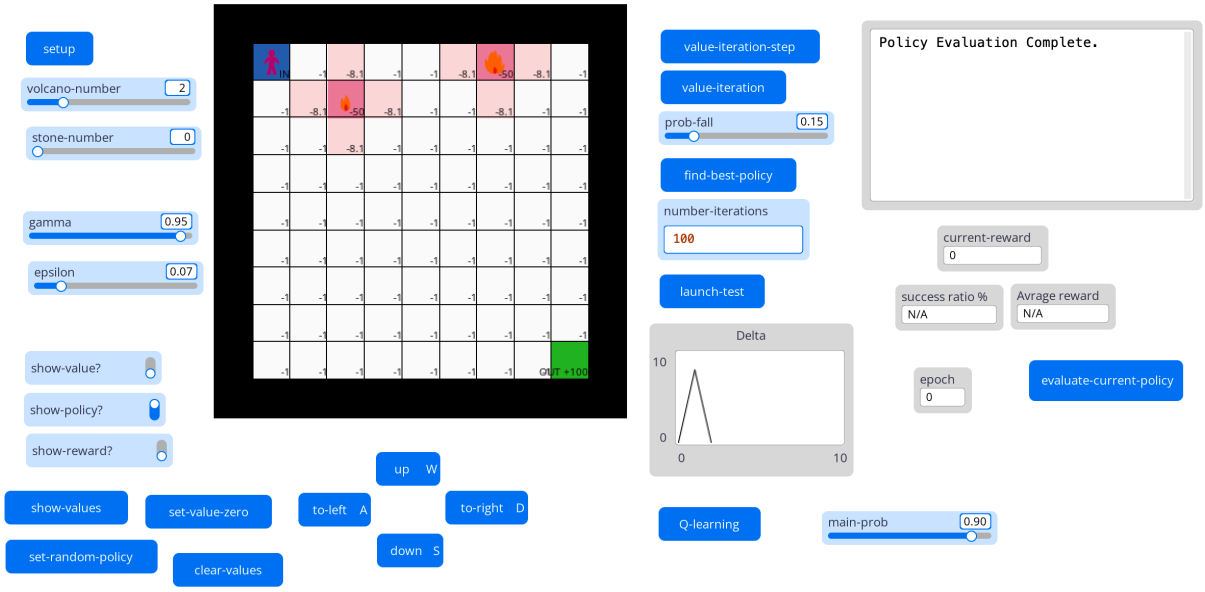
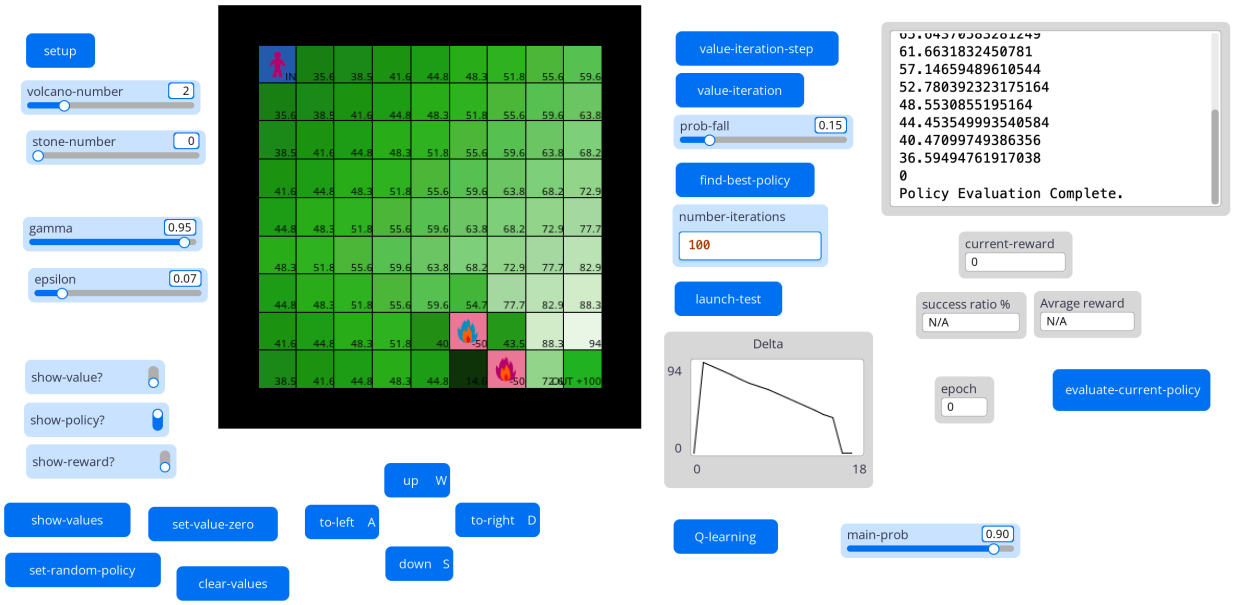
HW5 Reinforcement learning

Task 1. Policy evaluation

Random policy



Solved policy

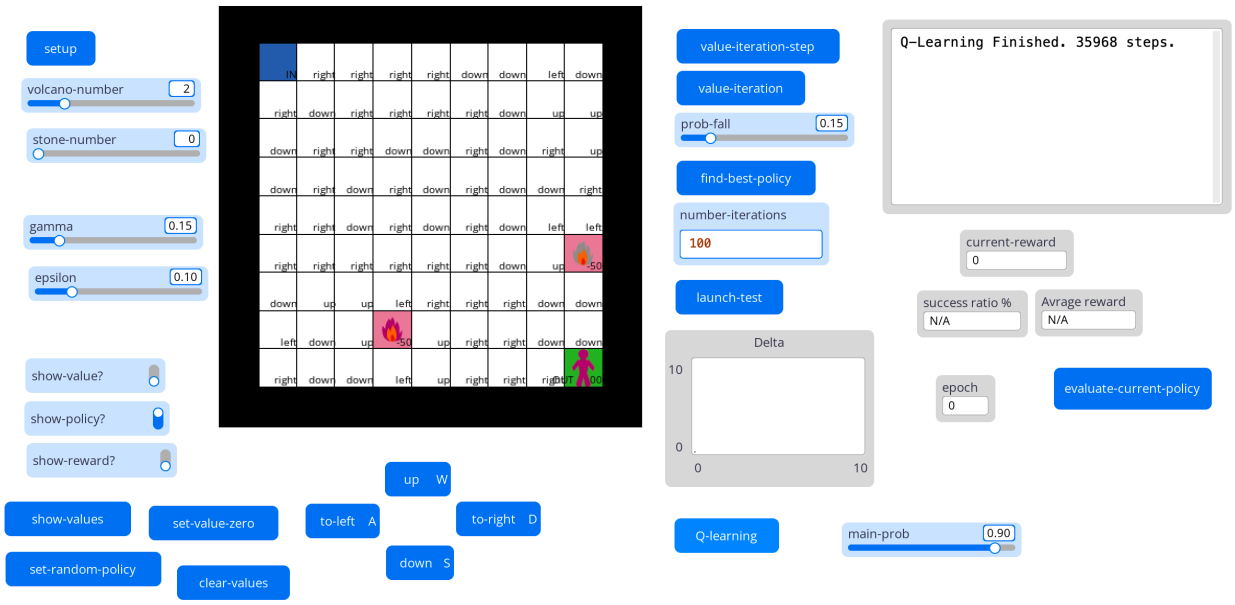


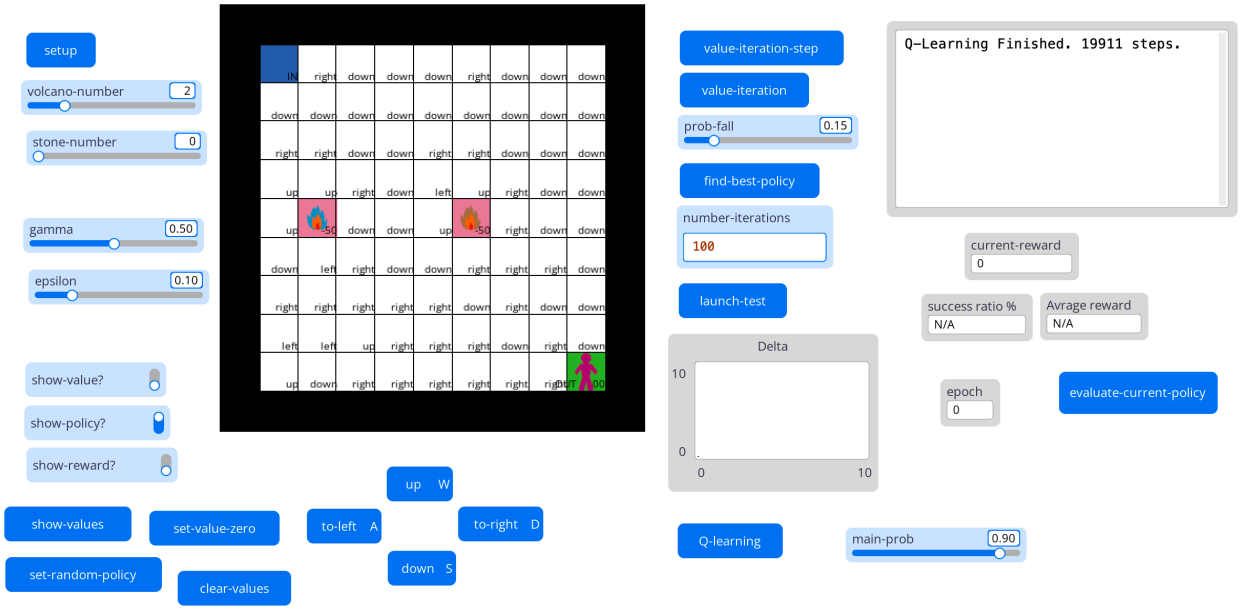
I implemented a function `evaluate-current-policy` that calculates Bellman expectation function for each cell.

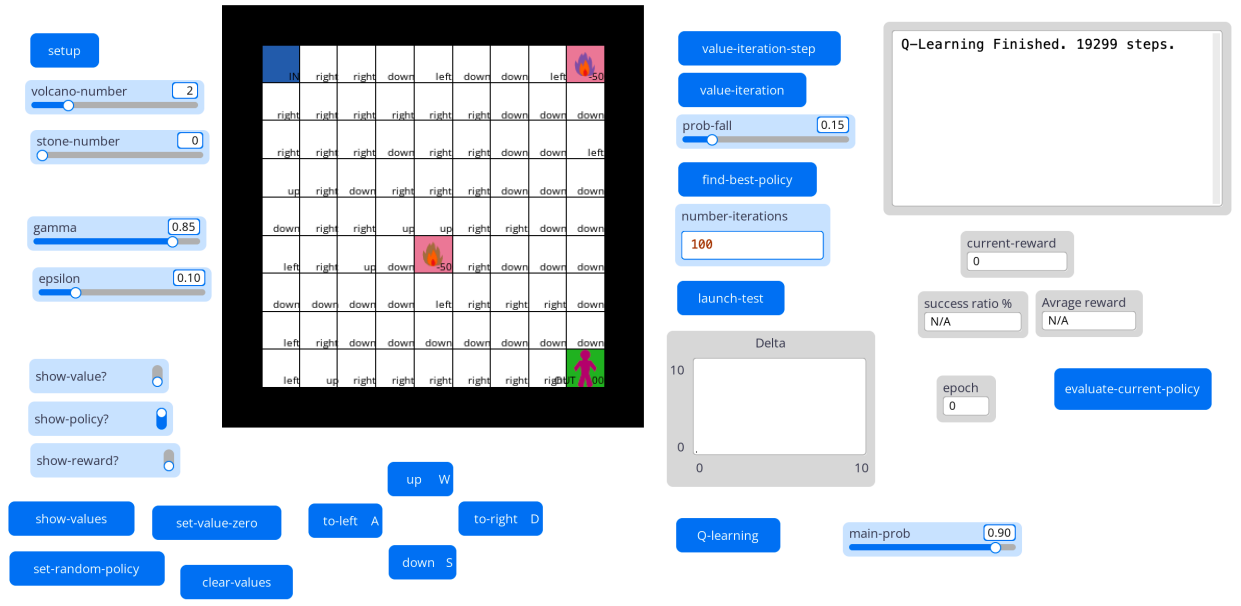
I noticed that, for random policy, cells near volcanos get lower values, which is expected.

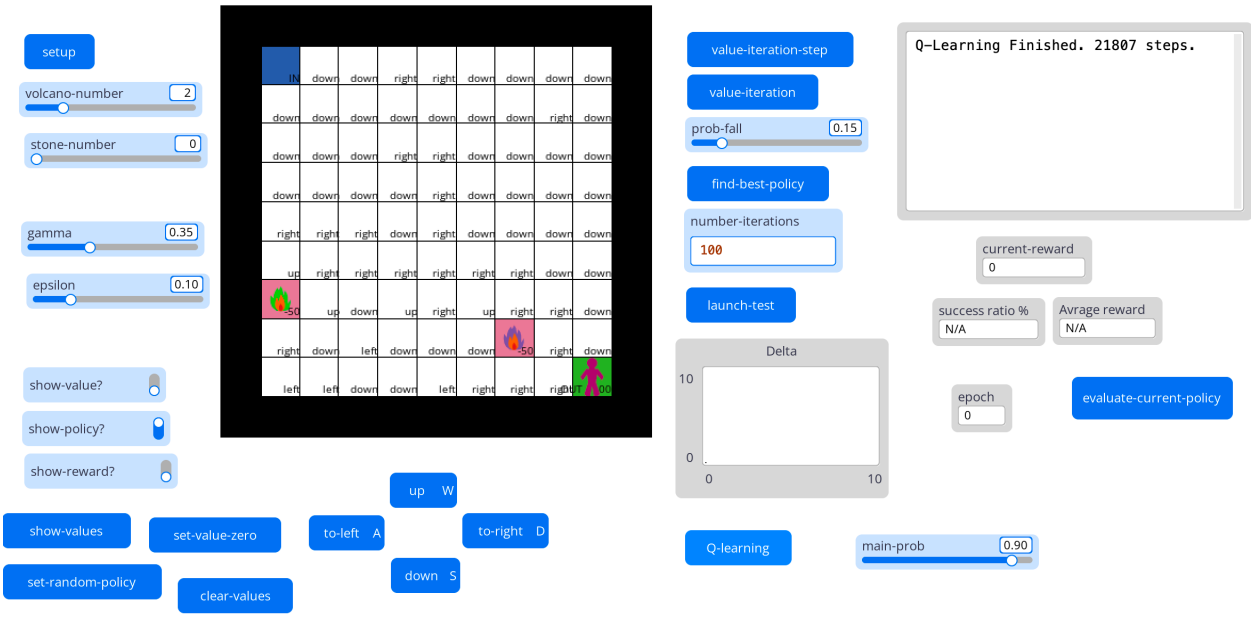
The same holds for solved policy. You can see that for solved policy almost all cells give high reward, it is explained by the fact, that there is only 2 volcanos and no obstacles, thus almost all routes lead to END.

