This report presents the performing the methods proposed in our study with different parameters: learning rate (μ) and discount factor (γ). We prepared this report to show the conclusions of the study will remain the same by changing the parameters of the temporal difference learning algorithm. Here we follow a grid search strategy by fixing the discount factor and searching over a set of learning rate. We confirm that a similar result can also be achieved by searching different adjustment factor.

Steps	Normalized reward	Standard deviation	Wrong actions	Correct actions	X	C	X-C-3-5	Others
200	0.347	16.3	4.5%	95.5%	5	2	1	2
300	0.362	13.2	4.5%	95.5%	3	1	4	2
400	0.409	15.5	1.5%	98.5%	2	–	7	1
500	0.481	12.8	1.5%	98.5%	2	5	3	–
600	0.444	25.7	1.0%	99.0%	1	3	6	–
1000	0.479	35.3	–	100%	4	1	5	–

Table 1: Simulated agent experiment results for all iteration with 10 repetitions. In these experiments, the μ and γ variables are set to 0.7 and 0.4.

Steps	Normalized reward	Standard deviation	Wrong actions	Correct actions	X	C	X-C-3-5	Others
200	0.323	16.1	20.5%	79.5%	2	2	5	1
300	0.425	21.8	18.5%	81.5%	3	1	6	–
400	0.401	12.7	19.5%	80.5%	4	–	6	–
500	0.447	23.3	17.0%	83.0%	5	2	3	–
600	0.466	29.7	17.5%	82.5%	6	1	3	–
1000	0.489	41.7	16.0%	84.0%	2	3	5	–

Table 2: Simulated agent experiment results for all iteration with 10 repetitions. In these experiments, the μ and γ variables are set to 0.5 and 0.4.

Steps	Normalized reward	Standard deviation	Wrong actions	Correct actions	X	C	X-C-3-5	Others
200	0.339	17.3	24.0%	76.0%	2	1	4	3
300	0.368	14.4	18.0%	82.0%	1	2	5	2
400	0.420	19.5	21.0%	79.0%	1	–	4	5
500	0.399	24.5	12.5%	87.5%	1	3	6	–
600	0.450	19.1	19.0%	81.0%	5	1	4	–
1000	0.485	26.9	14.0%	86.0%	–	2	5	3

Table 3: Simulated agent experiment results for all iteration with 10 repetitions. In these experiments, the μ and γ variables are set to 0.9 and 0.4.