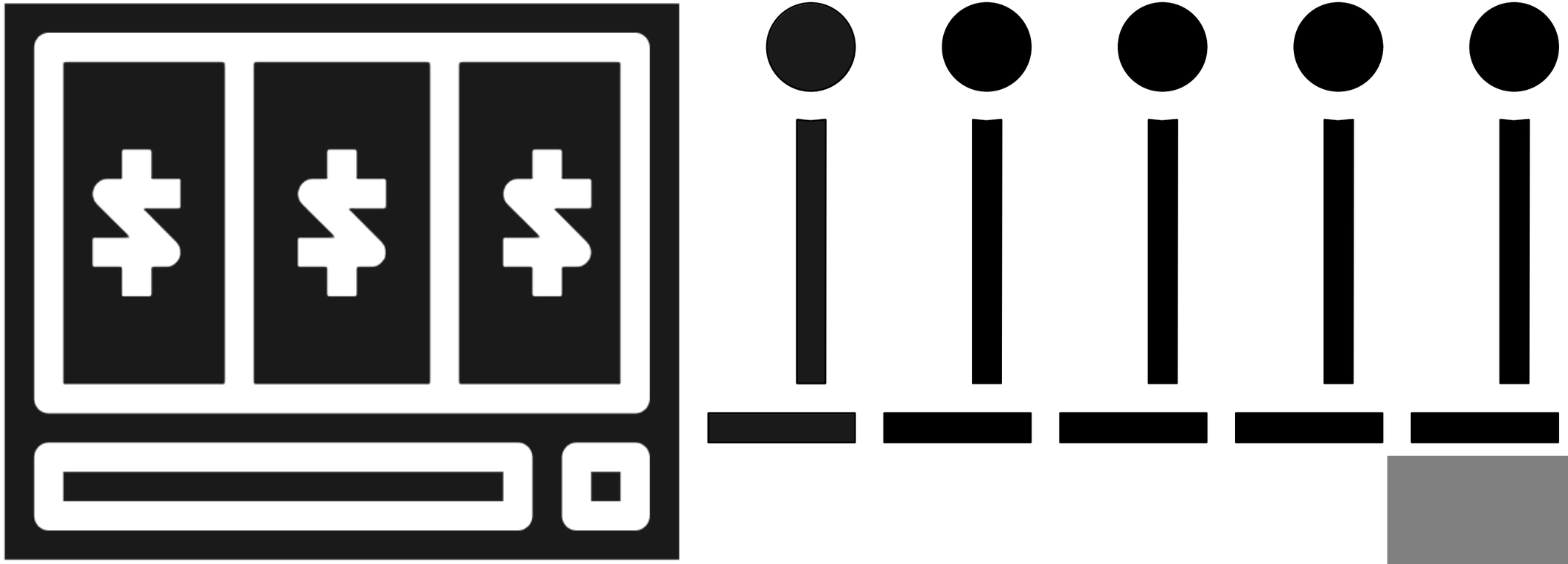
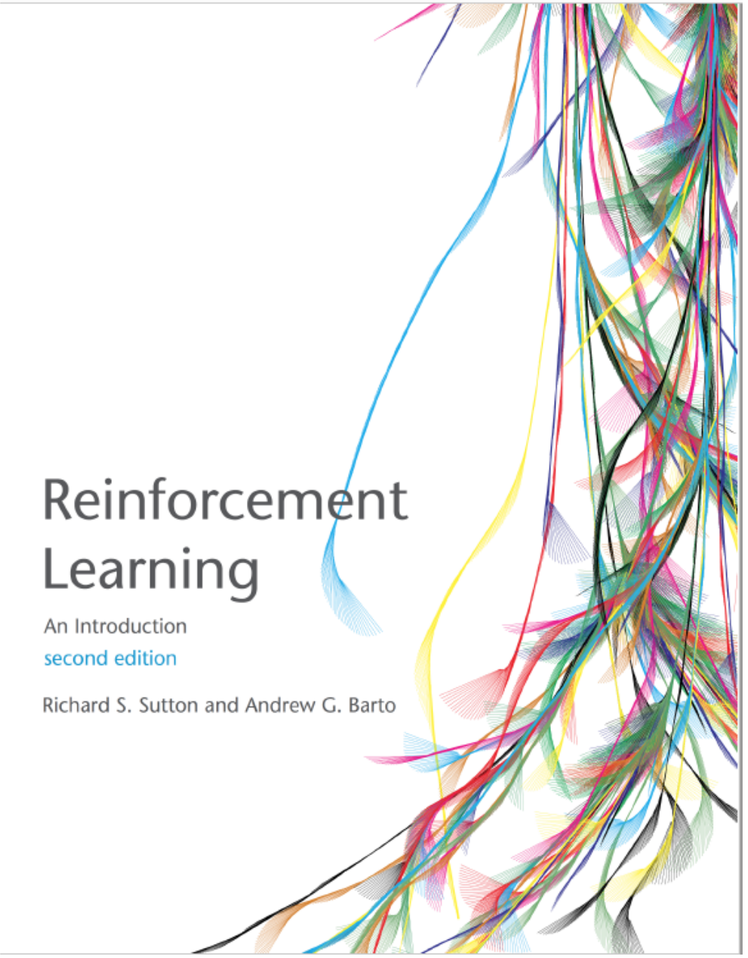
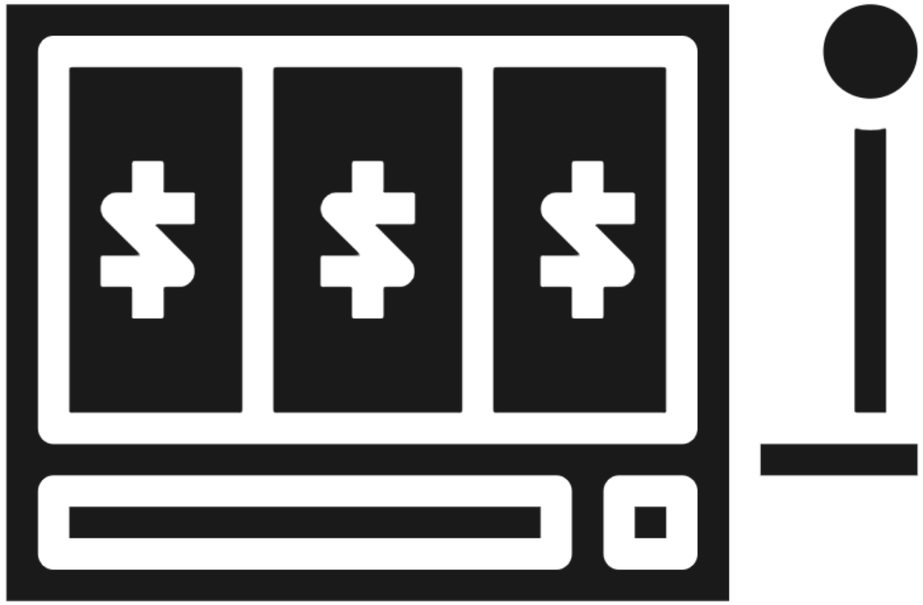


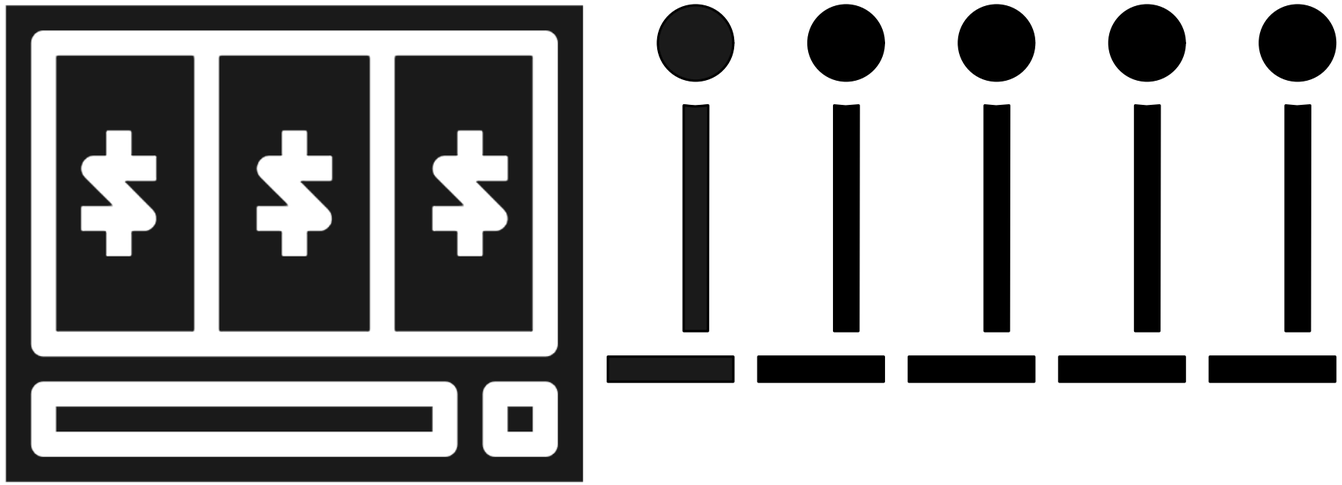

Reinforcement
Learning
An Introduction
second edition
Richard S. Sutton and Andrew G. Barto

