COMPUTATIONAL APPLICATIONS TO POLICY AND STRATEGY (**CAPS**)

Session 4 – Model Deployment Evaluation

Leo Klenner, Henry Fung, Cory Combs

Outline

1. Auditing Algorithms
2. Case: Auditing Google‘s Search Algorithms
3. Auto-Complete and Recommender Algorithms
4. Reinforcement Learning
5. Failure Modes and Human-in-the-Loop Learning
6. Short Case: Determining the Agency of an Aerial Vehicle

1. Auditing Algorithms

> Primer's Chief executive, Sean Gourley, said vetting the behavior of this new technology would become so important, it will spawn a whole new industry, where companies pay specialists to audit their algorithms for all kinds of bias and other unexpected behavior.
>
> "This is probably a billion-dollar industry," he said.

1.1 Determining the Type of Audit

	Interpretable	Not interpretable
Access to the code	Easy	Difficult
No access to the code	Difficult	Hard

1.2 Challenges of Auditing in Non-Cooperative Environments

- Methods of inquiry are inherently fuzzy
- Results of audit will be imperfect
- Have to make judgement about algorithm based on imperfect information

1.3 Black-Box Testing

```
# example of simple black-box test
# we want to audit the algorithm MathOp that takes two numbers (x, y) as input
and
# performs an unknown mathematical operation on them

MathOp(3, 3)
6  # operation could be x + y or x + 3, or multiple other alternatives
MathOp(3, 2)
5  # operation seems like x + y
MathOp(1, 0)
Error  # unclear what the source of the error is, needs further investigation
```

2 Case: Auditing Google‘s Search Algorithms

WSJ INVESTIGATION

How Google Interferes With Its Search Algorithms and Changes Your Results

The internet giant uses blacklists, algorithm tweaks and an army of contractors to shape what you see

2.1 Comparison of Search Results A

Joe Biden is

GOOGLE	
done	100%
how old	100%
from	99%
running for president	79%
he democrat	78%
he running for president	76%
toast	71%
a democrat	70%

SHOW BING

DUCKDUCKGO	
an idiot	100%
creepy	100%
from what state	100%
too old to run for president	100%
a moron	94%
a liar	84%
a joke	78%
done	22%
a creep	22%

2.1 Comparison of Search Results B

Joe Biden is

SHOW DUCKDUCKGO

GOOGLE		BING	
done	100%	donald trump	100%
how old	100%	a sen	78%
from	99%	he done	78%
running for president	79%	a reclamation project	71%
he democrat	78%	presidential	64%
he running for president	76%	a grouper	58%
toast	71%	going to cure cancer	58%
a democrat	70%	issues	58%

2.1 Comparison of Search Results C

Immigrants are

SHOW DUCKDUCKGO

GOOGLE	
a blessing not a burden	100%
entrepreneurs	100%
good for the environment	98%
important	98%
more likely to be entrepreneurs	98%
net contributors	98%
treated unfairly	77%
coming from what countries	76%

BING	
given shelter	100%
increasing taxes	100%
law abiding	100%
net contributors	100%
tax exempt	100%
entrepreneurs pew	85%
funding social security	71%
a burden to taxpayers	43%

2.1 Comparison of Search Results C

Immigrants are

SHOW BING

GOOGLE	
a blessing not a burden	100%
entrepreneurs	100%
good for the environment	98%
important	98%
more likely to be entrepreneurs	98%
net contributors	98%
treated unfairly	77%
coming from what countries	76%

DUCKDUCKGO	
animals	100%
dangerous	100%
less likely to commit crimes	100%
ruining america	100%
taking our jobs	100%
good	84%
an infestation	77%
not criminals	77%

3. Auto-Complete and Recommender Algorithms

How do auto-completion algorithms work?

