower
Trial

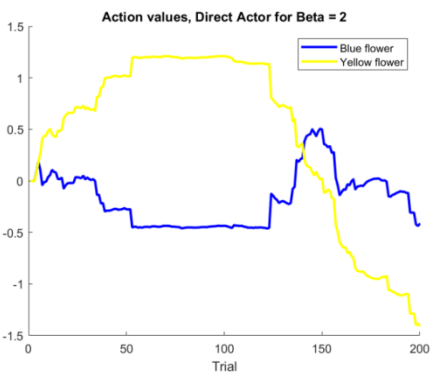
(e)


Action values, Direct Actor for Beta = 0.1
Blue flower
Yellow flower
Trial

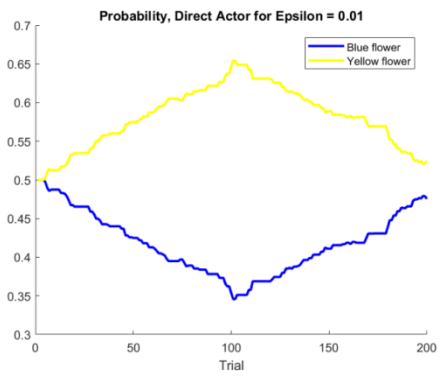
(b)


Probability, Direct Actor for Beta = 2
Blue flower
Yellow flower
Trial

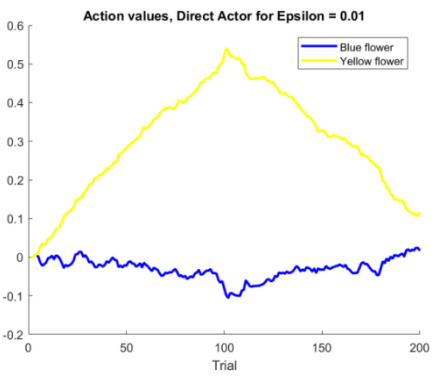
(f)


Action values, Direct Actor for Beta = 2
Blue flower
Yellow flower
Trial

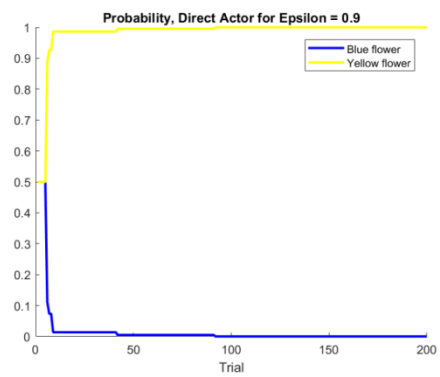
(c)


Probability, Direct Actor for Epsilon = 0.01
Blue flower
Yellow flower
Trial

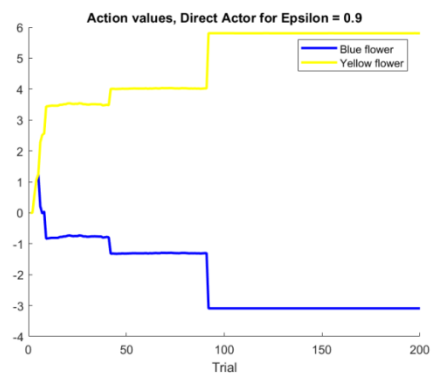
(g)


Action values, Direct Actor for Epsilon = 0.01
Blue flower
Yellow flower
Trial

(d)


Probability, Direct Actor for Epsilon = 0.9
Blue flower
Yellow flower
Trial

(h)


