

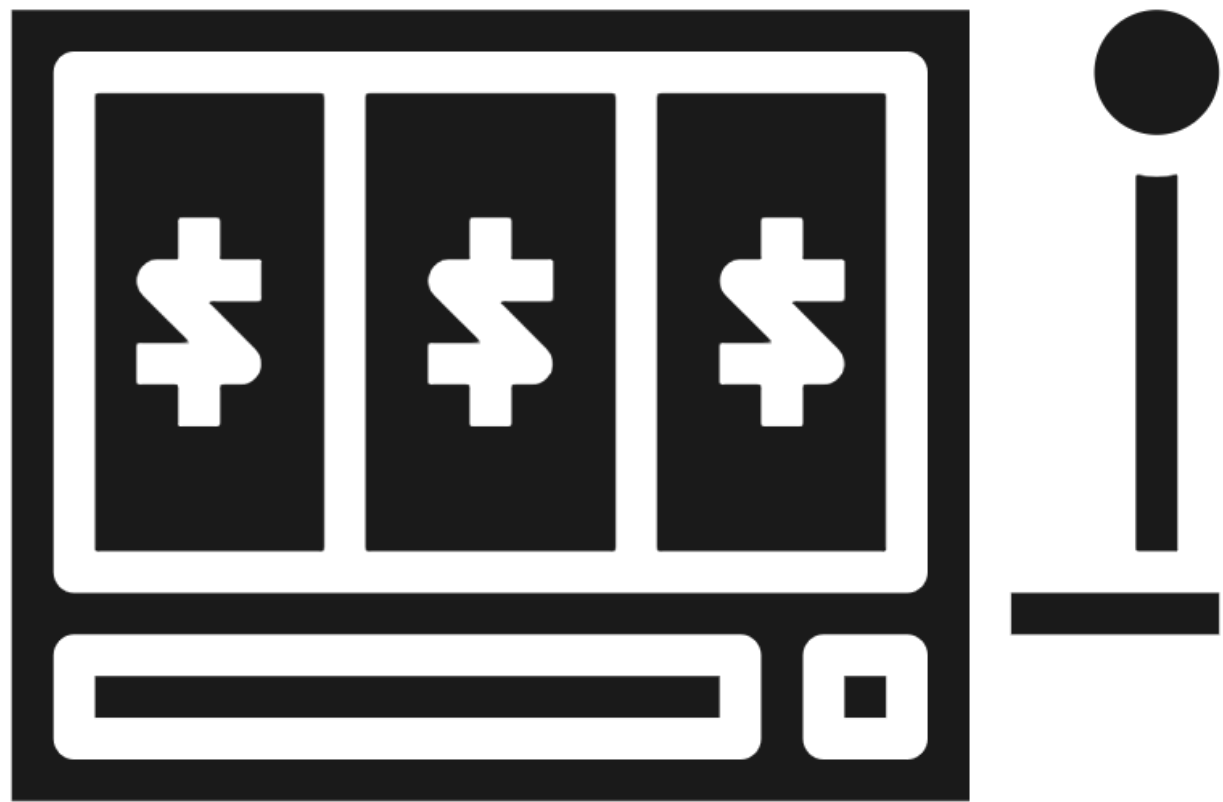


Reinforcement Learning

An Introduction
second edition

Richard S. Sutton and Andrew G. Barto

Chapter 2

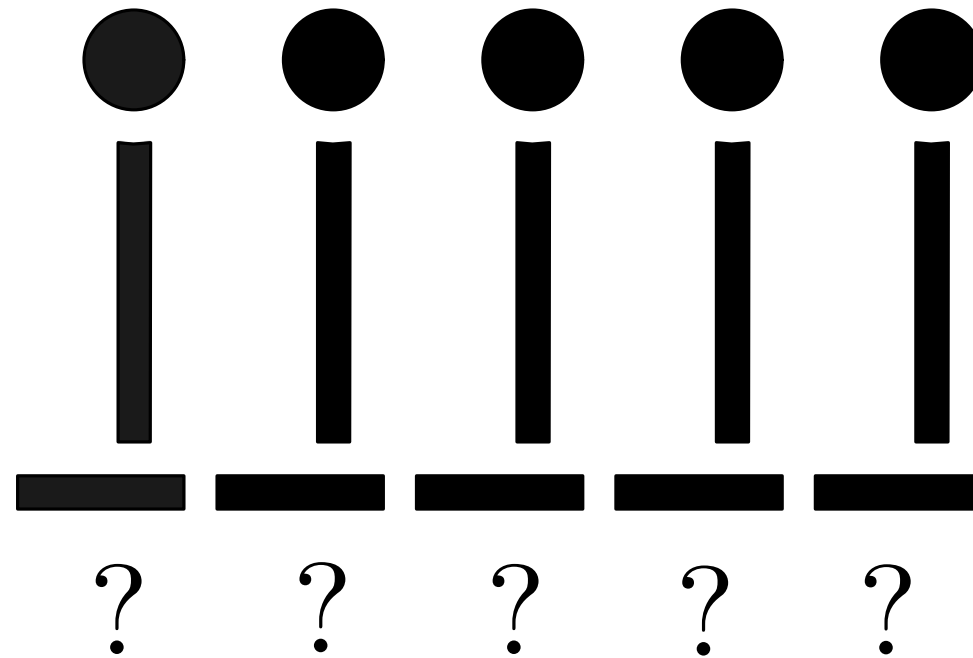
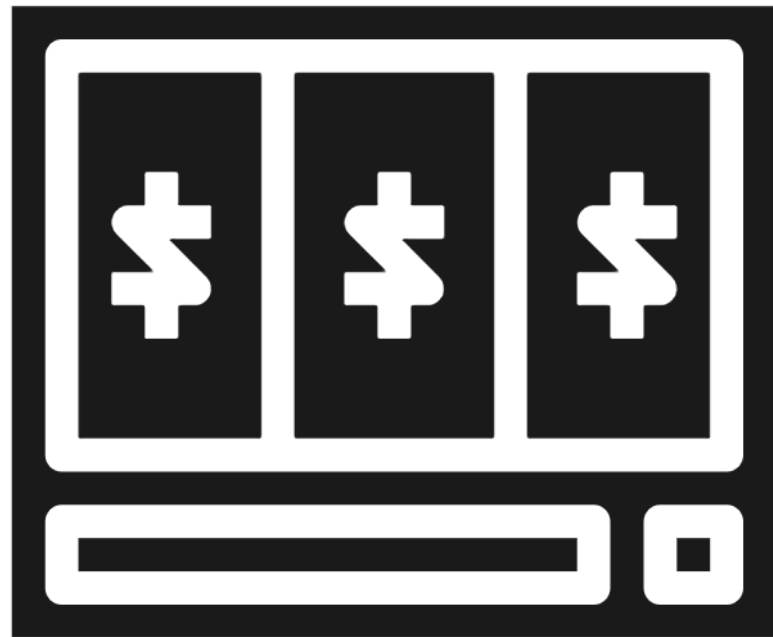


slot machine
one-armed bandit

k -armed Bandit Problem



k -armed Bandit Problem

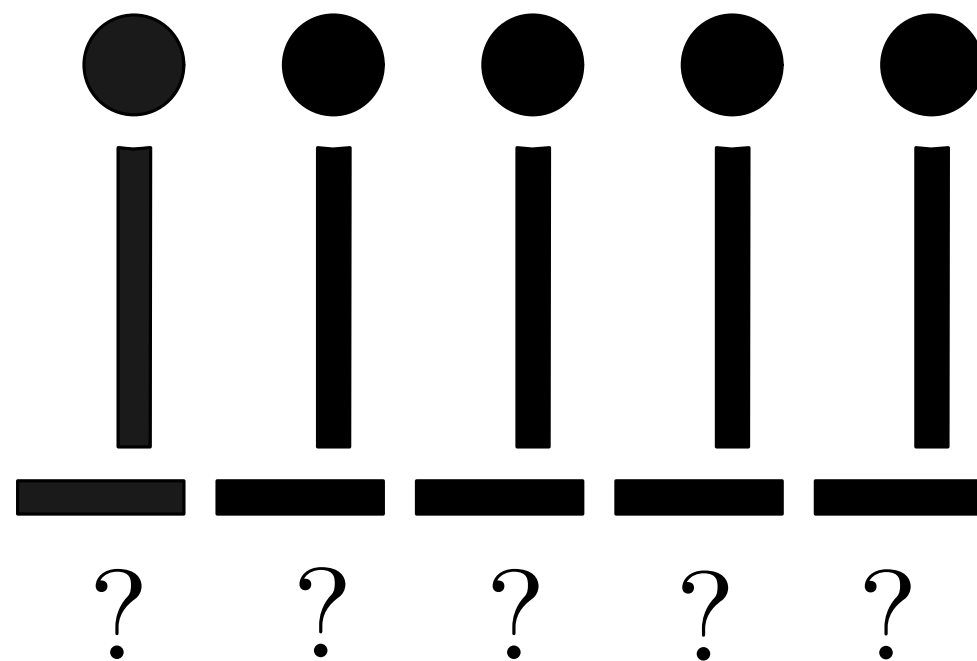


reward



probabilistic

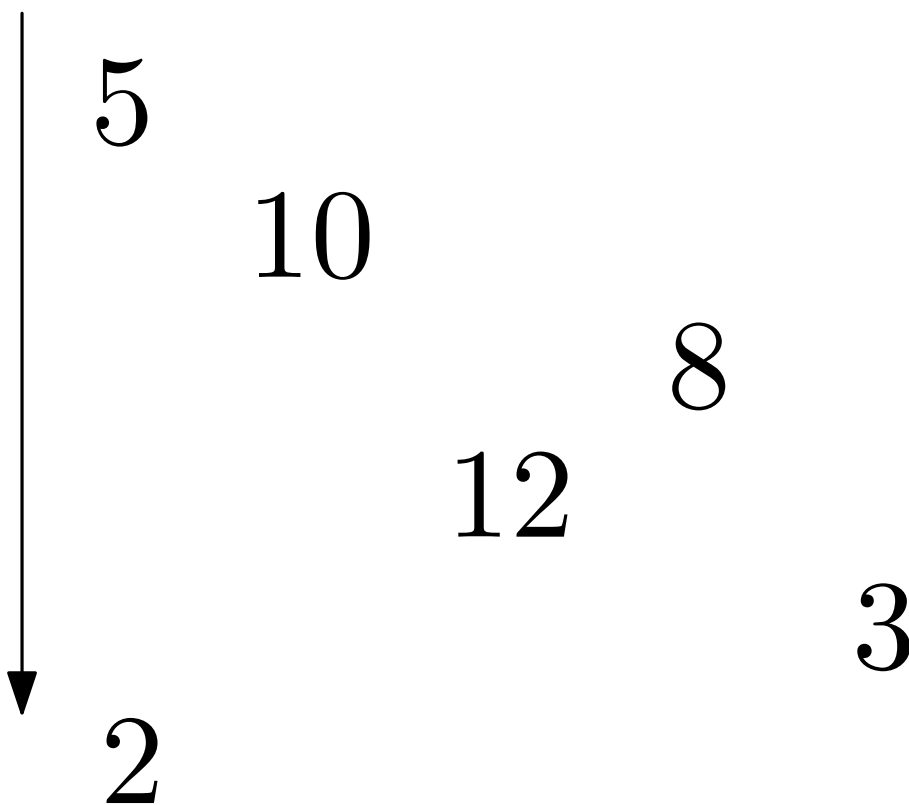
k -armed Bandit Problem



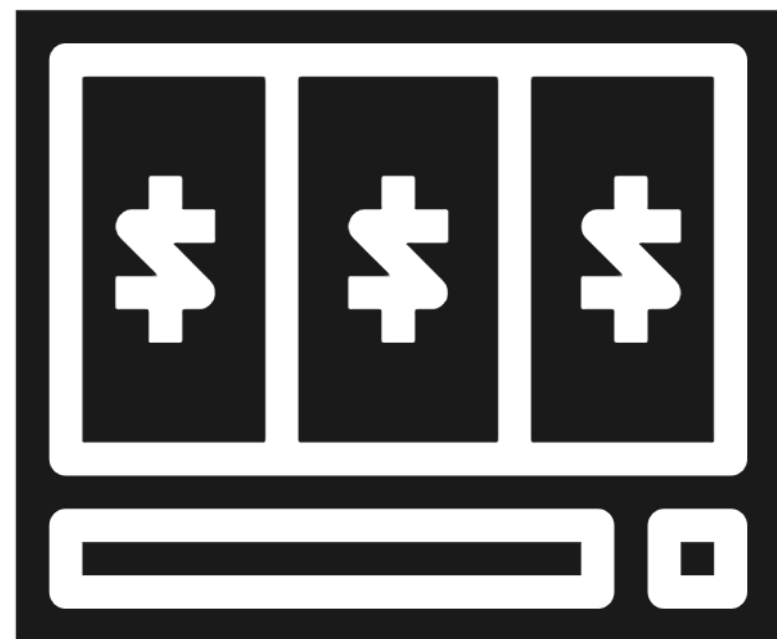
reward



probabilistic



k -armed Bandit Problem



? ? ? ? ?

5

10

8

12

3

2



reward



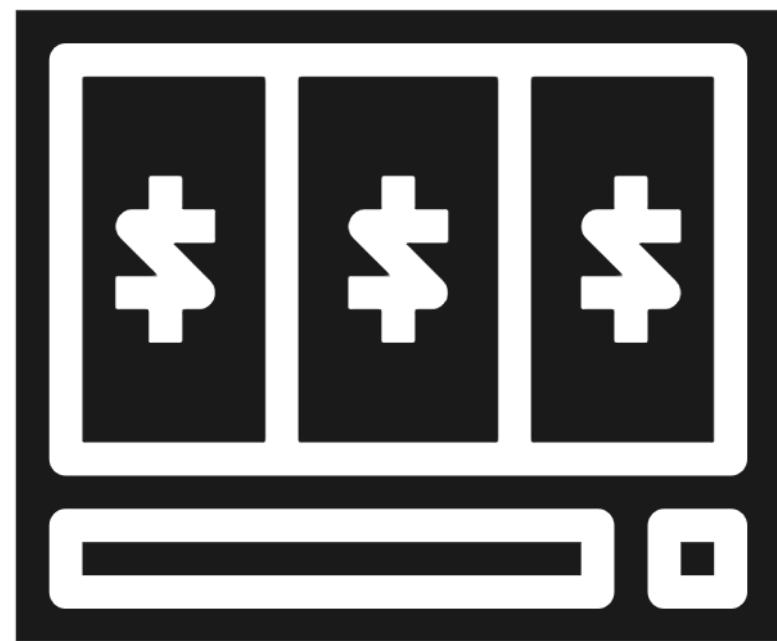
probabilistic



value



k -armed Bandit Problem



? ? ? ? ?

5

10

8

12

3

2



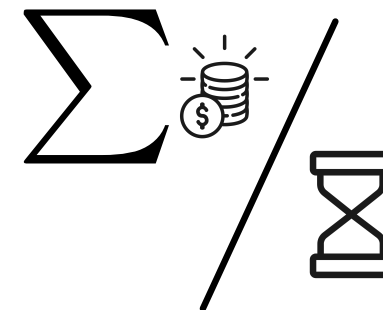
reward



probabilistic



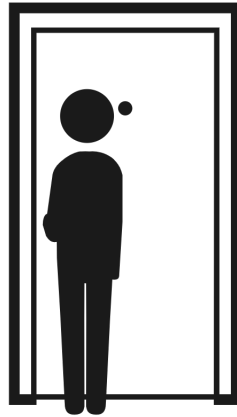
value

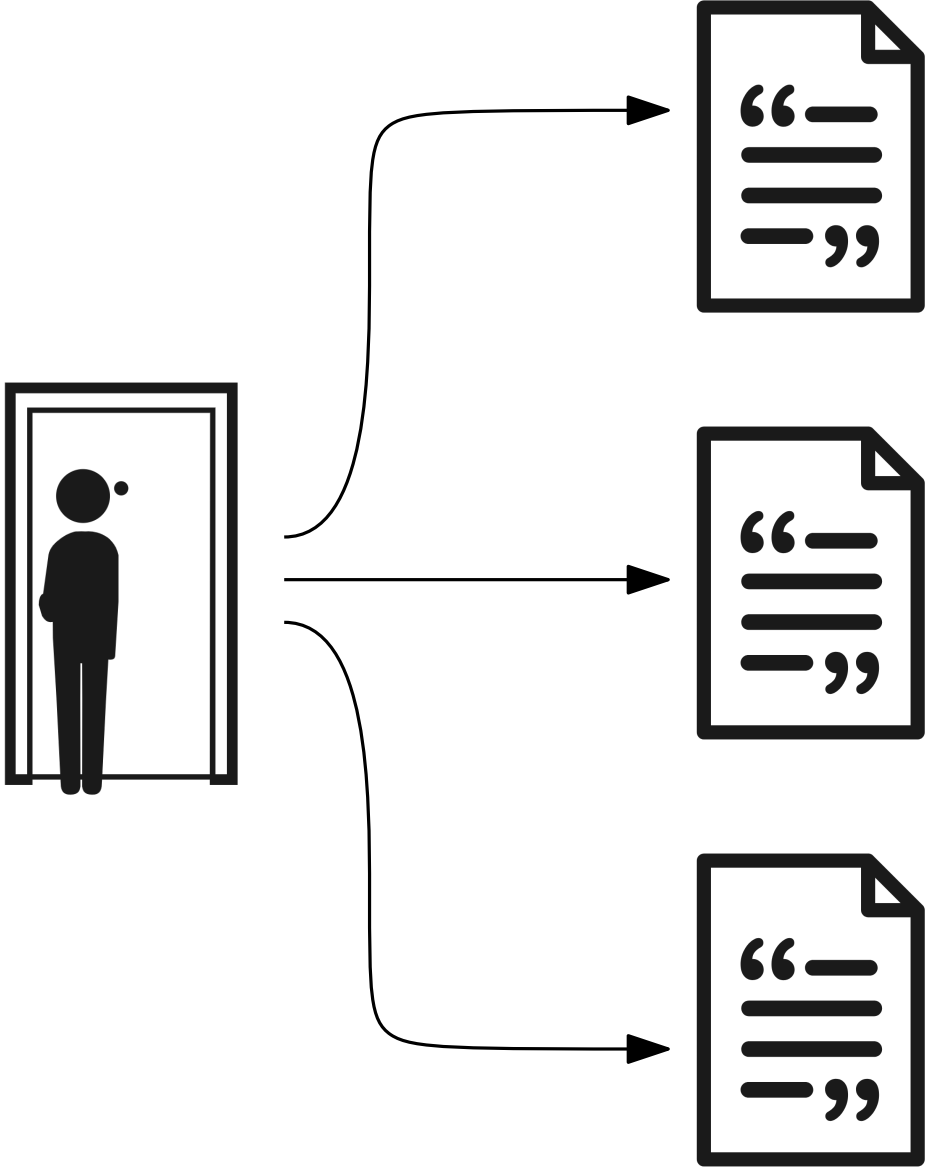
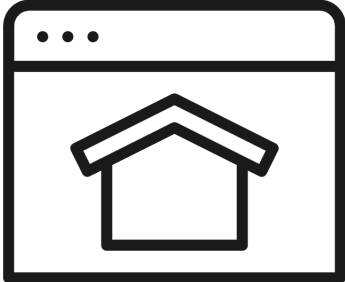