- A user provides the beginning of a search query and the auto-complete algorithm provides the user with a number of suggested alternatives for completing the query

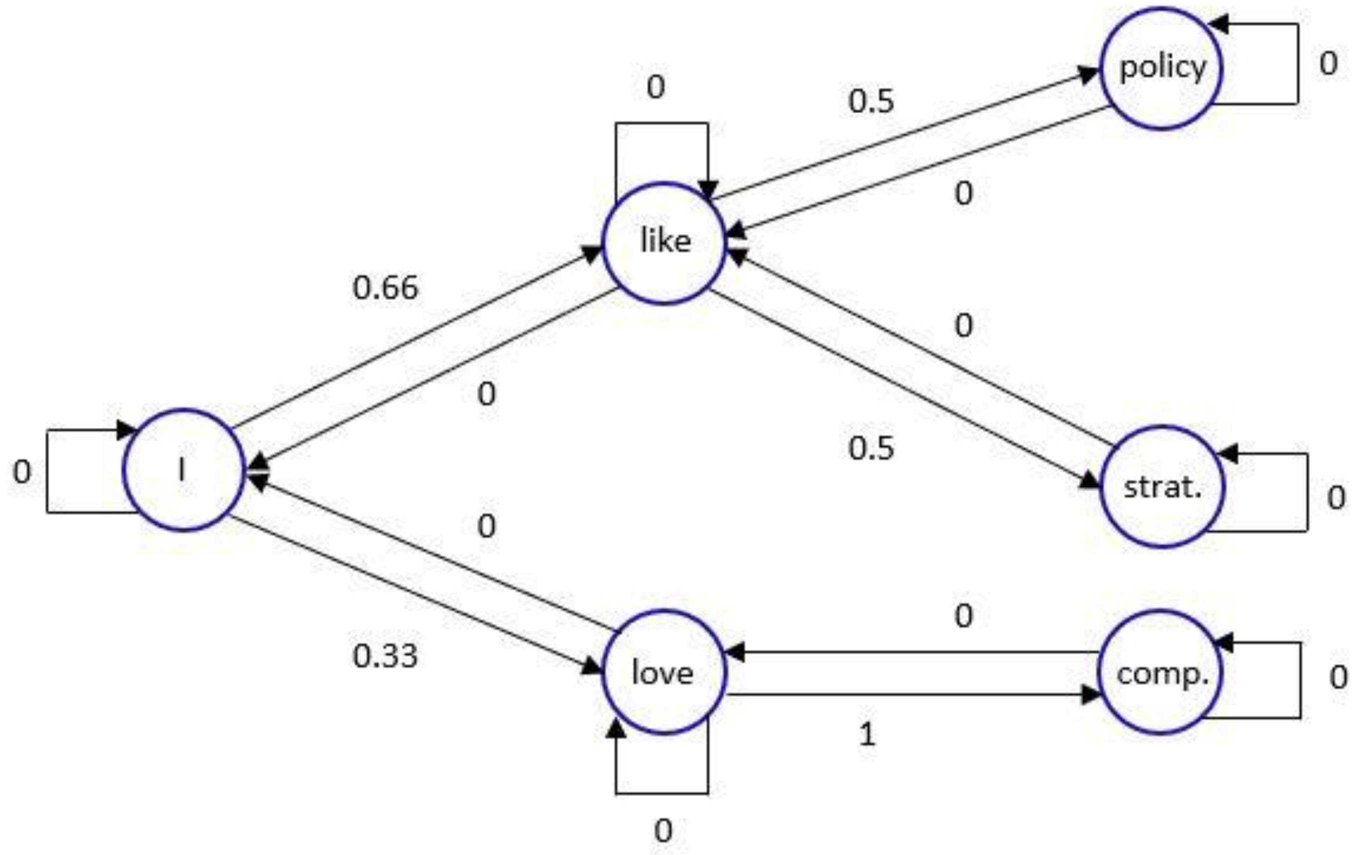
How do recommender algorithms work?

- A user makes a choice among alternatives (movies, items, etc.) and, based on features of the choice, the recommender algorithm generates new alternatives for the user

3.2 Sequential Probabilistic State Transitions

How do auto-completion algorithms work?