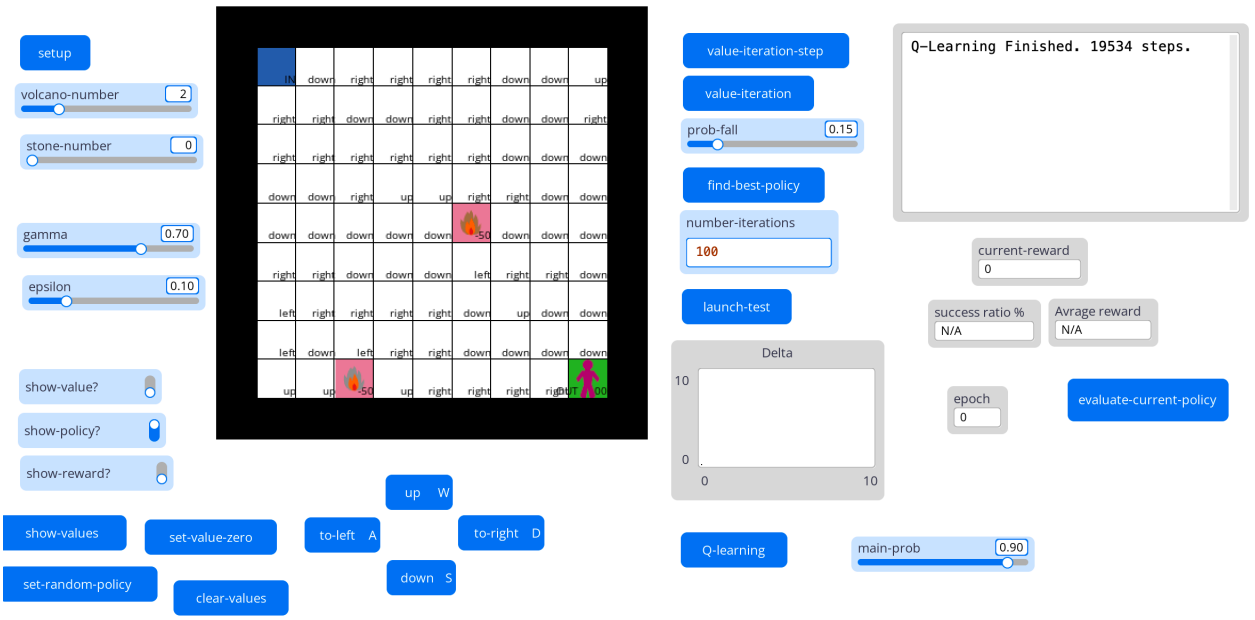
Q-learning



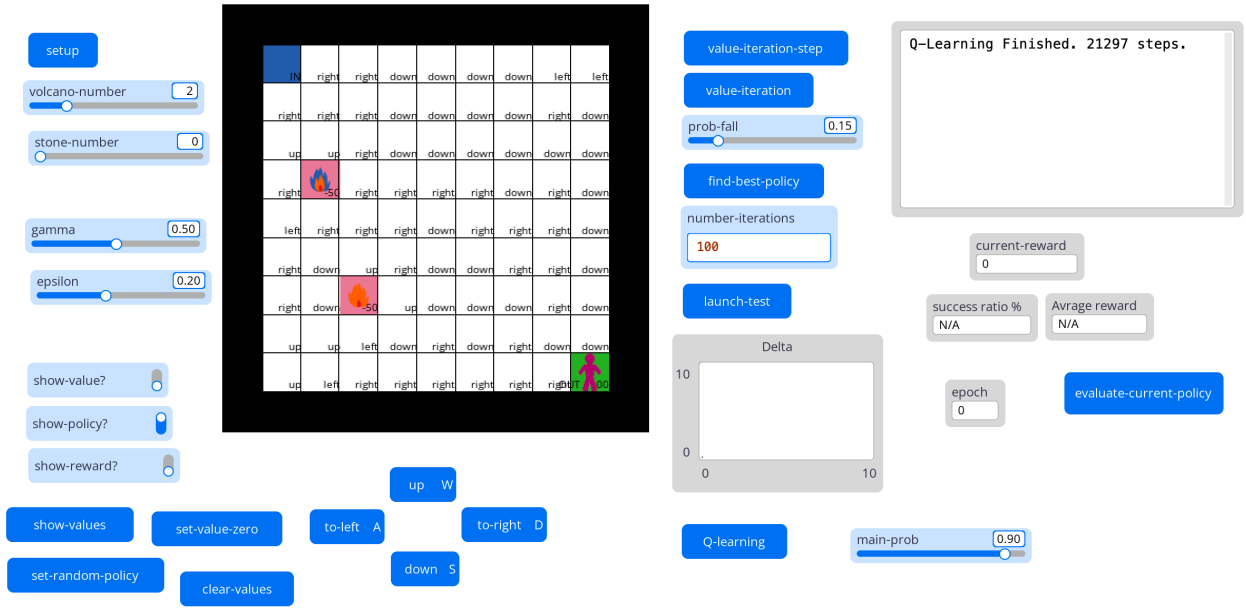

setup
volcano-number 2
stone-number 0
gamma 0.50
epsilon 0.10
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
to-left A
up W
down S
to-right D
value-iteration-step
value-iteration
prob-fall 0.15
find-best-policy
number-iterations 100
launch-test
Delta
10
0
0
10
Q-Learning
main-prob 0.90
Q-Learning Finished. 19911 steps.
current-reward 0
success ratio % N/A
Avrage reward N/A
epoch 0
evaluate-current-policy

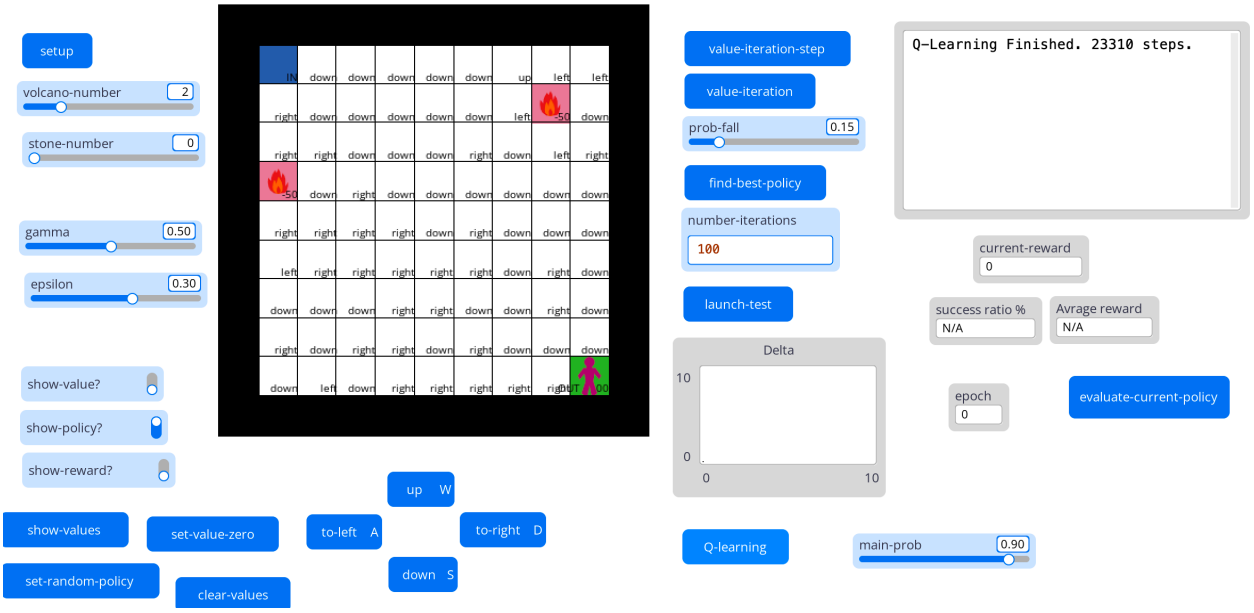

setup
volcano-number 2
stone-number 0
gamma 0.85
epsilon 0.10
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
to-left A
up W
down S
to-right D
value-iteration-step
value-iteration
prob-fall 0.15
find-best-policy
number-iterations 100
launch-test
Delta
10
0
0
10
Q-Learning
main-prob 0.90
Q-Learning Finished. 19299 steps.
current-reward 0
success ratio % N/A
Avrage reward N/A
epoch 0
evaluate-current-policy

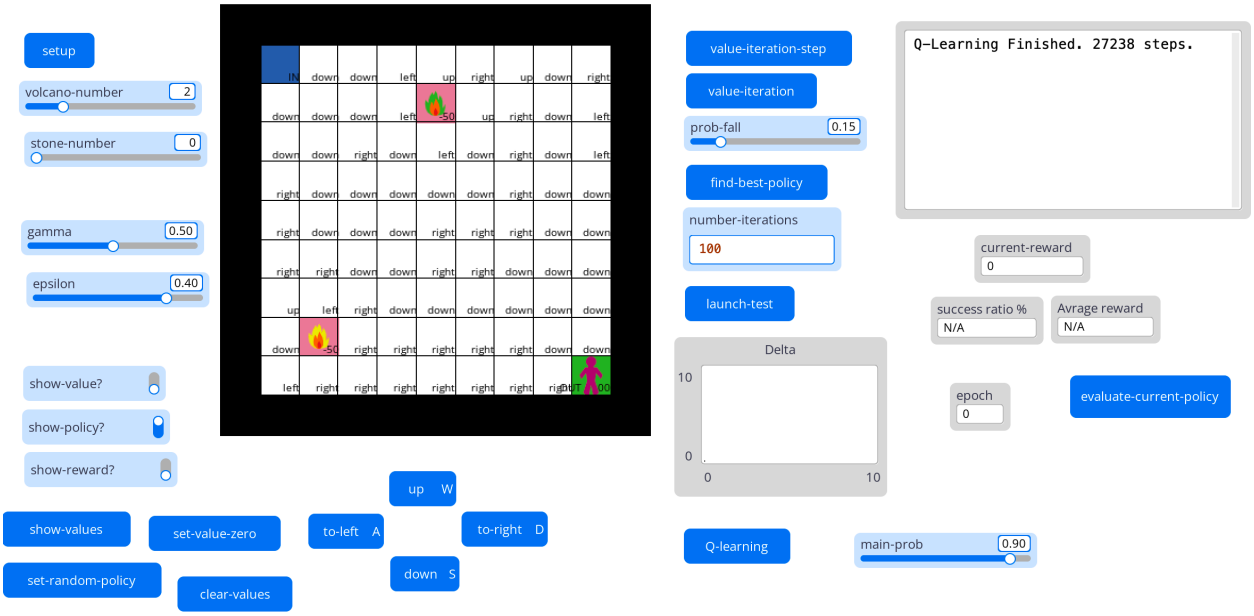

setup
volcano-number 2
stone-number 0
gamma 0.35
epsilon 0.10
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
up W
to-left A
to-right D
down S
value-iteration-step
value-iteration
prob-fall 0.15
find-best-policy
number-iterations 100
launch-test
Delta
Q-learning
main-prob 0.90
Q-Learning Finished. 21807 steps.
current-reward 0
success ratio % N/A
Avrage reward N/A
epoch 0
evaluate-current-policy

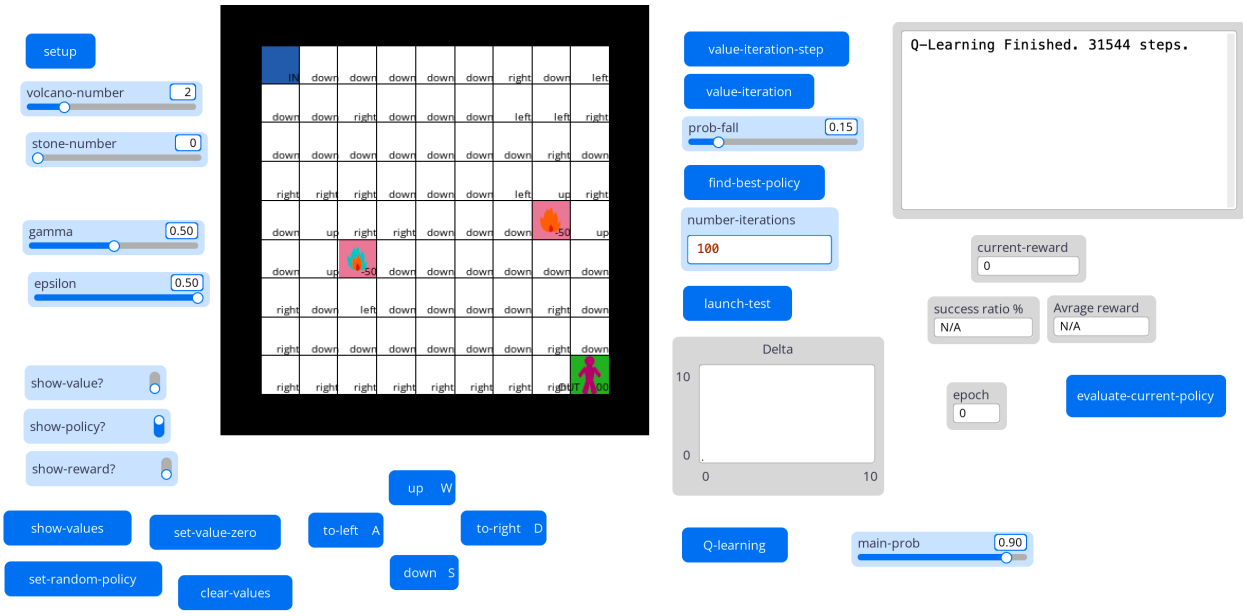

setup
volcano-number 2
stone-number 0
gamma 0.70
epsilon 0.10
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
up W
to-left A
to-right D
down S
value-iteration-step
value-iteration
prob-fall 0.15
find-best-policy
number-iterations 100
launch-test
Delta
Q-learning
main-prob 0.90
Q-Learning Finished. 19534 steps.
current-reward 0
success ratio % N/A
Avrage reward N/A
epoch 0
evaluate-current-policy