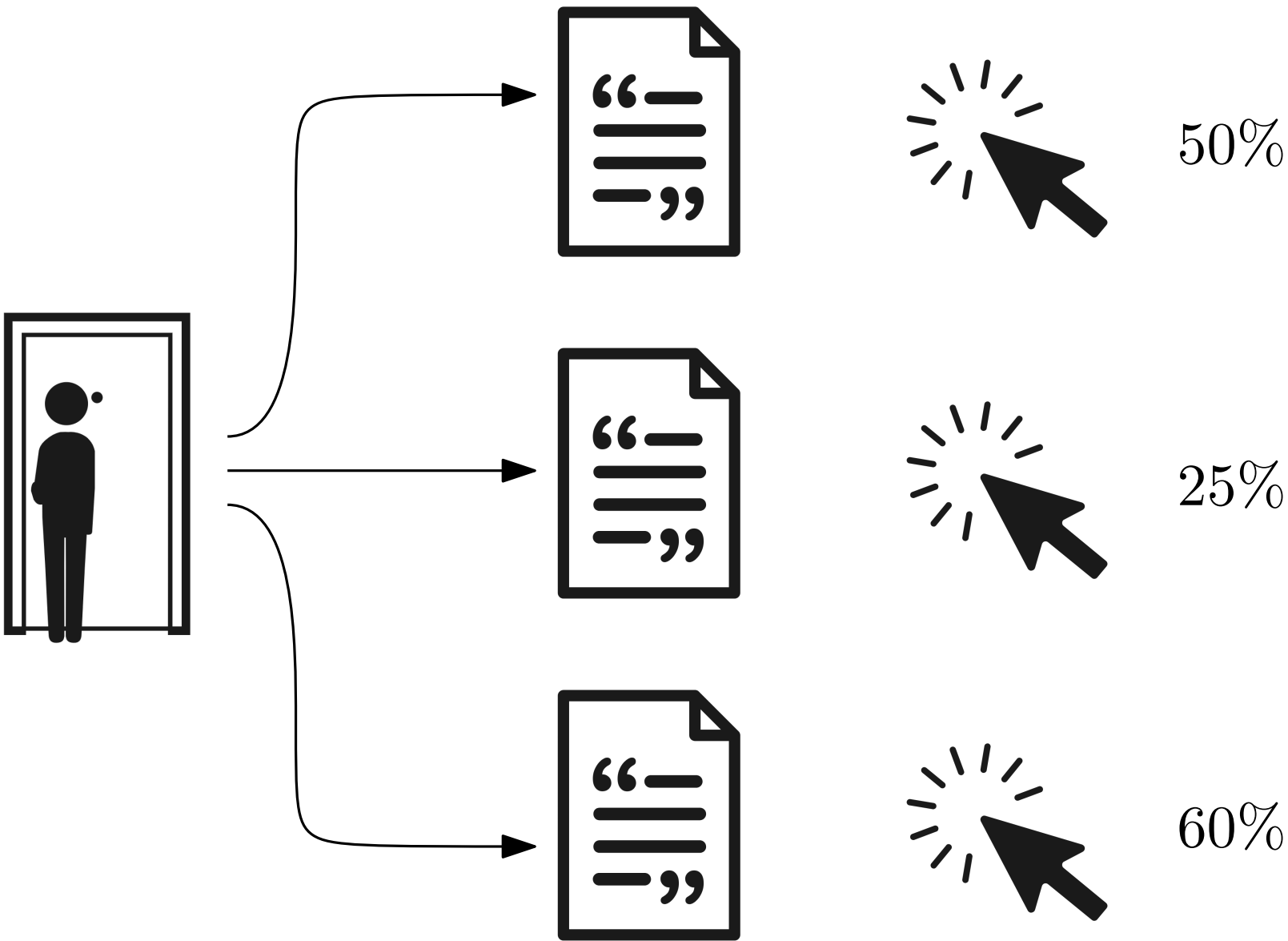
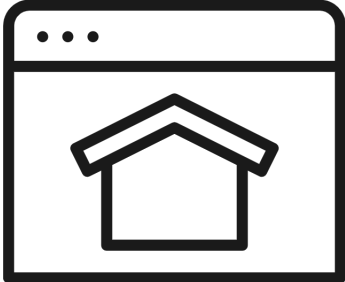
approximate value

Example









Assignment



groupsize $\in [2, 3]$



20 minutes

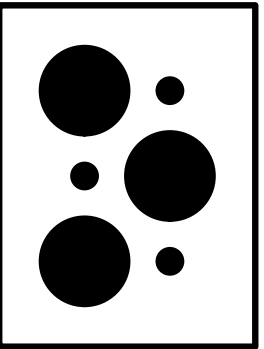
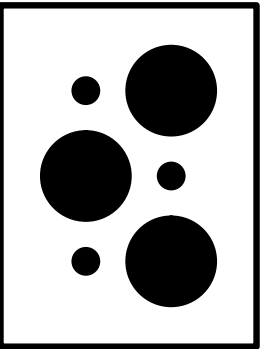
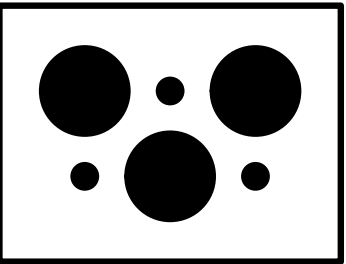
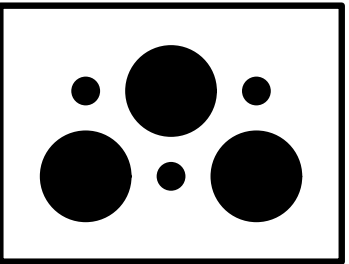
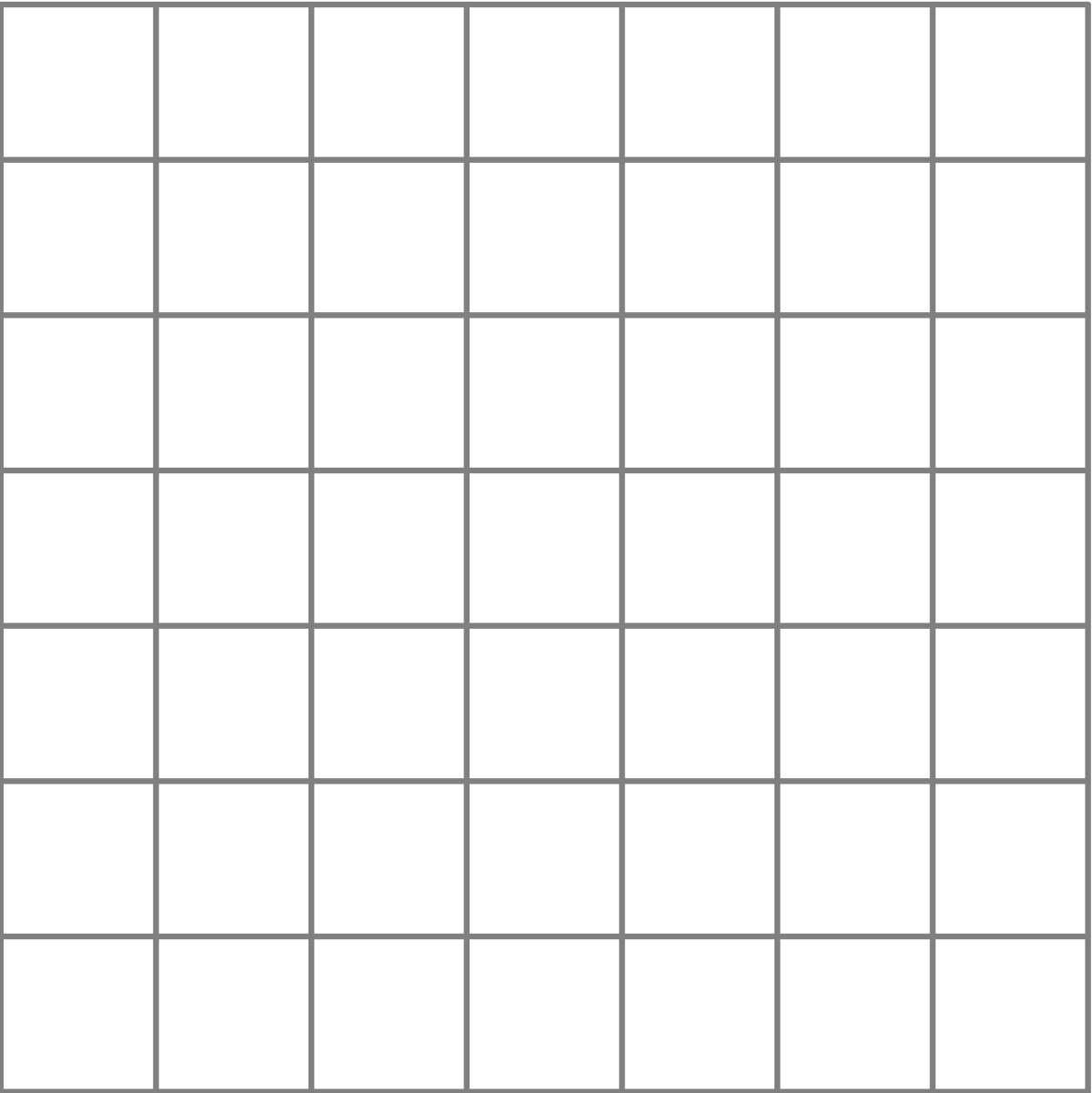
examples?

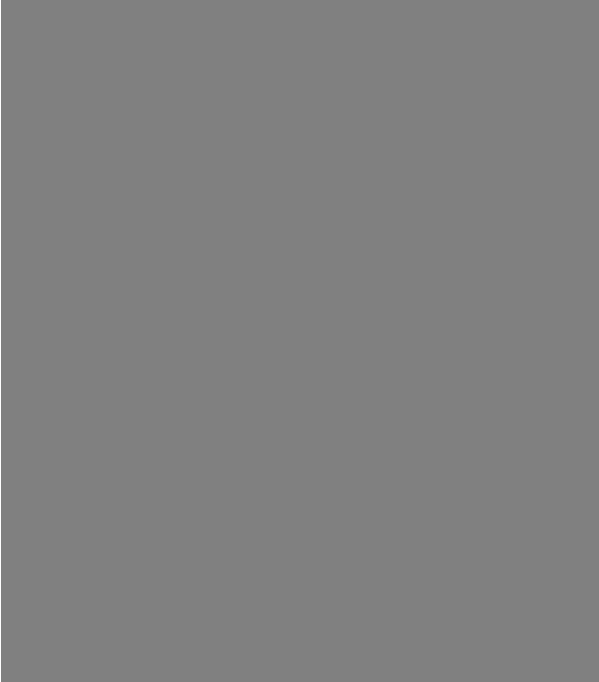
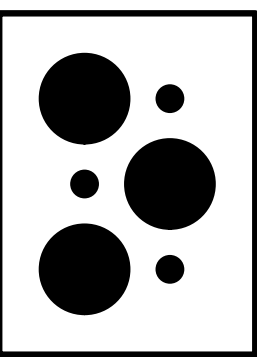
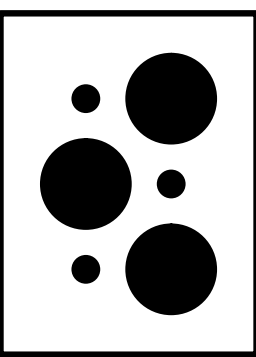
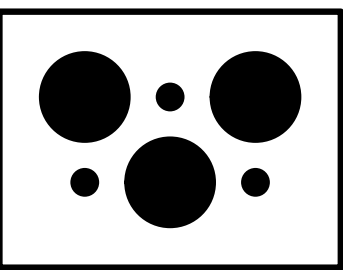
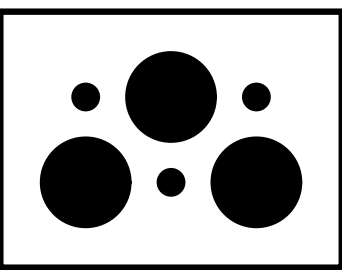
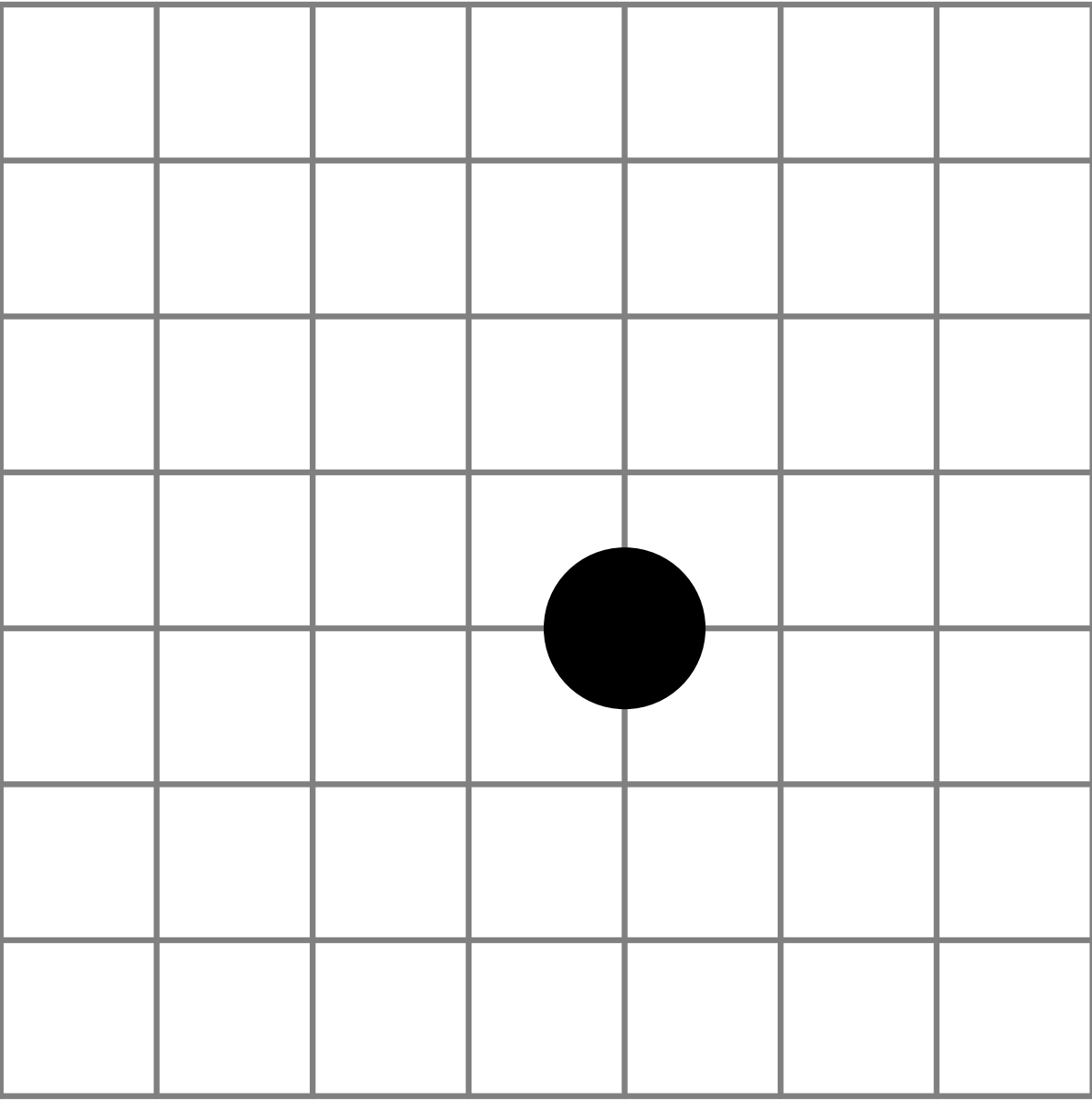
- involving games
- involving farming
- options / distribution / reward / value