- A user provides the beginning of a search query and the auto-complete algorithm provides the user with a number of suggested alternatives for completing the query



3.3 Optimal Action Selection Under Uncertainty

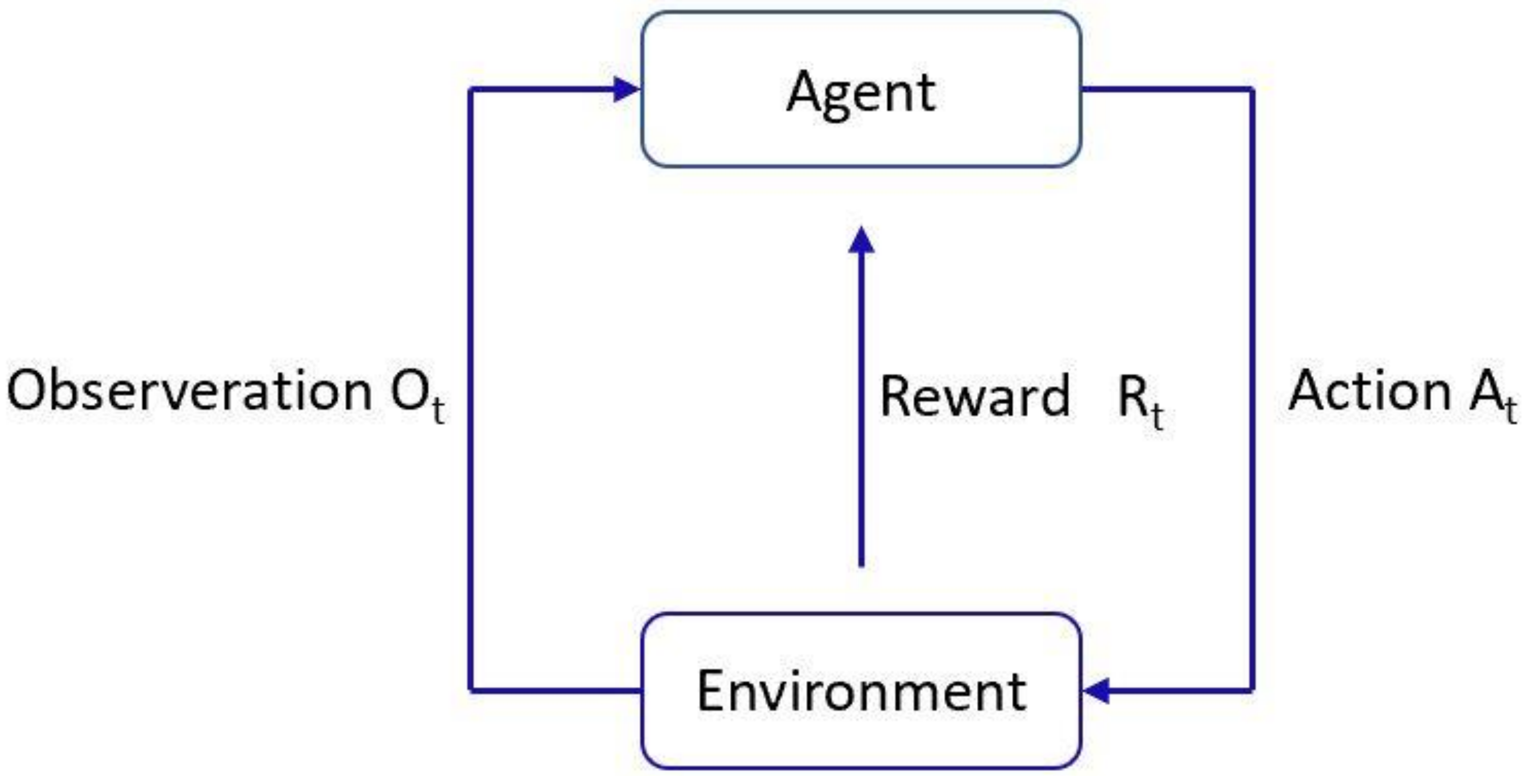
How do recommender algorithms work?