setup
volcano-number 2
stone-number 0
gamma 0.50
epsilon 0.2
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
up W
to-left A
to-right D
down S
value-iteration-step
value-iteration
prob-fall 0.15
find-best-policy
number-iterations
100
launch-test
Delta
10
0
0
10
Q-Learning
main-prob 0.90
Q-Learning Finished. 18382 steps.
current-reward 0
success ratio % N/A
Avrage reward N/A
epoch 0
evaluate-current-policy


setup
volcano-number 2
stone-number 0
gamma 0.50
epsilon 0.20
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
up W
to-left A
to-right D
down S
value-iteration-step
value-iteration
prob-fall 0.15
find-best-policy
number-iterations
100
launch-test
Delta
10
0
0
10
Q-Learning
main-prob 0.90
Q-Learning Finished. 21297 steps.
current-reward 0
success ratio % N/A
Avrage reward N/A
epoch 0
evaluate-current-policy


setup
volcano-number 2
stone-number 0
gamma 0.50
epsilon 0.30
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
up W
to-left A
to-right D
down S
value-iteration-step
value-iteration
prob-fall 0.15
find-best-policy
number-iterations 100
launch-test
Q-learning
main-prob 0.90
Q-Learning Finished. 23310 steps.
current-reward 0
success ratio % N/A
Avrage reward N/A
epoch 0
evaluate-current-policy
Delta
10
0
0
10


setup
volcano-number 2
stone-number 0
gamma 0.50
epsilon 0.40
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
up W
to-left A
to-right D
down S
value-iteration-step
value-iteration
prob-fall 0.15
find-best-policy
number-iterations 100
launch-test
Q-learning
main-prob 0.90
Q-Learning Finished. 27238 steps.
current-reward 0
success ratio % N/A
Avrage reward N/A
epoch 0
evaluate-current-policy
Delta
10
0
0
10



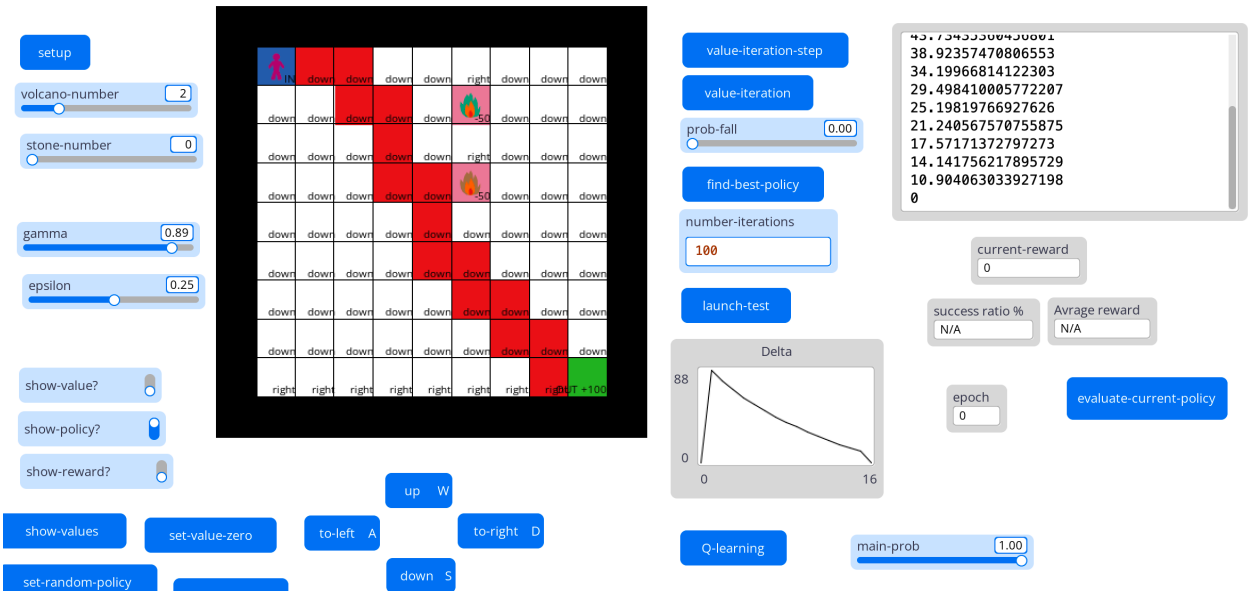
It tried different different learning rates γ and different e-e ratio epsilon. Turns out that the quickest convergence is achieved at $\gamma = 0.5, \varepsilon = 0.3$.

Task after Q-learning

Take MazeMDPDeterministic and add current change in value function plot and find-best-policy button.

How is it working?

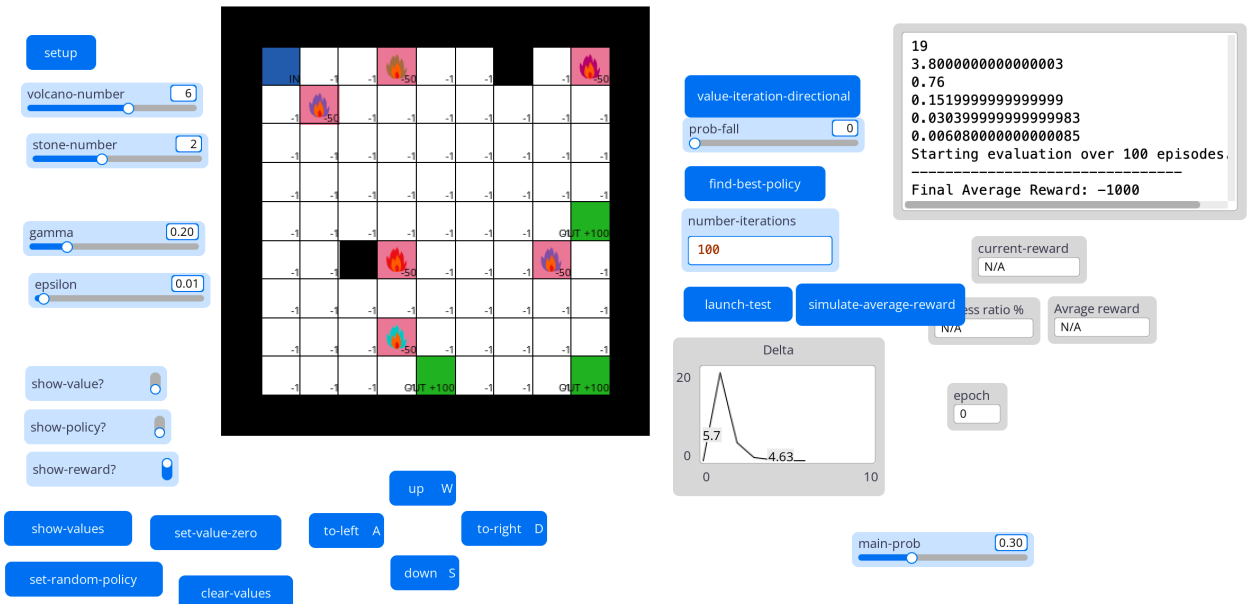
Works well. Converges

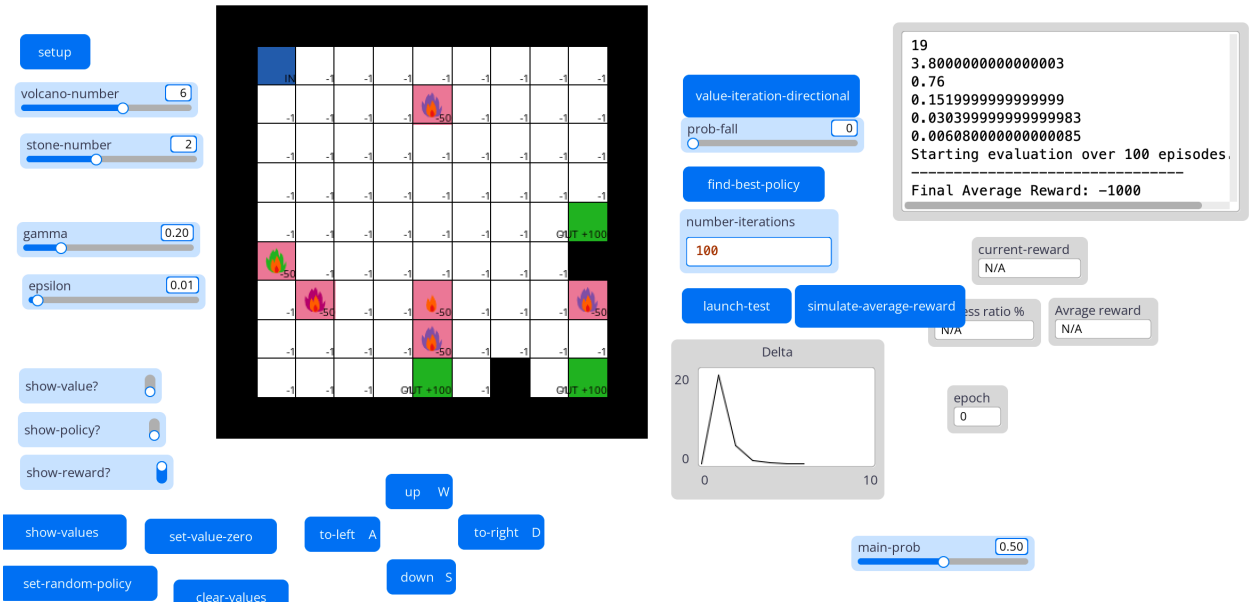


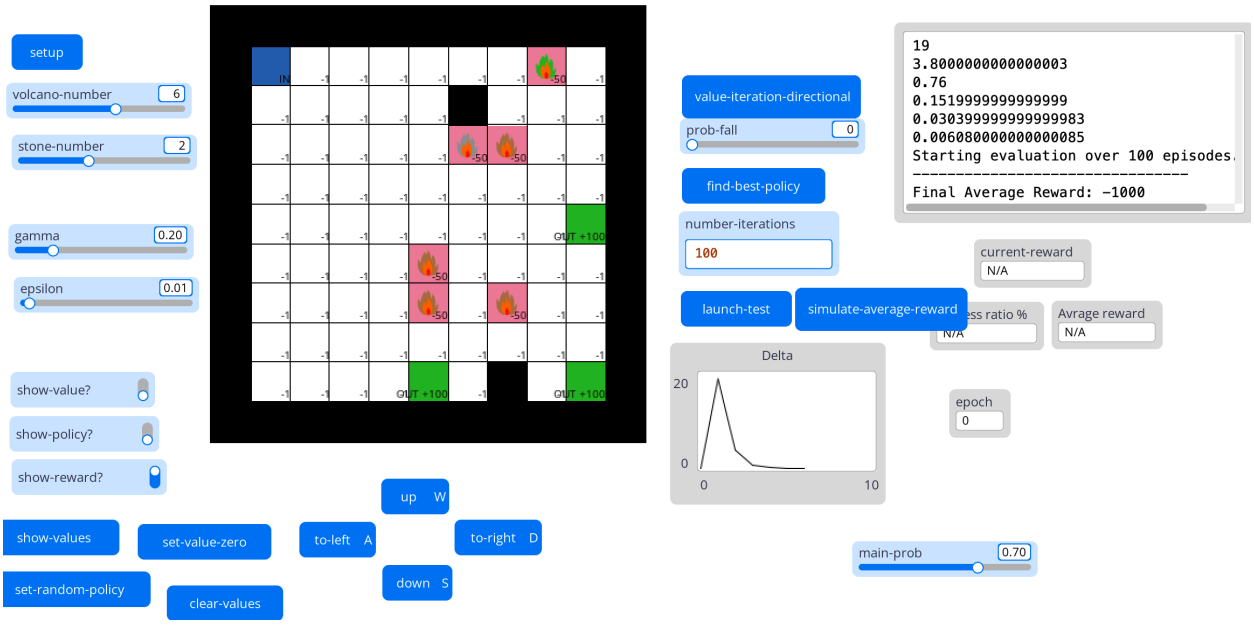
Task probabilistic case (new action model)

The state are now coordinate + heading (one out of four possible). Actions are now rotation to the left/right or step, each having reward (or rather punishment) -1. To implement this model I also need to store list of 4 values for each possible direction.