Chapter 2

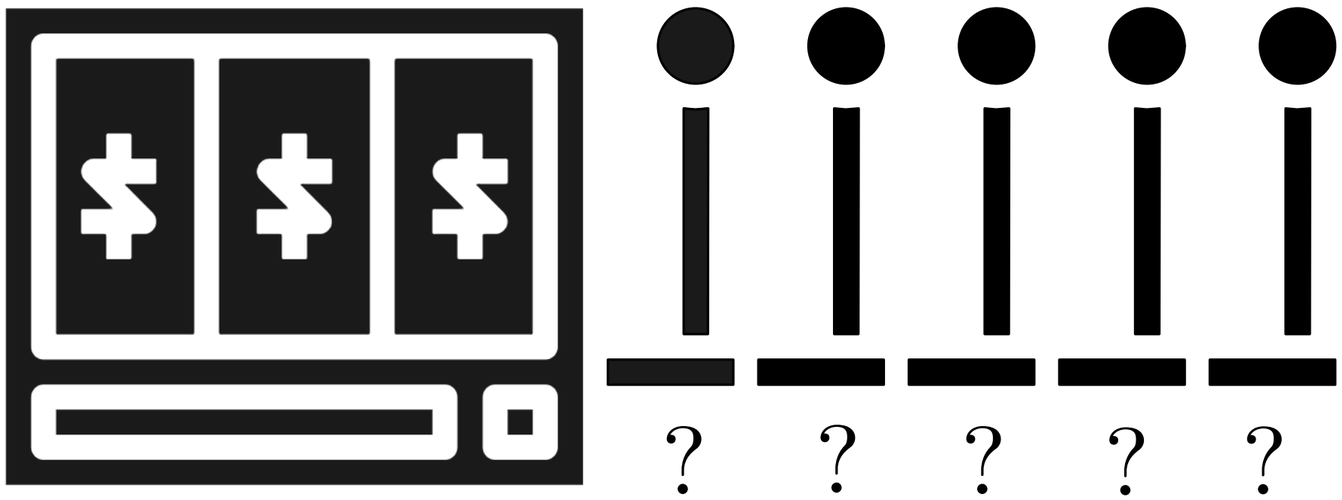


slot machine
one-armed bandit

k-armed Bandit Problem



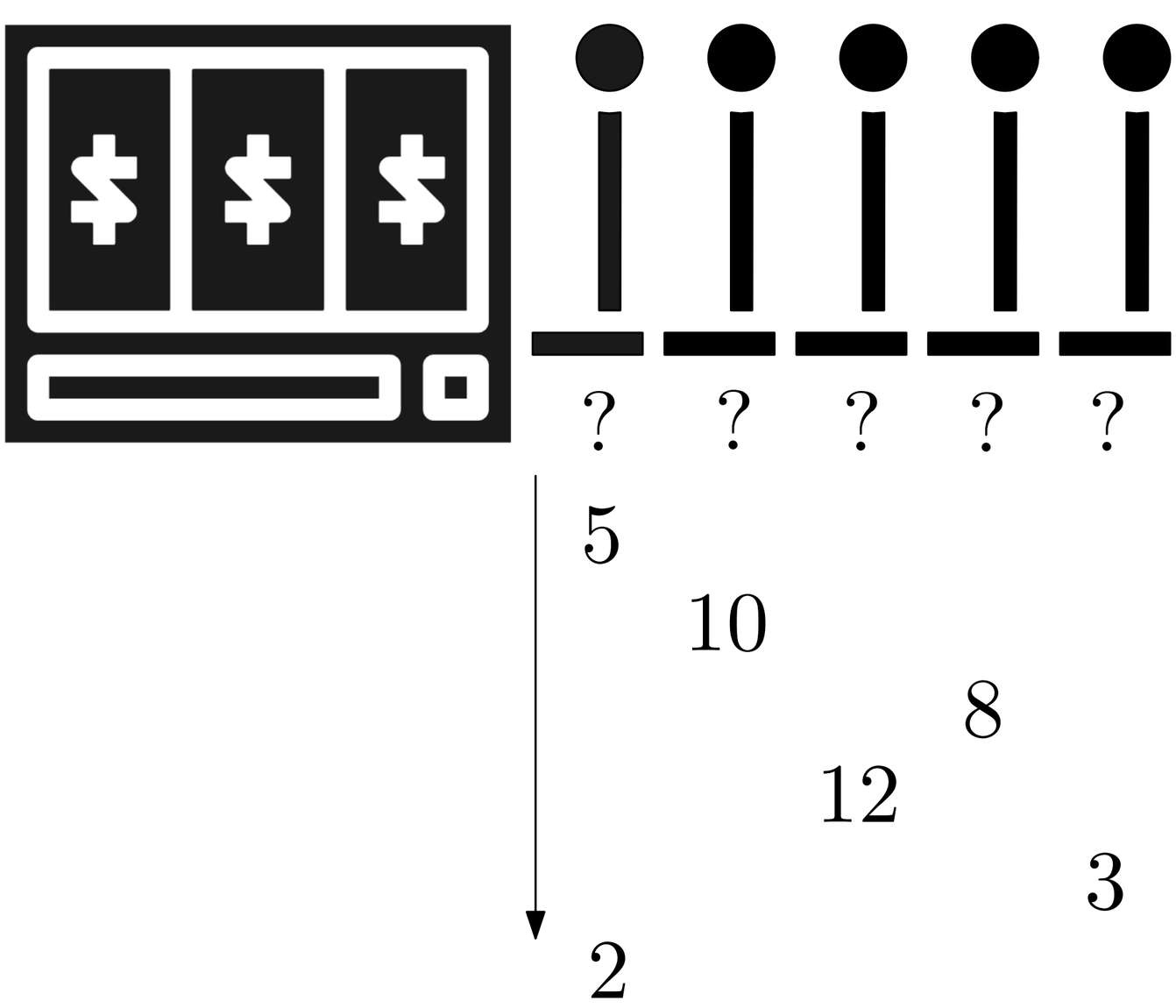
k-armed Bandit Problem



reward

probabilistic

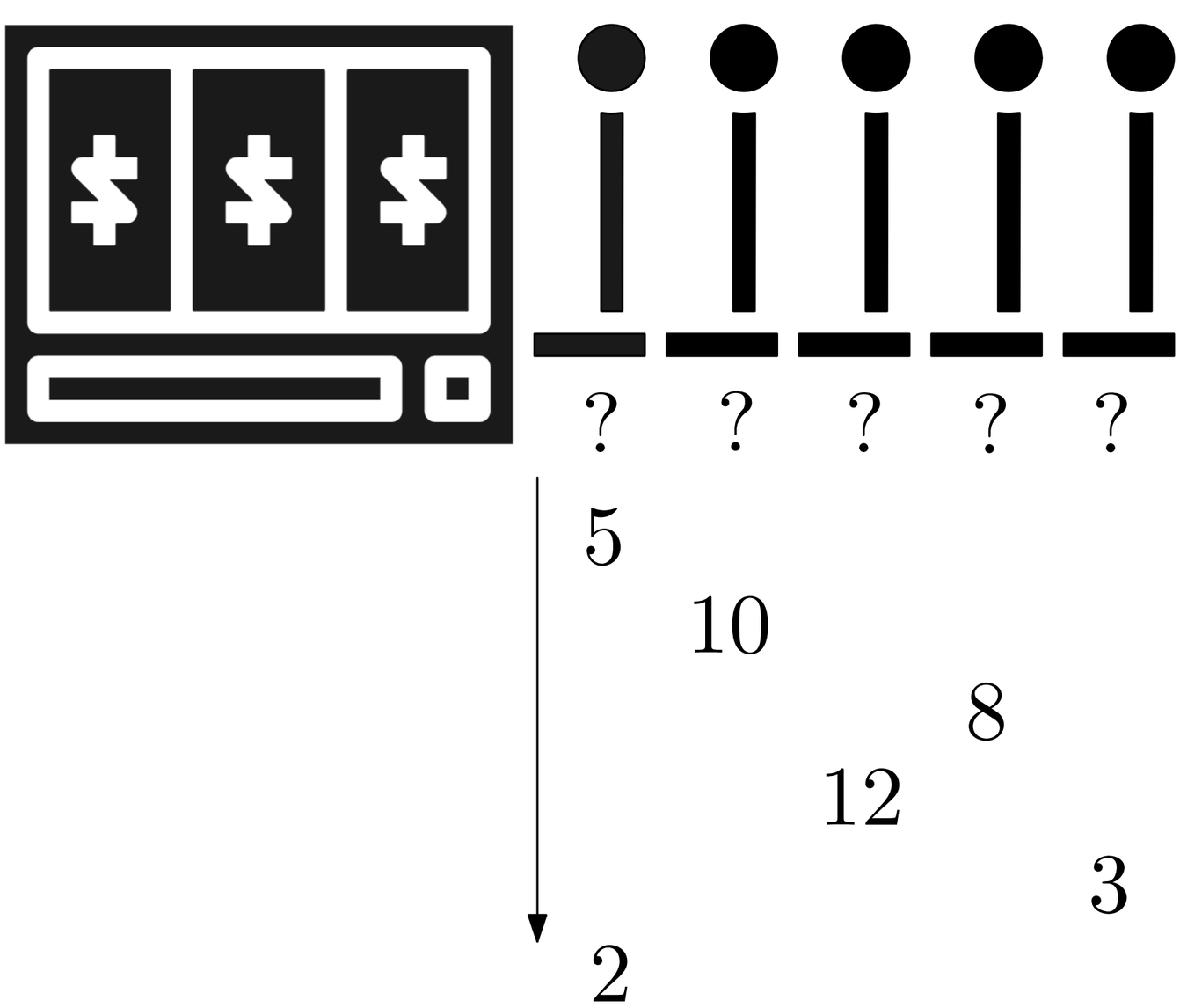
k-armed Bandit Problem



reward

probabilistic

k-armed Bandit Problem

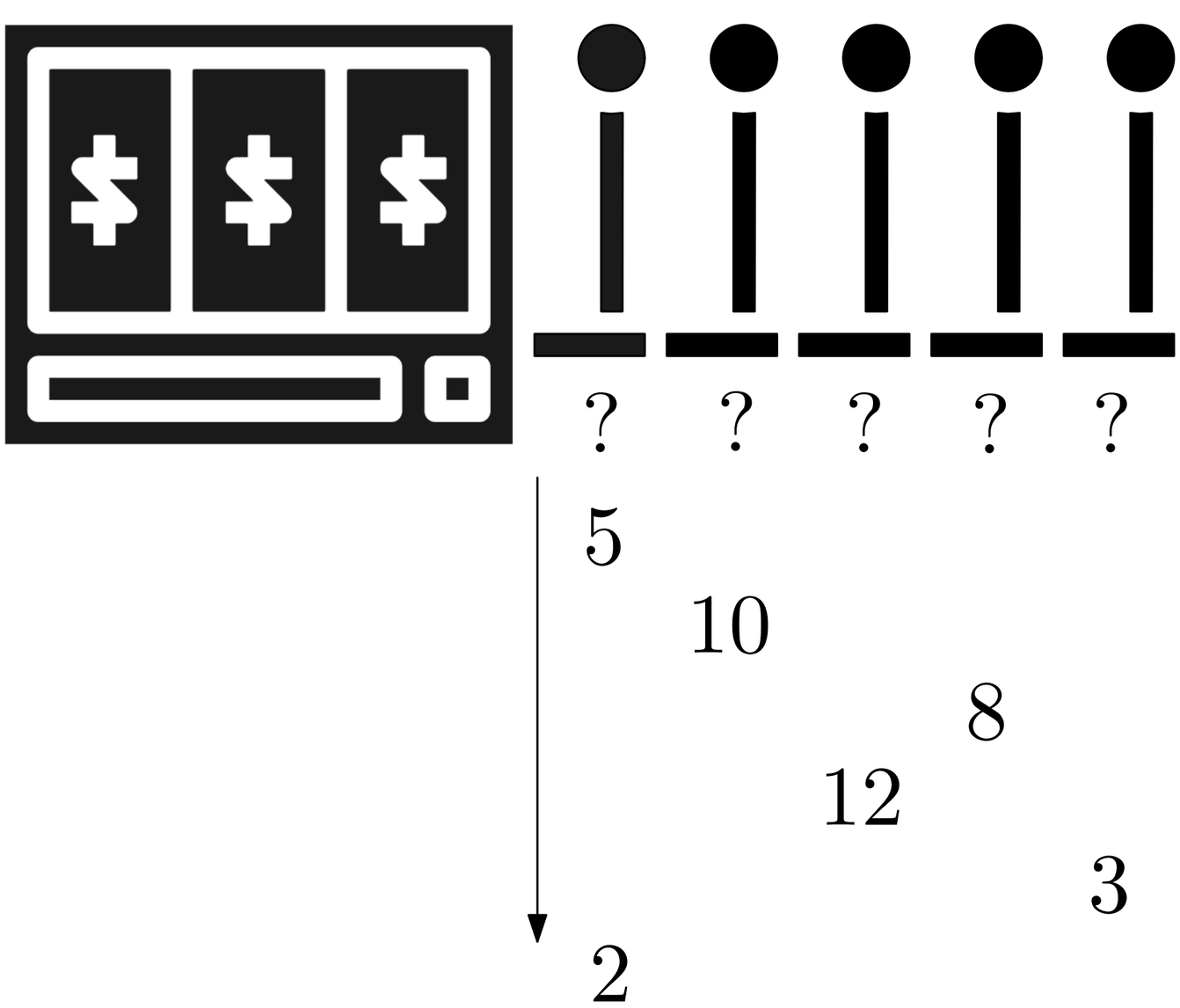


reward

probabilistic

value

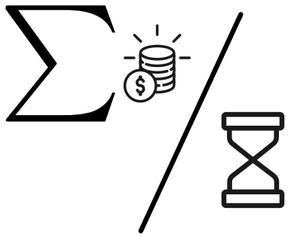
k-armed Bandit Problem



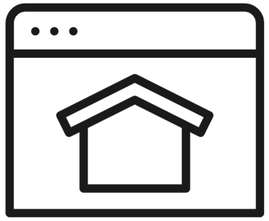

reward

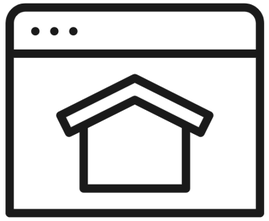

probabilistic

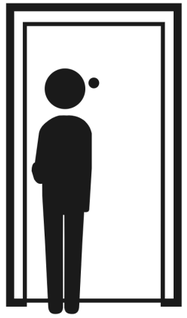

value

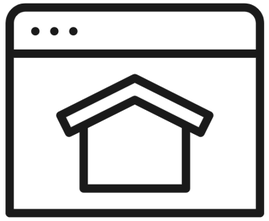

approximate value

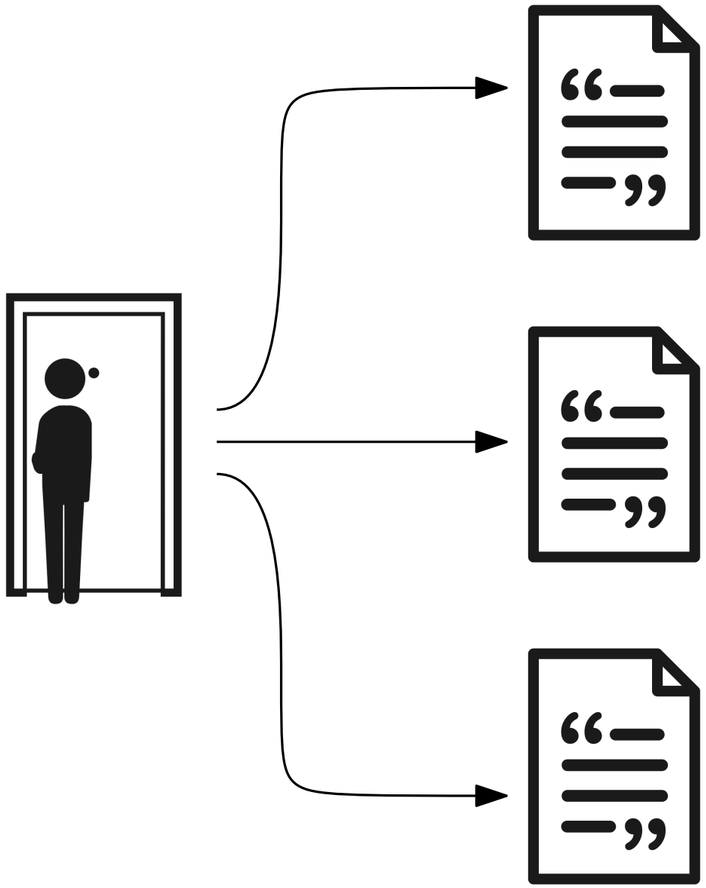
Example

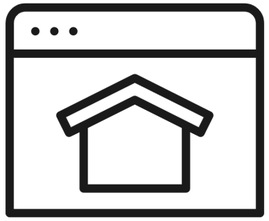


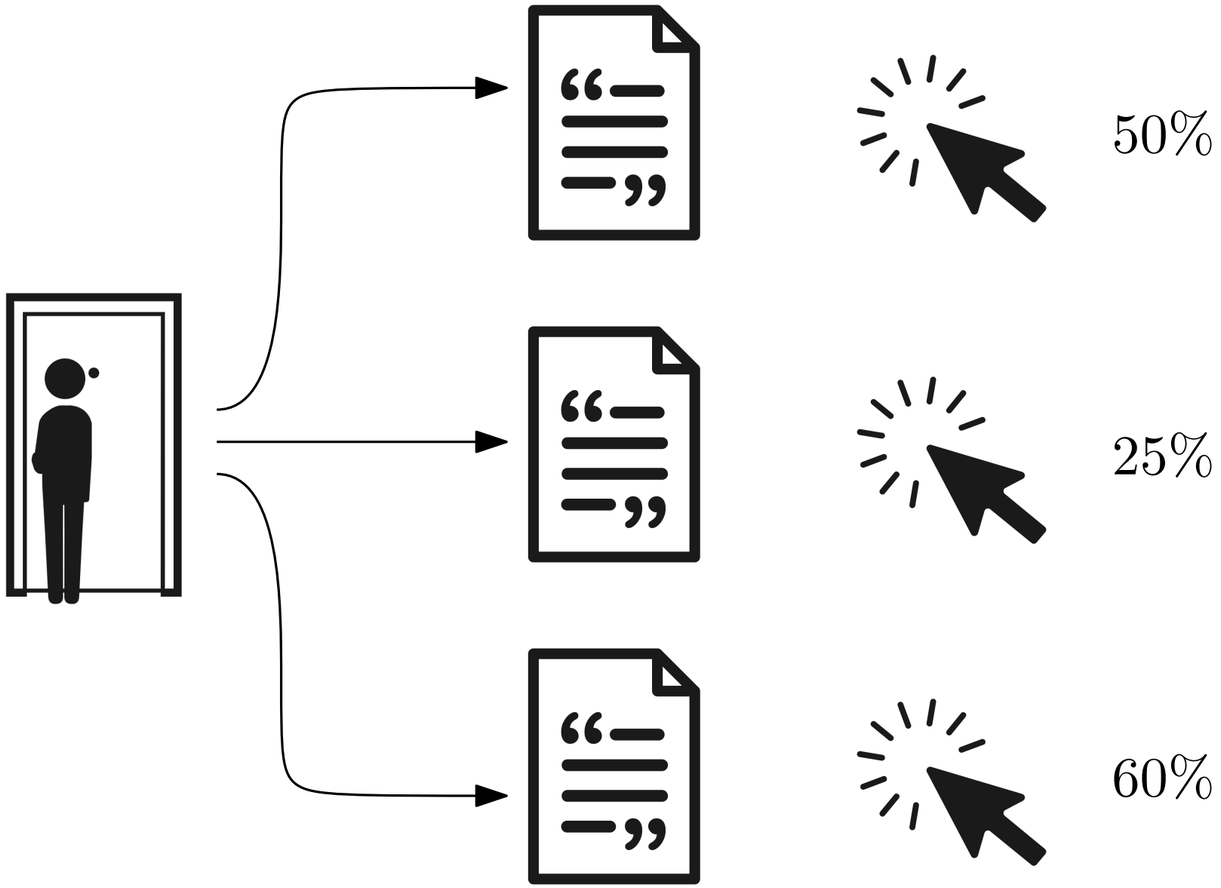











50%
25%
60%

Assignment

groupsize $\in [2, 3]$

20 minutes

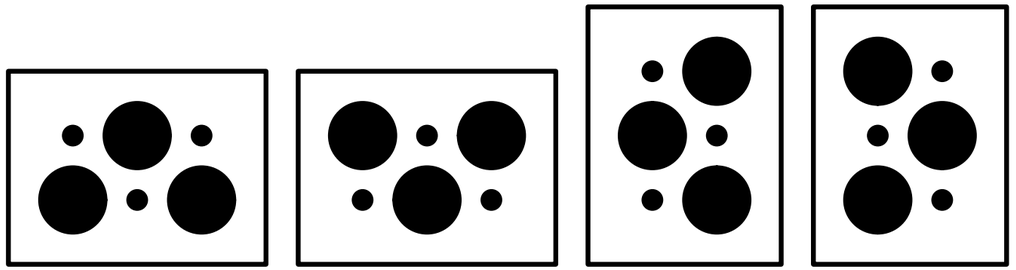
examples?