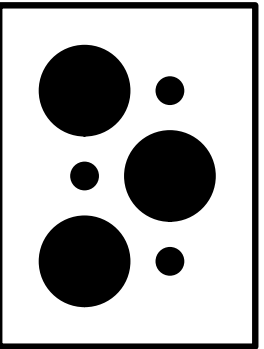
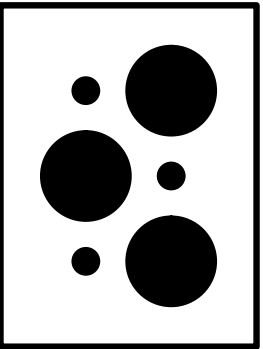
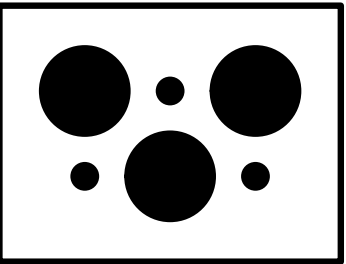
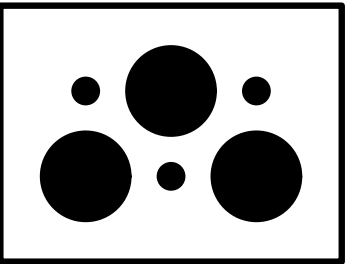
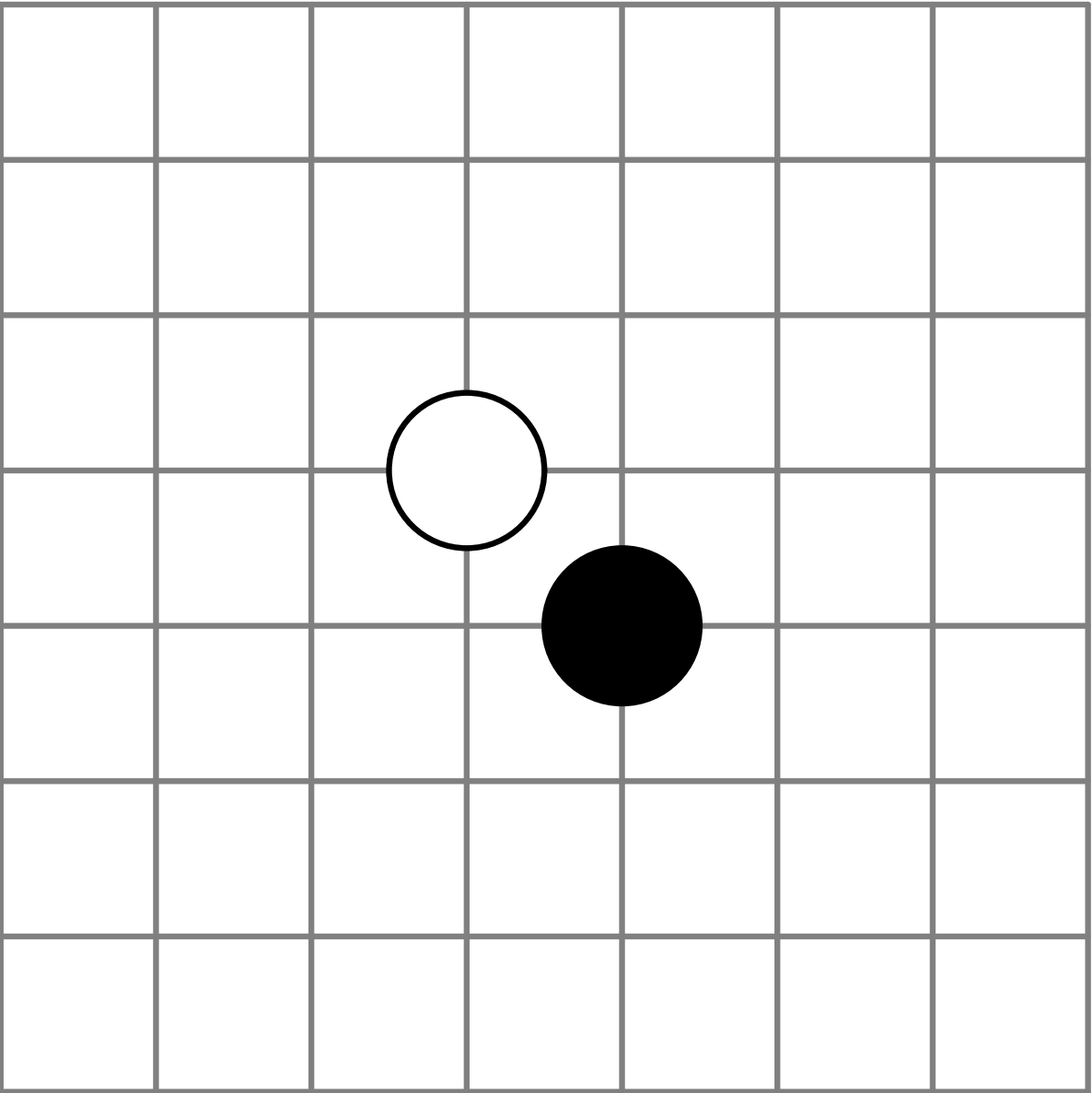
non-examples

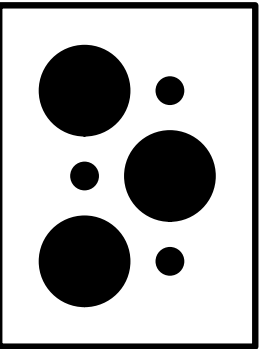
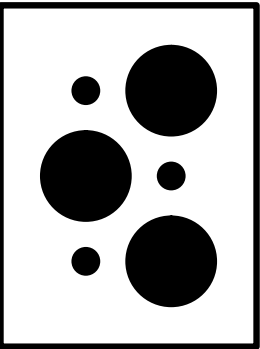
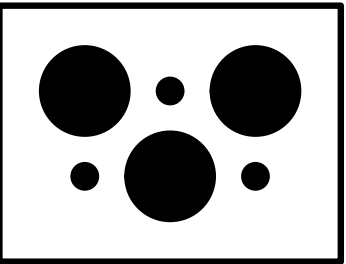
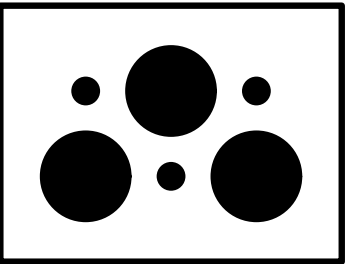
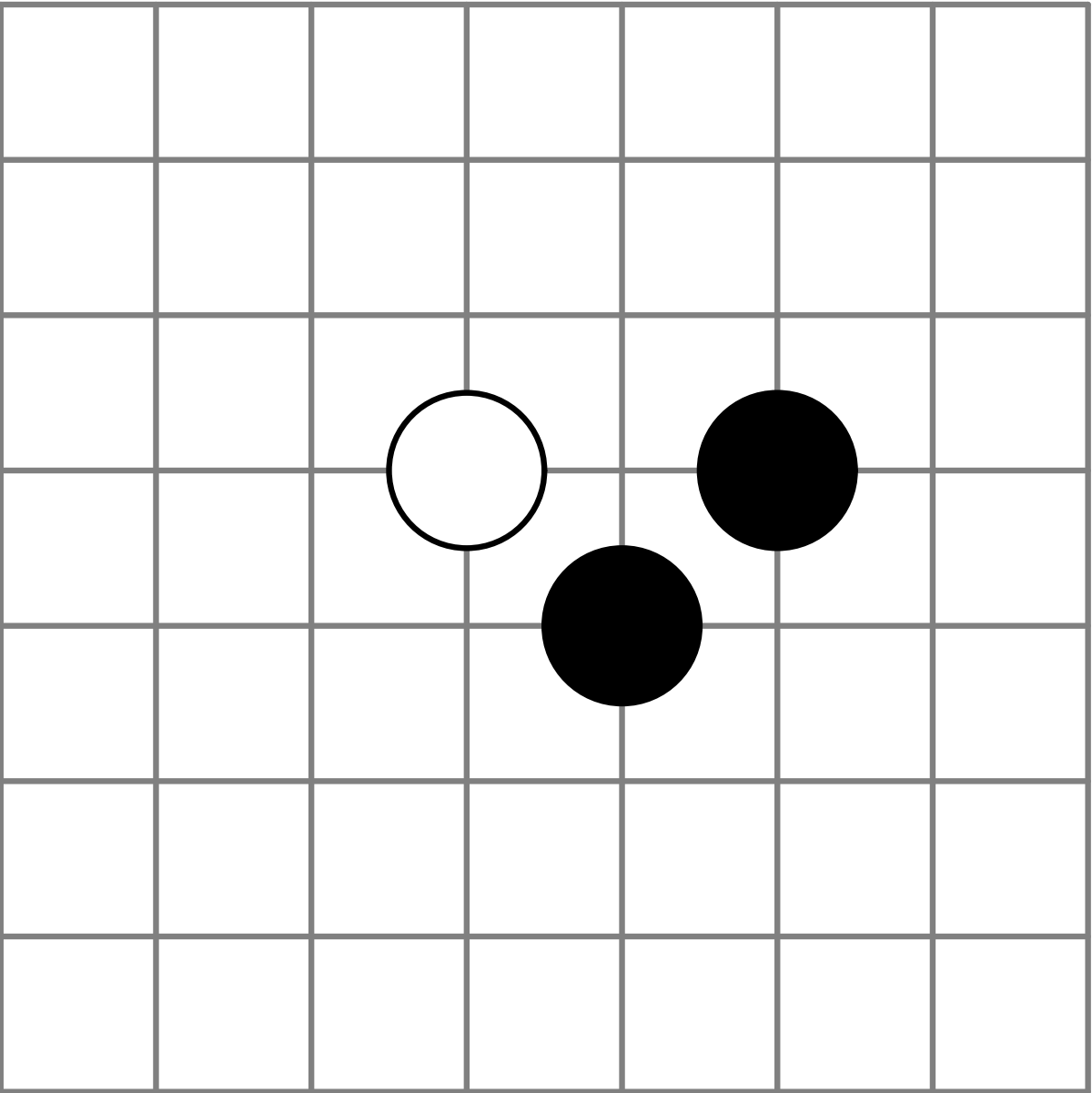
- properties?

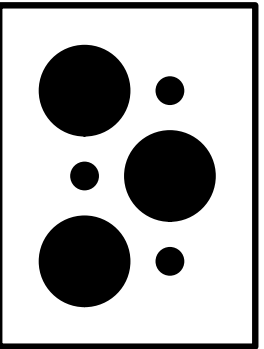
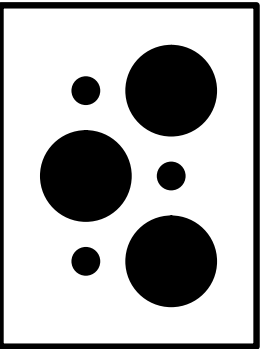
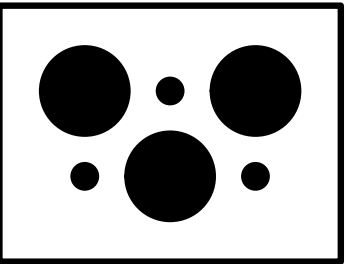
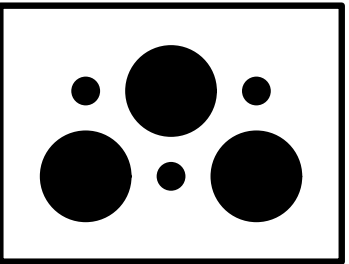
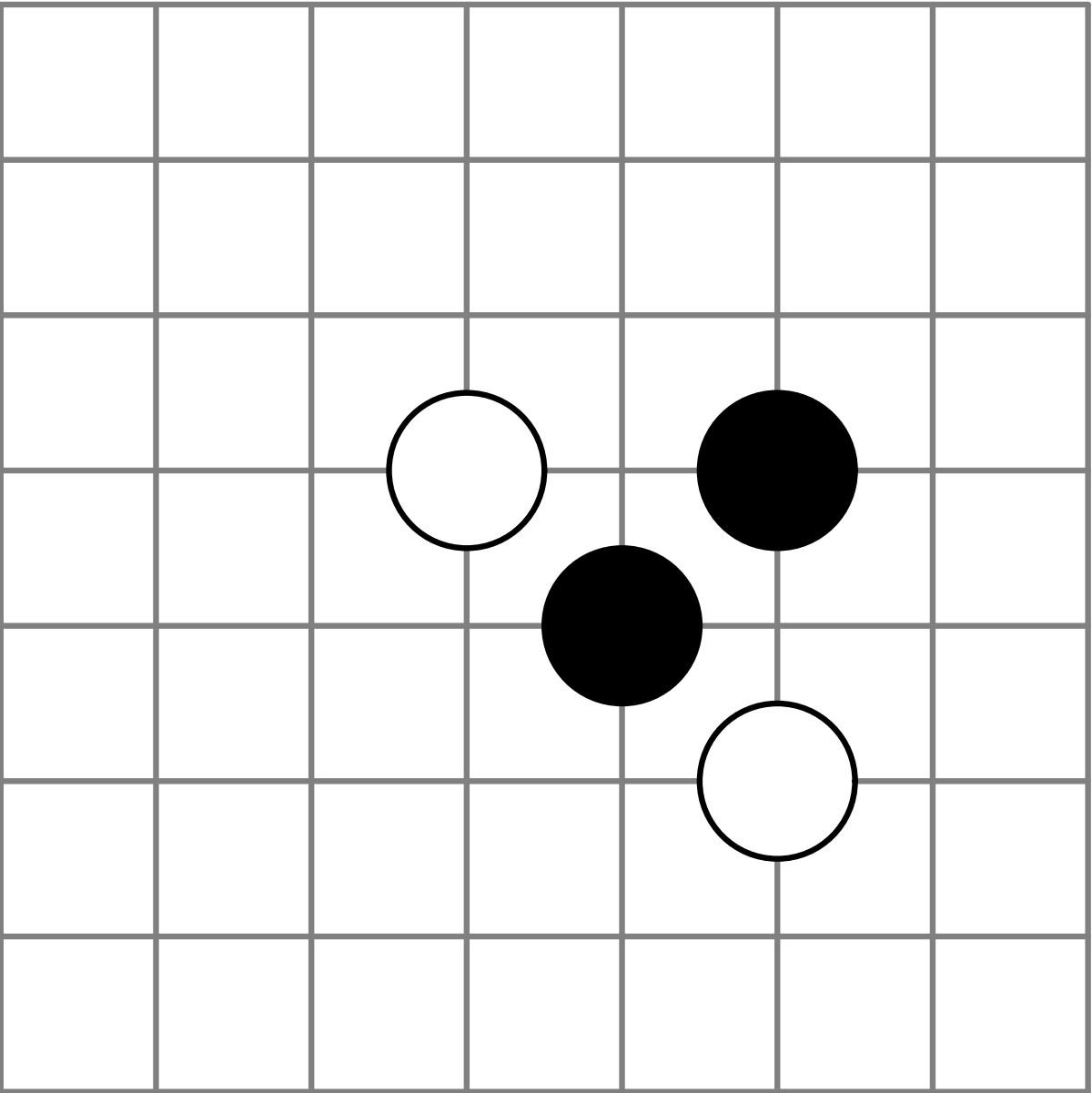
Application

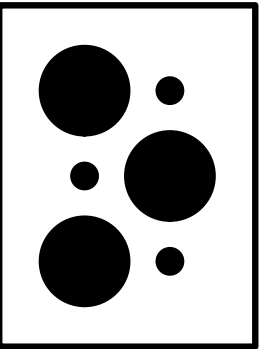
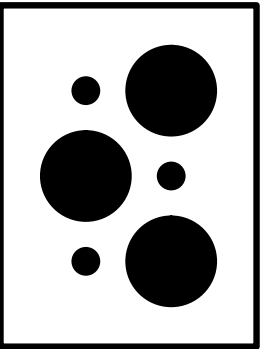
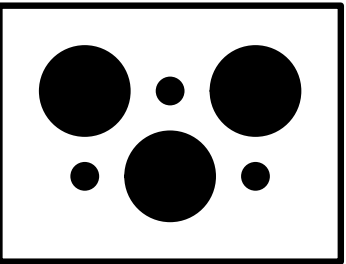
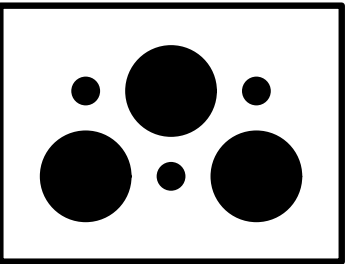
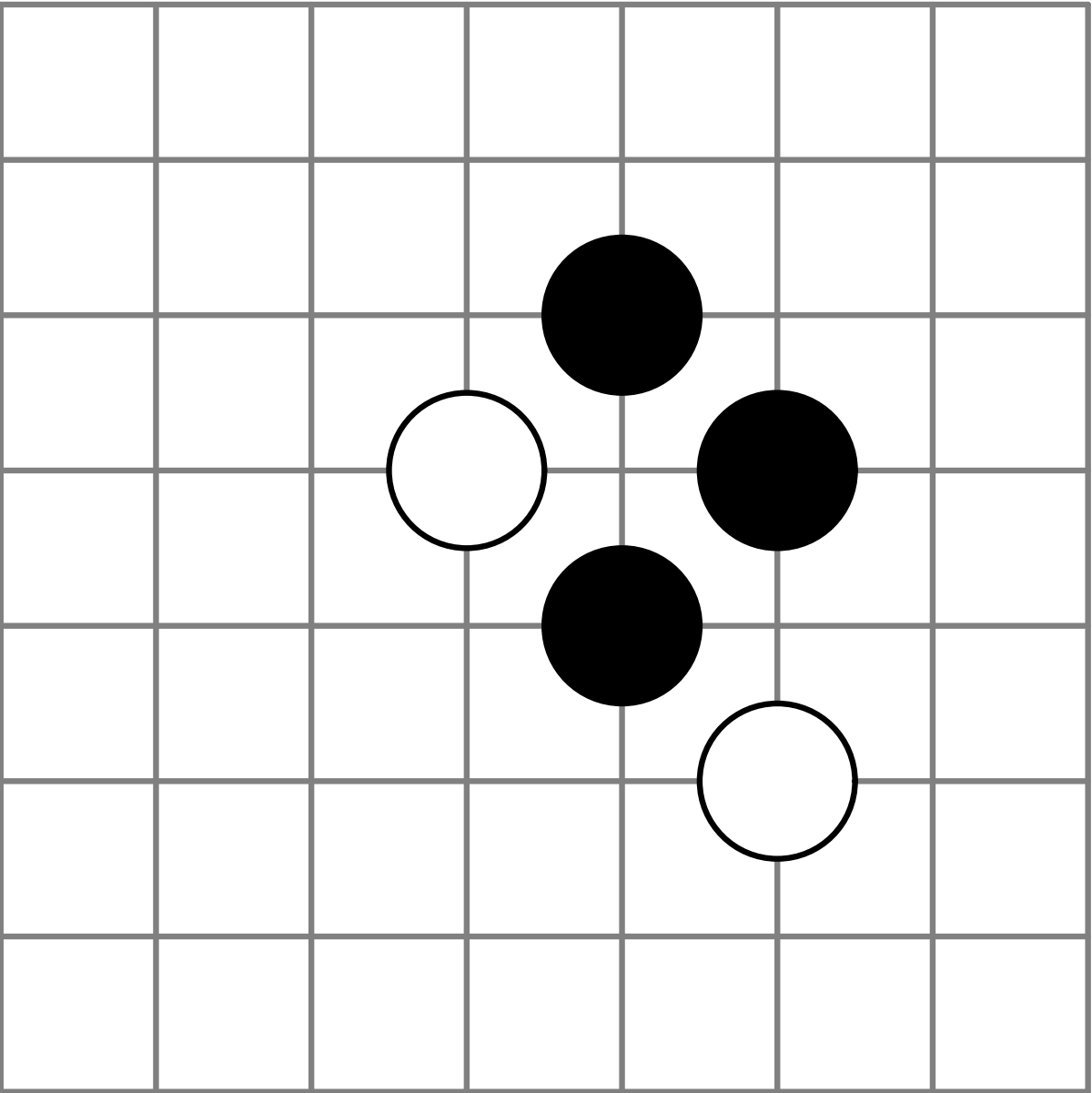