- A user makes a choice among alternatives (movies, items, etc.) and, based on features of the choice, the recommender algorithm generates new alternatives for the user

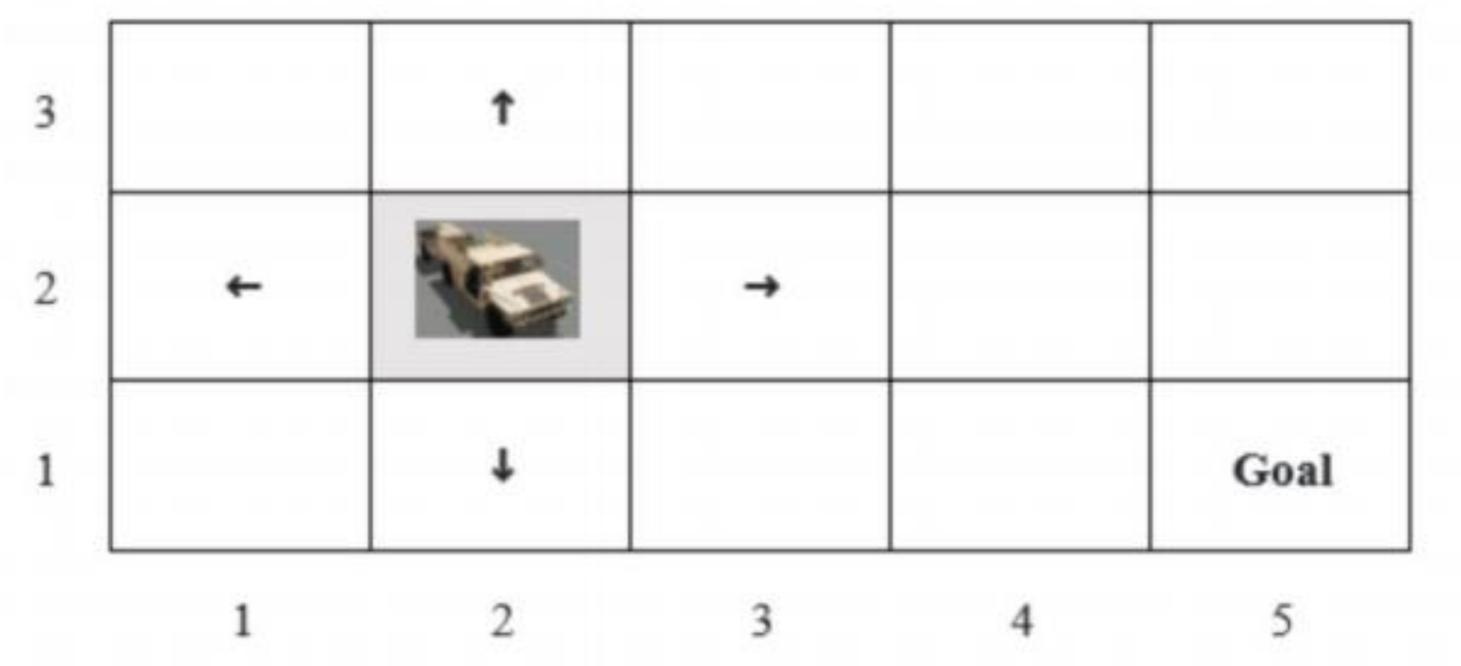
4. Reinforcement Learning

> Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards. These two characteristics —trial-and-error search and delayed reward—are the two most important distinguishing features of reinforcement learning.