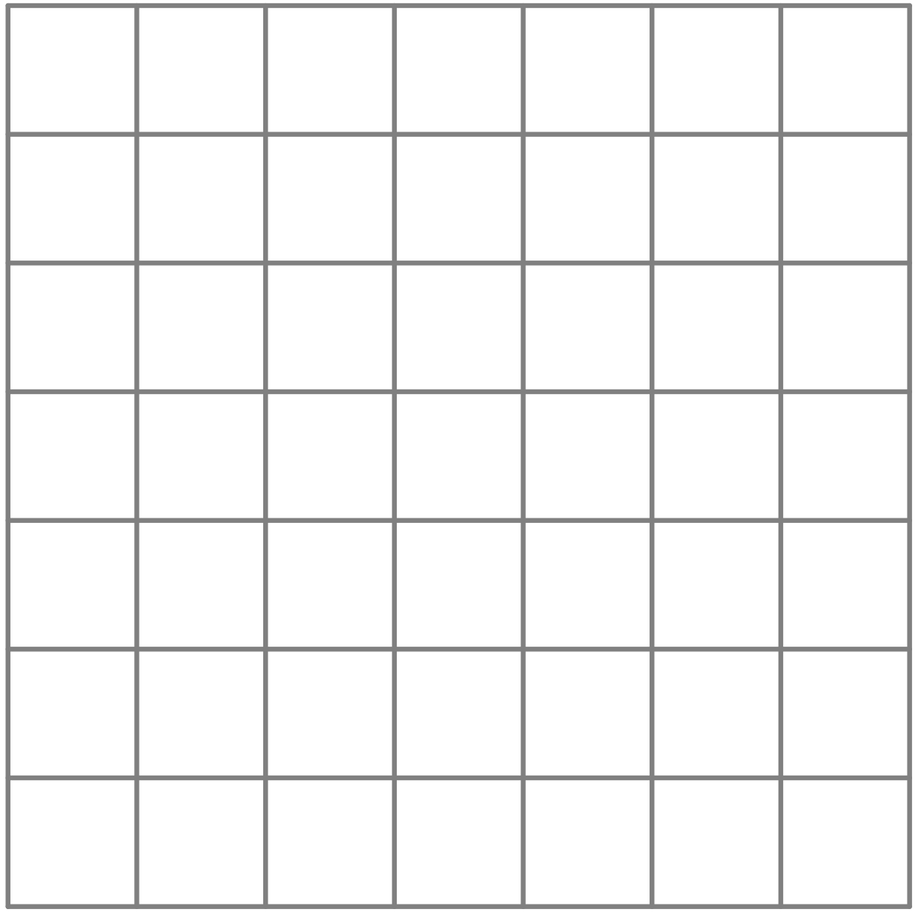
- involving games
- involving farming
- options / distribution / reward / value

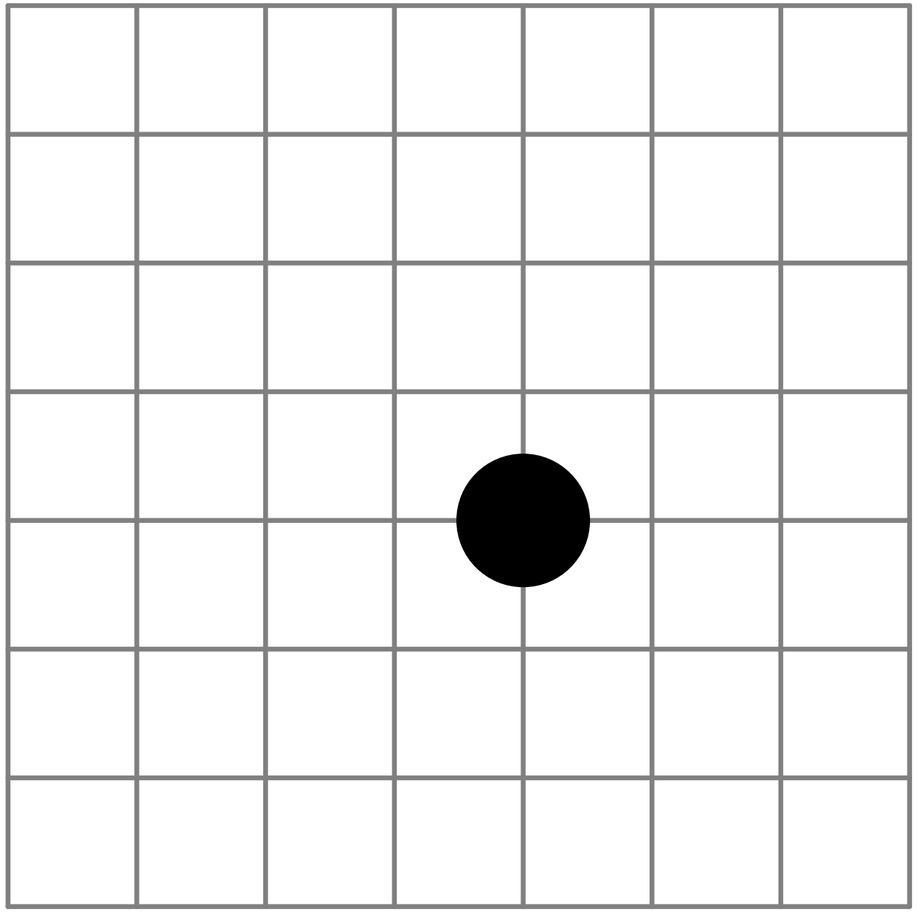
non-examples

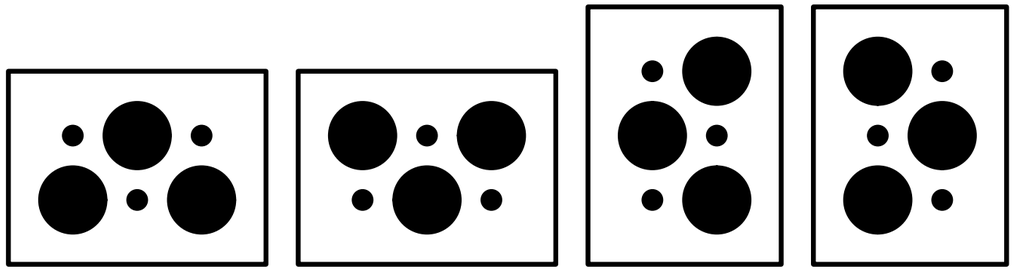
- properties?