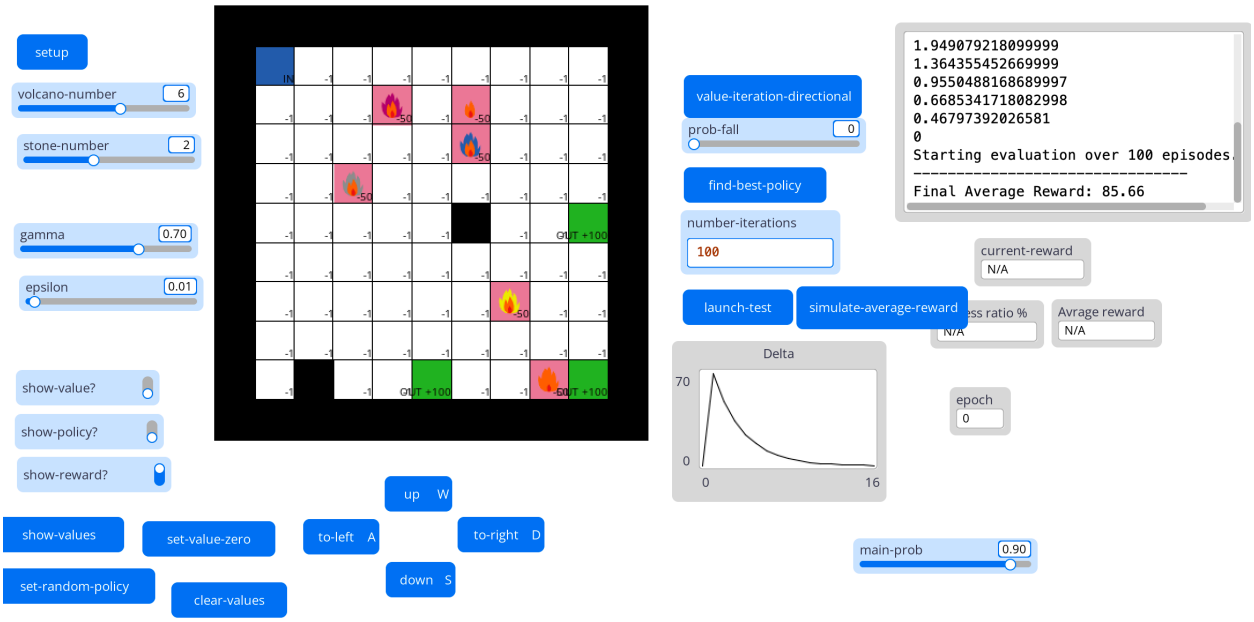
I created a function that runs solved policy for n-times and calculated average total reward

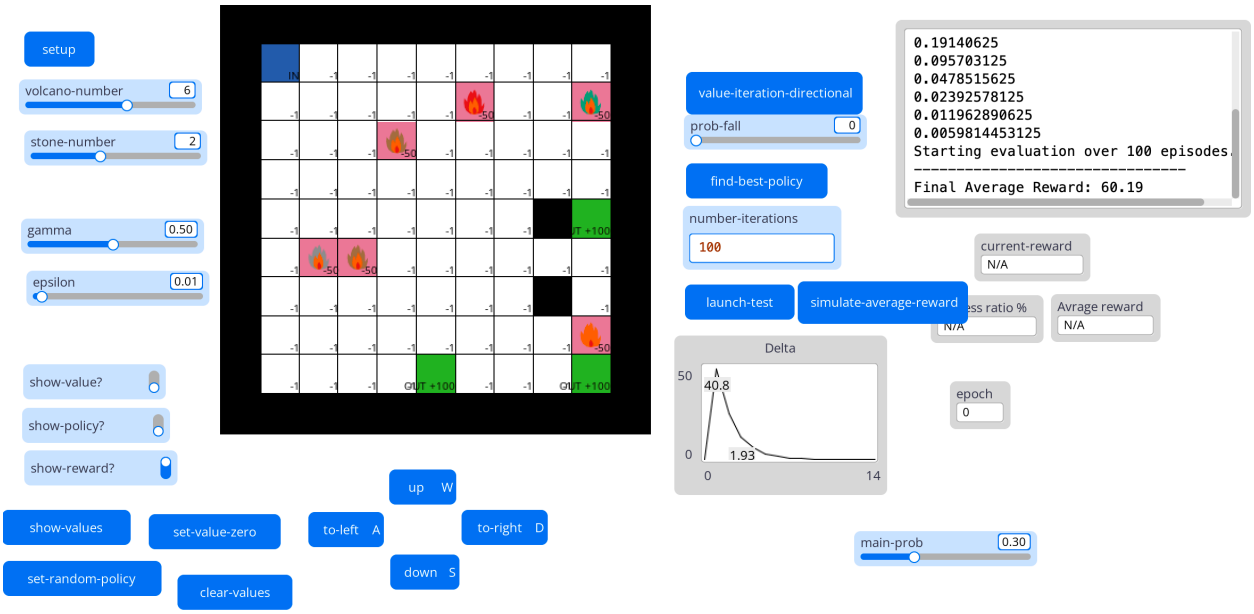


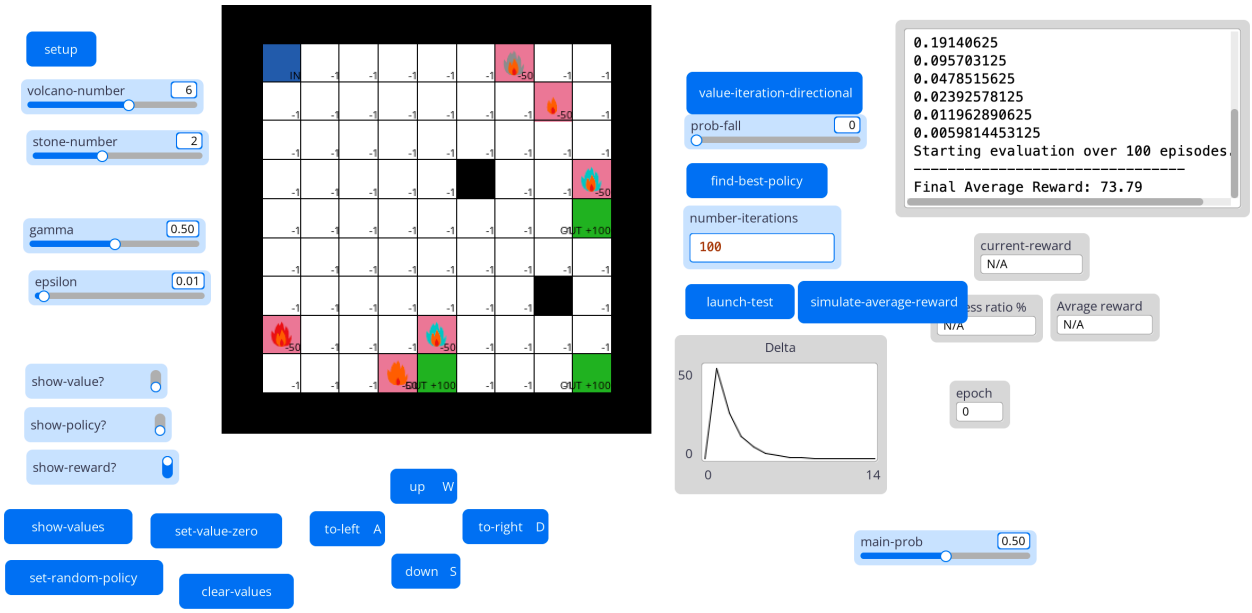


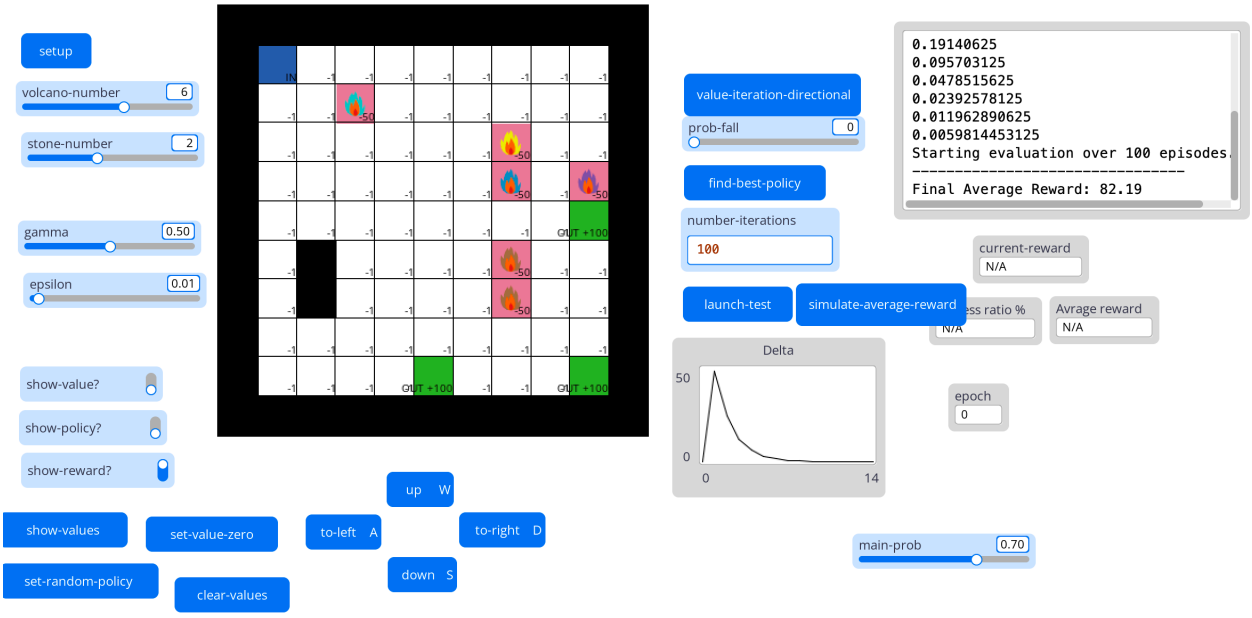


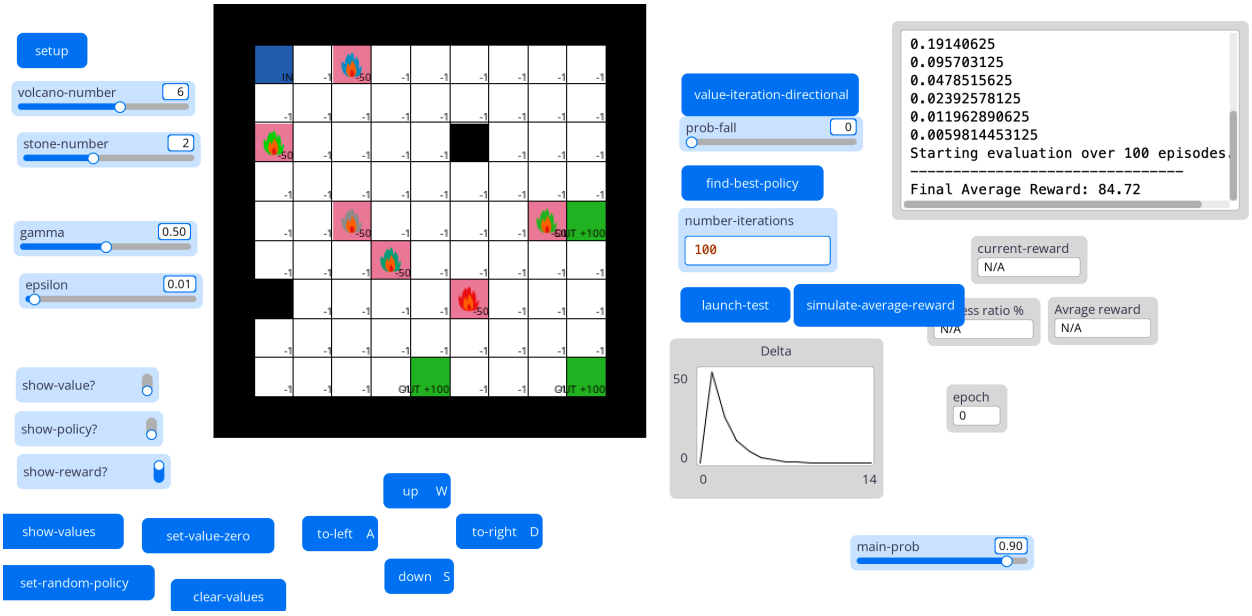
We can notice, that for low learning rate γ policy cannot be solved

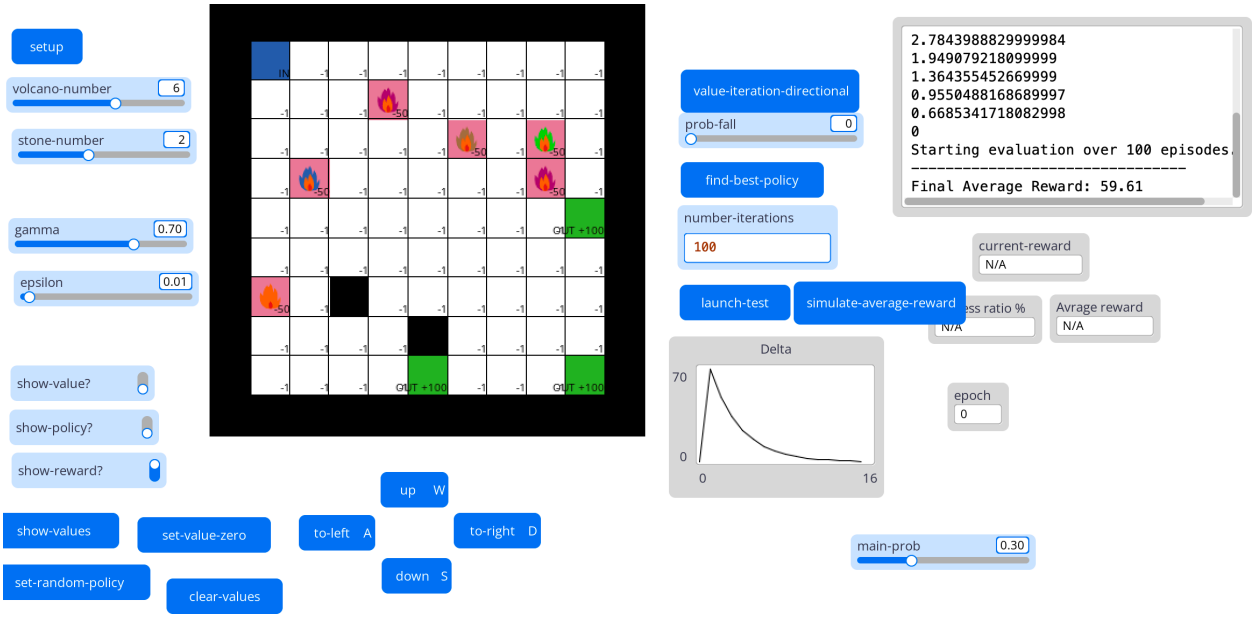

setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
clear-values
set-random-policy
up W
to-left A
to-right D
down S
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
current-reward N/A
Avrage reward N/A
Delta
70
0
0
16
epoch 0
main-prob 0.90
1.949079218099999
1.364355452669999
0.9550488168689997
0.6685341718082998
0.46797392026581
0
Starting evaluation over 100 episodes.
Final Average Reward: 85.66