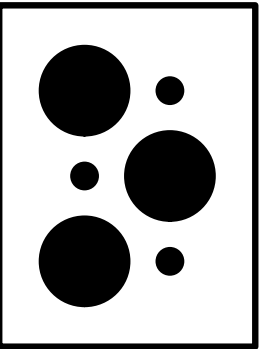
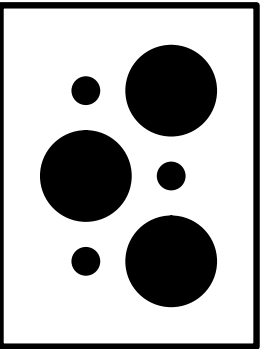
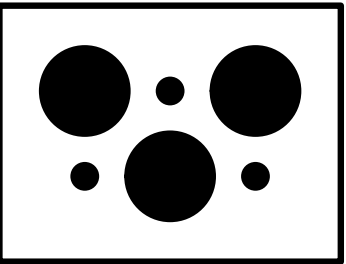
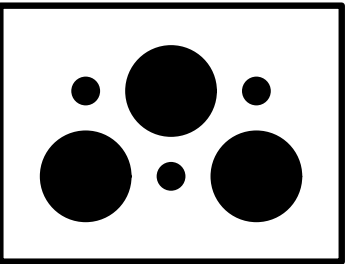
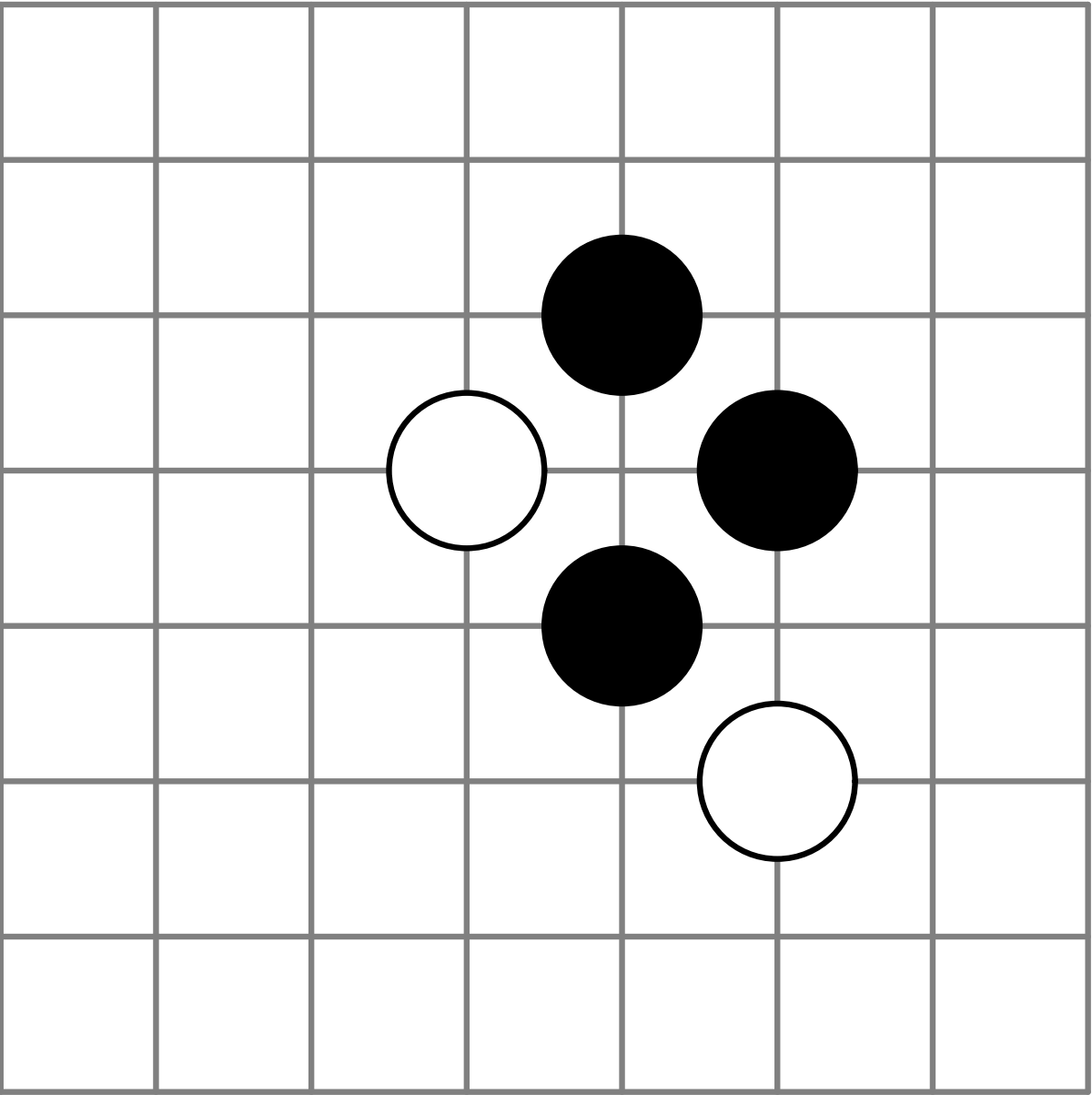