Action values, Direct Actor for Epsilon = 0.9
Blue flower
Yellow flower
Trial

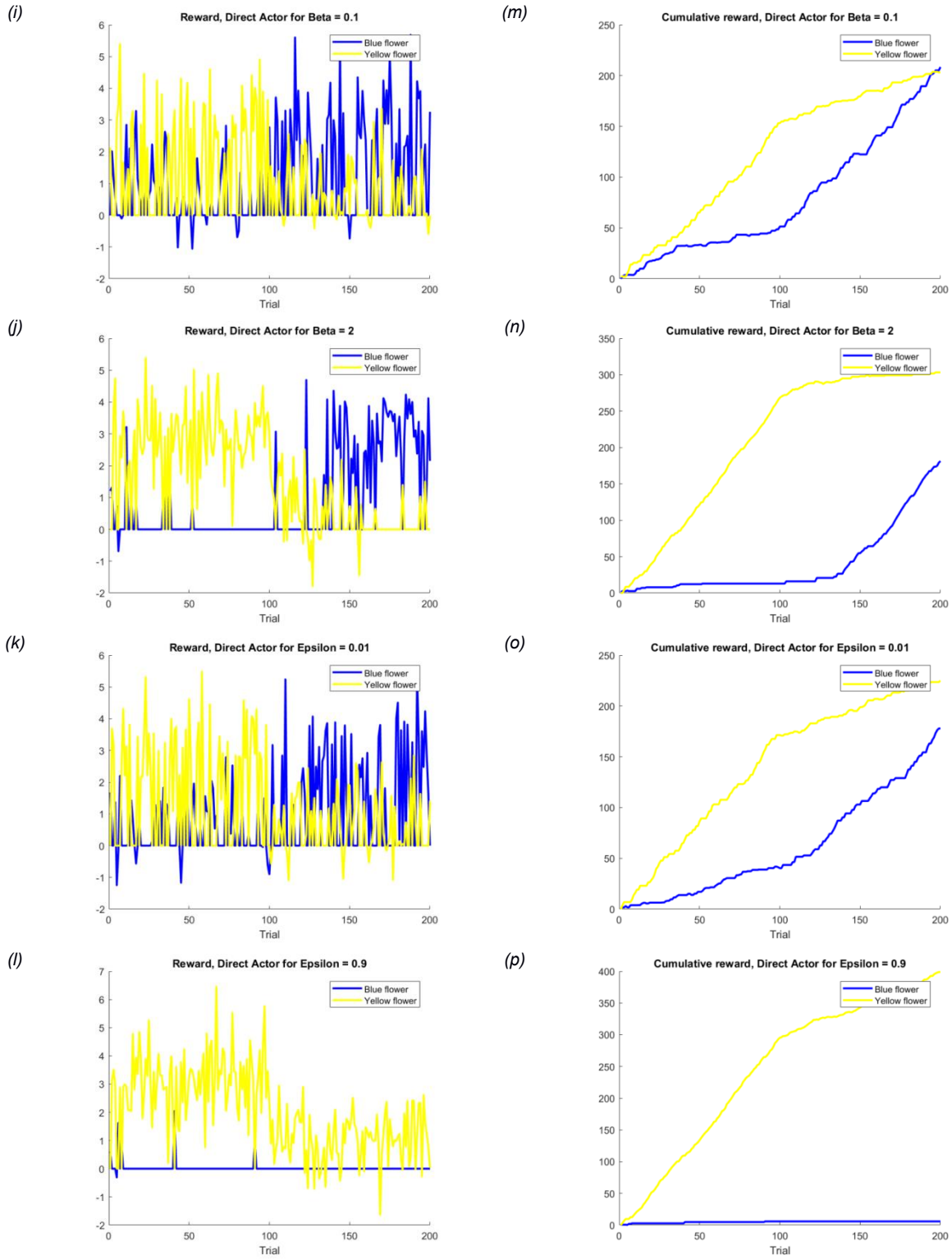


Fig. 4 Effect of β and ϵ on policy learning with direct actor scheme. Using different value of *β* or *ϵ*, the same scheme was used to study their effects on different parameters, namely the probabilities *(a)-(d)*, the action values *(e)-(h)*, the instantaneous rewards *(i)-(l)* and the cumulative rewards *(m)-(p)* for the blue and yellow flowers (the blue and yellow lines). The values of variables changed were as shown in the title.

Parameters		Indirect Actor		Total Indirect Actor	Direct Actor		Total Direct Actor
Epsilon(ϵ)	**Beta** (β)	**Blue flower**	**Yellow flower**		**Blue flower**	**Yellow flower**	
0,1	1	268	259	527	25	369	394
0,9	1	263	285	548	5	398	403
0,01	1	100	305	405	177	225	402
0,1	2	227	322	549	181	303	484
0,1	0,1	224	192	416	208	203	411

Table 1. Cumulative rewards depended on β and ϵ.