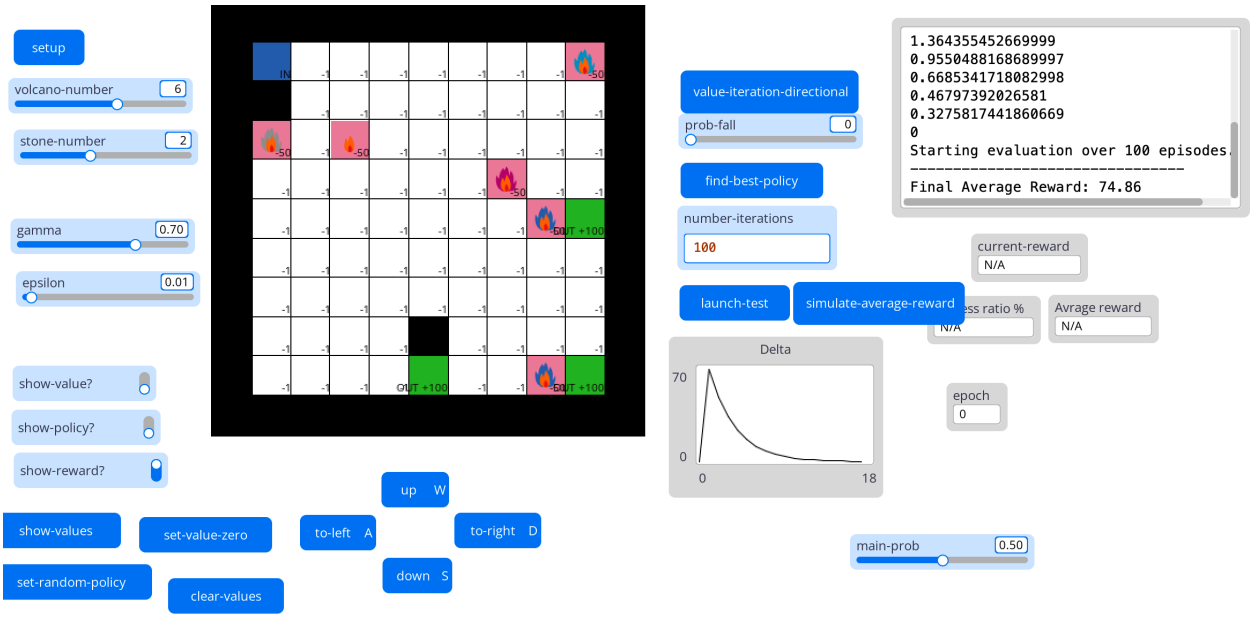

setup
volcano-number 6
stone-number 2
gamma 0.50
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
clear-values
set-random-policy
up W
to-left A
to-right D
down S
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
current-reward N/A
Avrage reward N/A
Delta
50
0
40.8
1.93
0
14
epoch 0
main-prob 0.30
0.19140625
0.095703125
0.0478515625
0.02392578125
0.011962890625
0.0059814453125
Starting evaluation over 100 episodes.
Final Average Reward: 60.19

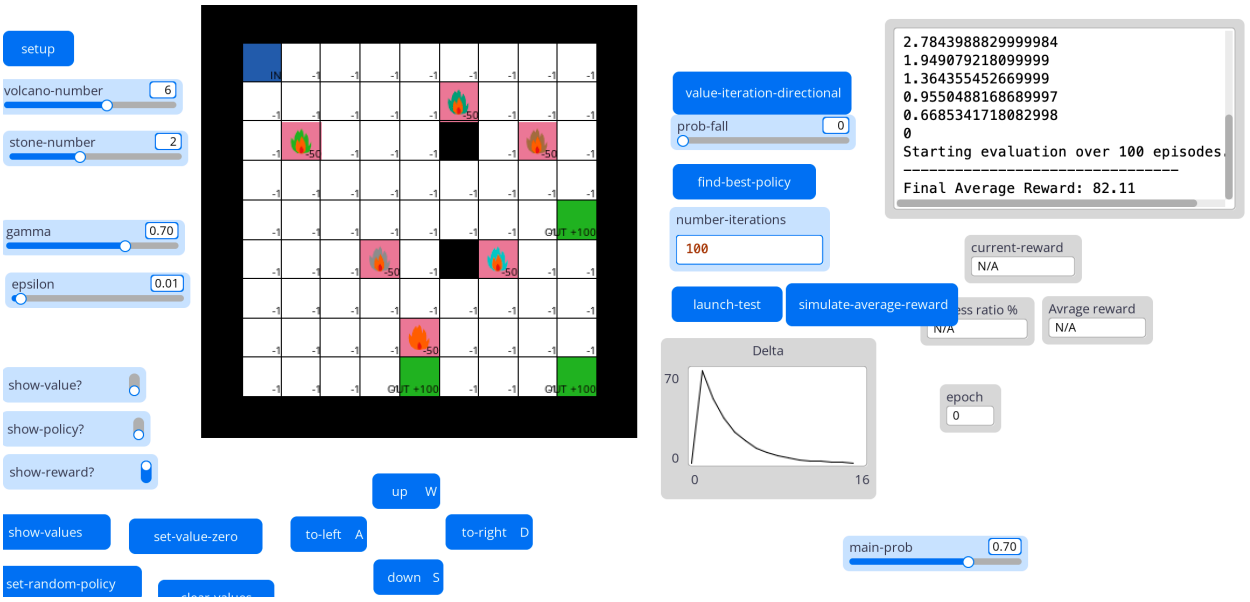

setup
volcano-number 6
stone-number 2
gamma 0.50
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
to-left A
up W
down S
to-right D
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
current-reward N/A
Avrage reward N/A
Delta
50
0
0
14
epoch 0
main-prob 0.50
0.19140625
0.095703125
0.0478515625
0.02392578125
0.011962890625
0.0059814453125
Starting evaluation over 100 episodes.
Final Average Reward: 73.79


setup
volcano-number 6
stone-number 2
gamma 0.50
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
to-left A
up W
down S
to-right D
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
current-reward N/A
Avrage reward N/A
Delta
50
0
0
14
epoch 0
main-prob 0.70
0.19140625
0.095703125
0.0478515625
0.02392578125
0.011962890625
0.0059814453125
Starting evaluation over 100 episodes.
Final Average Reward: 82.19


setup
volcano-number 6
stone-number 2
gamma 0.50
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
up W
to-left A
to-right D
down S
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
0.19140625
0.095703125
0.0478515625
0.02392578125
0.011962890625
0.0059814453125
Starting evaluation over 100 episodes.
Final Average Reward: 84.72
current-reward N/A
Avrage reward N/A
Delta
50
0
0
14
epoch 0
main-prob 0.90

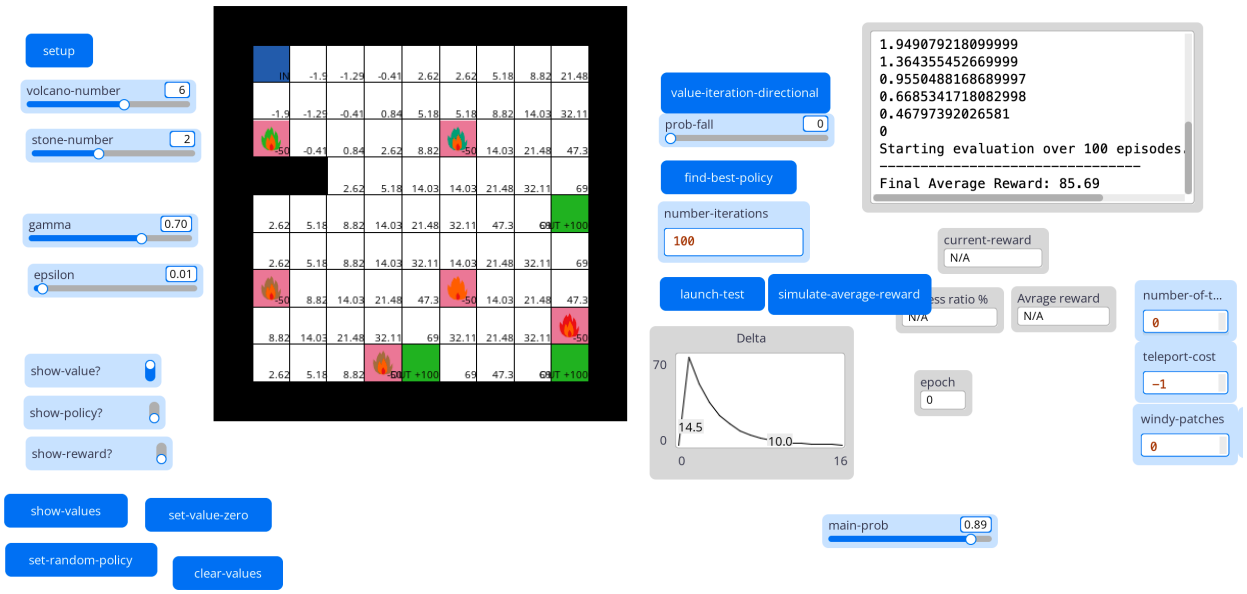

setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
up W
to-left A
to-right D
down S
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
2.7843988829999984
1.949079218099999
1.364355452669999
0.9550488168689997
0.6685341718082998
0
Starting evaluation over 100 episodes.
Final Average Reward: 59.61
current-reward N/A
Avrage reward N/A
Delta
70
0
0
16
epoch 0
main-prob 0.30

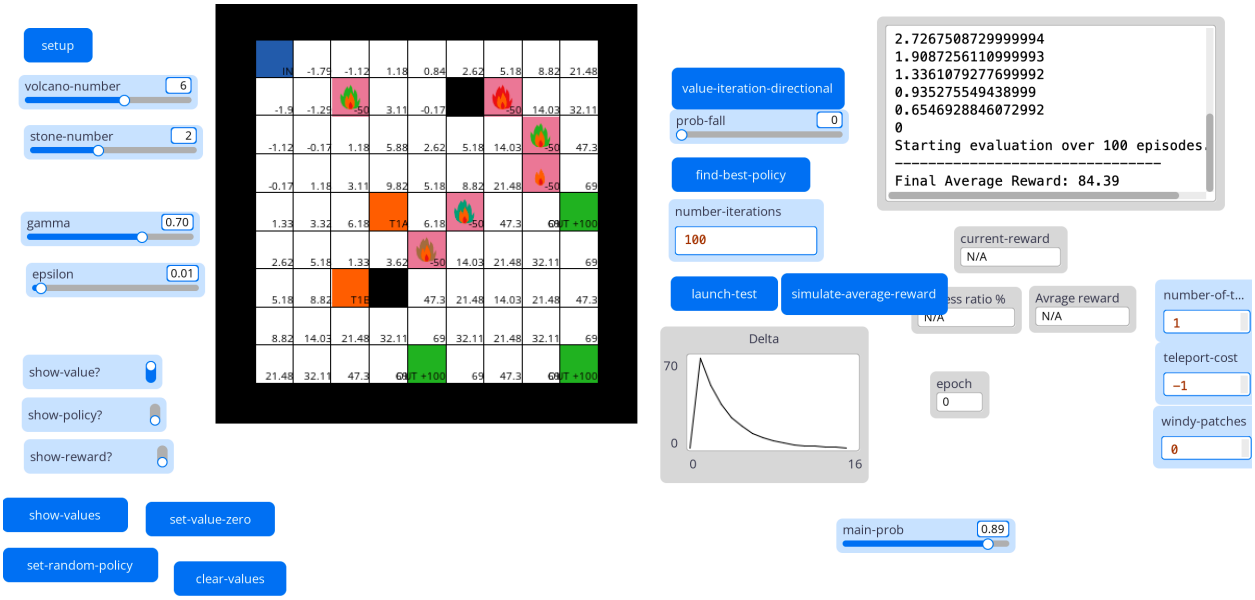


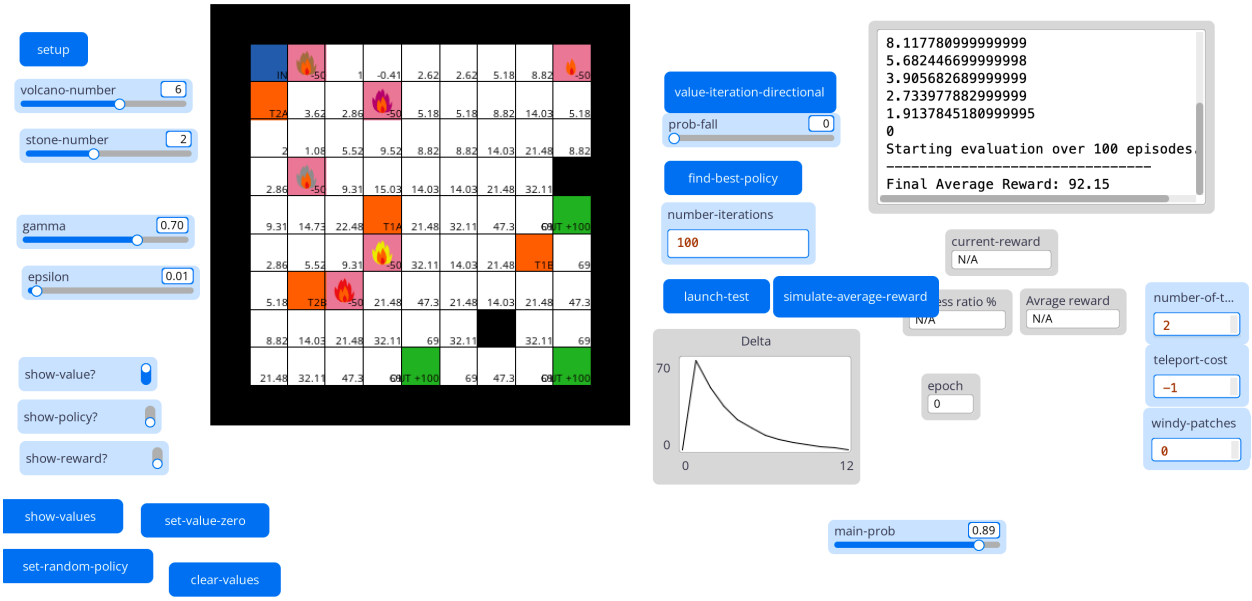


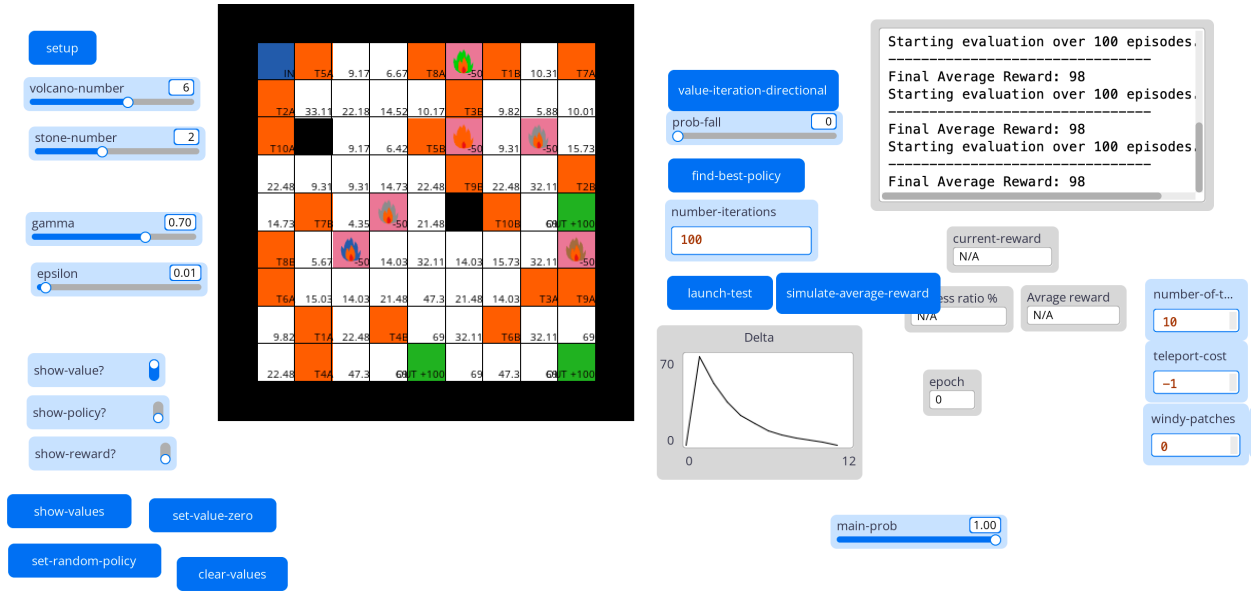
We can see that optimal learning rate is ≈ 0.5. Obviously, the higher probability of success of given action the better final reward.