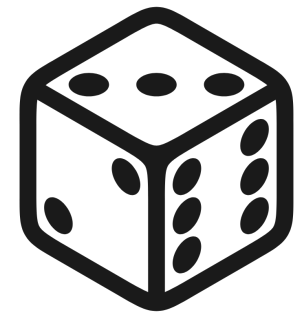










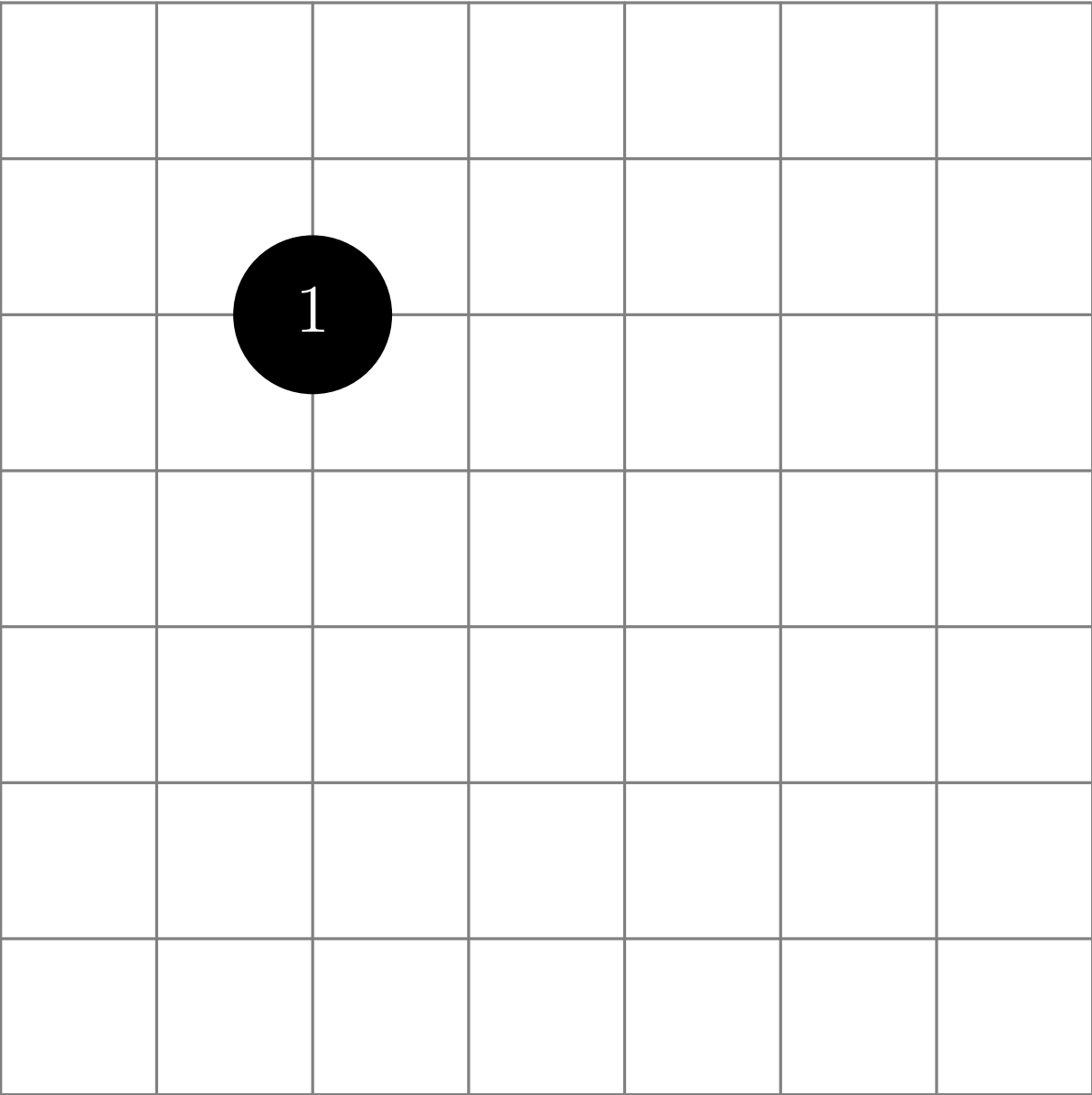


random



agent



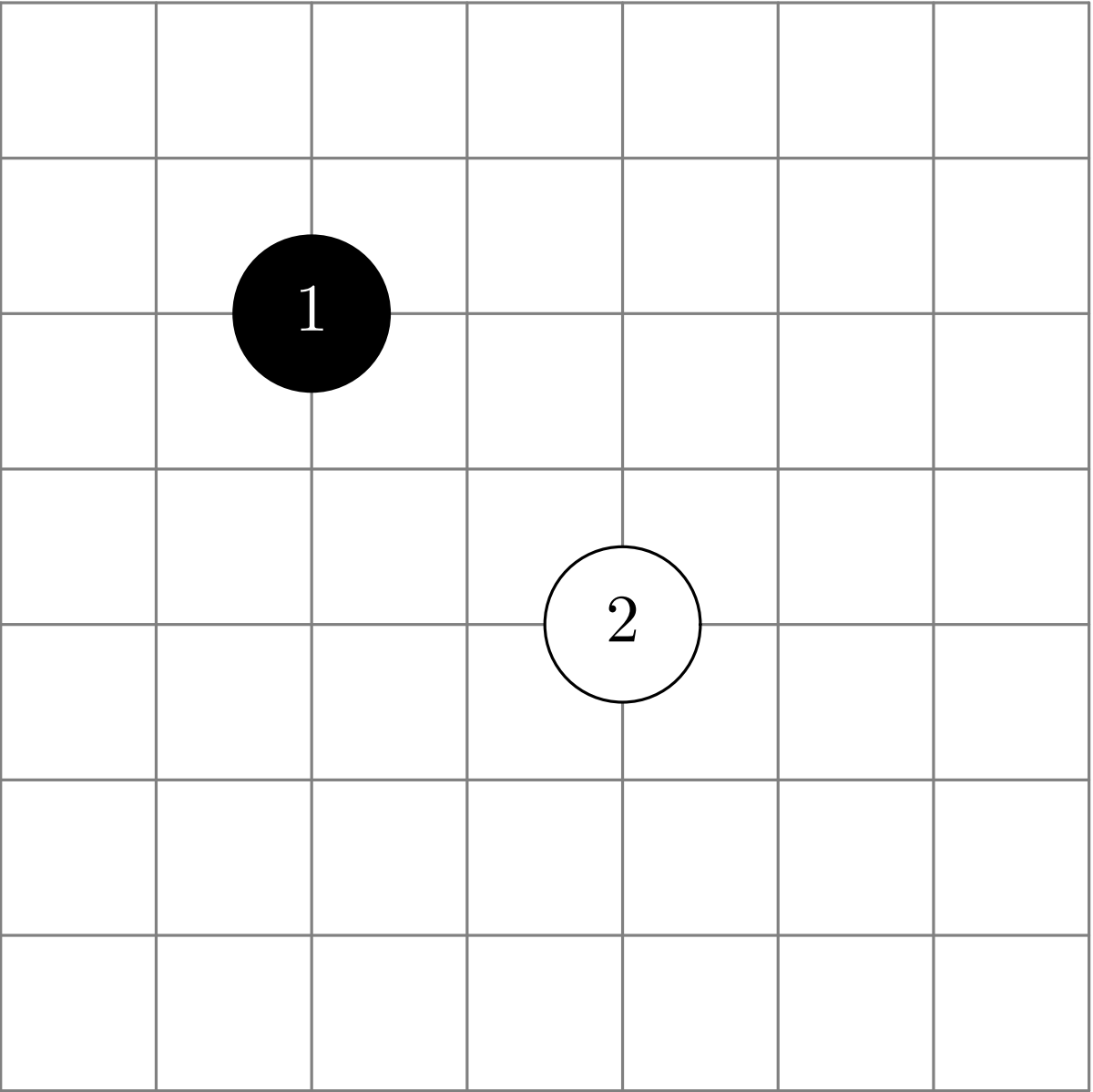


random



agent



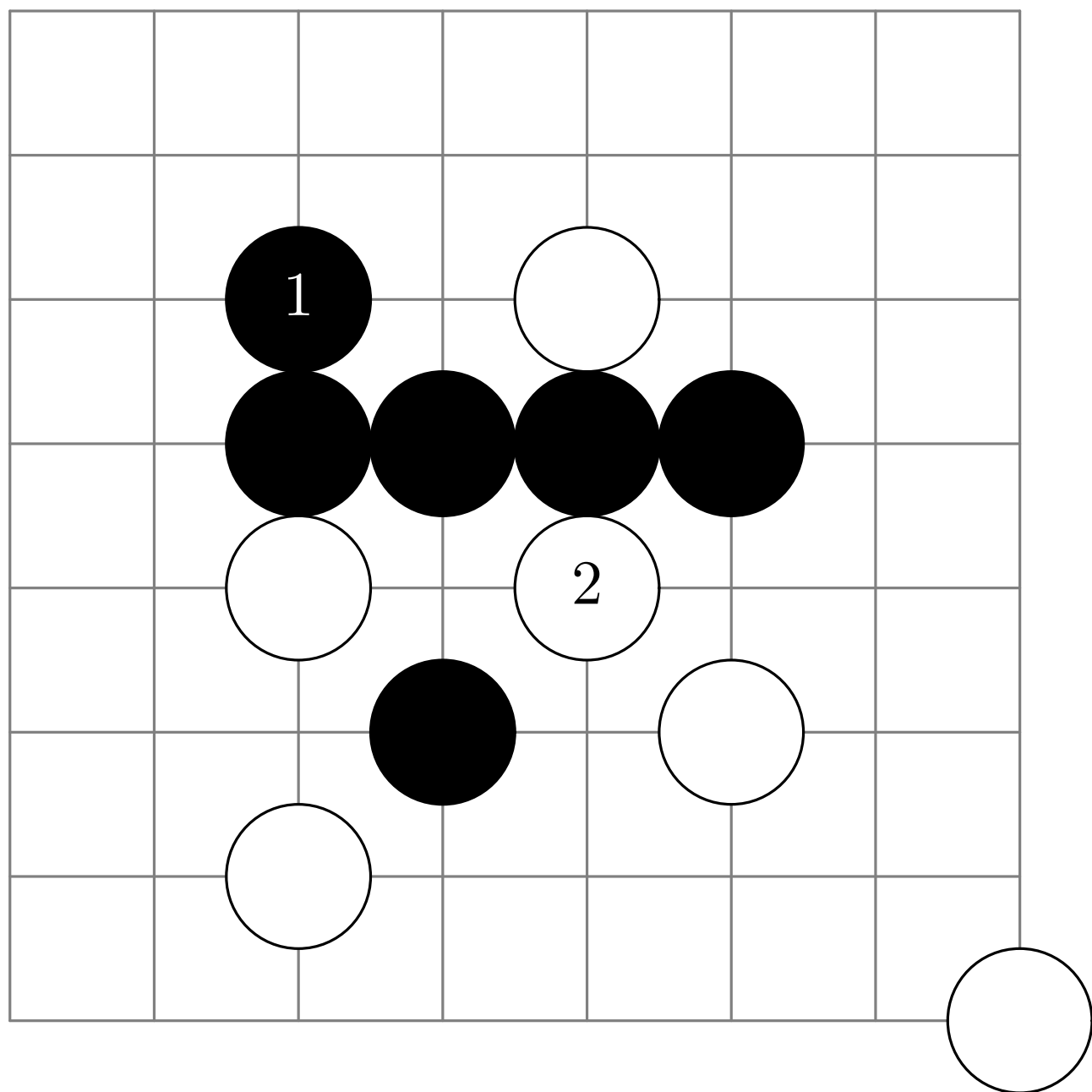


random



agent

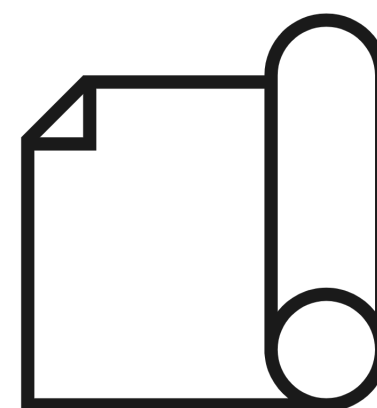




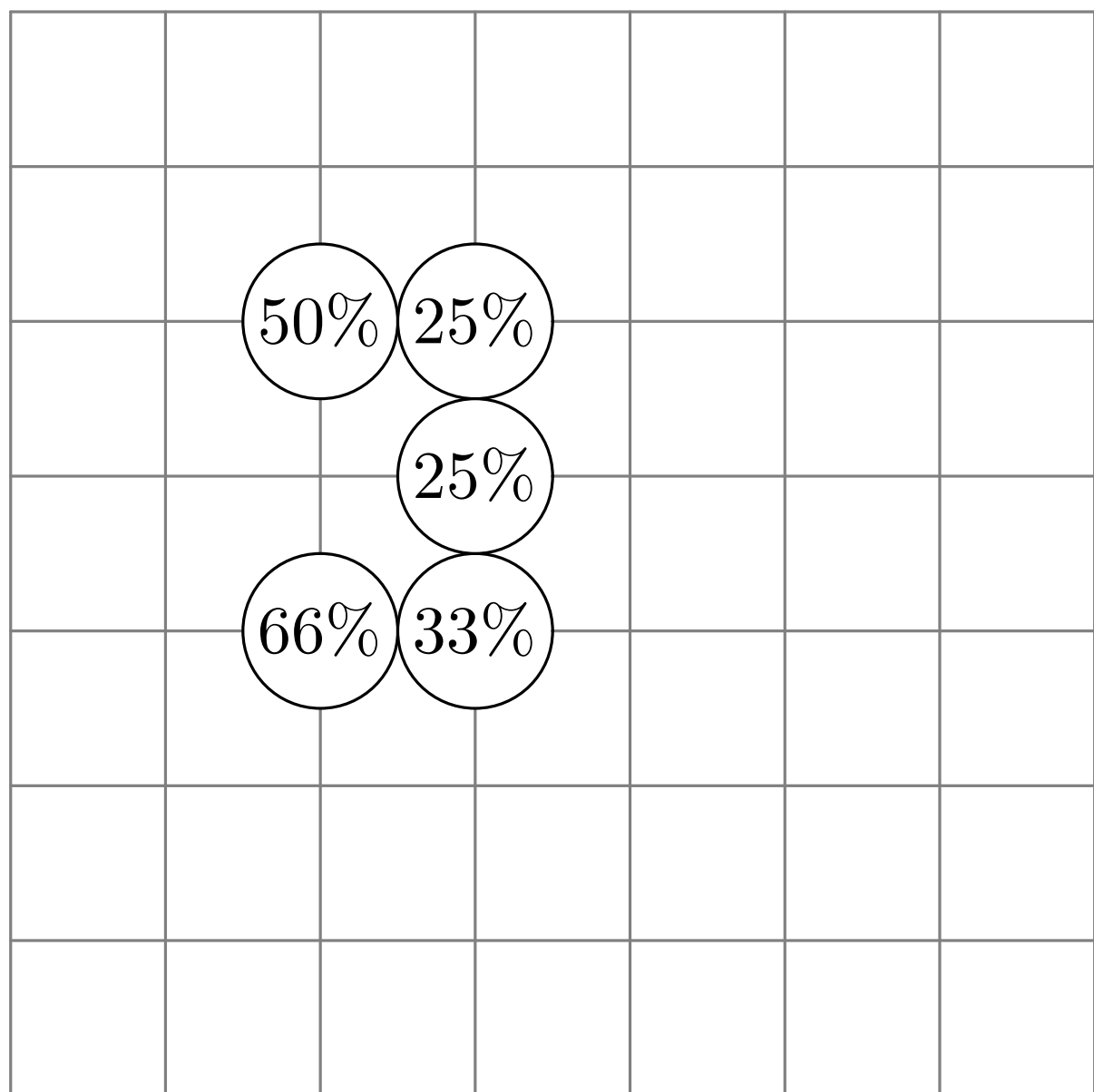
random



agent



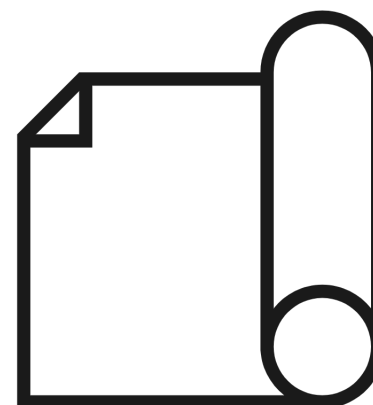
roll out



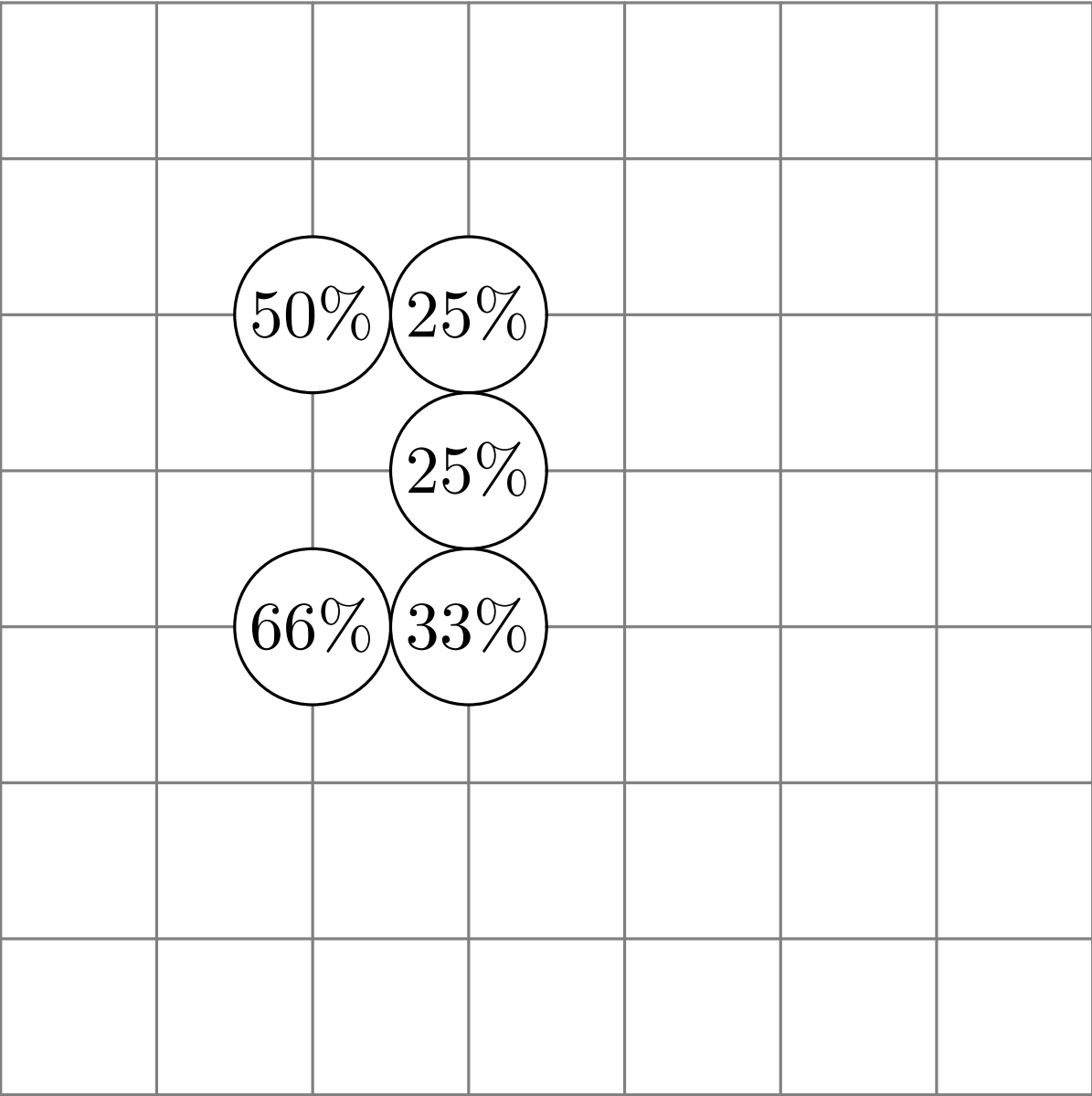
random



agent



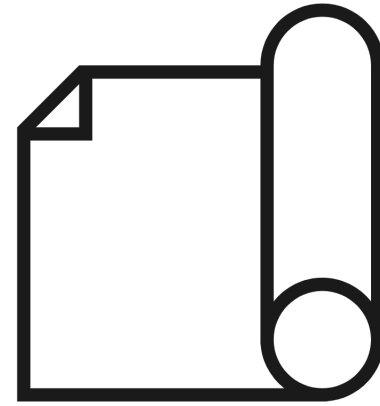
roll out



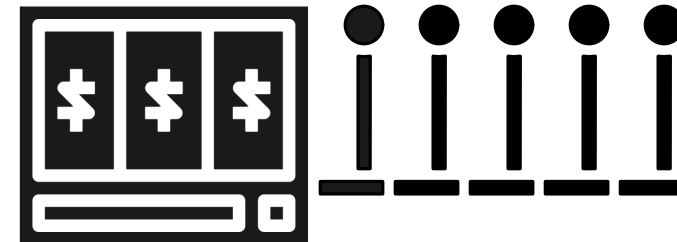
random

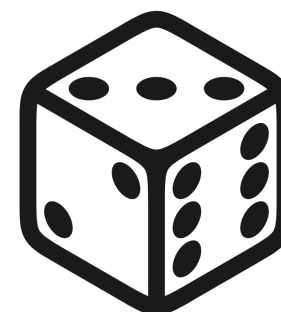
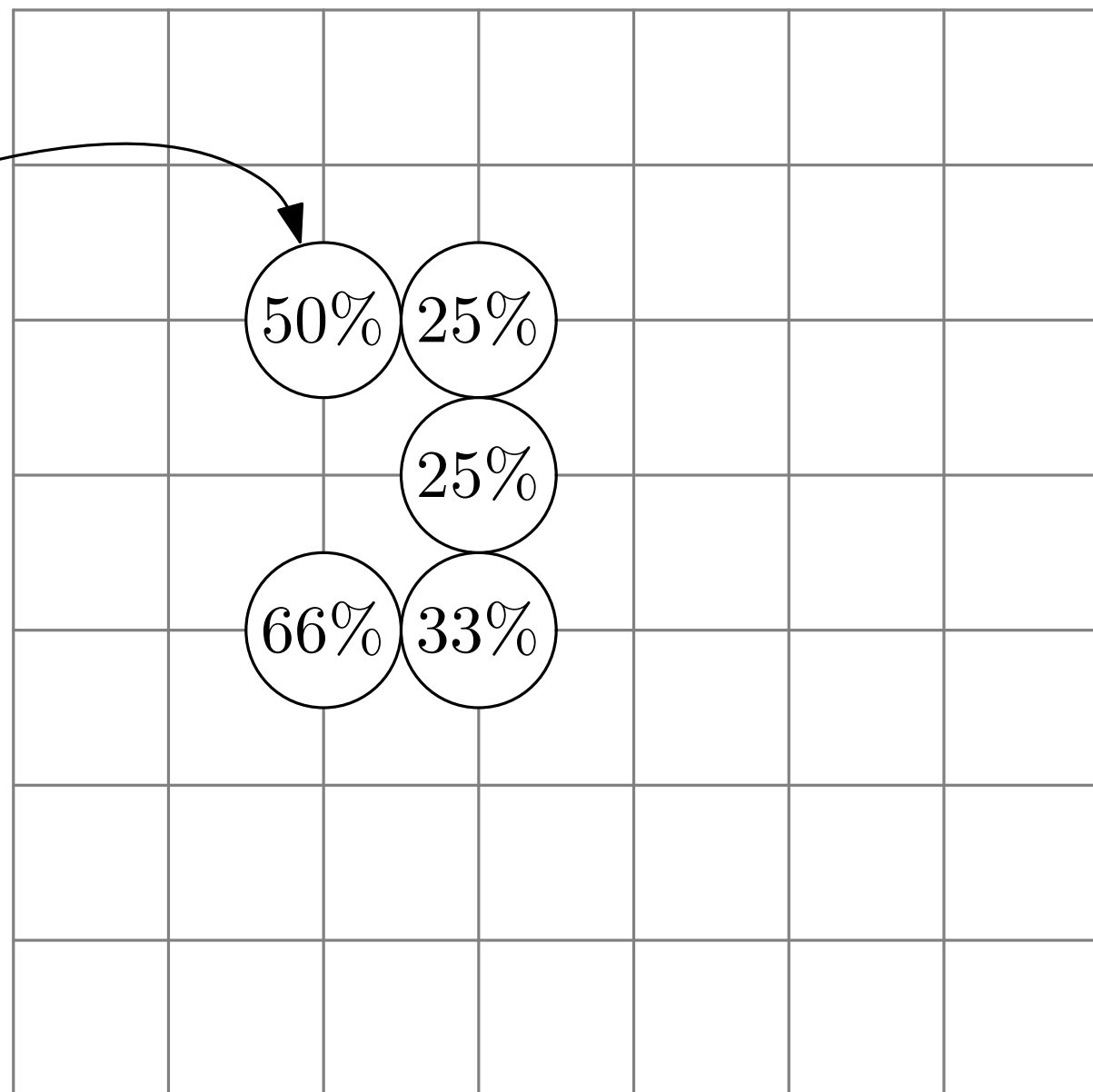


agent



roll out

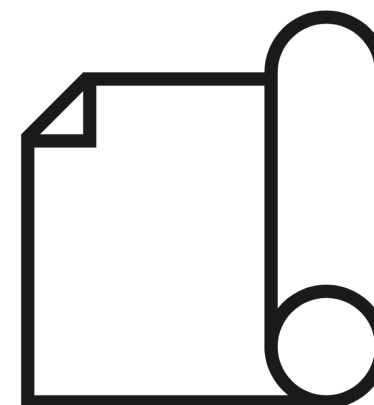




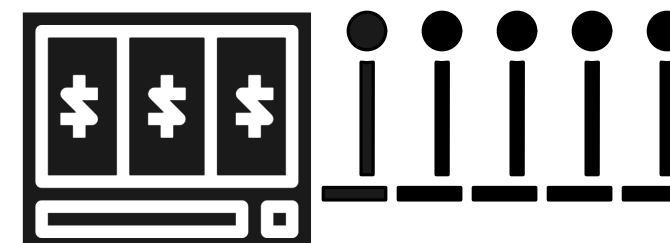
random

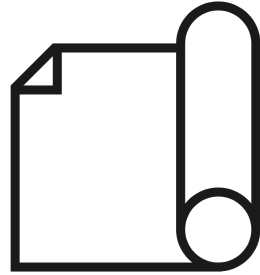
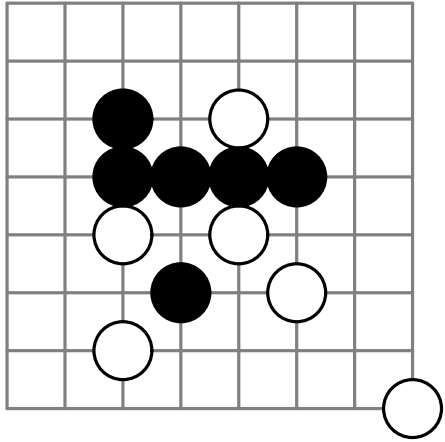


agent

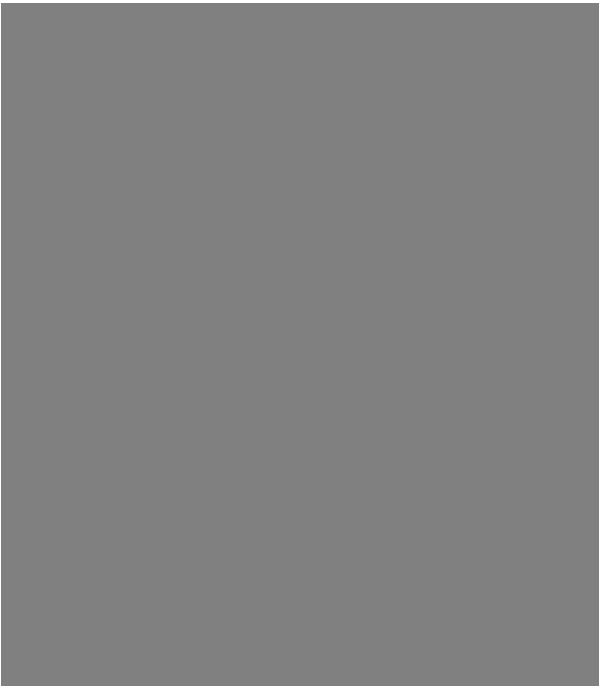
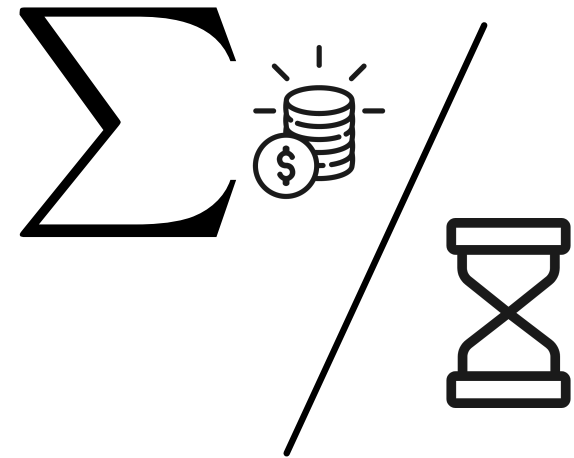
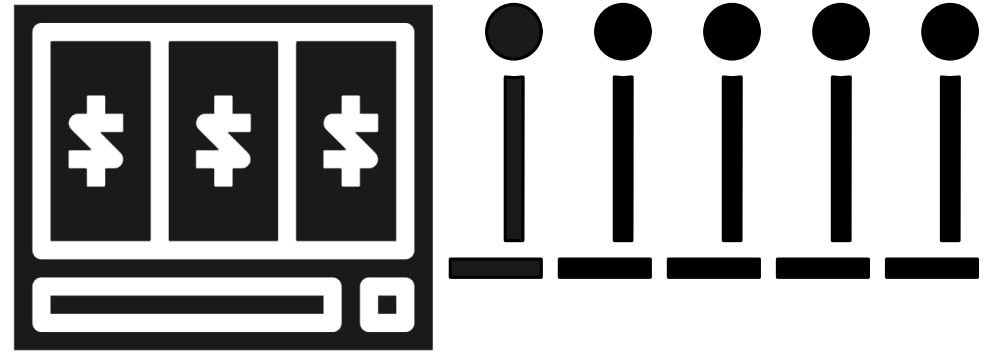


