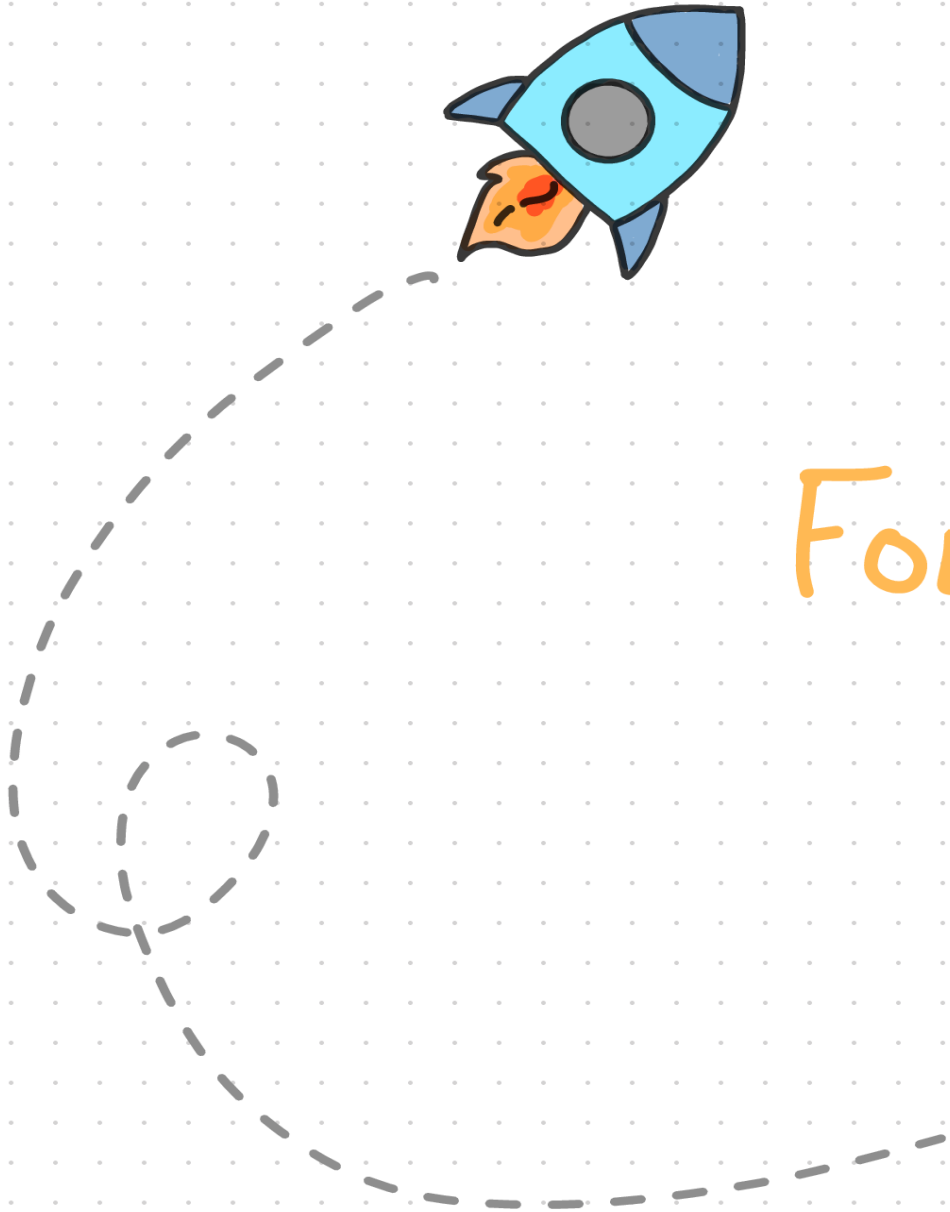


IARCS VSS
10 Feb 2026

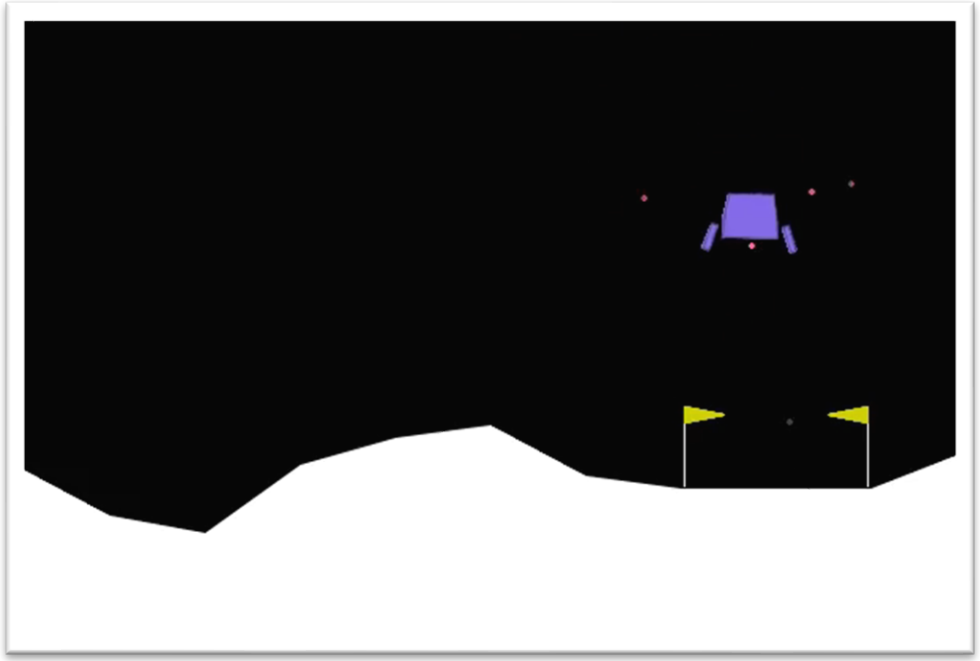
Following The Stars

Shielding Learnt Policies

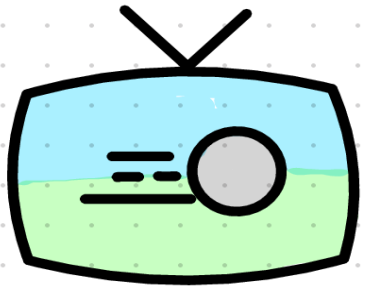
Ashwani Anand
MPI-SWS

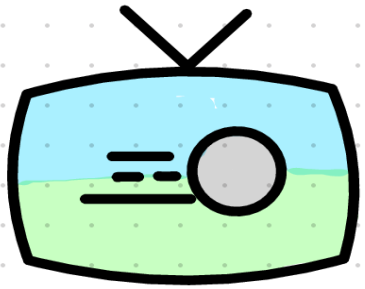
with S. A. Nayak, R. Raha and A.-K. Schmuck