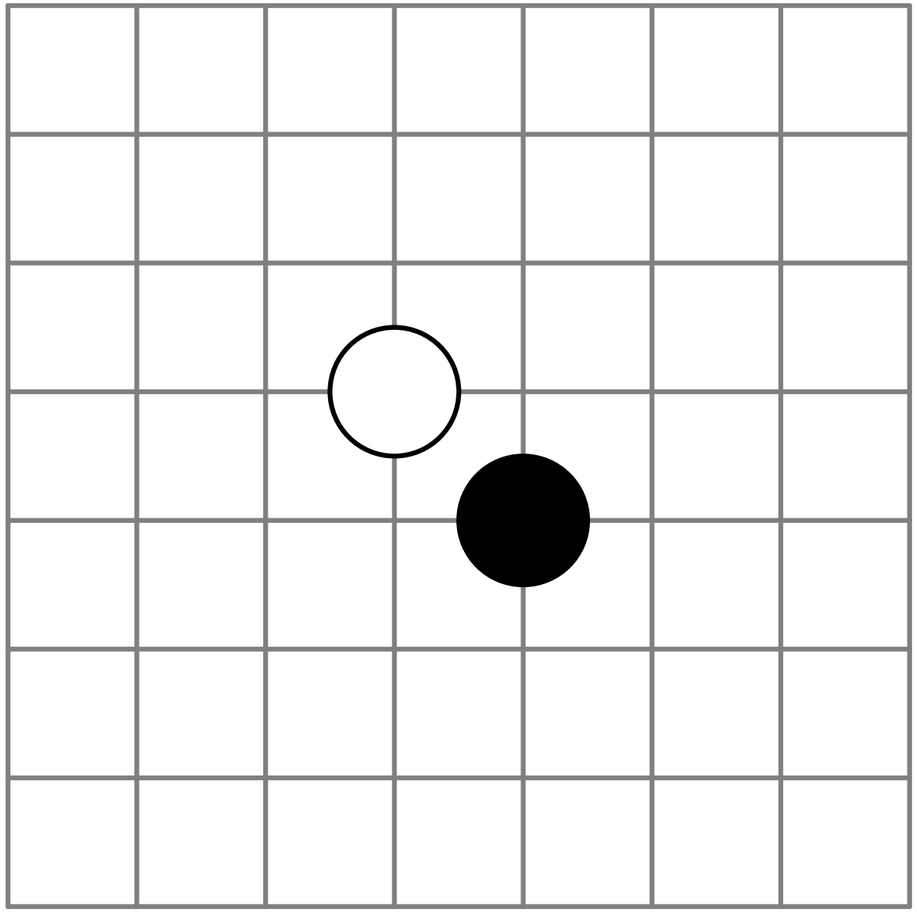
Application

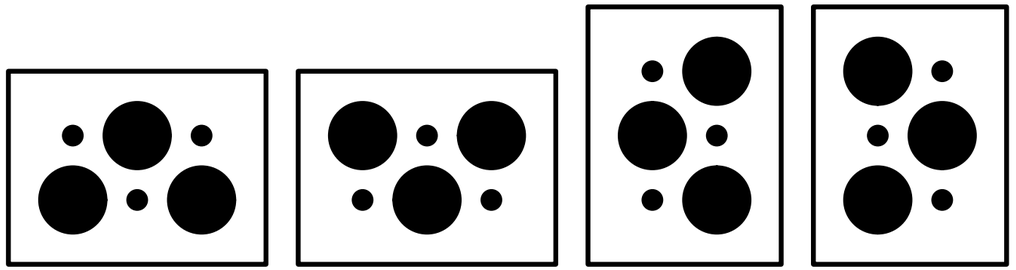


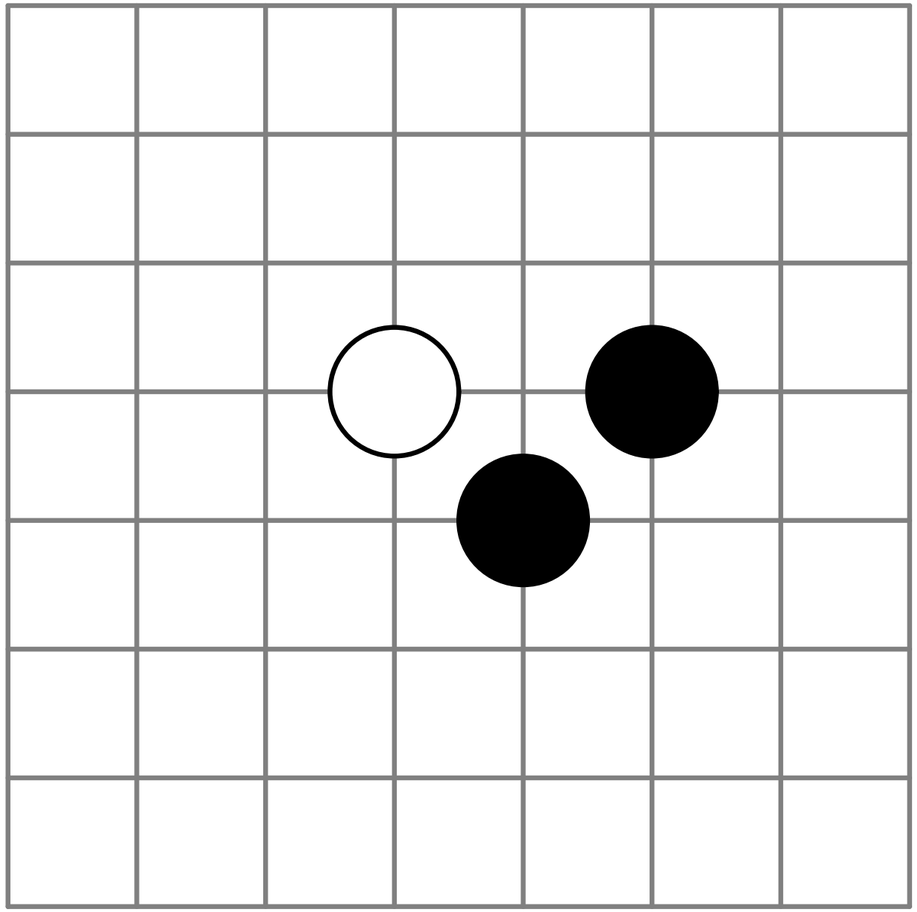


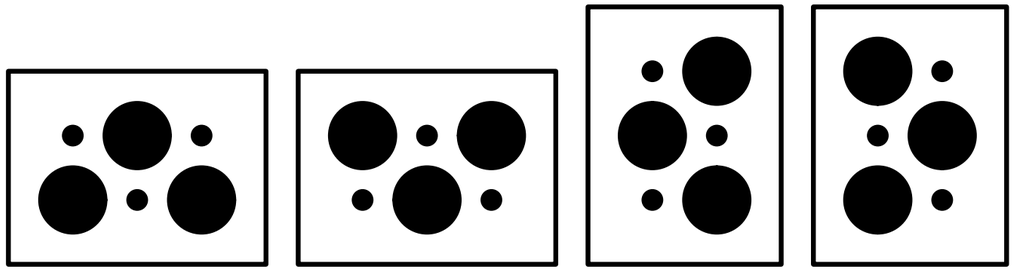


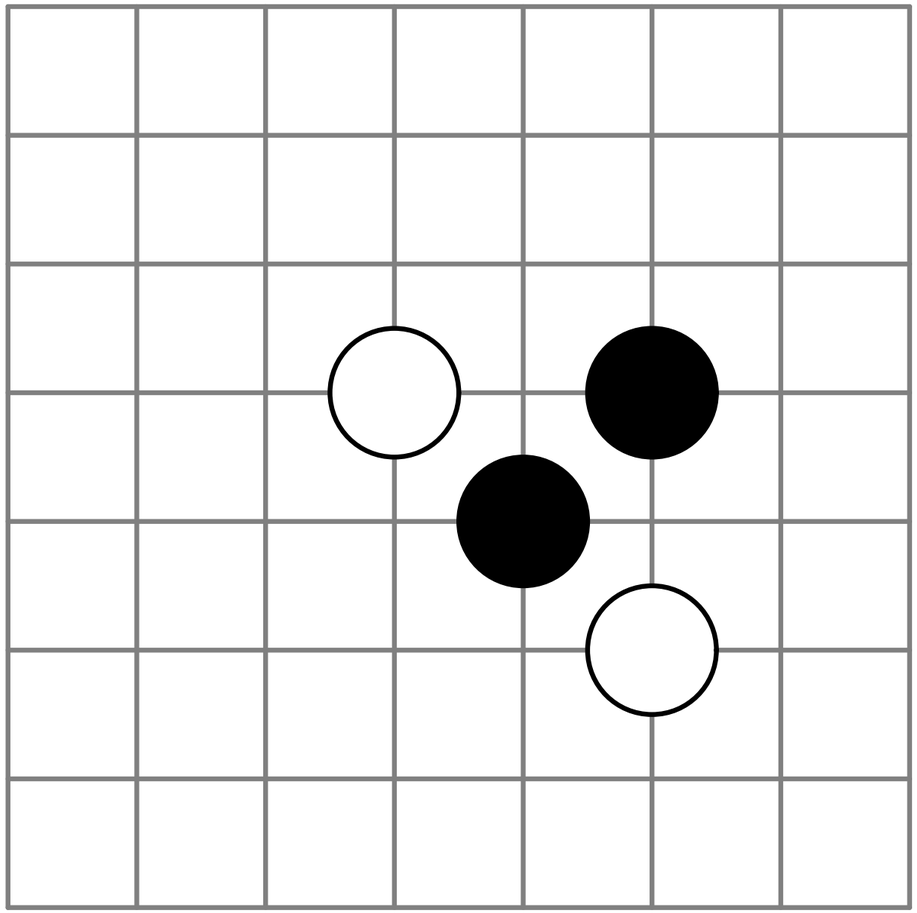


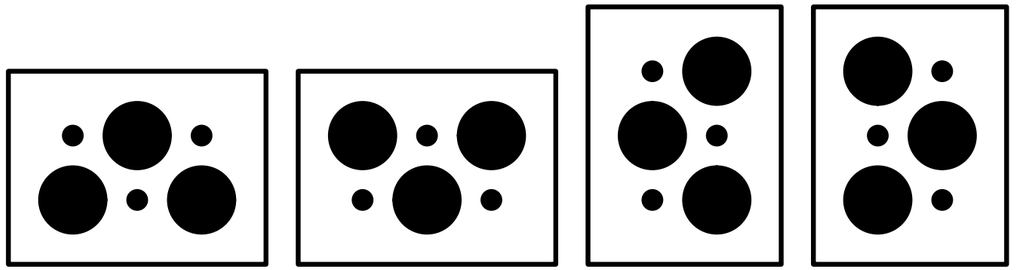


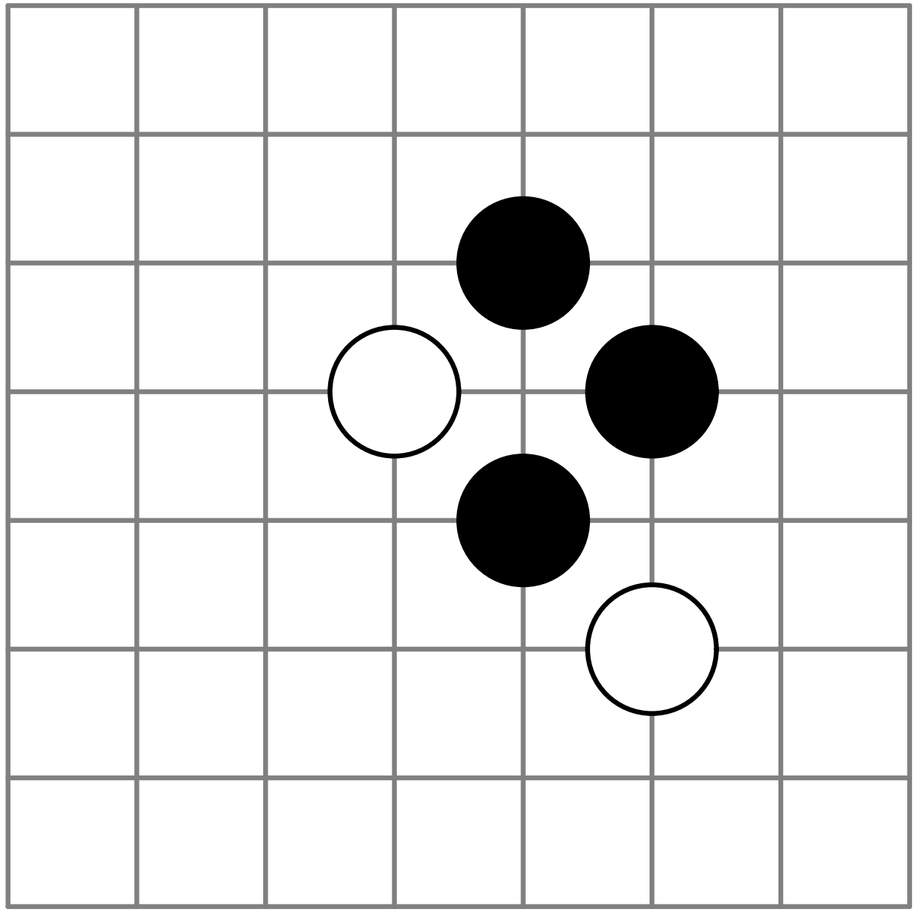


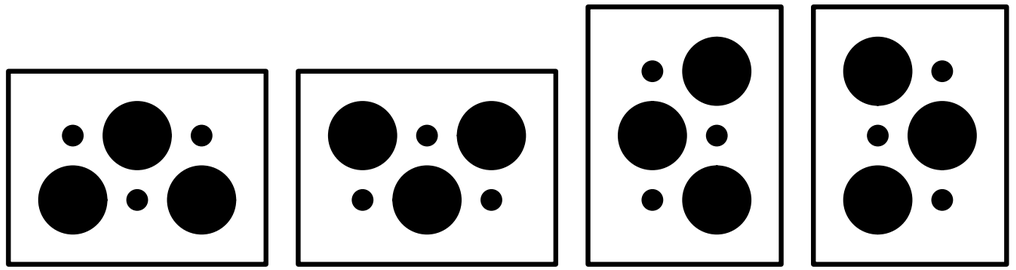


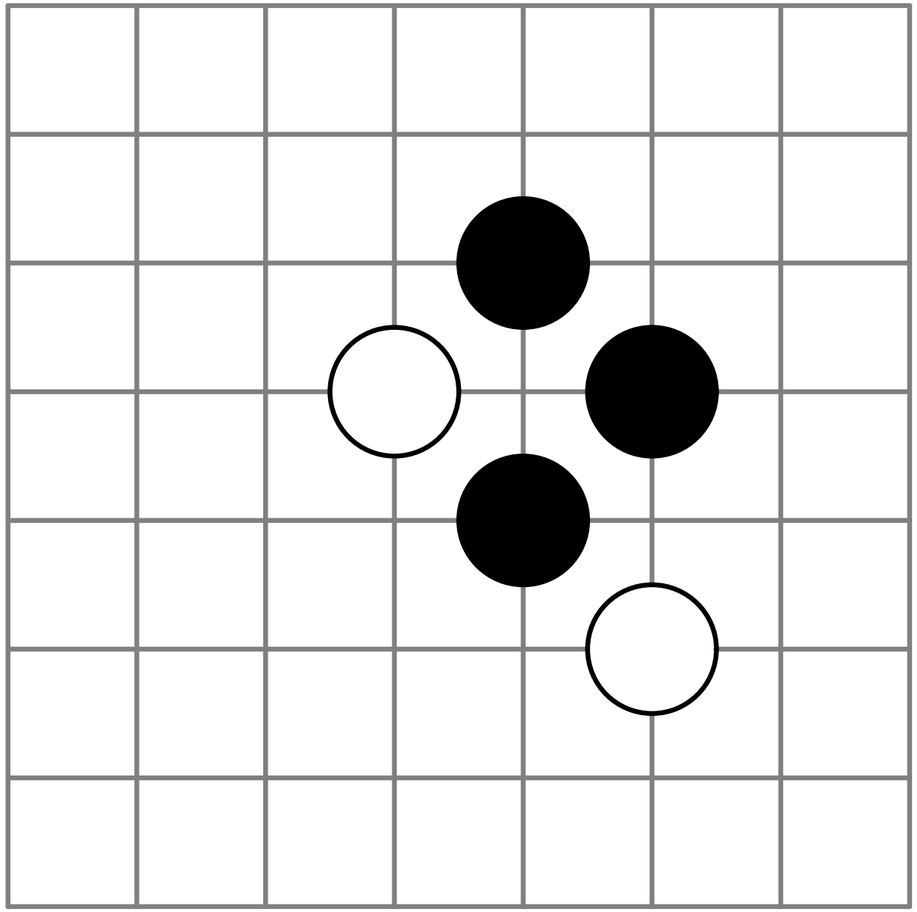


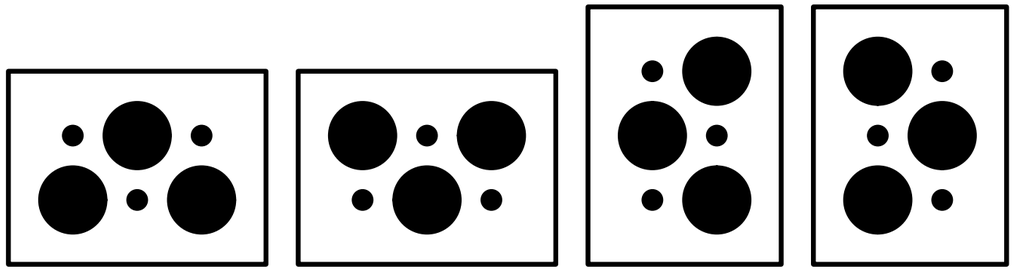


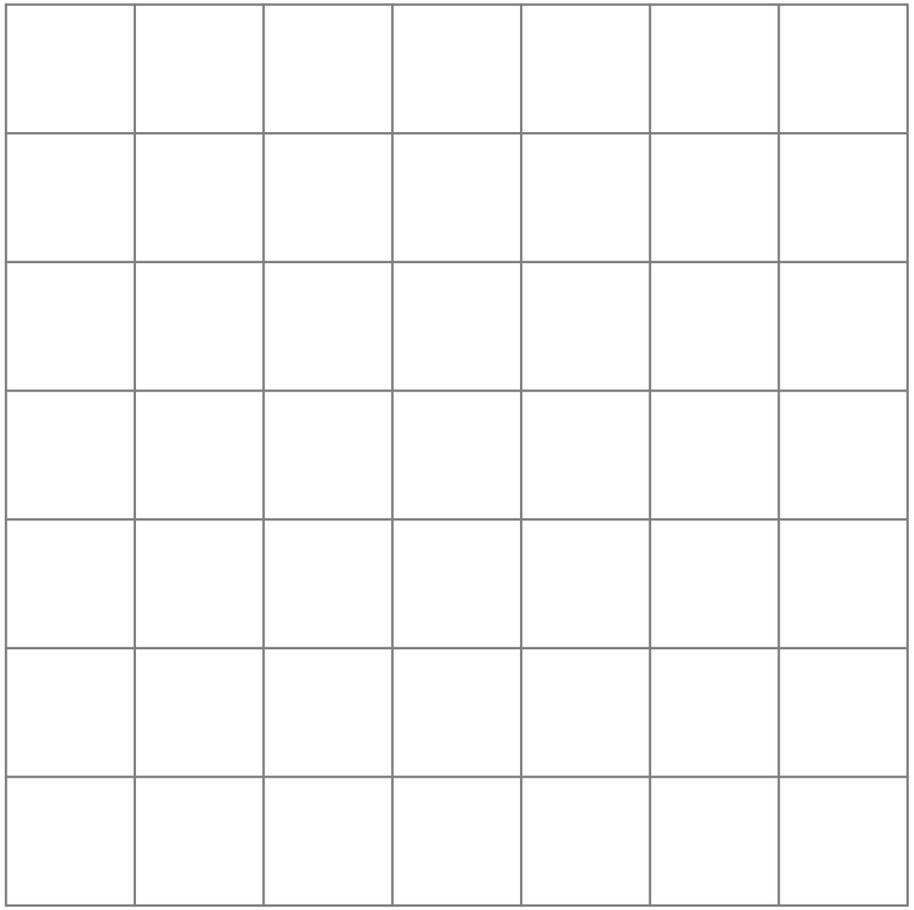






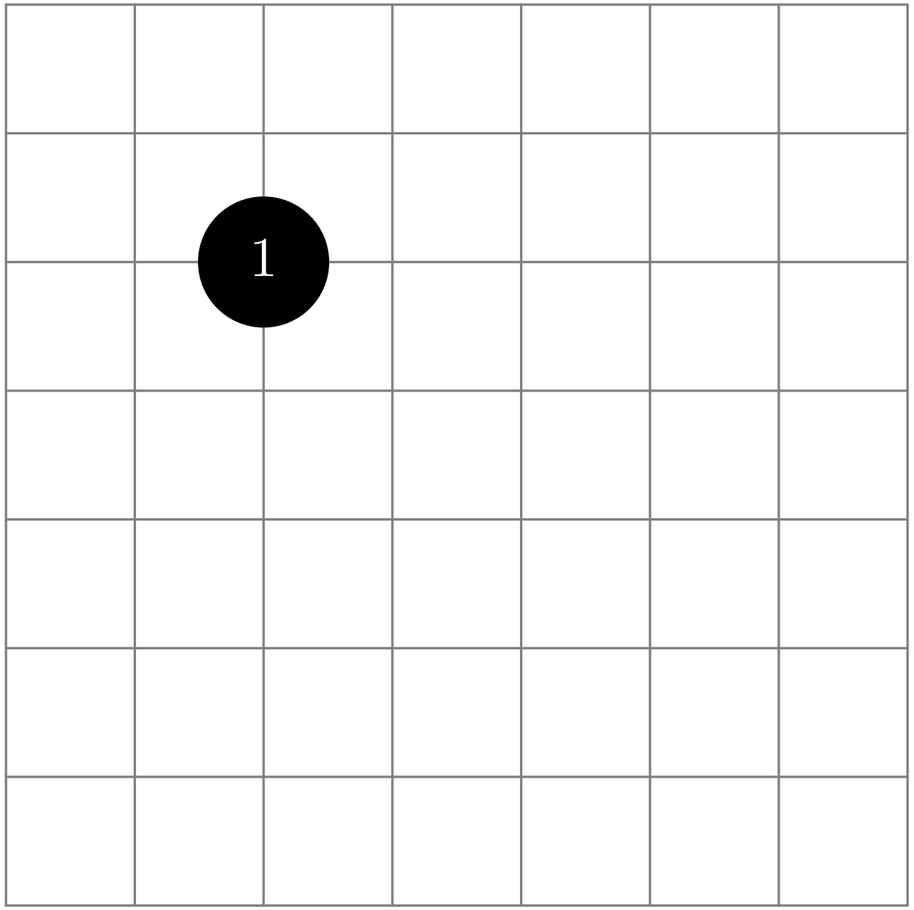






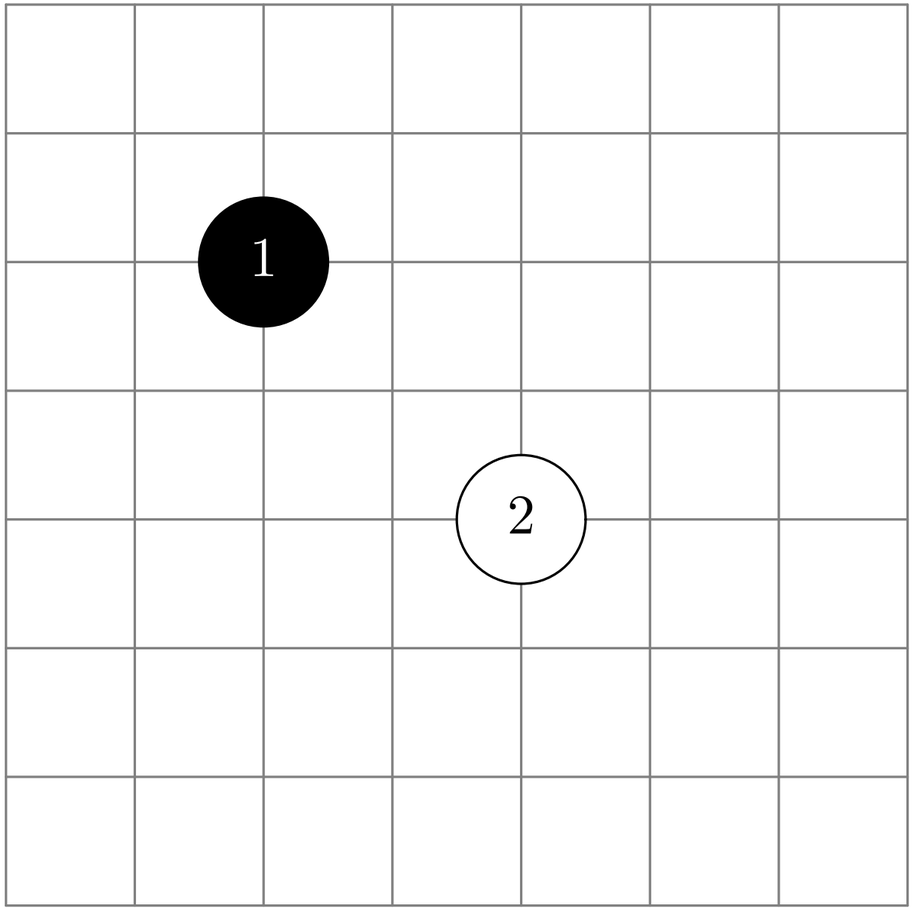
random

agent



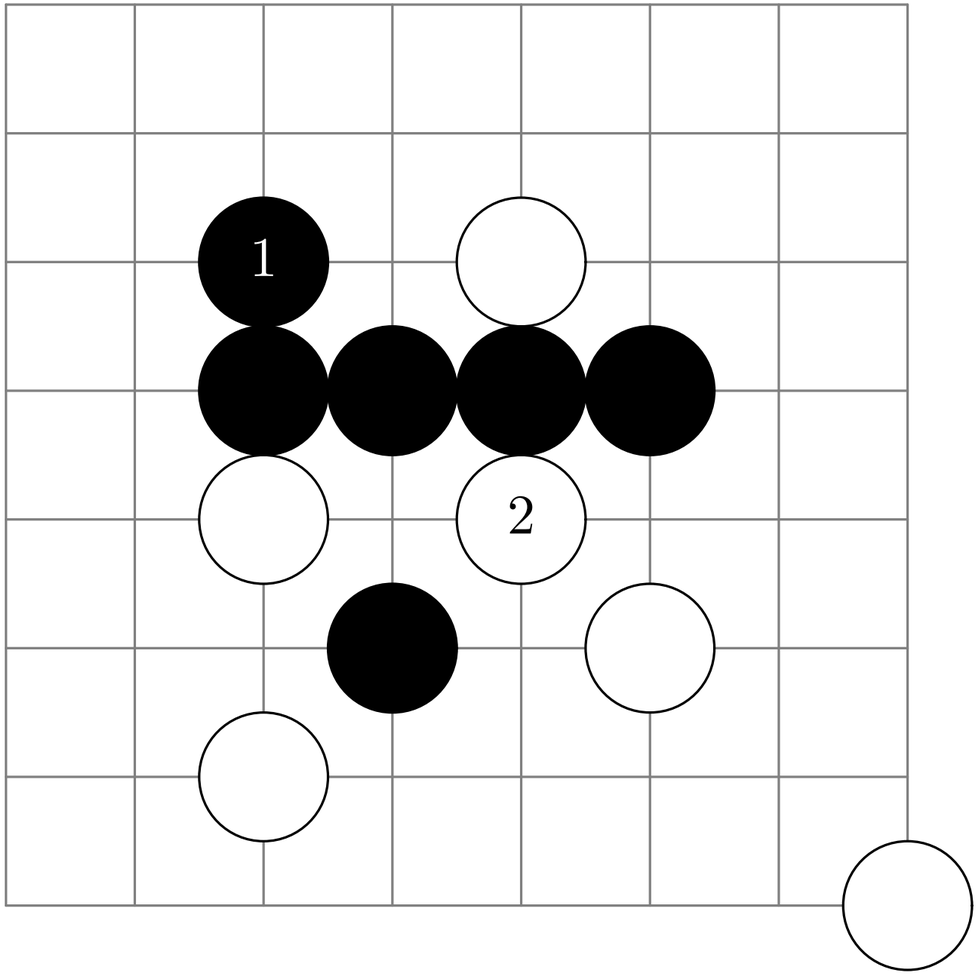
random

agent



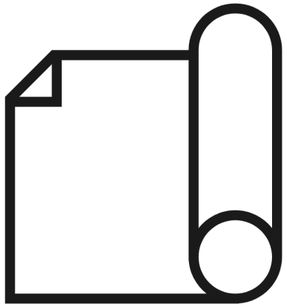
random

agent

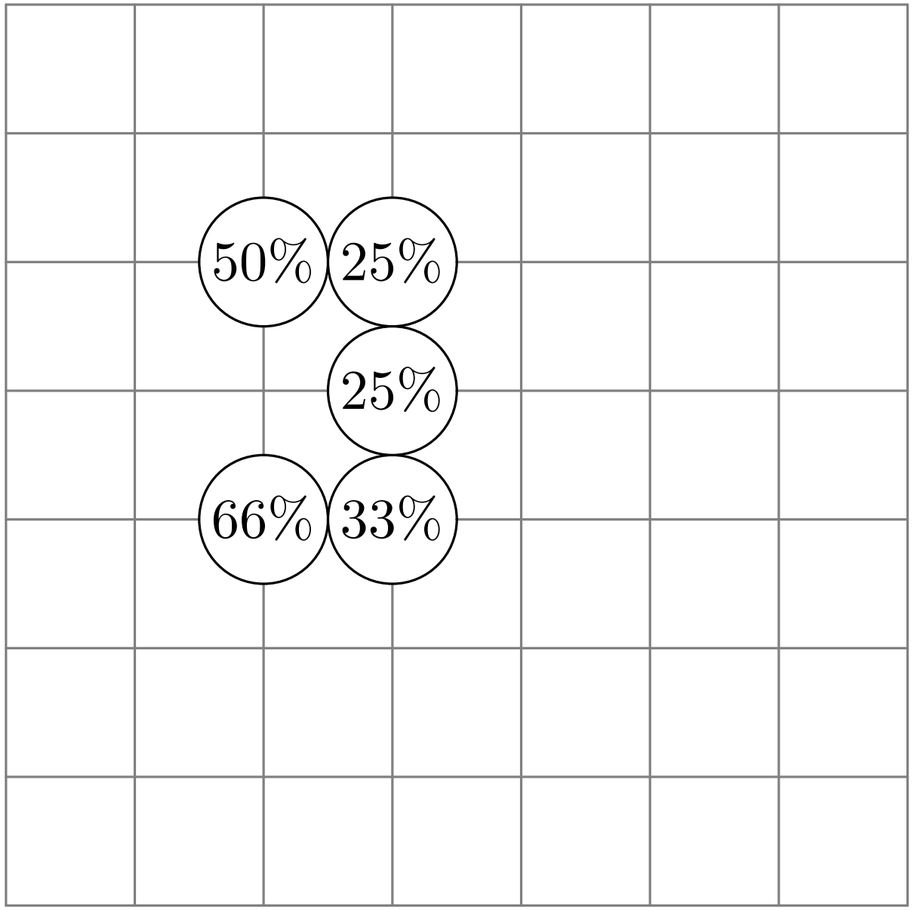


random

agent

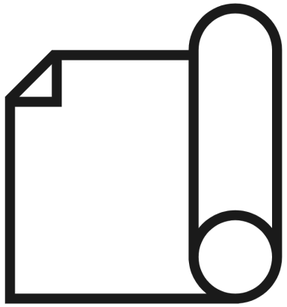


roll out

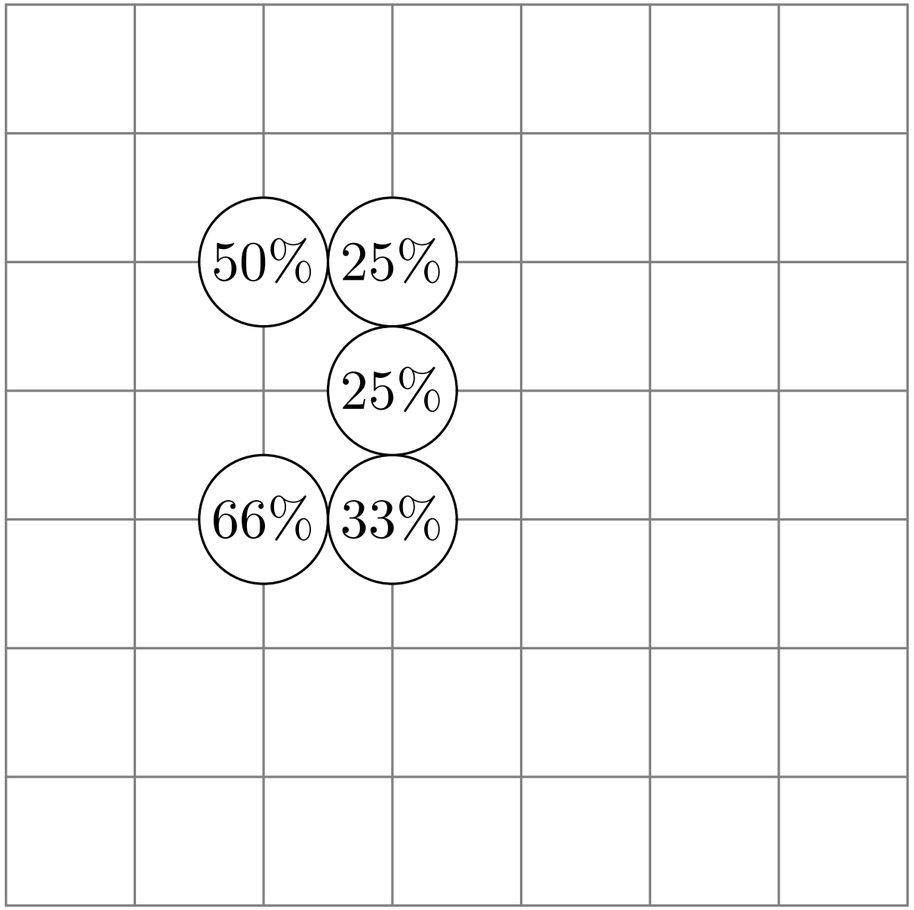


random

agent

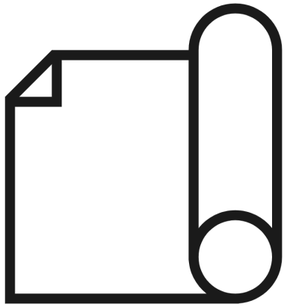


roll out

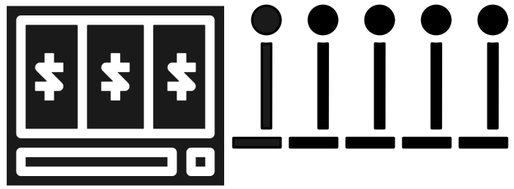