roll out

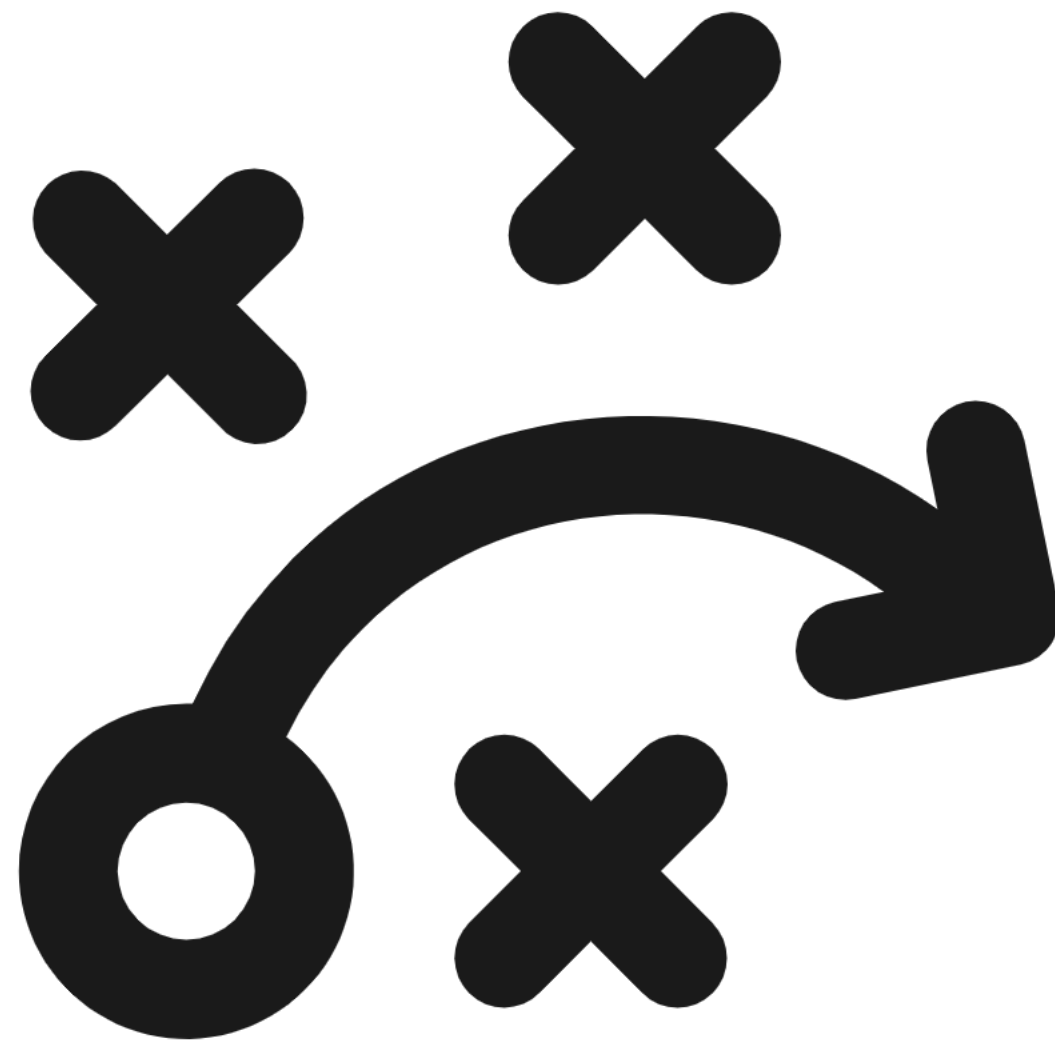




max 



Strategy 1 ε -greedy



ϵ -greedy











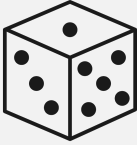































ϵ -greedy

<i>A</i>	3
<i>B</i>	4
<i>C</i>	2
<i>D</i>	6
<i>E</i>	3

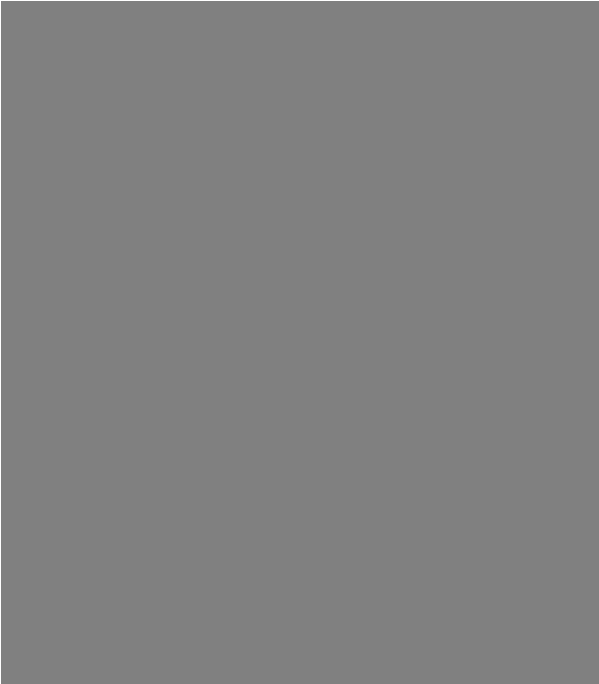


ϵ -greedy

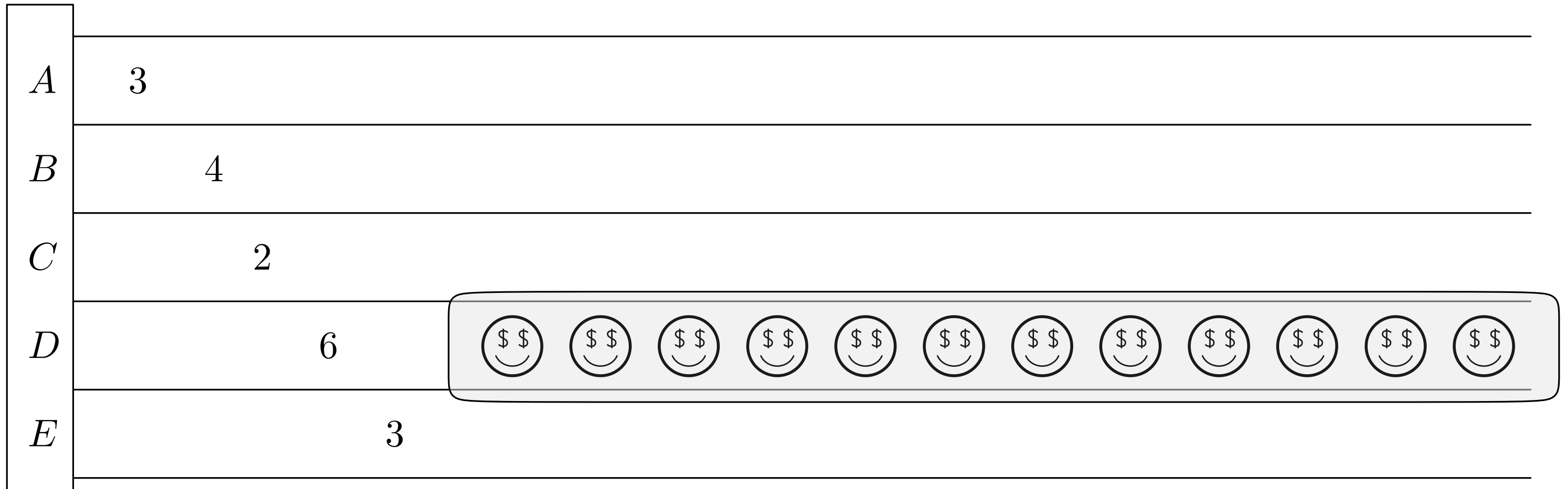
<i>A</i>	3								
<i>B</i>	4								
<i>C</i>	2								
<i>D</i>	6								
<i>E</i>	3								



exploration



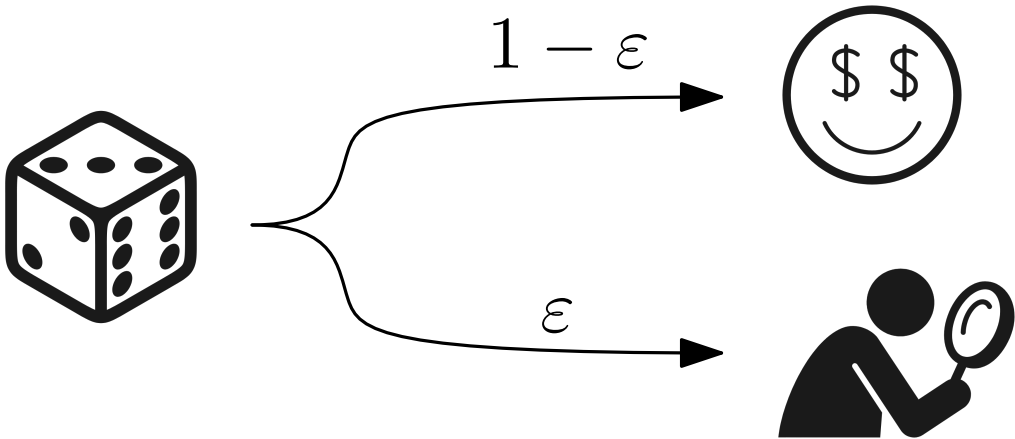
ϵ -greedy



exploitation = greedy


ϵ -greedy

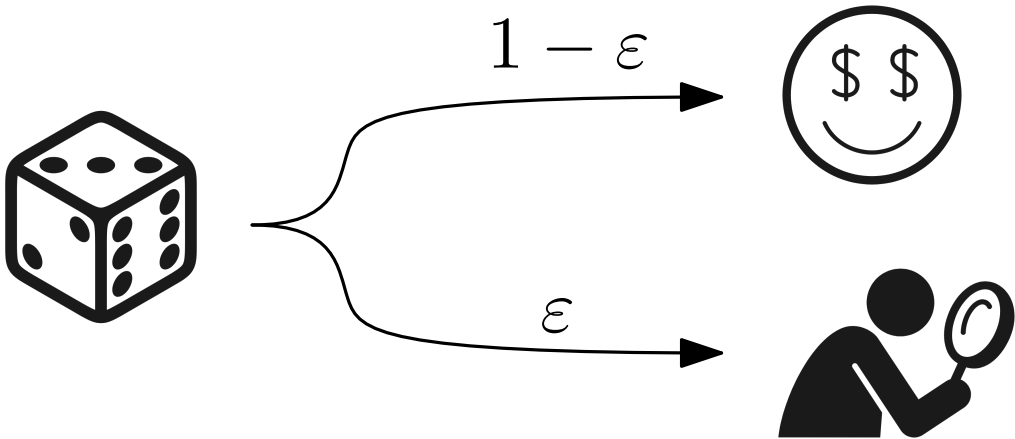
<i>A</i>	3
<i>B</i>	4
<i>C</i>	2
<i>D</i>	6
<i>E</i>	3



ϵ -greedy

<i>A</i>	3
<i>B</i>	4
<i>C</i>	2
<i>D</i>	6
<i>E</i>	3

 = ?



$\epsilon = 0.3$



ϵ -greedy

<i>A</i>	3
<i>B</i>	4
<i>C</i>	2
<i>D</i>	6
<i>E</i>	3



$= 0.76$



$1 - \epsilon$



ϵ



$\epsilon = 0.3$

Assignment



groupsize $\in [2, 3]$



20 minutes

What happens if $\varepsilon = 0$, $\varepsilon = 1$?



Why balance?



Plot return over time

$\varepsilon = 0$, $\varepsilon = 1$, $\varepsilon = 0.2$, $\varepsilon = 0.1$.



Details



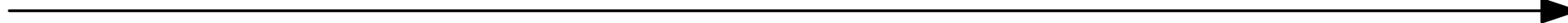
?

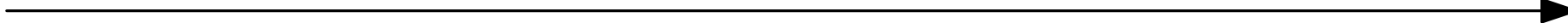
?

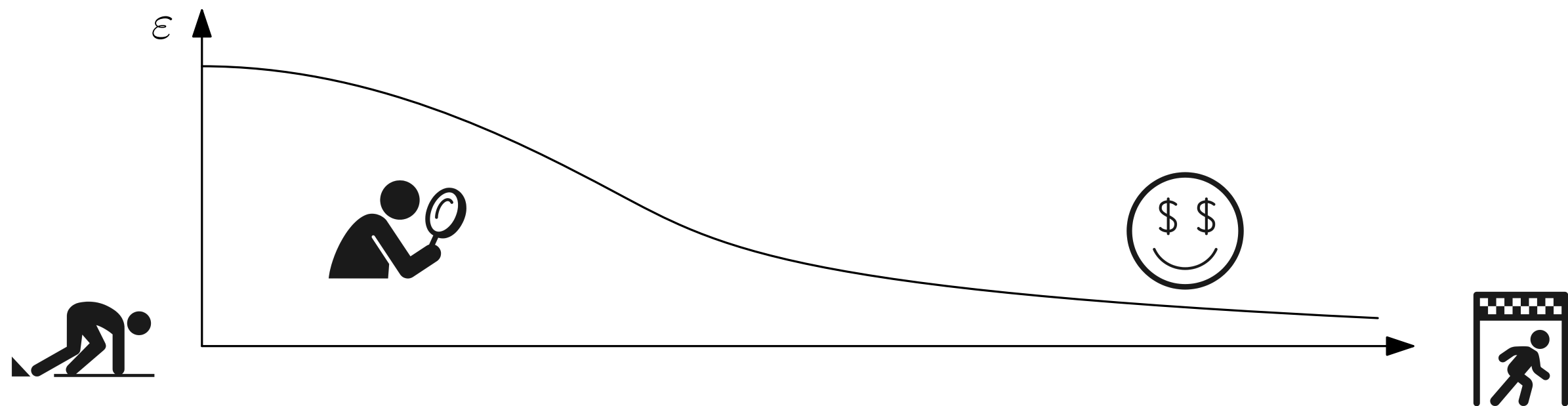
?

?

?