Teleports


setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
1.949079218099999
1.364355452669999
0.9550488168689997
0.6685341718082998
0.46797392026581
0
Starting evaluation over 100 episodes.
Final Average Reward: 85.69
current-reward N/A
Avrage reward N/A
number-of-t... 0
teleport-cost -1
windy-patches 0
epoch 0
Delta
70
0
14.5
10.0
0
16
main-prob 0.89

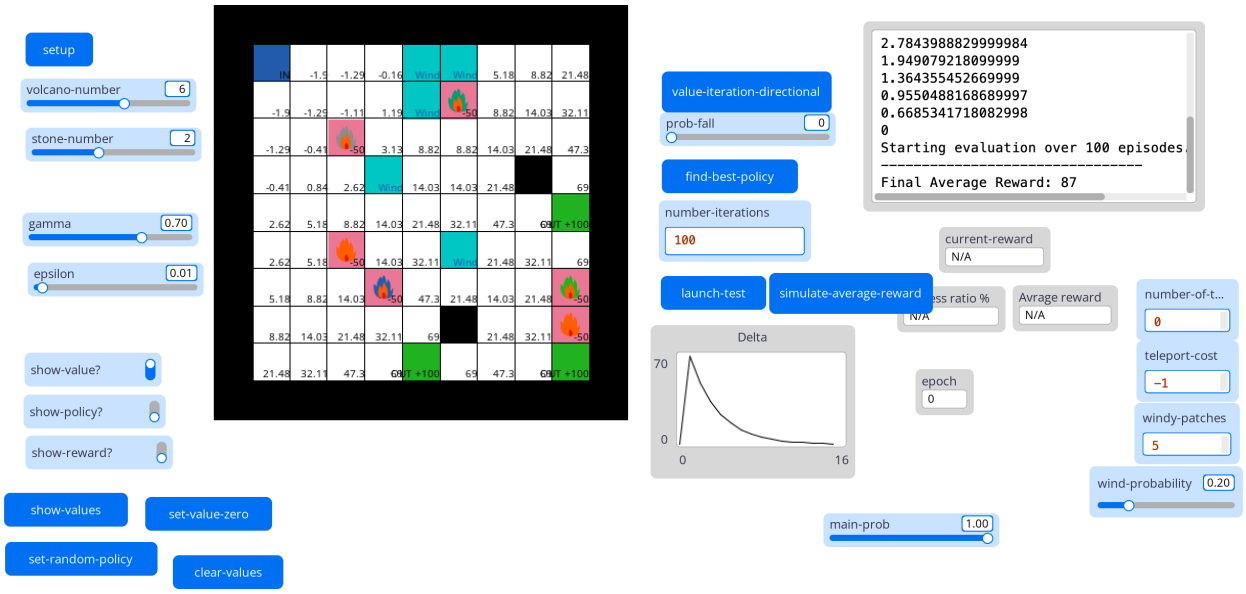

setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
2.7267508729999994
1.9087256110999993
1.3361079277699992
0.935275549438999
0.6546928846072992
0
Starting evaluation over 100 episodes.
Final Average Reward: 84.39
current-reward N/A
Avrage reward N/A
number-of-t... 1
teleport-cost -1
windy-patches 0
epoch 0
Delta
70
0
0
16
main-prob 0.89

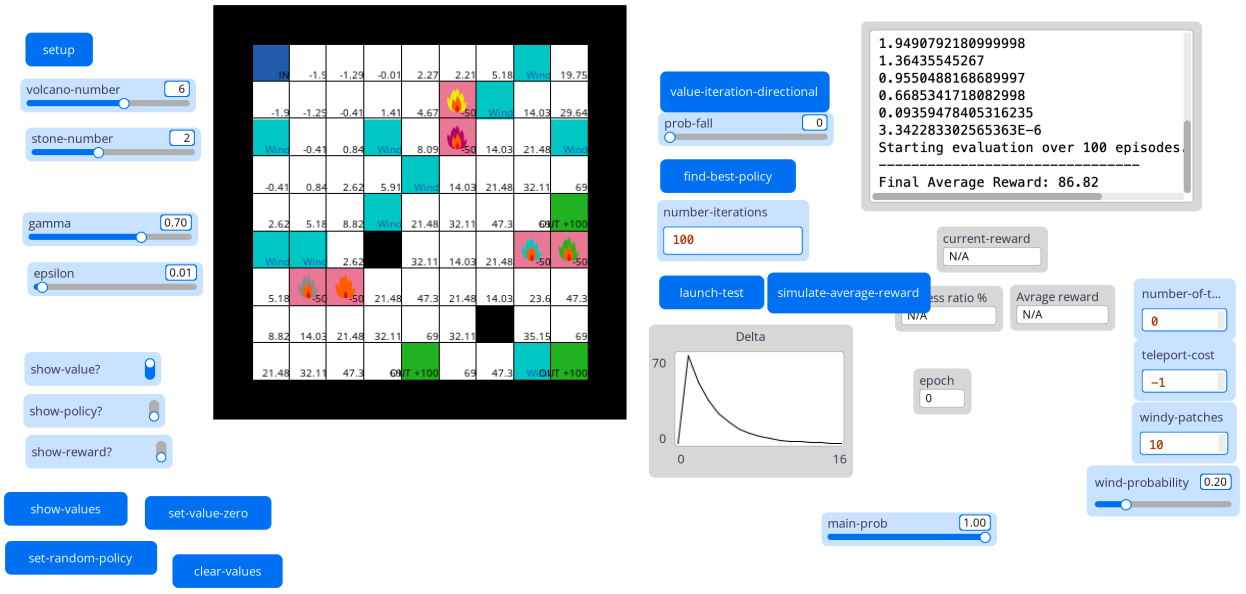


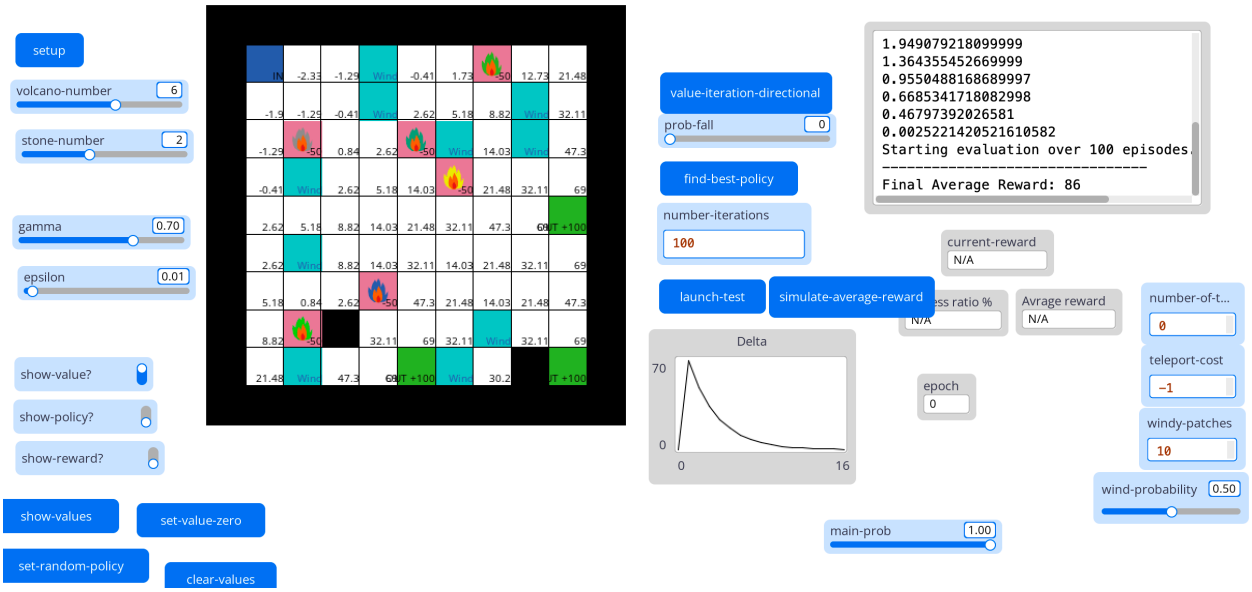


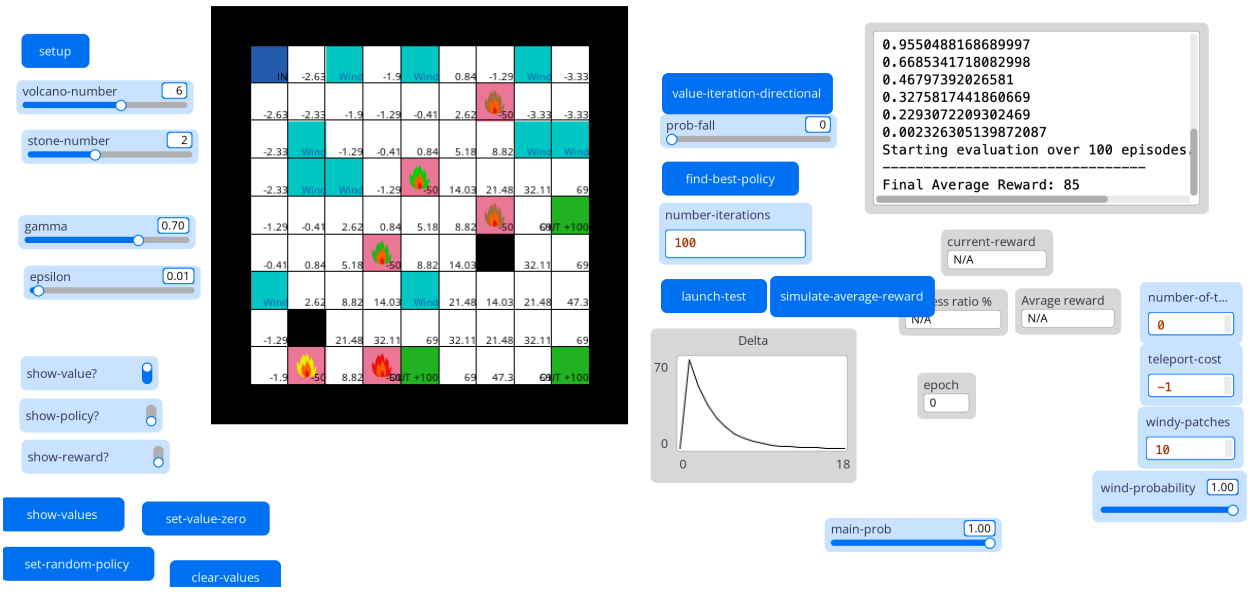
We can see, that average reward growth as the number of teleports growth. This is explained by the fact that agent learns position of teleport, which create a shortcut.