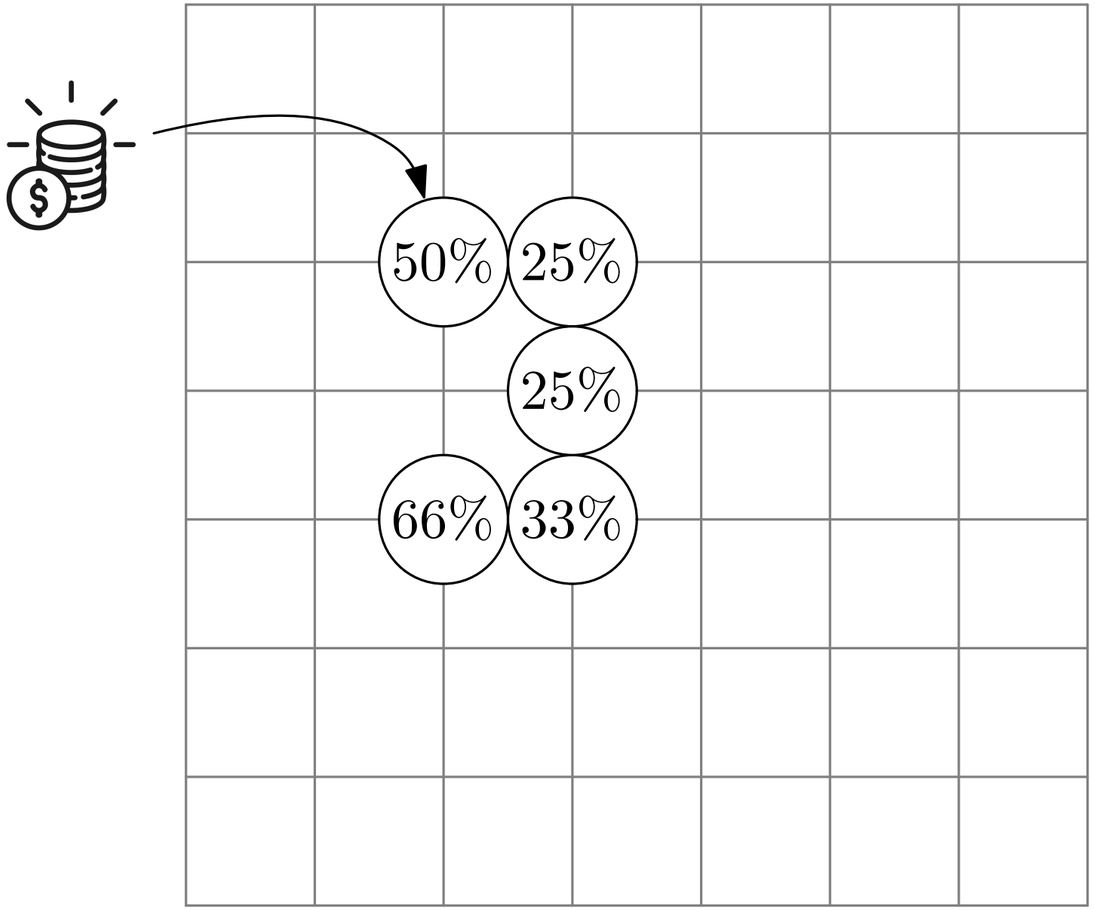
random

agent



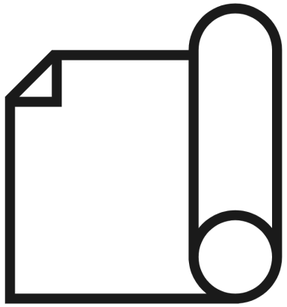
roll out



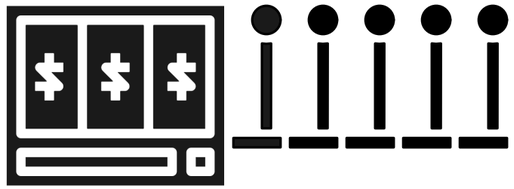


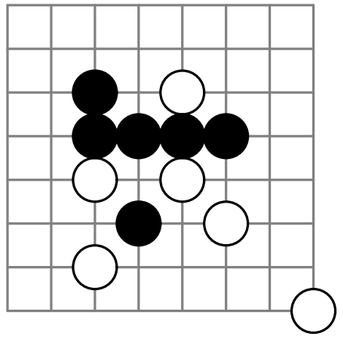
random

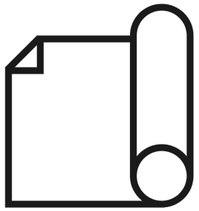
agent

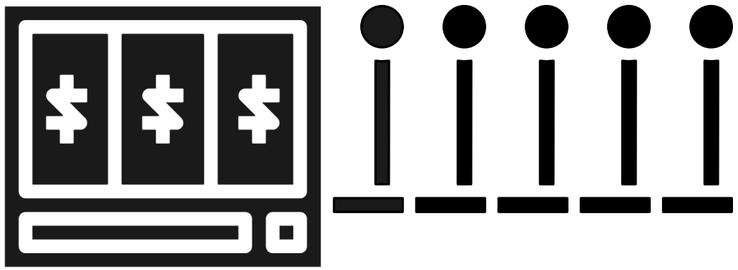


roll out

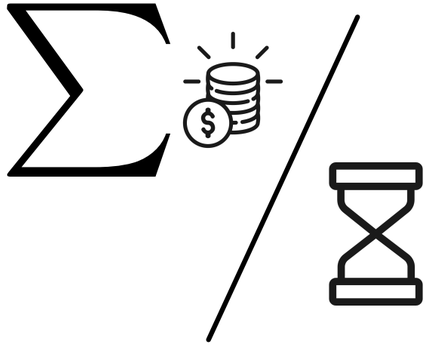




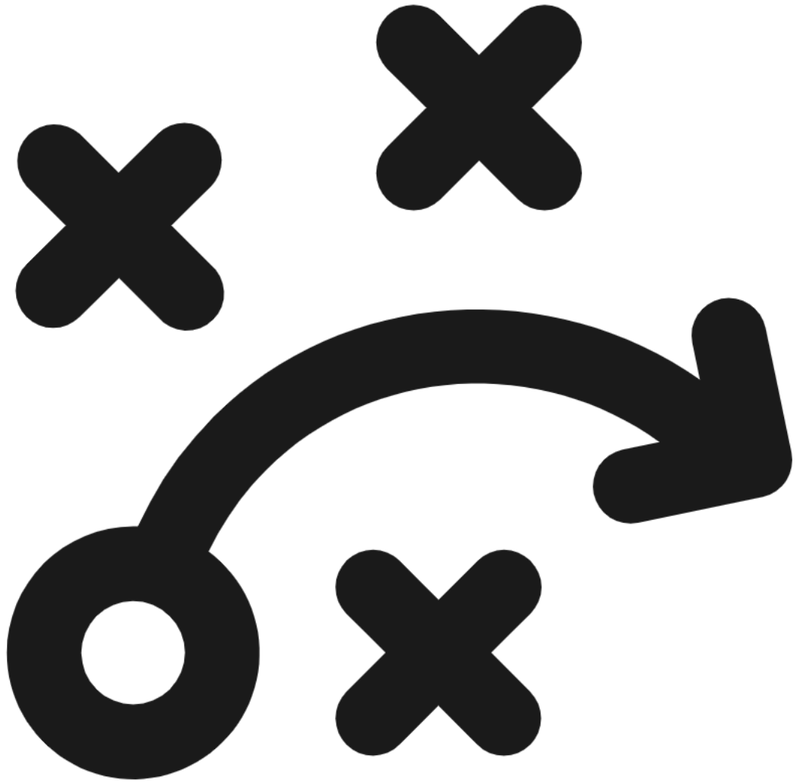




max 



Strategy 1 ε-greedy

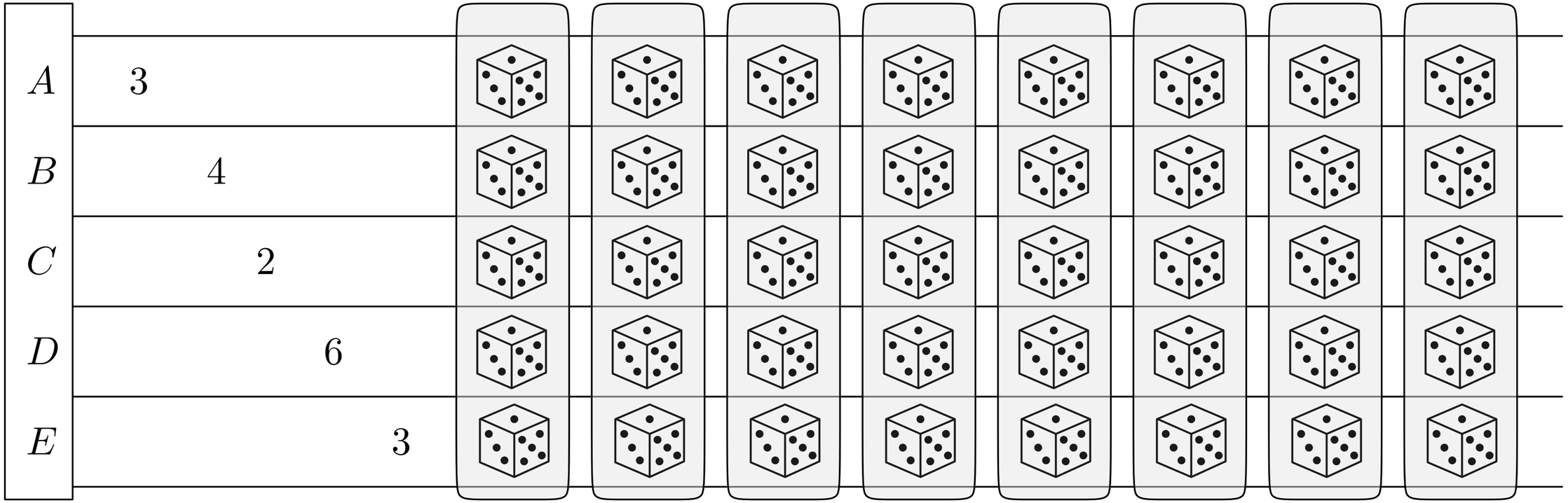


ε-greedy

ε-greedy

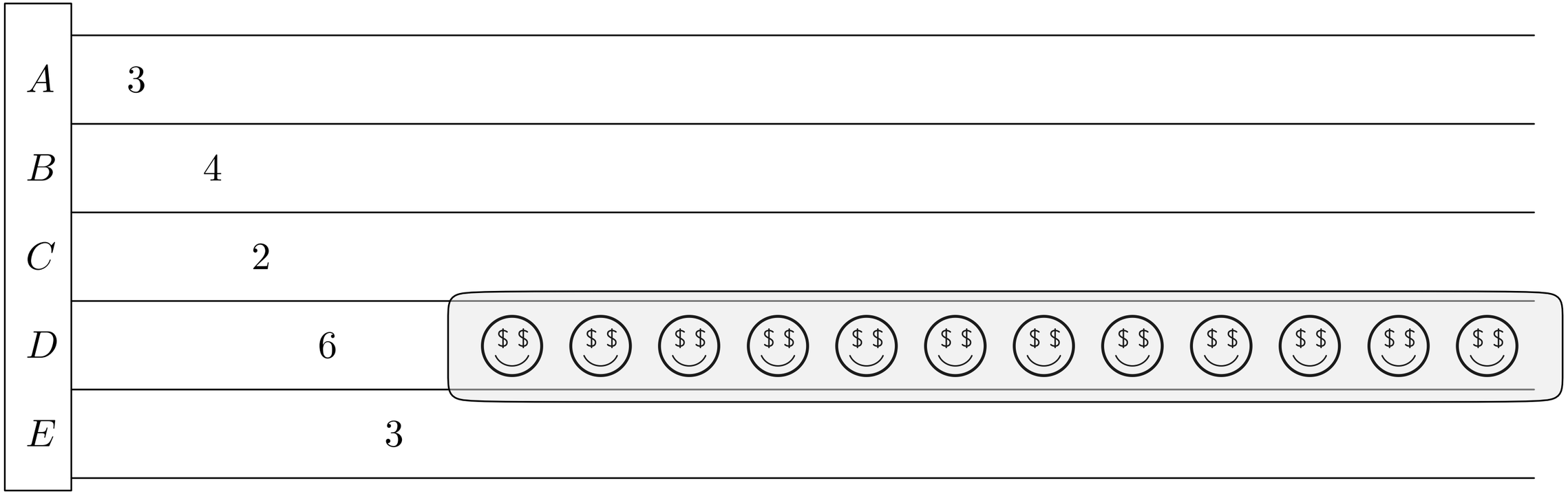
A	3
B	4
C	2
D	6
E	3

ε-greedy



exploration

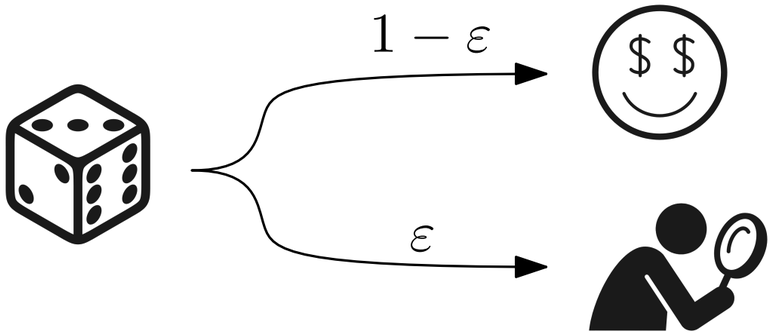
ε-greedy



exploitation = greedy

ε-greedy

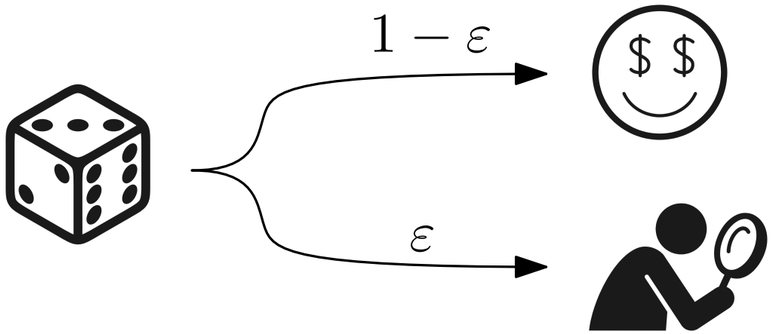
A	3
B	4
C	2
D	6
E	3



ε-greedy

A	3
B	4
C	2
D	6
E	3

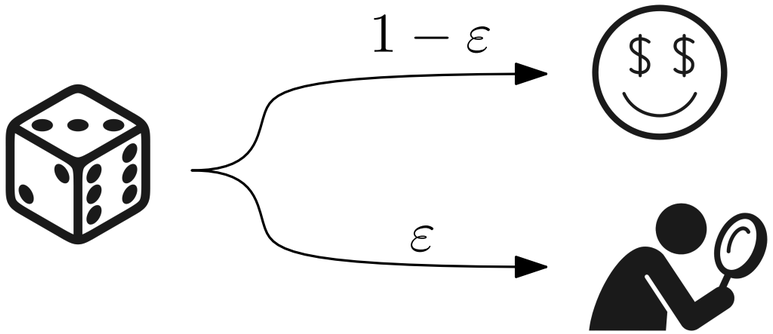
 $= ?$



$\varepsilon = 0.3$

ε-greedy

A	3
B	4
C	2
D	6  $= 0.76$
E	3



$\varepsilon = 0.3$

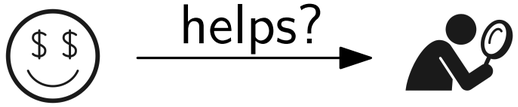
Assignment