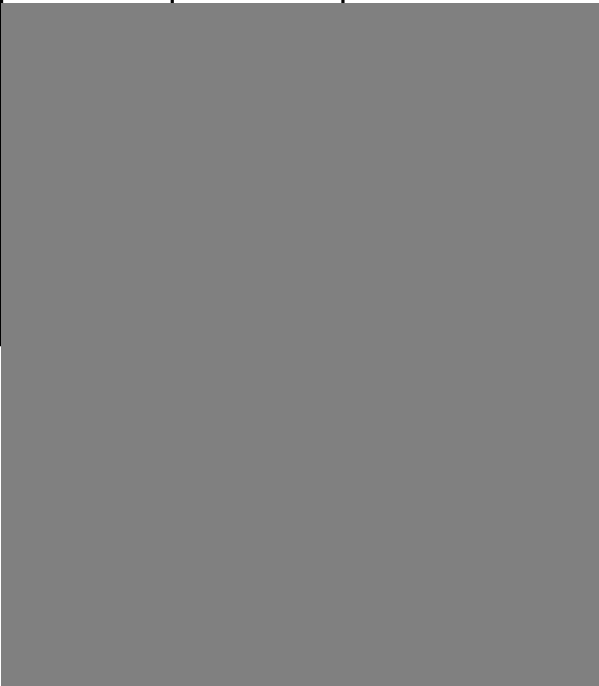


ϵ -decay





<div><div>A</div><div>B</div><div>C</div><div>D</div><div>E</div></div>	<div><div><div>Σ</div><div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div></div></div></div>														
	<div><div><div>Σ</div><div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div></div></div></div>														
	<div><div><div>Σ</div><div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div></div></div></div>														
	<div><div><div>Σ</div><div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div></div></div></div>														
	<div><div><div>Σ</div><div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div><div><div><div></div><div></div><div></div></div></div></div></div></div>														

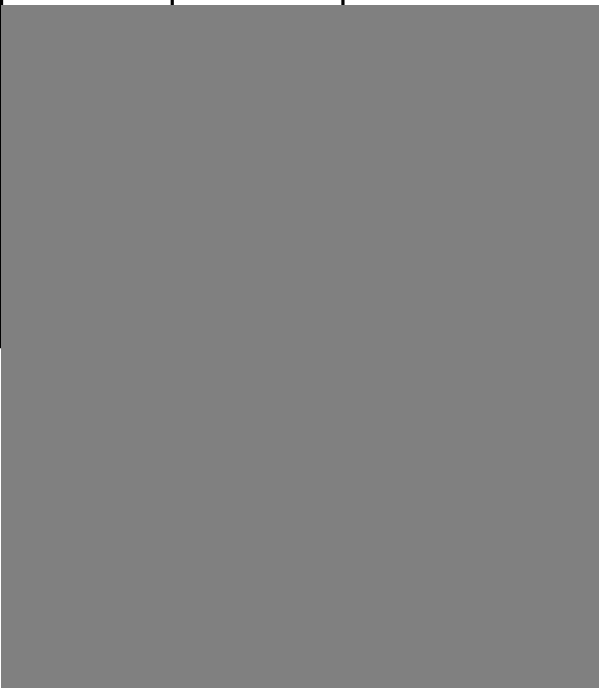




<i>A</i> <i>B</i> <i>C</i> <i>D</i> <i>E</i>	$\Sigma \text{💡} / \text{Ⓢ} \quad 0$															
	$\Sigma \text{💡} / \text{Ⓢ} \quad 0$															
	$\Sigma \text{💡} / \text{Ⓢ} \quad 0$															
	$\Sigma \text{💡} / \text{Ⓢ} \quad 0$															
	$\Sigma \text{💡} / \text{Ⓢ} \quad 0$															



pessimistic initial values





<i>A</i> <i>B</i> <i>C</i> <i>D</i> <i>E</i>	$\sum \text{💡} / \text{Ⓢ} \quad 100$														
	$\sum \text{💡} / \text{Ⓢ} \quad 100$														
	$\sum \text{💡} / \text{Ⓢ} \quad 100$														
	$\sum \text{💡} / \text{Ⓢ} \quad 100$														
	$\sum \text{💡} / \text{Ⓢ} \quad 100$														



optimistic initial values



<i>A</i> <i>B</i> <i>C</i> <i>D</i> <i>E</i>	$\sum \text{💡} / \text{Ⓢ} \quad 100$	52 4														
	$\sum \text{💡} / \text{Ⓢ} \quad 100$															
	$\sum \text{💡} / \text{Ⓢ} \quad 100$															
	$\sum \text{💡} / \text{Ⓢ} \quad 100$															
	$\sum \text{💡} / \text{Ⓢ} \quad 100$															



optimistic initial values



A

$$\sum \text{💡} / \text{Ⓢ} \quad 100$$

52
4

B

$$\sum \text{💡} / \text{Ⓢ} \quad 100$$

54
8

C

$$\sum \text{💡} / \text{Ⓢ} \quad 100$$

D







$$\sum \text{💡} / \text{Ⓢ} \quad 100$$

E

$$\sum \text{💡} / \text{Ⓢ} \quad 100$$

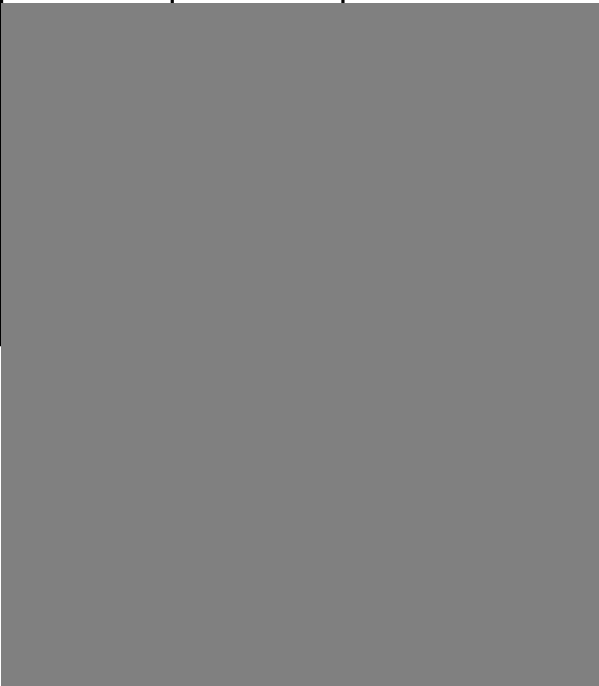


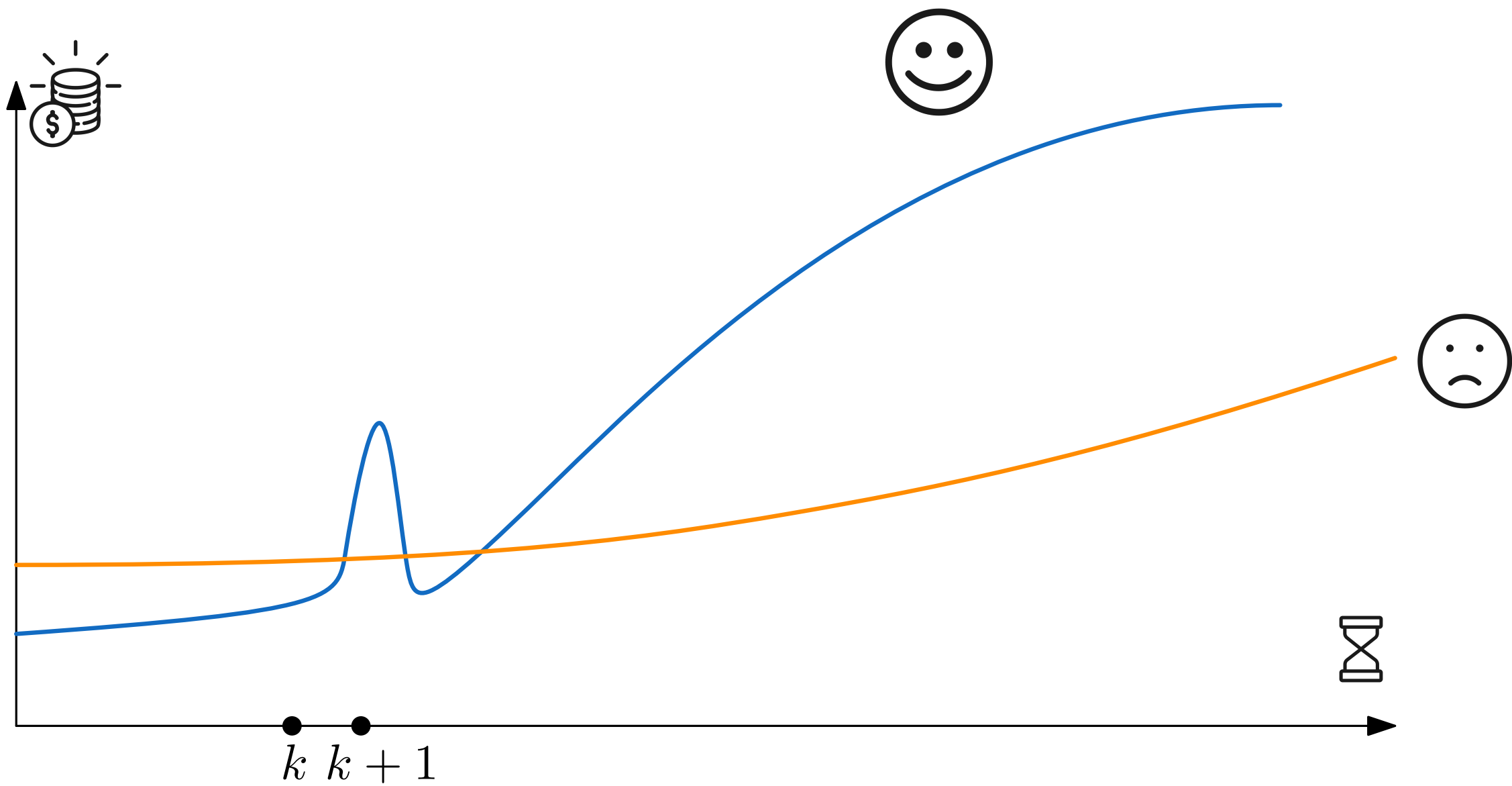
optimistic initial values

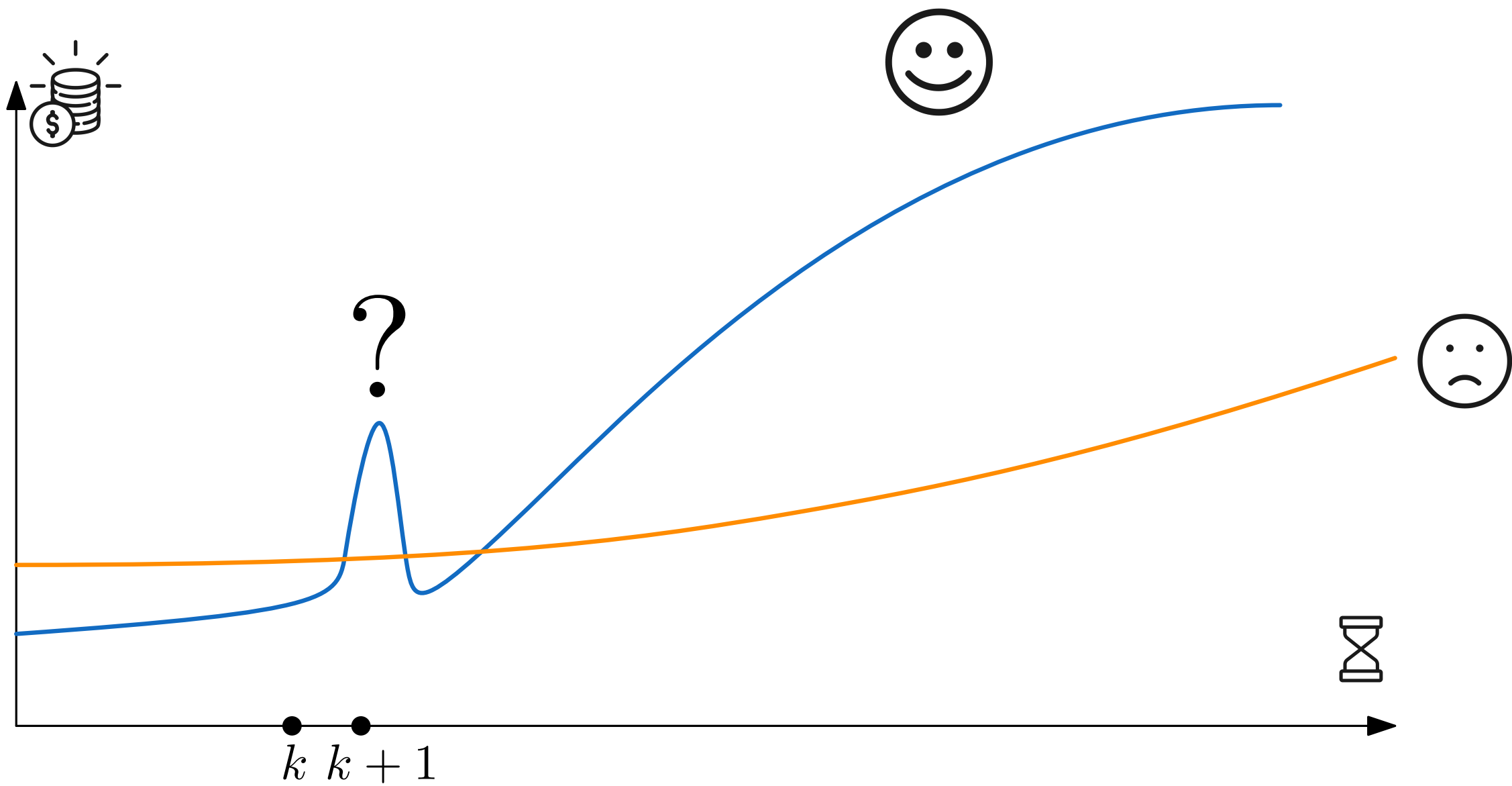
																
<i>A</i>	$\sum \text{💡}/\text{Ⓢ} \ 100$	52 4														
<i>B</i>	$\sum \text{💡}/\text{Ⓢ} \ 100$		54 8			40 12										
<i>C</i>	$\sum \text{💡}/\text{Ⓢ} \ 100$															
<i>D</i>	$\sum \text{💡}/\text{Ⓢ} \ 100$				50 0											
<i>E</i>	$\sum \text{💡}/\text{Ⓢ} \ 100$			57 14												



optimistic initial values

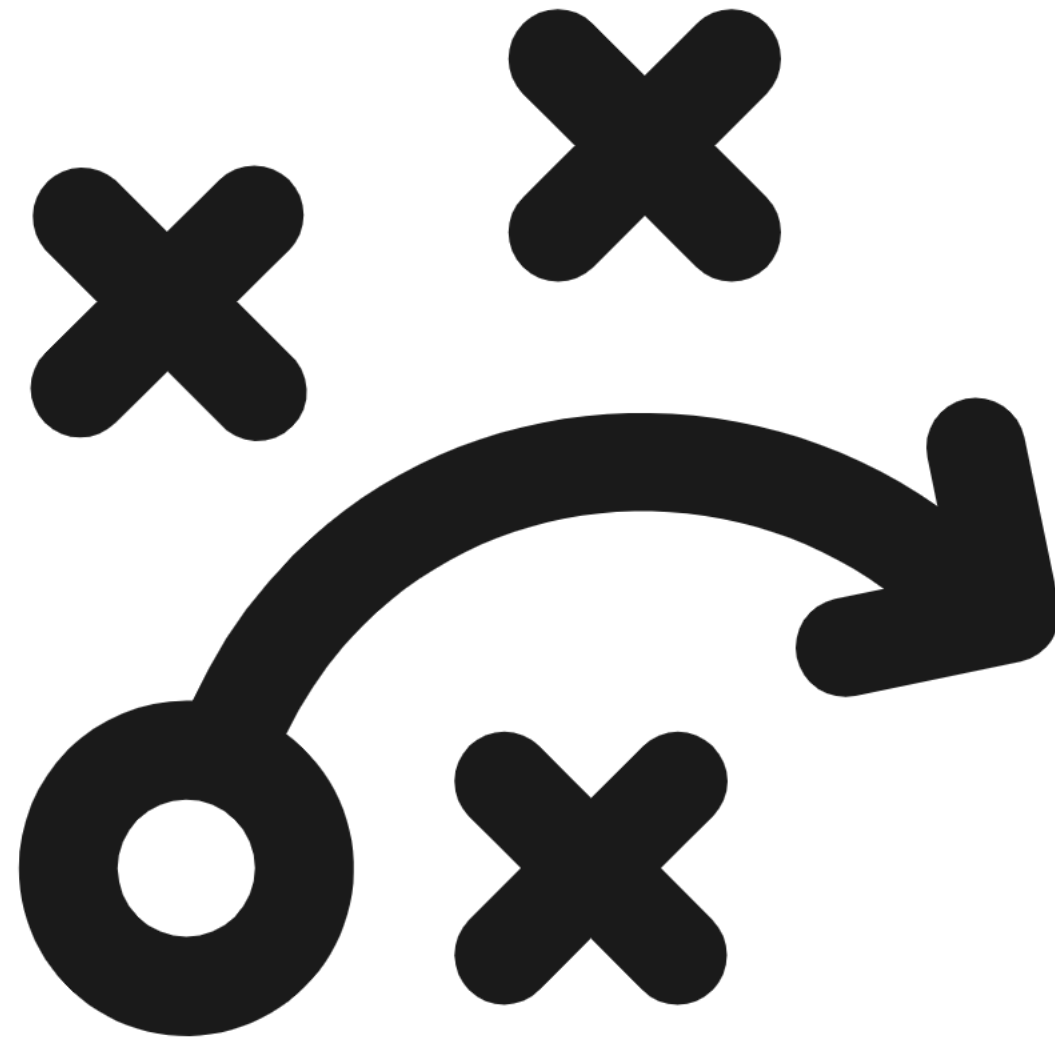






Strategy 2

Upper Confidence Bound




Upper Confidence Bound

<i>A</i>	0										0	
<i>B</i>	5										7	
<i>C</i>	1	3	0									
<i>D</i>	6 6	7 6	7 6	5	7 7 5			7 6	5 6	5		
<i>E</i>	3	1 2				1						



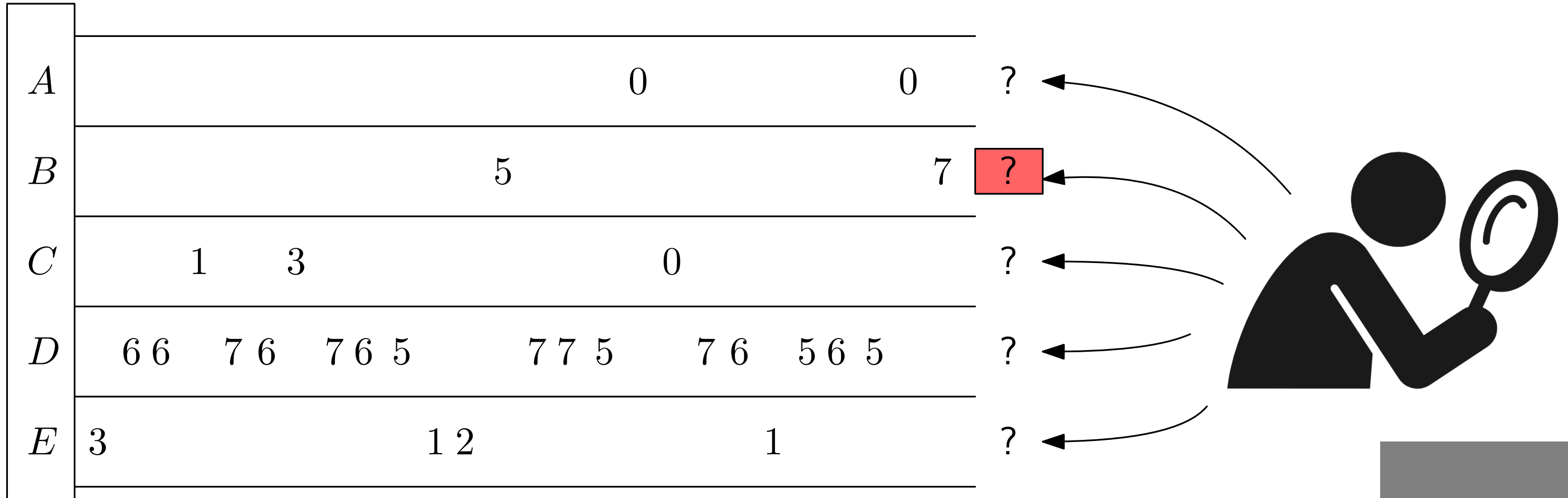
Upper Confidence Bound

<i>A</i>	0										0	?	
<i>B</i>	5										7	?	
<i>C</i>	1	3	0										?
<i>D</i>	6 6	7 6	7 6 5	7 7 5			7 6	5 6 5	?				
<i>E</i>	3	1 2				1				?			