4.1 Markov Decision Processes



4.2 Simple Reinforcement Learning Example

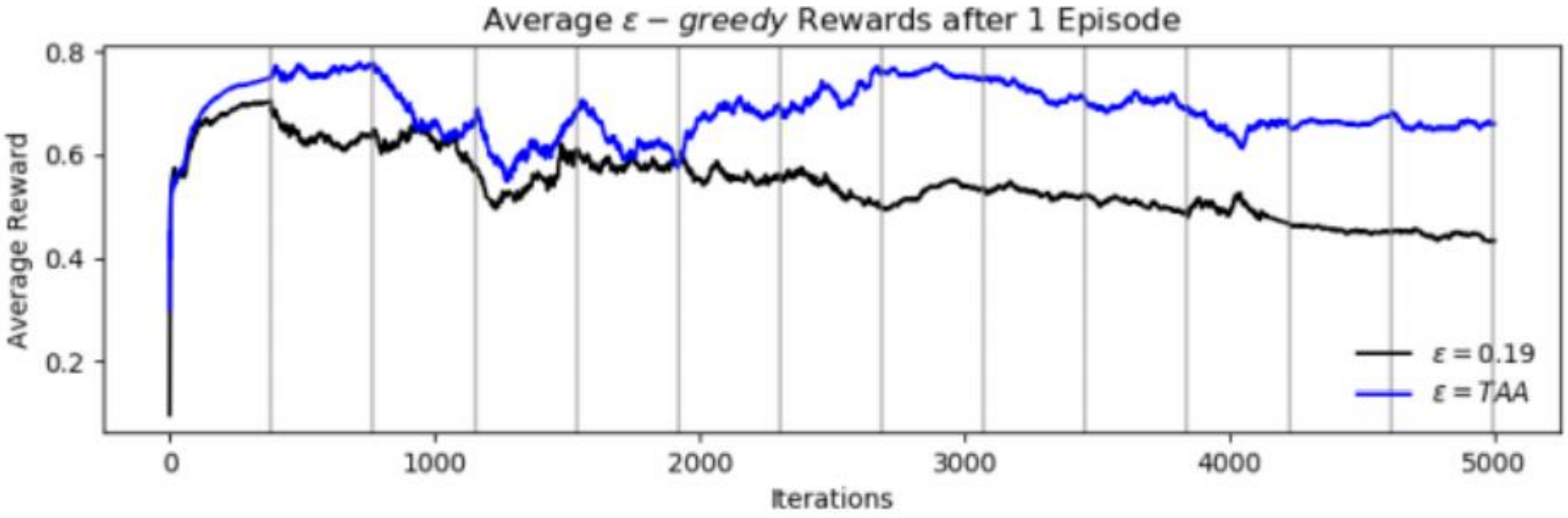


4.3 Specifying a Multi-Armed Bandit

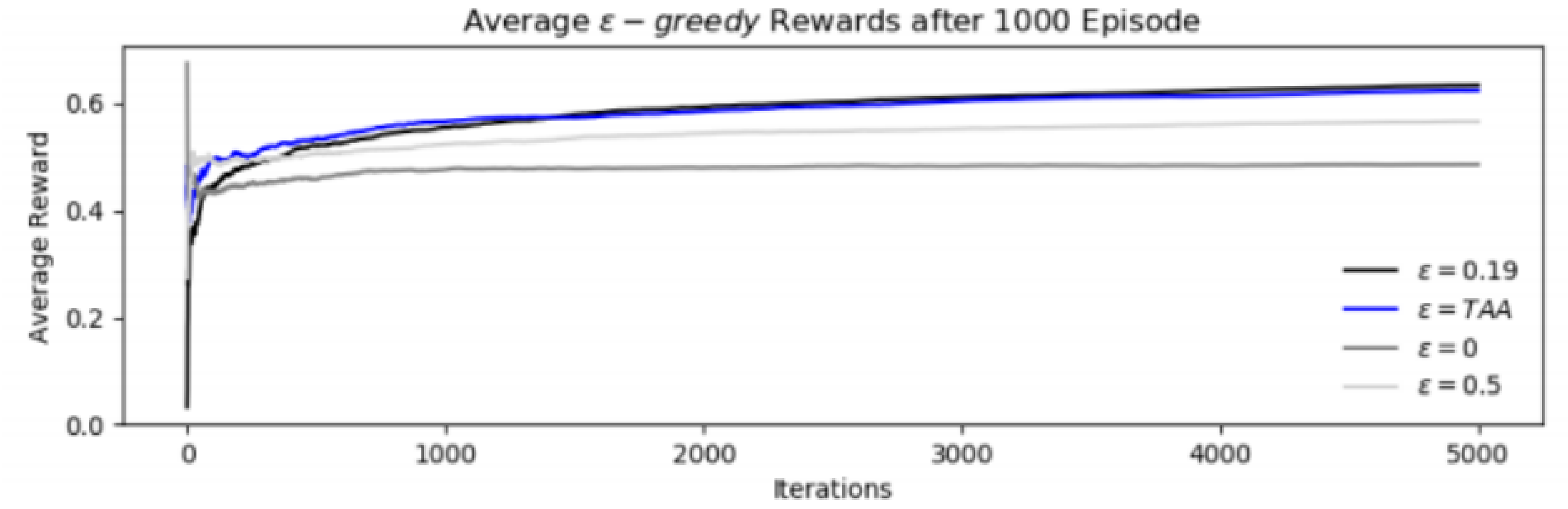
Action	0	1	2	3	4	5	6	7
Mean	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Variance	2	2	2	2	2	2	2	2

4.4 Epsilon-Greedy Action Selection

4.4 Epsilon-Greedy Rewards



4.4 Epsilon-Greedy Rewards after 5000 Episodes



5. Failure Modes and Human-in-the Loop Learning

Reinforcement learning is a powerful technique but it comes with many unexpected failure modes that we often need human supervisors to fix.