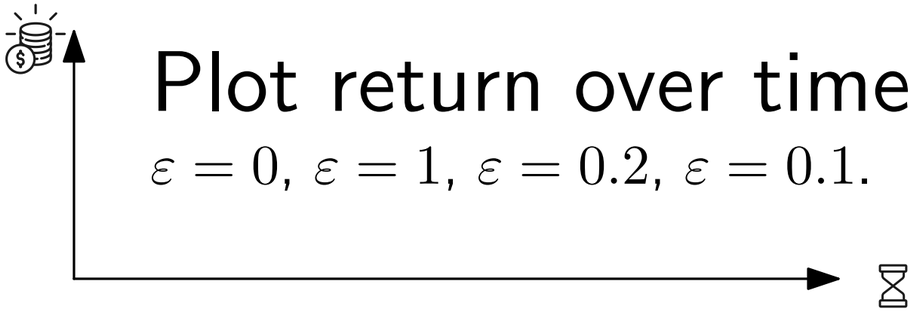
groupsize $\in [2, 3]$

20 minutes

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

Why balance?

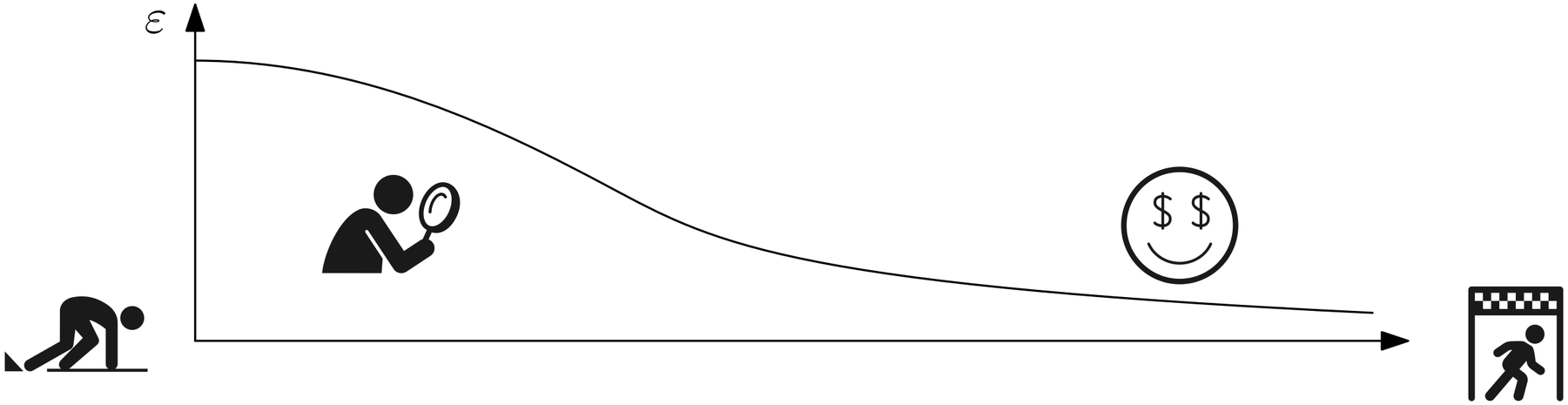




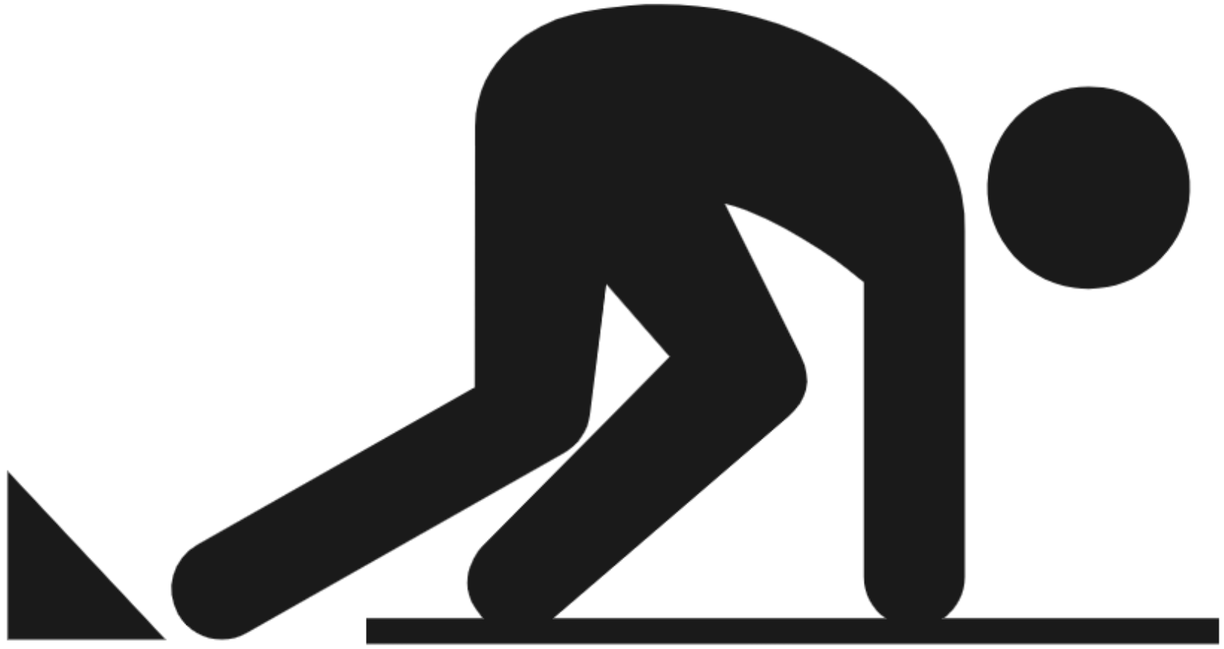
Plot return over time

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

Details



ε-decay



A	$\sum$														
B	$\sum$														
C	$\sum$														
D	$\sum$														
E	$\sum$														

A	$\sum$	0														
B	$\sum$	0														
C	$\sum$	0														
D	$\sum$	0														
E	$\sum$	0														

pessimistic initial values

A	$\sum$ 💡/# 100															
B	$\sum$ 💡/# 100															
C	$\sum$ 💡/# 100															
D	$\sum$ 💡/# 100															
E	$\sum$ 💡/# 100															

optimistic initial values

A	$\sum$ / # 100	52 4														
B	$\sum$ / # 100															
C	$\sum$ / # 100															
D	$\sum$ / # 100															
E	$\sum$ / # 100															