A stylized black silhouette of a person holding a magnifying glass, positioned to the right of the table. Five curved arrows point from the magnifying glass towards the question marks in the rightmost column of the table, indicating a focus on the unknown values.

Upper Confidence Bound




Upper Confidence Bound

<i>A</i>	0										0	
<i>B</i>	5										7	
<i>C</i>	1	3	0									
<i>D</i>	6 6	7 6	7 6	5	7 7 5			7 6	5 6	5		
<i>E</i>	3	1 2				1						



Upper Confidence Bound

													Σ  $/ \textcircled{\#}$	$\textcircled{\#}$
A	00												0	2
B	57												6	2
C	1	3	0										1.33	3
D	6 6	7 6	7 6 5	7 7 5				7 6	5 6 5				6.066	15
E	3	1 2					1					1.75	4	

Upper Confidence Bound

$\Sigma \frac{\text{💡💰}}{\text{Ⓢ}}$ Ⓢ

<i>A</i>	0										0	
<i>B</i>	5										7	
<i>C</i>	1	3	0									
<i>D</i>	6 6	7 6	7 6	5	7 7 5			7 6	5 6	5		
<i>E</i>	3	1 2					1					

0 2

6 2

1.33 3

6.066 15

1.75 4

$\Sigma \frac{\text{💡💰}}{\text{Ⓢ}} \uparrow$

$\text{Ⓢ} \downarrow$



Upper Confidence Bound

$\sum \text{💰} / \text{Ⓢ}$ Ⓢ

<i>A</i>											0	0	0	2				
											5	7	6	2				
											1	3	0	1.33	3			
											6 6	7 6	7 6 5	7 7 5	7 6	5 6 5	6.066	15
											3	1 2			1	1.75	4	

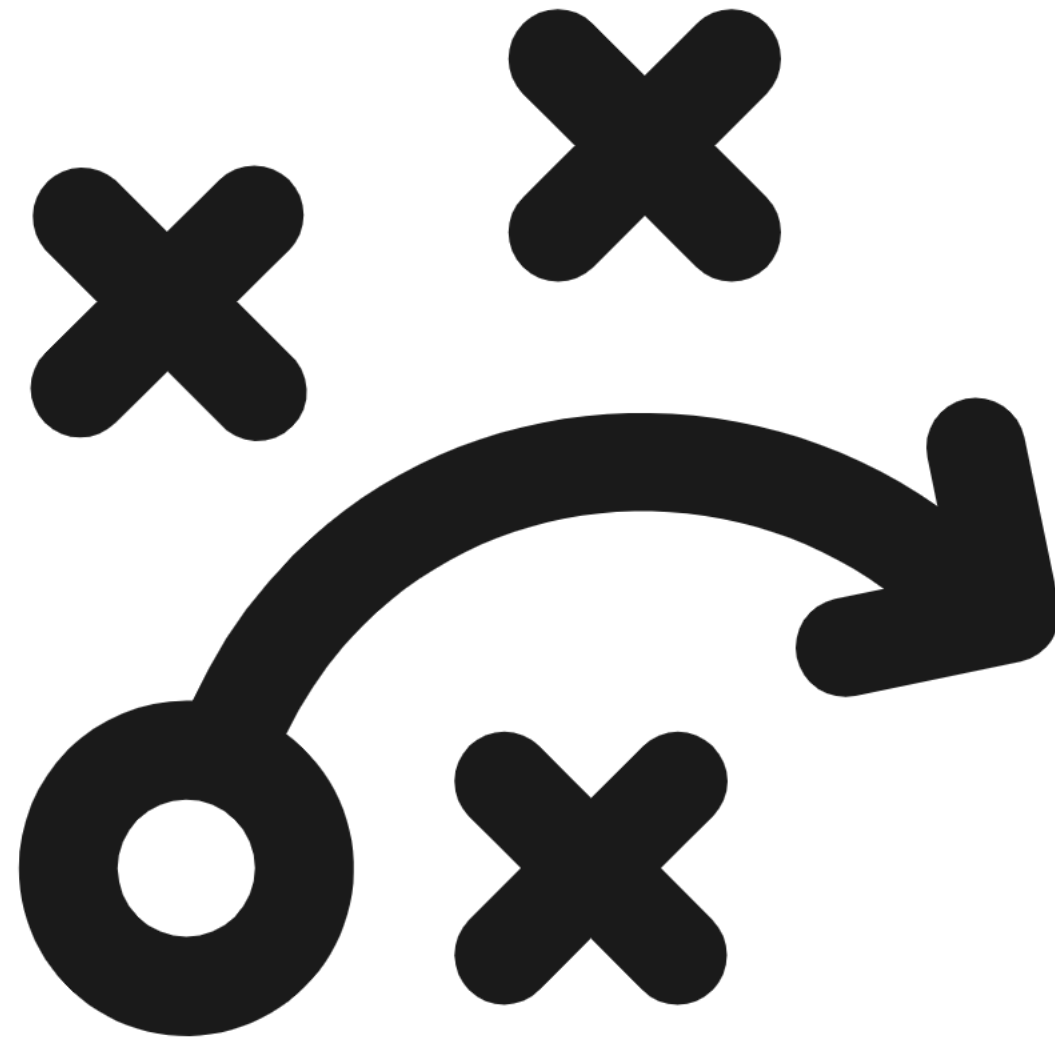
$\sum \text{💰} / \text{Ⓢ} \uparrow$ $\text{Ⓢ} \downarrow$

$\max \frac{\sum \text{💰}_i}{\text{Ⓢ}_i} + c \sqrt{\log \frac{\text{Ⓢ}}{\text{Ⓢ}_i}}$



Strategy 3

Explicit Probabilities





Strategy 4

Domain Knowledge

Assignment



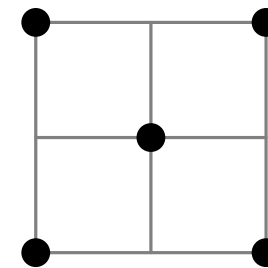
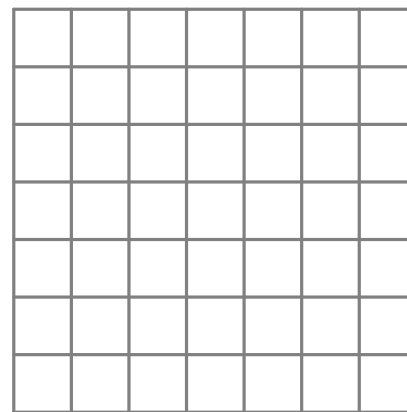


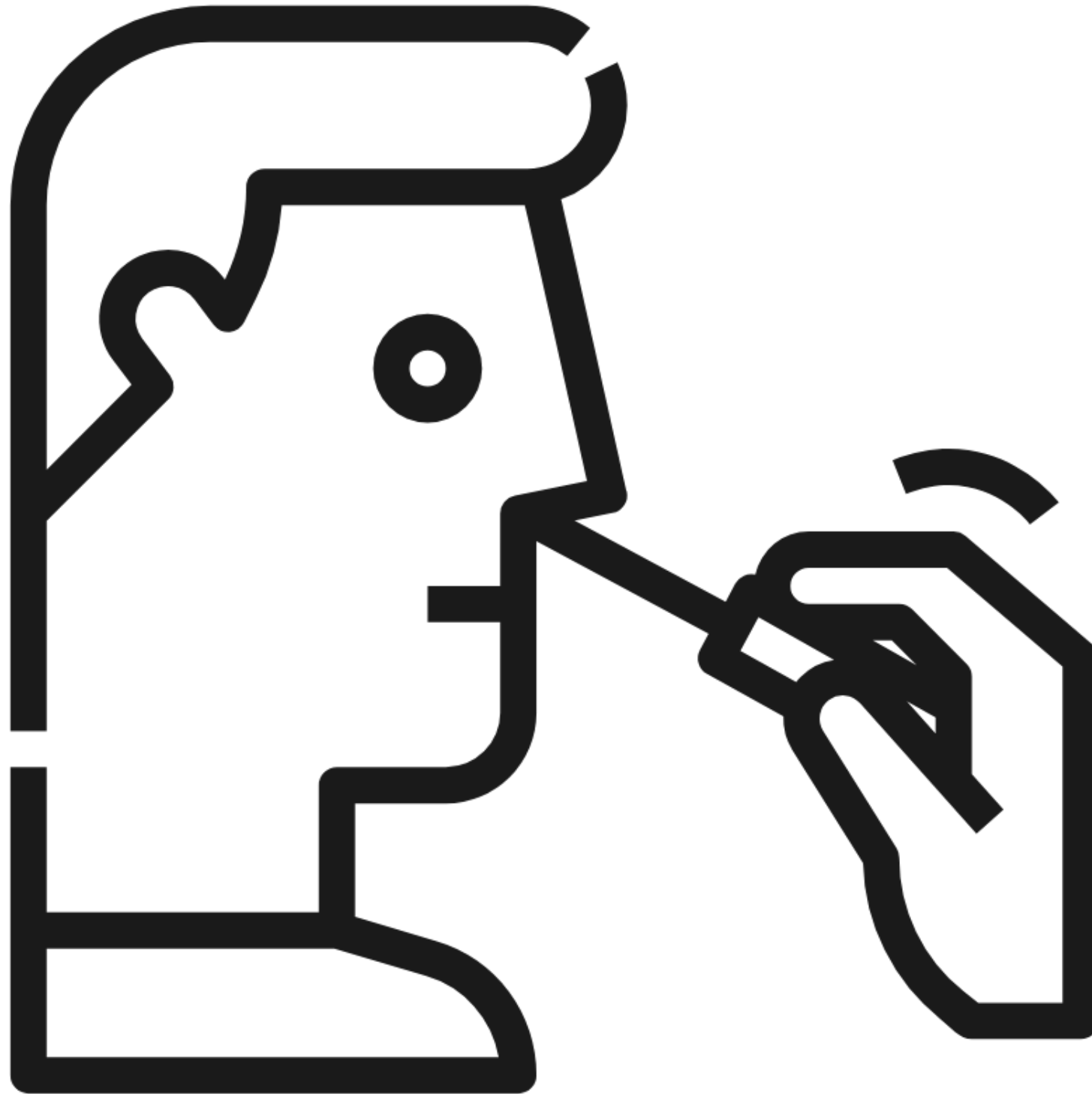
groupsize $\in [2, 3]$

- Using domain knowledge describe an agent
- Advantages?
- Downsides?



20 minutes

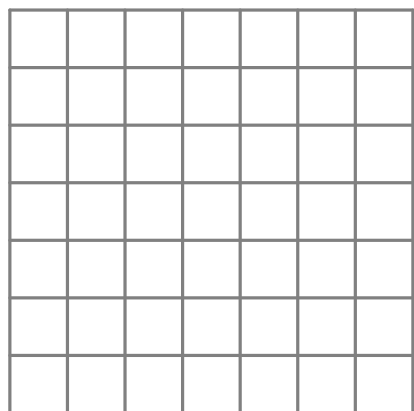




Testing

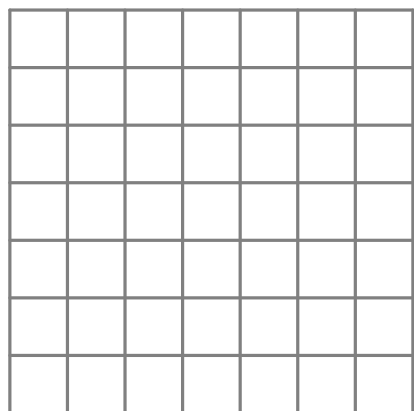




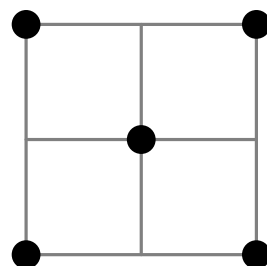


boardsize



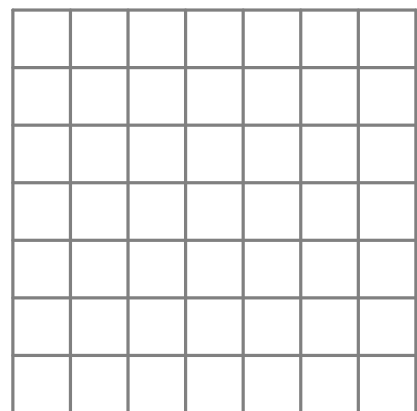


boardsize

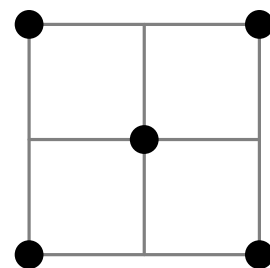


patterns





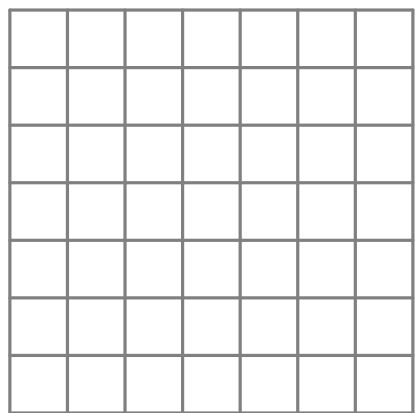
boardsize



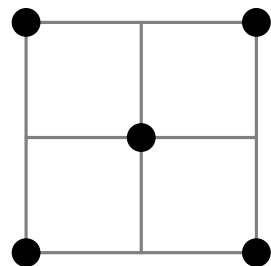
patterns

number of rollouts
time





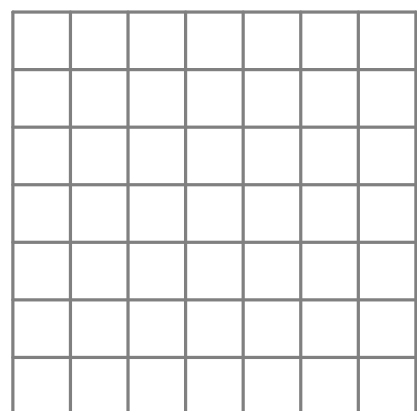
boardsize



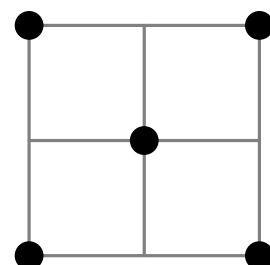
patterns

number of rollouts
time

parameters: ϵ, c



boardsize

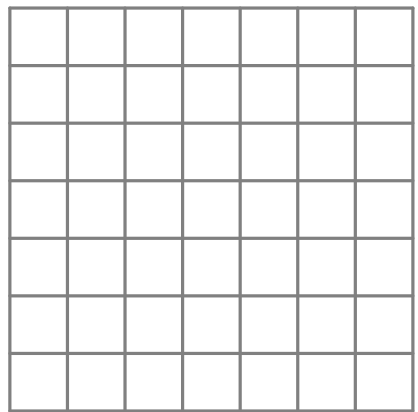


patterns

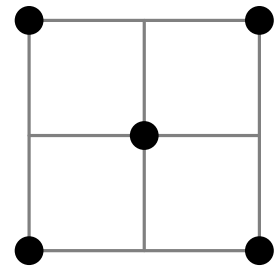
number of rollouts
time

parameters: ϵ, c





boardsize

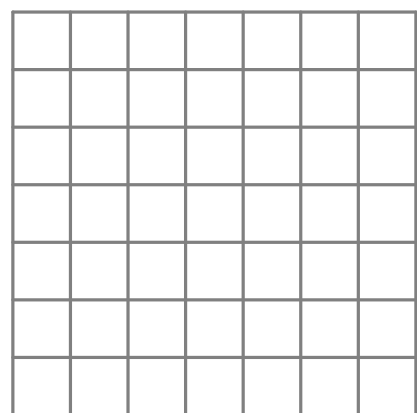


patterns

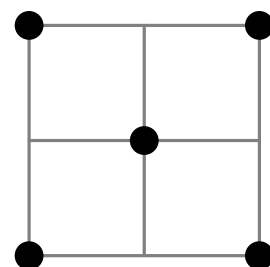
number of rollouts
time

parameters: ϵ, c





boardsize



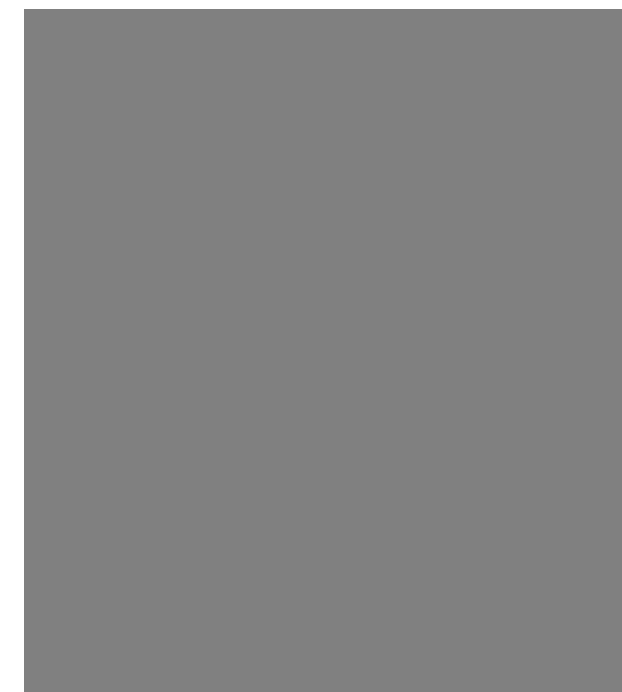
patterns



optimization

number of rollouts
time

parameters: ϵ, c



Assignment



groupsize $\in [2, 3]$



20 minutes

- simplest test?
- favorite agent?
- sophisticated test?
- optimization ideas?