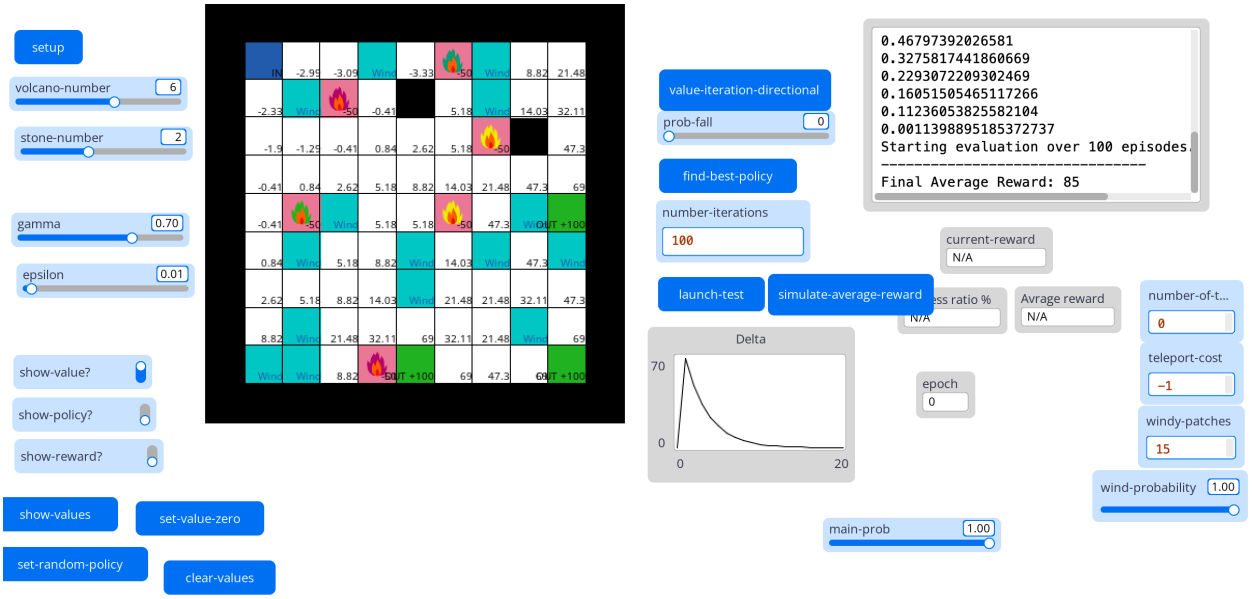
Wind

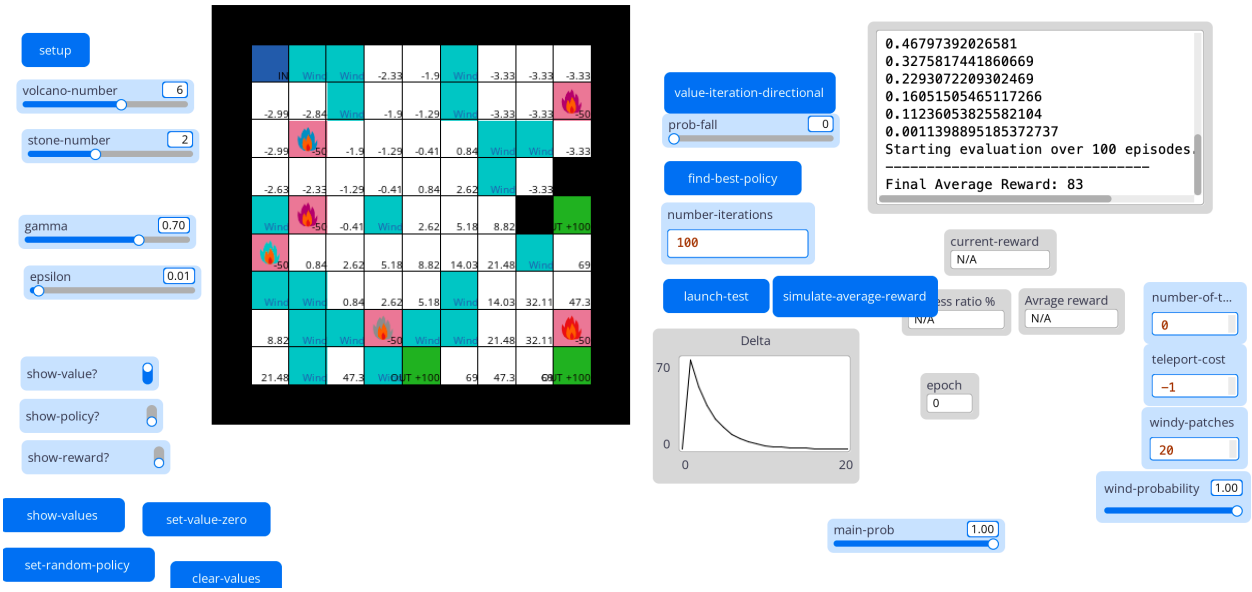

setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
current-reward N/A
Avrage reward N/A
Delta
epoch 0
number-of-t... 0
teleport-cost -1
windy-patches 5
wind-probability 0.20
main-prob 1.00
2.7843988829999984
1.949079218099999
1.364355452669999
0.9550488168689997
0.6685341718082998
0
Starting evaluation over 100 episodes.
Final Average Reward: 87


setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
current-reward N/A
Avrage reward N/A
Delta
epoch 0
number-of-t... 0
teleport-cost -1
windy-patches 10
wind-probability 0.20
main-prob 1.00
1.9490792180999998
1.36435545267
0.9550488168689997
0.6685341718082998
0.09359478405316235
3.342283302565363E-6
Starting evaluation over 100 episodes.
Final Average Reward: 86.82


setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
current-reward N/A
Avrage reward N/A
Delta
70
0
0
16
epoch 0
number-of-t... 0
teleport-cost −1
windy-patches 10
wind-probability 0.50
main-prob 1.00
1.949079218099999
1.364355452669999
0.9550488168689997
0.6685341718082998
0.46797392026581
0.0025221420521610582
Starting evaluation over 100 episodes.
Final Average Reward: 86


setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-value-zero
set-random-policy
clear-values
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
current-reward N/A
Avrage reward N/A
Delta
70
0
0
18
epoch 0
number-of-t... 0
teleport-cost −1
windy-patches 10
wind-probability 1.00
main-prob 1.00
0.9550488168689997
0.6685341718082998
0.46797392026581
0.3275817441860669
0.2293072209302469
0.002326305139872087
Starting evaluation over 100 episodes.
Final Average Reward: 85



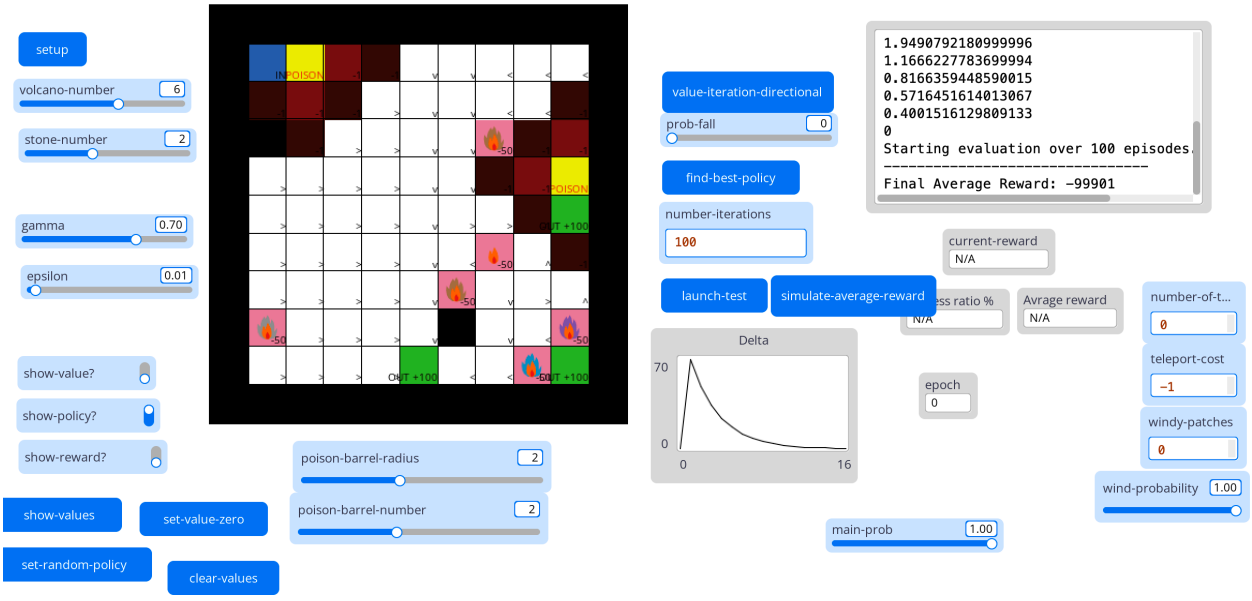


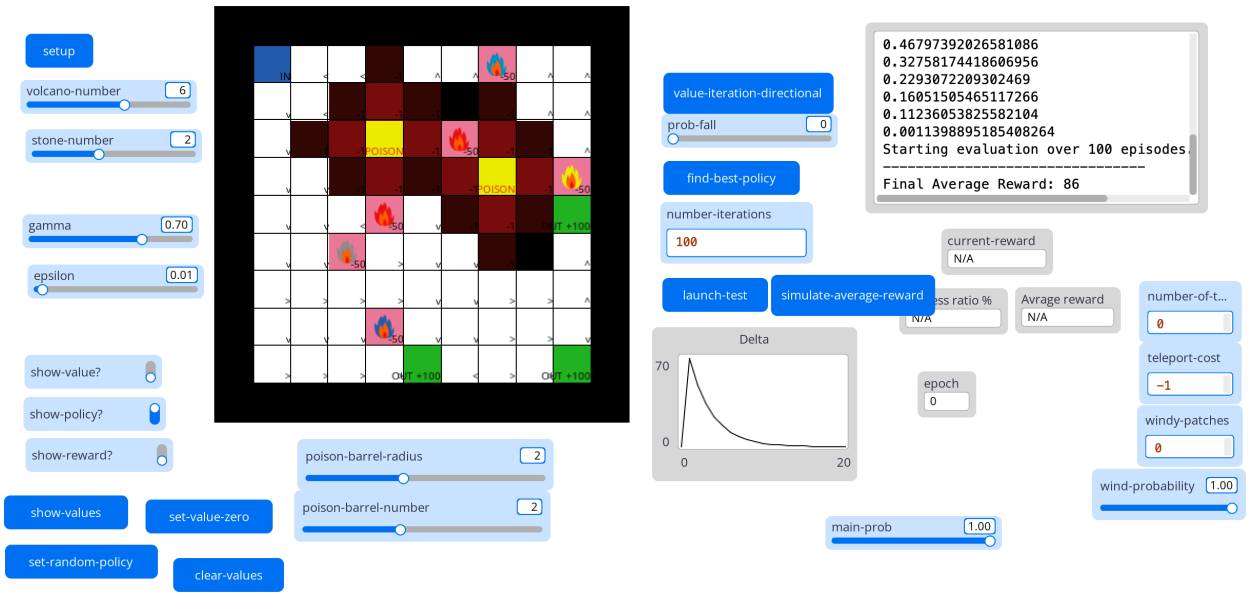
In general, I do not see that average reward decreases as number of windy blocks and probability increases. My assumption is that wind blows in the direction of OUT, thus agent learns how to escape “risky” windy blocks, and how to abuse a handful ones.

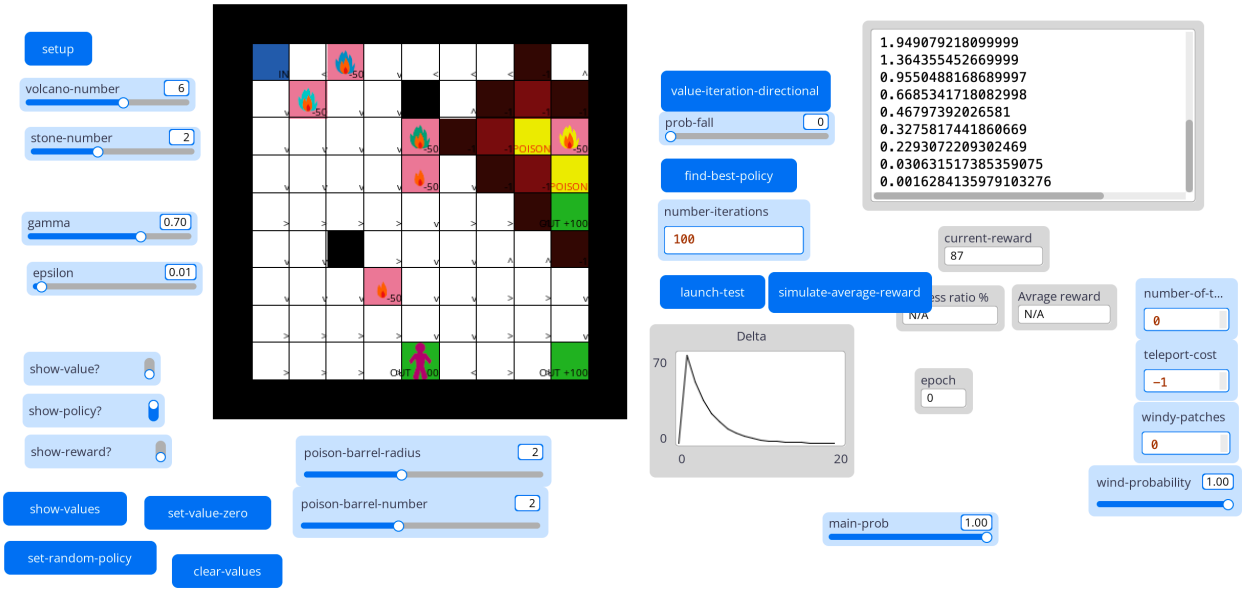
Poison

I implemented poison as described in the task. To include it in the VIA model, I add probability of being poisoned * punishment cost to expected probability.