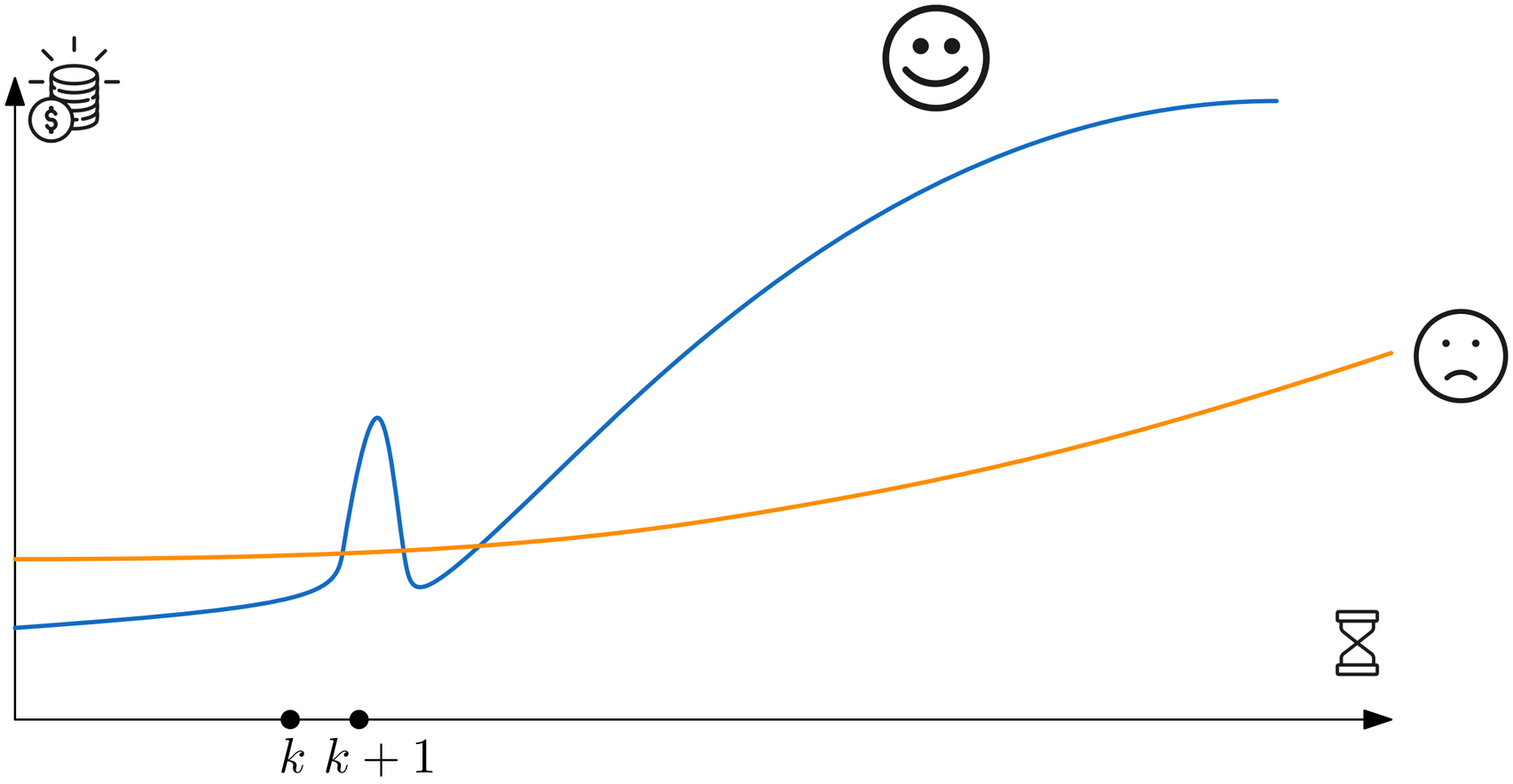
optimistic initial values

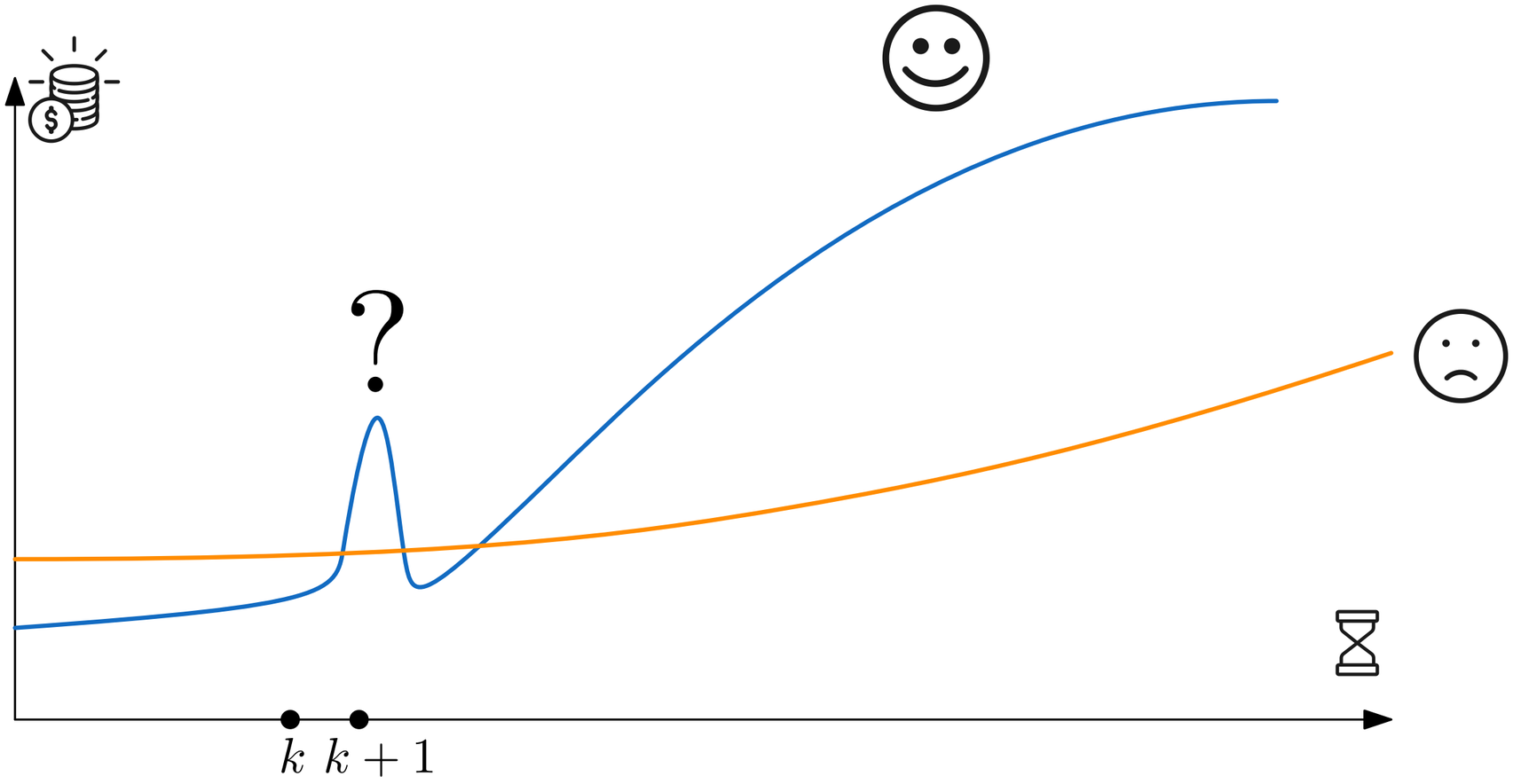
A	$\sum$ 💰/# 100	52 4														
B	$\sum$ 💰/# 100		54 8													
C	$\sum$ 💰/# 100															
D	$\sum$ 💰/# 100															
E	$\sum$ 💰/# 100															

optimistic initial values

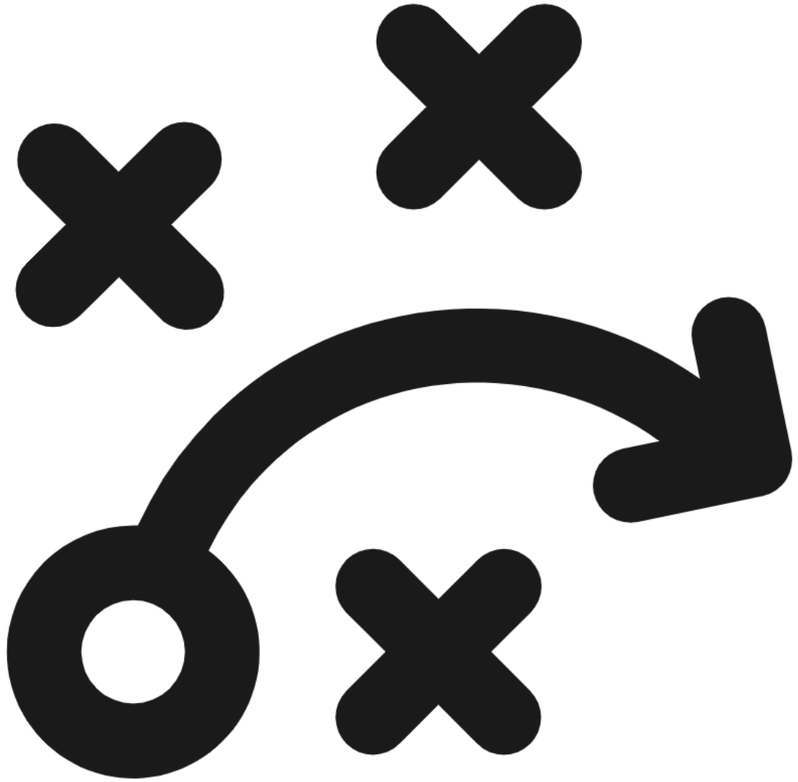
A	$\sum$ / # 100	52 4														
B	$\sum$ / # 100		54 8			40 12										
C	$\sum$ / # 100															
D	$\sum$ / # 100				50 0											
E	$\sum$ / # 100			57 14												

optimistic initial values

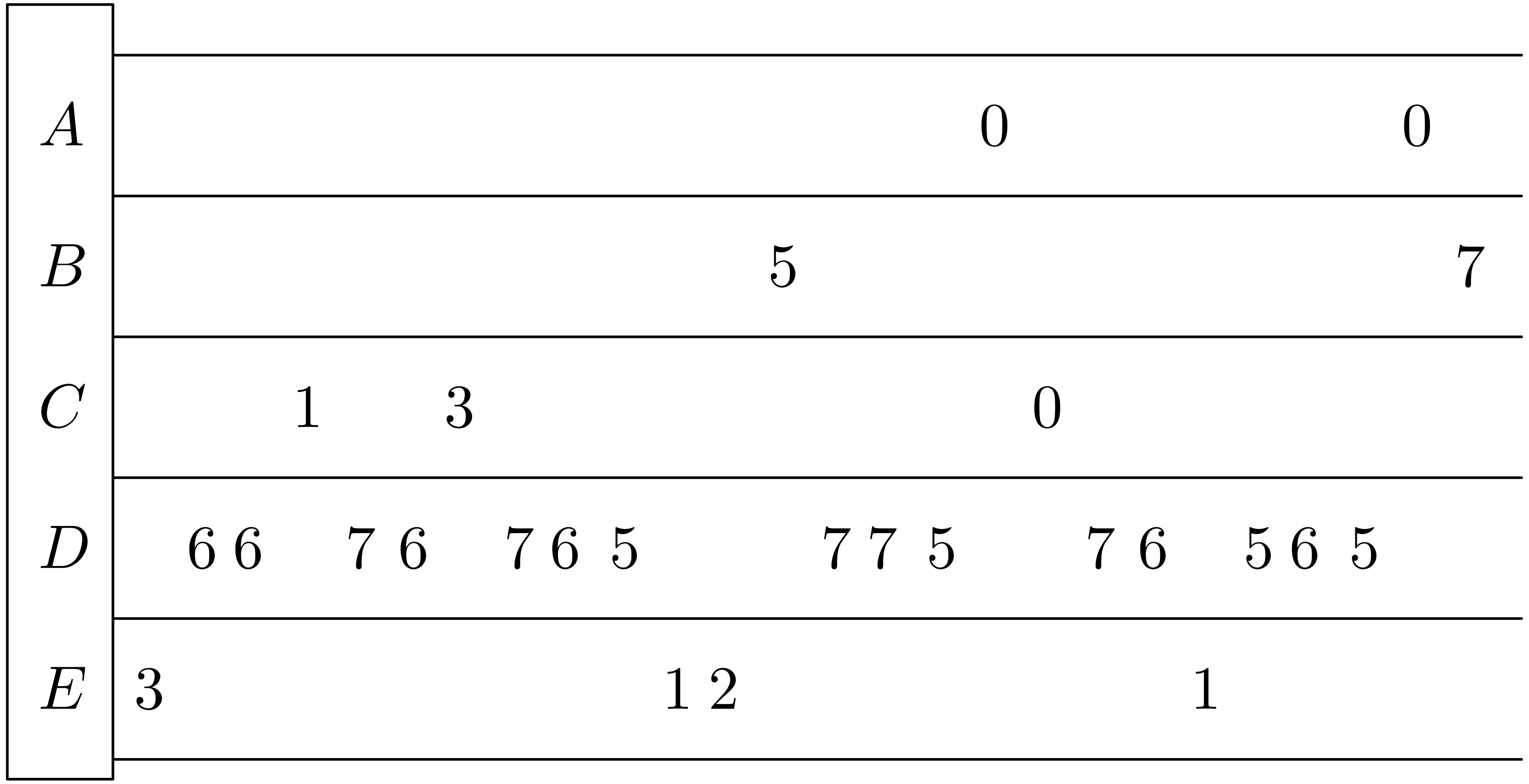

k
$k+1$


?
k $k+1$

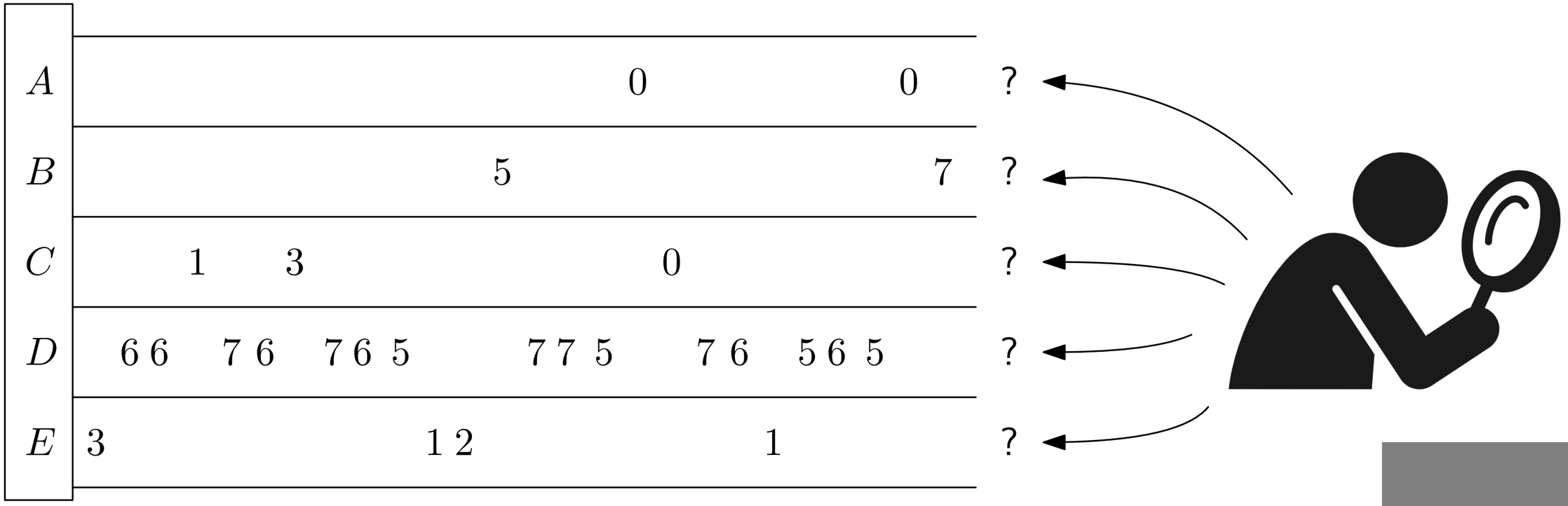
Strategy 2
Upper Confidence Bound



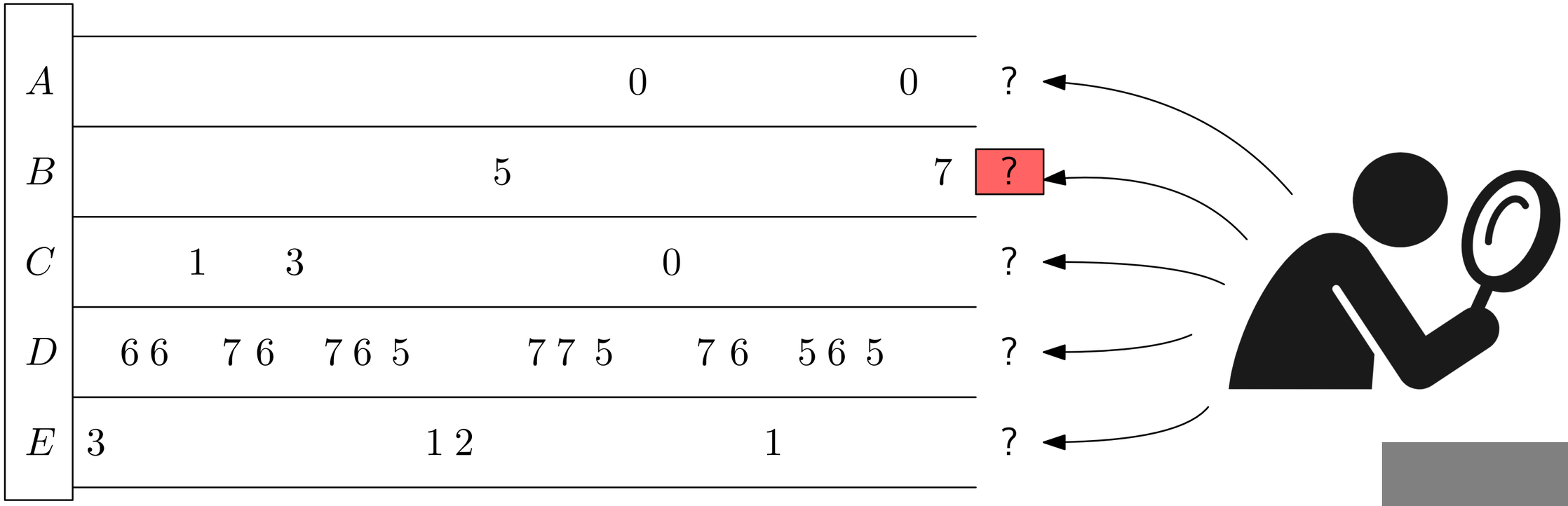
Upper Confidence Bound



Upper Confidence Bound



Upper Confidence Bound



Upper Confidence Bound

A																	0								0	
B													5													7
C				1			3											0								
D		6	6		7	6		7	6	5				7	7	5			7	6		5	6	5		
E	3										1	2									1					