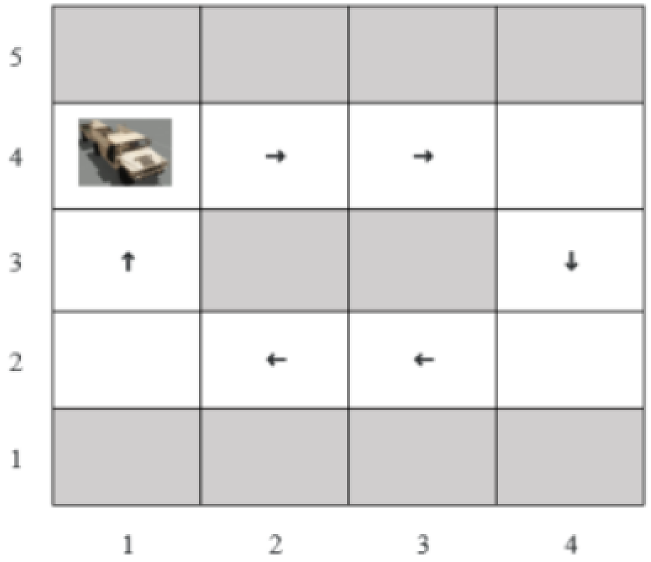
In this section, as an exercise, we look at two of these failure modes and discuss how humans can fix them.

We look at **reward gaming** and **negative side effects**.

5.1 Reward Gaming

Reward Gaming

> Agent exploits an unintended loophole in the reward specification, to get more reward than deserved



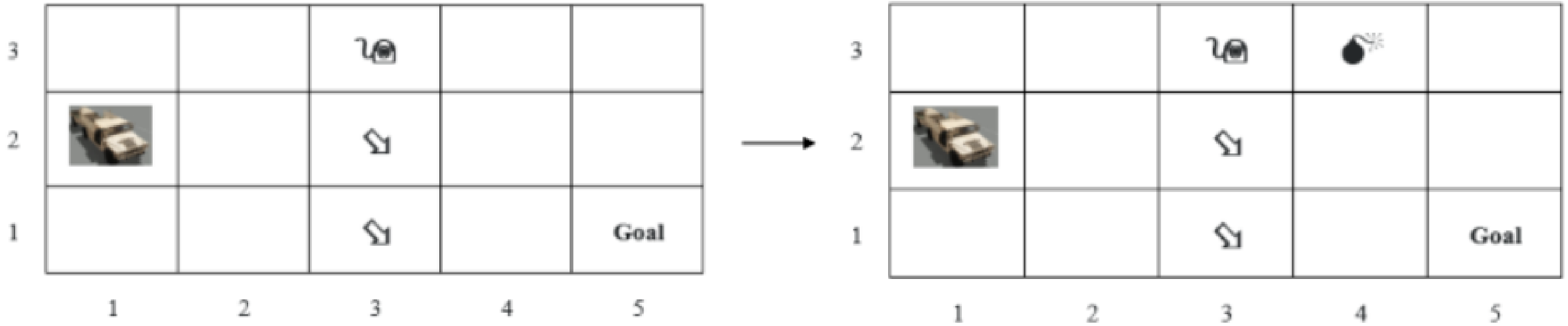
> Desired outcome: clockwise completion of race

> Arrows are checkpoints associated with a reward of 3

5.2 Negative Side Effects

Negative Side Effects

> Reward function does not fully capture all the properties of the test environment



Training environment *Test environment*

> Desired outcome: reach goal state

> (spotted by enemy) = -1, (bad terrain) = -3, (land mine) = - 100, **Goal** = 10

5. Short Case: Determining the Agency of an Unknown AV