It is interesting that agent cannot escape starting point after solving policy, if starting point is surrounded by poison

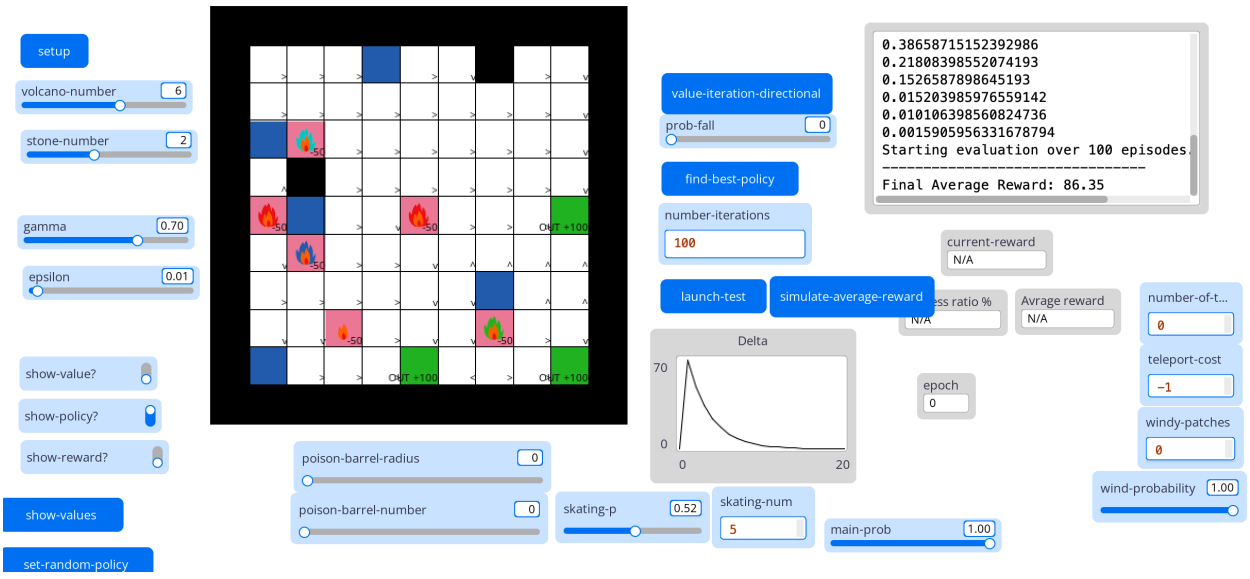


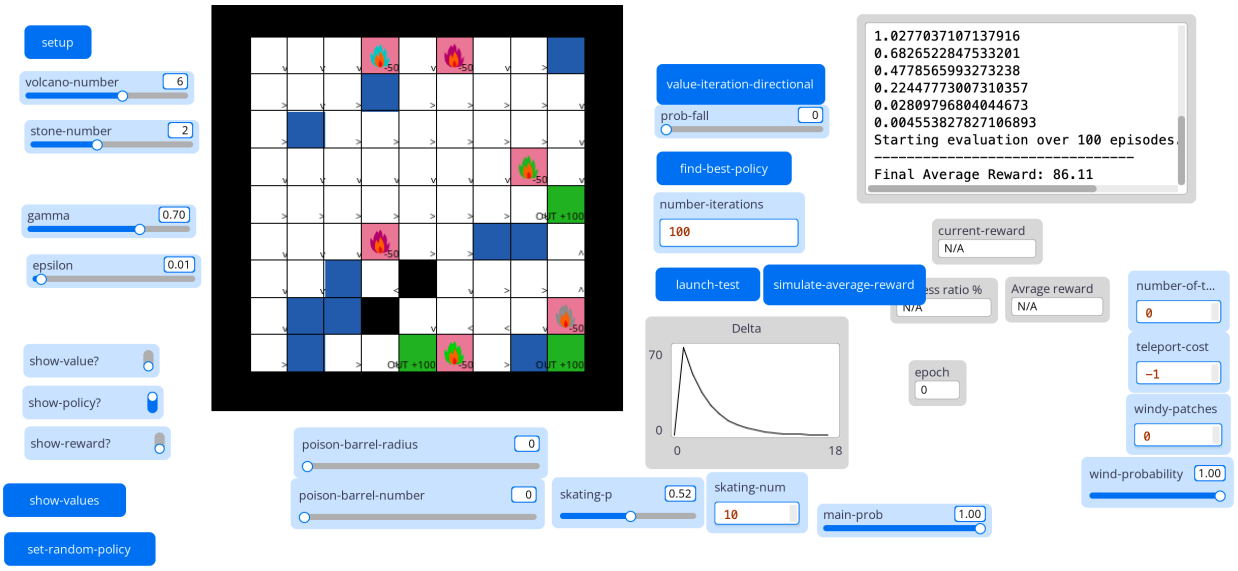


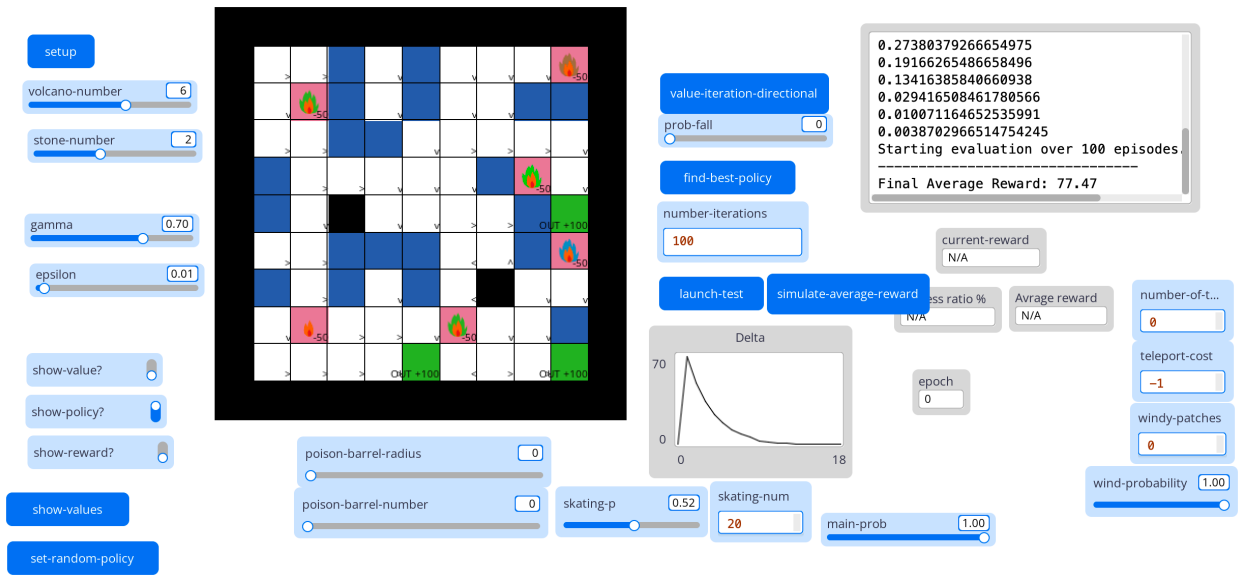


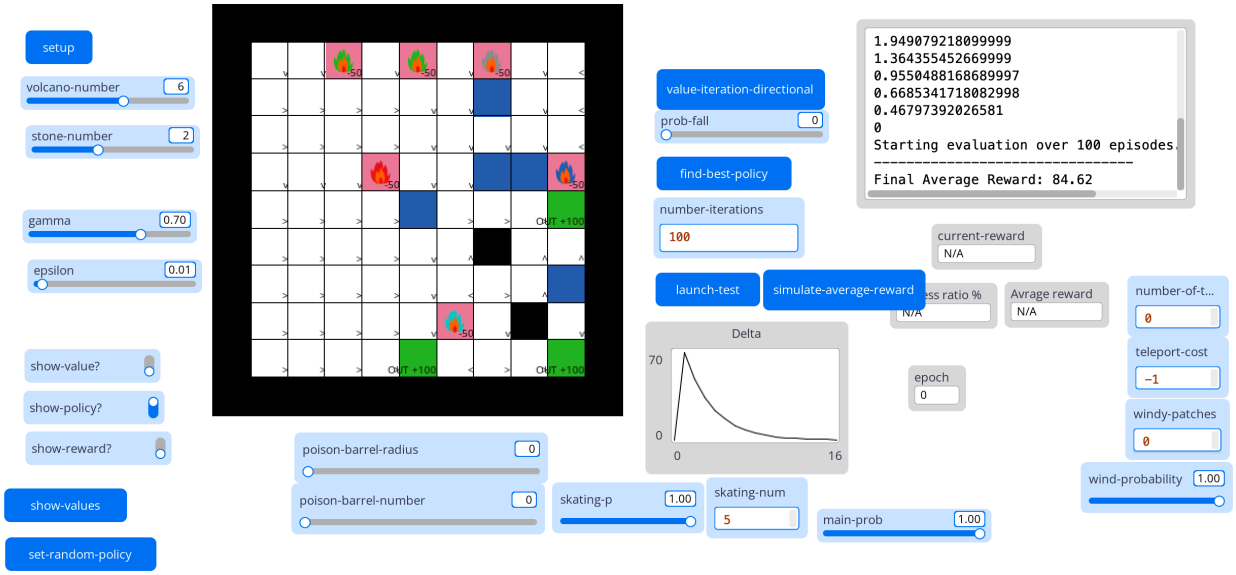
However, if starting point is not surrounded everything is fine.

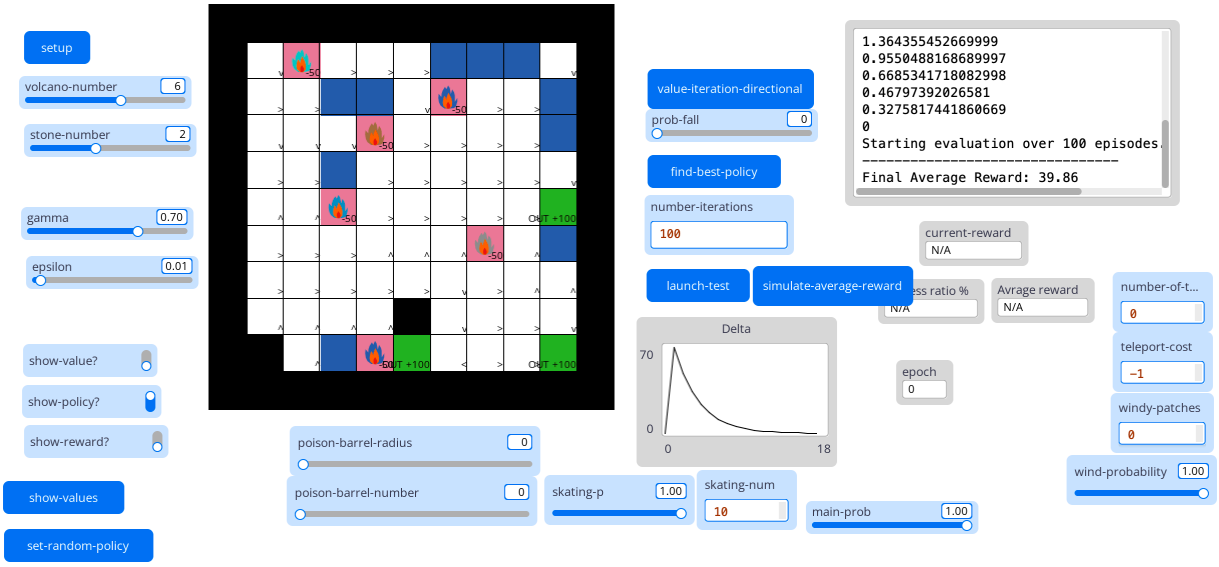
Ice skating

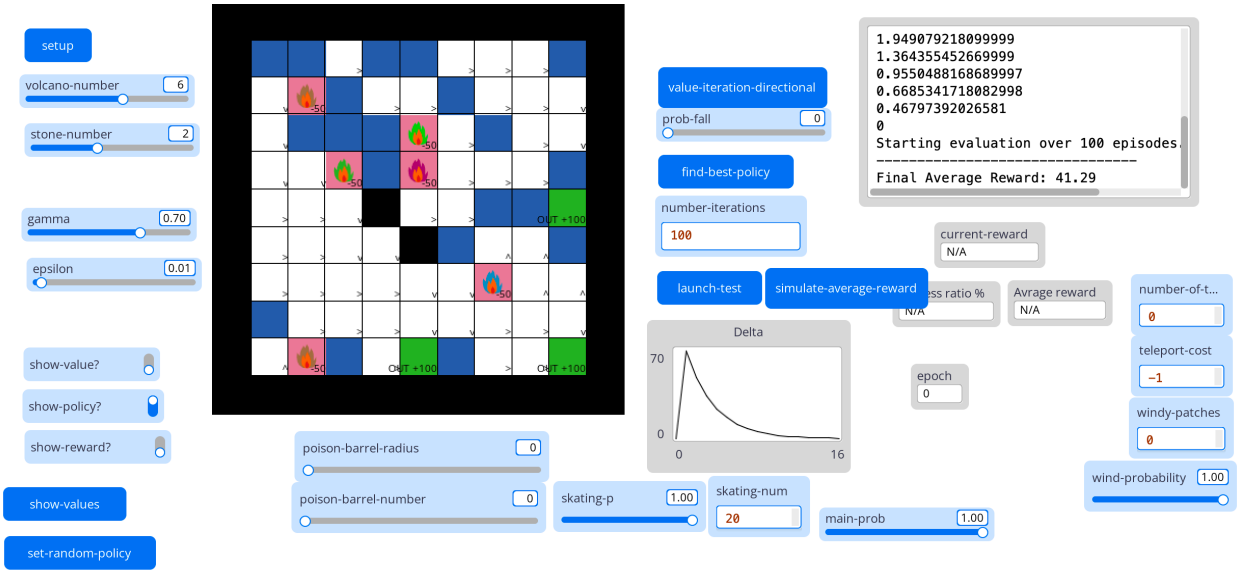



setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-random-policy
poison-barrel-radius 0
poison-barrel-number 0
skating-p 0.52
skating-num 10
main-prob 1.00
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
Delta
70
0
0
18
epoch 0
current-reward N/A
Avrage reward N/A
number-of-t... 0
teleport-cost -1
windy-patches 0
wind-probability 1.00
1.0277037107137916
0.6826522847533201
0.4778565993273238
0.22447773007310357
0.08209796804044673
0.004553827827106893
Starting evaluation over 100 episodes.
Final Average Reward: 86.11


setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-random-policy
poison-barrel-radius 0
poison-barrel-number 0
skating-p 0.52
skating-num 20
main-prob 1.00
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
Delta
70
0
0
18
epoch 0
current-reward N/A
Avrage reward N/A
number-of-t... 0
teleport-cost -1
windy-patches 0
wind-probability 1.00
0.27380379266654975
0.19166265486658496
0.13416385840660938
0.029416508461780566
0.010071164652535991
0.003870296514754245
Starting evaluation over 100 episodes.
Final Average Reward: 77.47


setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-random-policy
poison-barrel-radius 0
poison-barrel-number 0
skating-p 1.00
skating-num 5
main-prob 1.00
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
current-reward N/A
Avrage reward N/A
Delta
70
0
0
16
epoch 0
number-of-t... 0
teleport-cost -1
windy-patches 0
wind-probability 1.00
1.949079218099999
1.364355452669999
0.9550488168689997
0.6685341718082998
0.46797392026581
0
Starting evaluation over 100 episodes.
Final Average Reward: 84.62


setup
volcano-number 6
stone-number 2
gamma 0.70
epsilon 0.01
show-value?
show-policy?
show-reward?
show-values
set-random-policy
poison-barrel-radius 0
poison-barrel-number 0
skating-p 1.00
skating-num 10
main-prob 1.00
value-iteration-directional
prob-fall 0
find-best-policy
number-iterations 100
launch-test
simulate-average-reward
current-reward N/A
Avrage reward N/A
Delta
70
0
0
18
epoch 0
number-of-t... 0
teleport-cost -1
windy-patches 0
wind-probability 1.00
1.364355452669999
0.9550488168689997
0.6685341718082998
0.46797392026581
0.3275817441860669
0
Starting evaluation over 100 episodes.
Final Average Reward: 39.86



Increasing number of skating patches or skating probability certainly seems to decrease average reward.

Additional task. Moving blocks.