Upper Confidence Bound

		Σ 💰/#	#
A	0 0	0	2
B	5 7	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

		$\sum$ 💡/#	#
A	0 0	0	2
B	5 7	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

$\sum$ 💡/# ↑

↓

Upper Confidence Bound

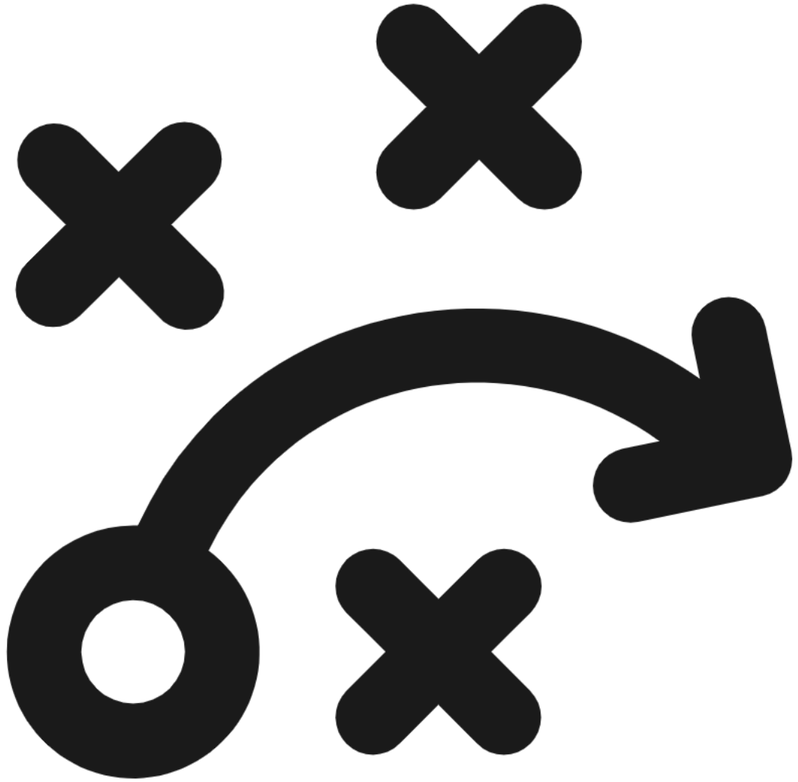
		$\sum \text{🪙} / \textcircled{\#}$	$\textcircled{\#}$
A	0 0	0	2
B	5 7	6	2
C	1 3 0	1.33	3
D	6 6 7 6 7 6 5 7 7 5 7 6 5 6 5	6.066	15
E	3 1 2 1	1.75	4

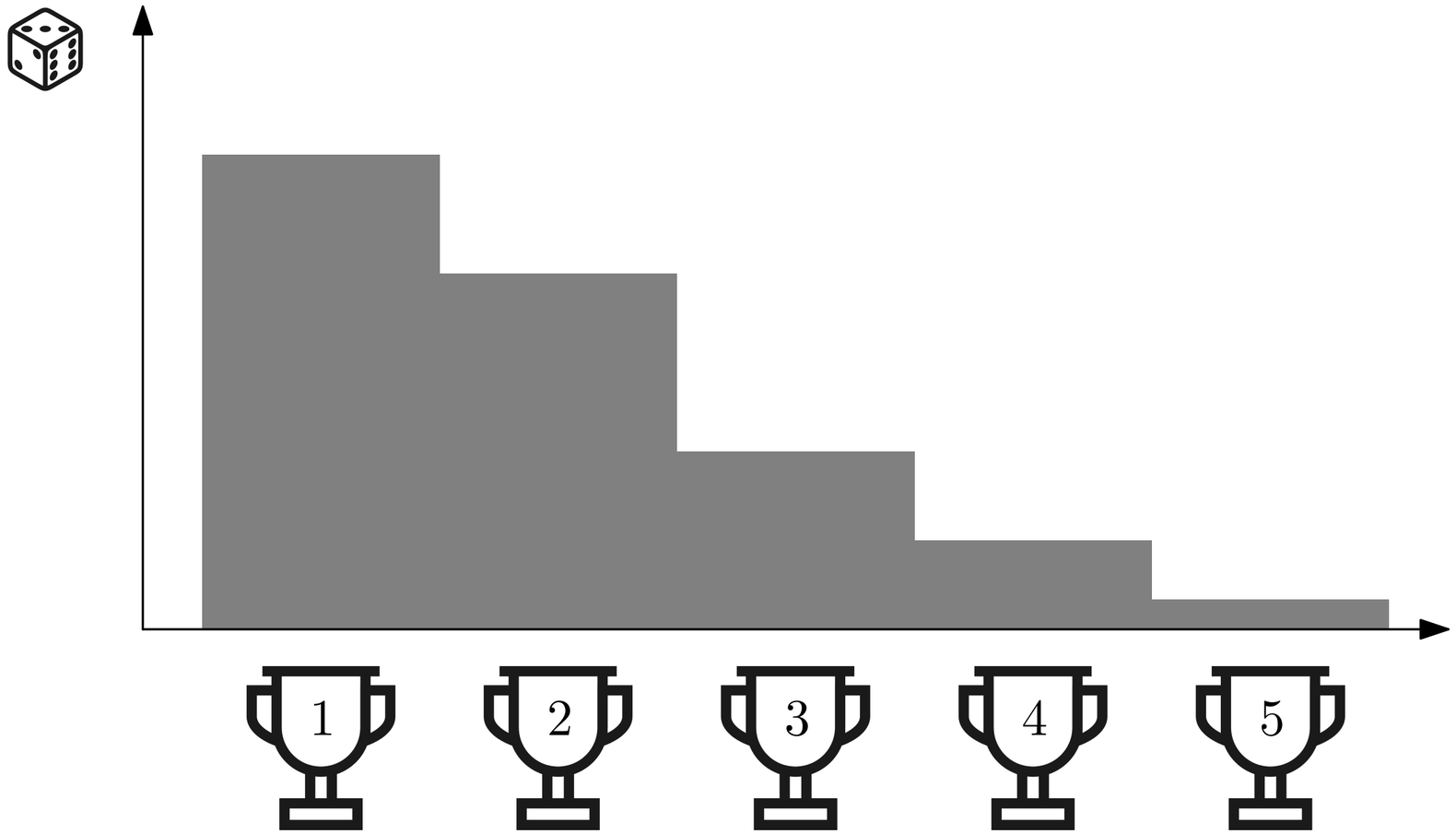
$\sum \text{🪙} / \textcircled{\#}$ ↑

$\textcircled{\#}$ ↓

$$\max \frac{\sum \text{🪙}_i}{\textcircled{\#}_i} + c\sqrt{\log \textcircled{\#} / \textcircled{\#}_i}$$

Strategy 3
Explicit Probabilities




1
2
3
4
5

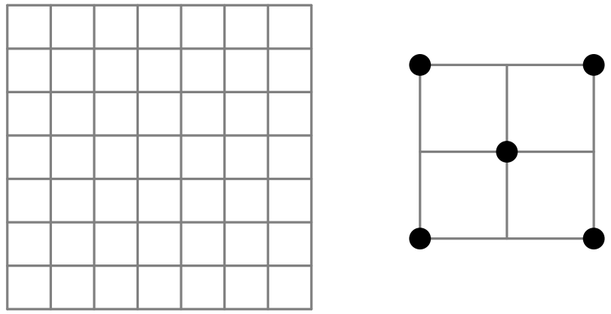
Strategy 4
Domain Knowledge

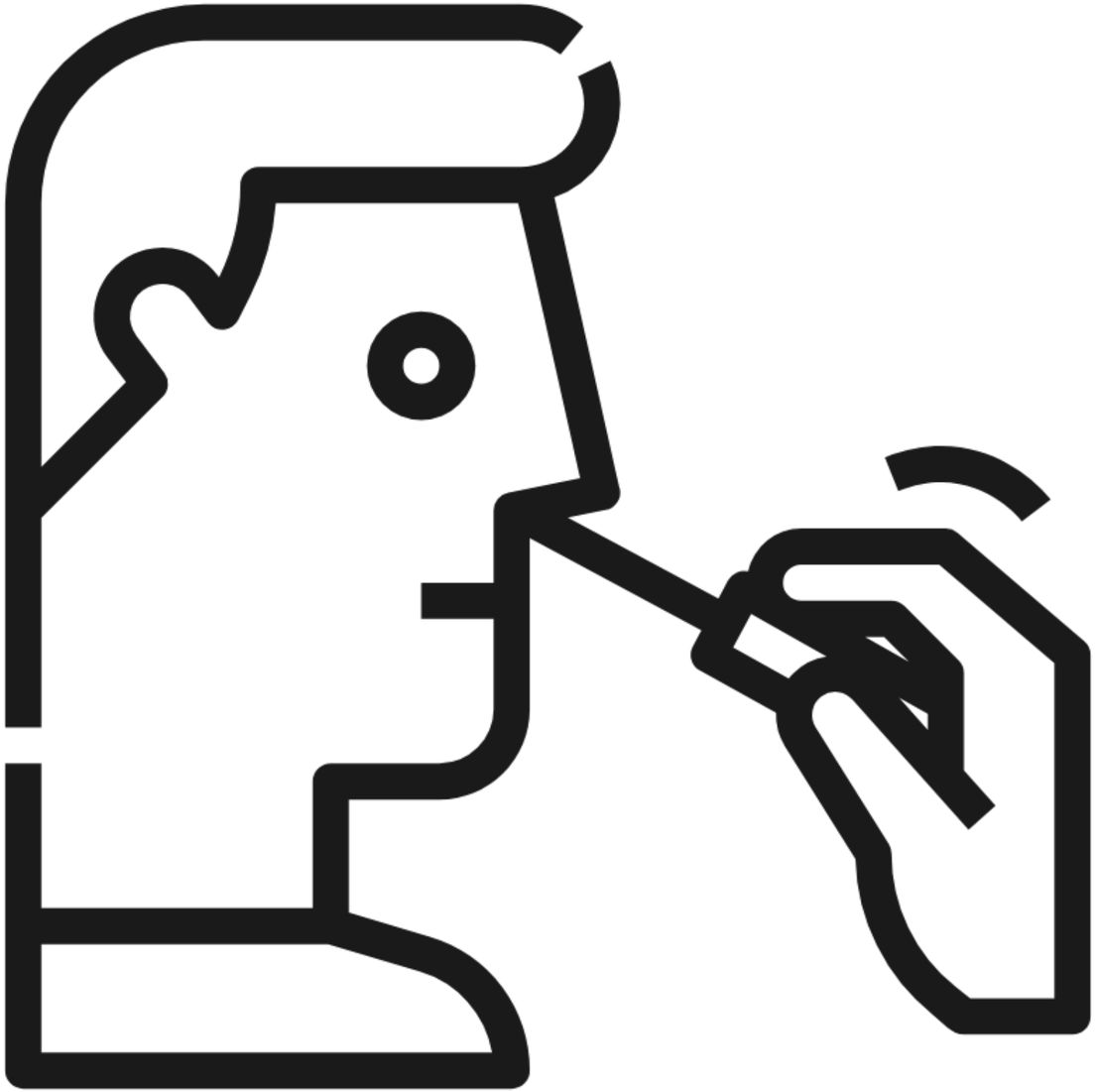
Assignment

groupsize $\in [2, 3]$

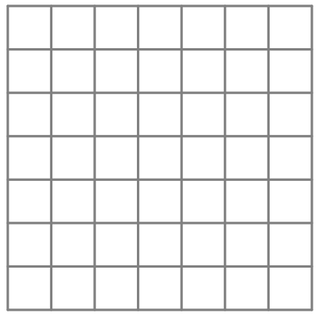
- Using domain knowledge describe an agent
- Advantages?
- Downsides?

20 minutes

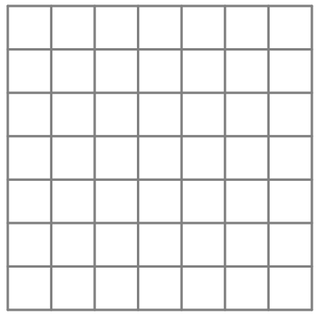




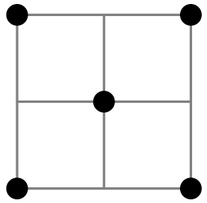
Testing



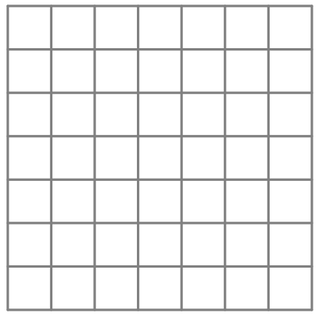
boardsize



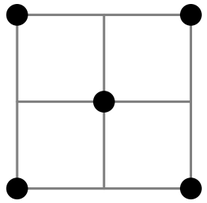
boardsize



patterns



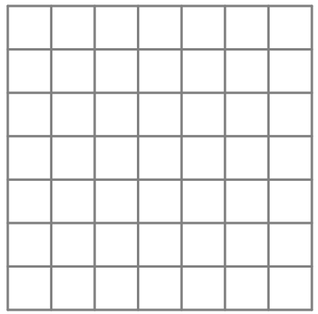
boardsize



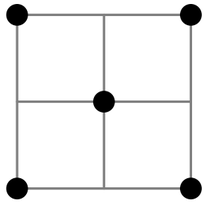
patterns

number of rollouts

time



boardsize

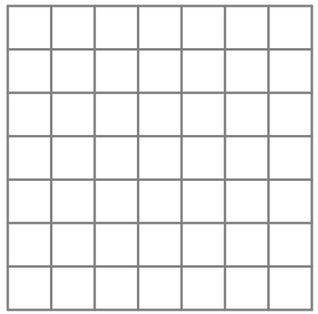


patterns

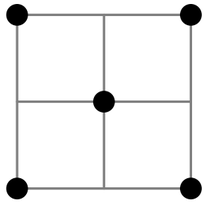
number of rollouts

time

parameters: ε, c



boardsize

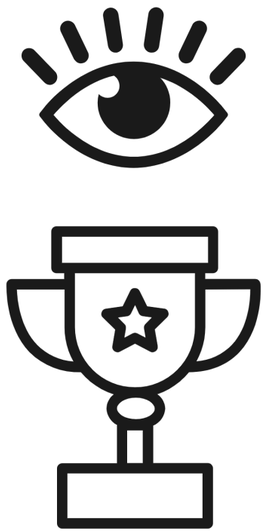


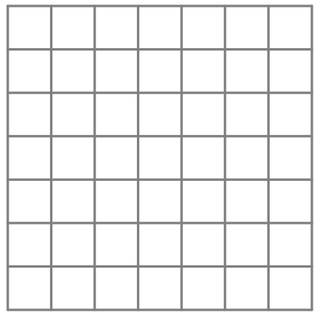
patterns

number of rollouts

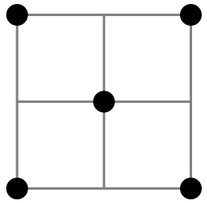
time

parameters: ε, c





boardsize

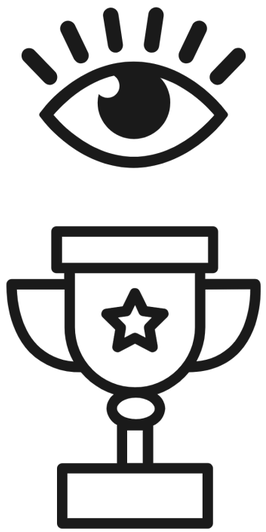


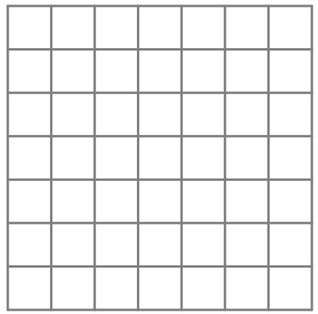
patterns

number of rollouts

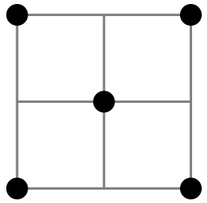
time

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?