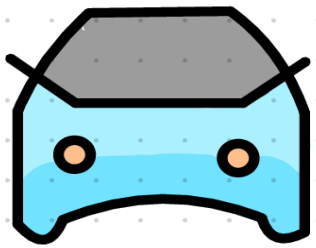
Before

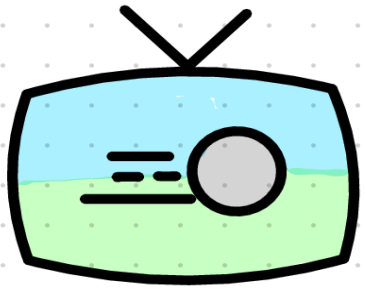


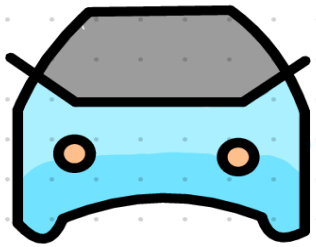
After

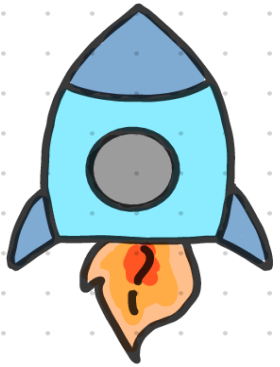


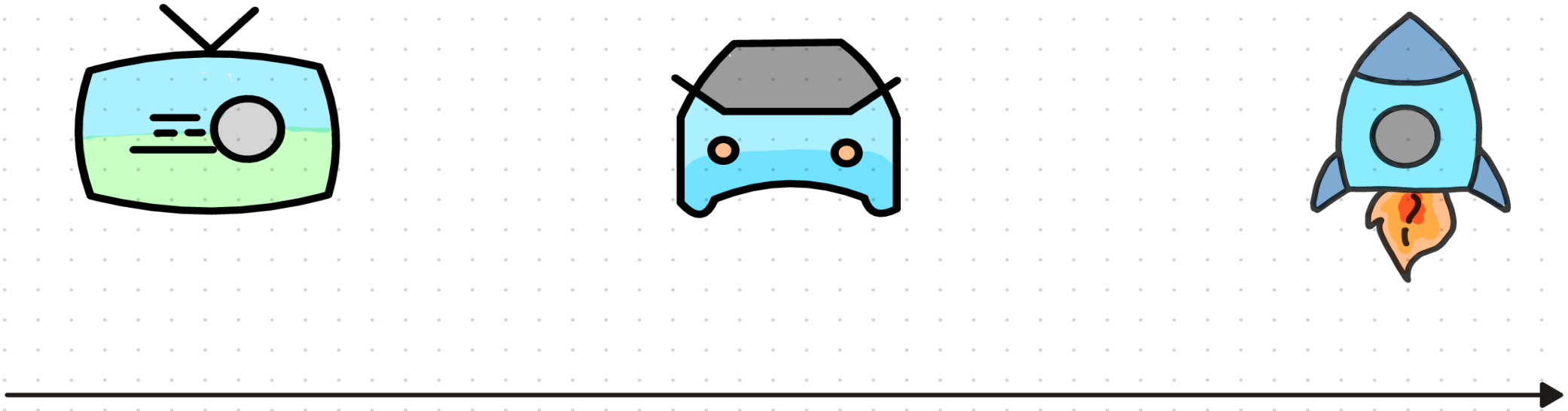


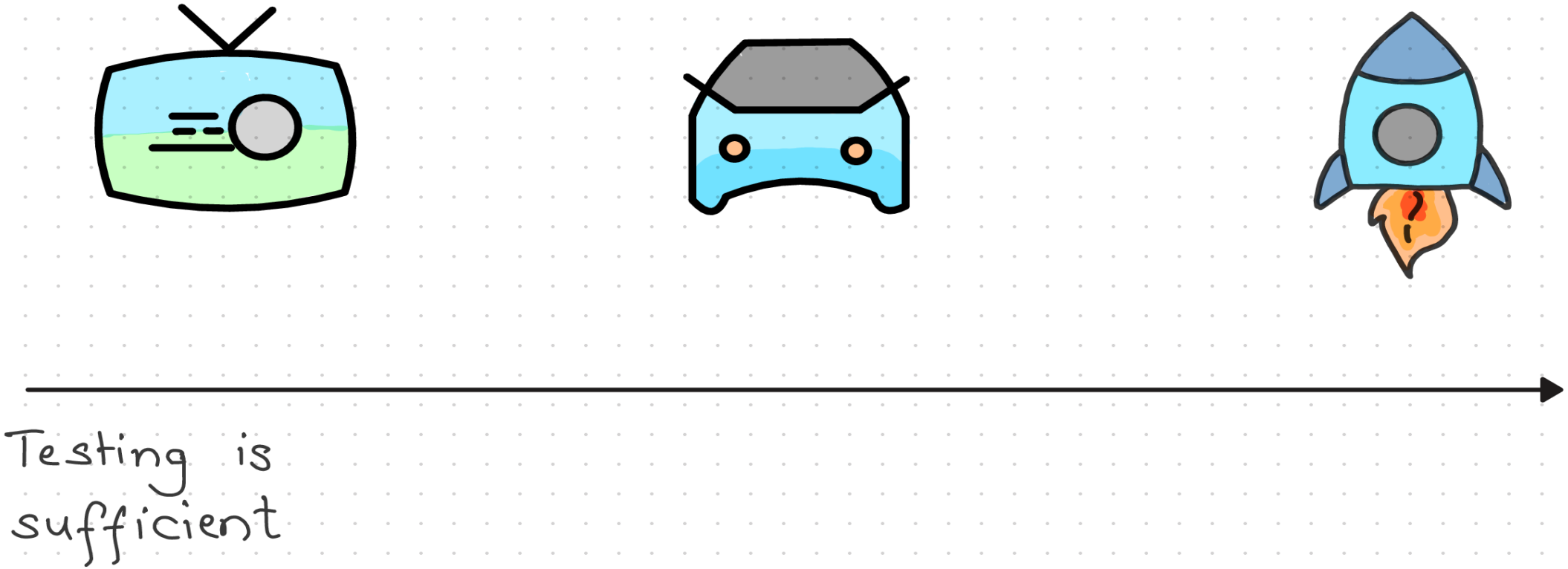


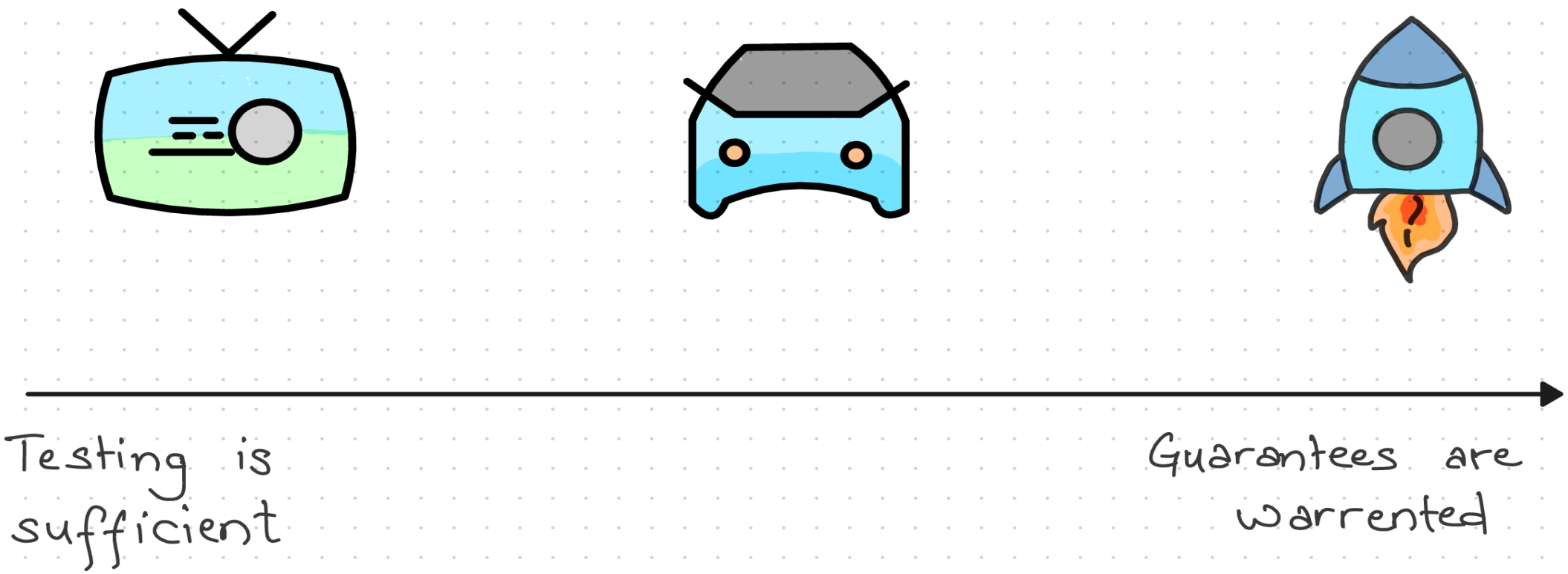






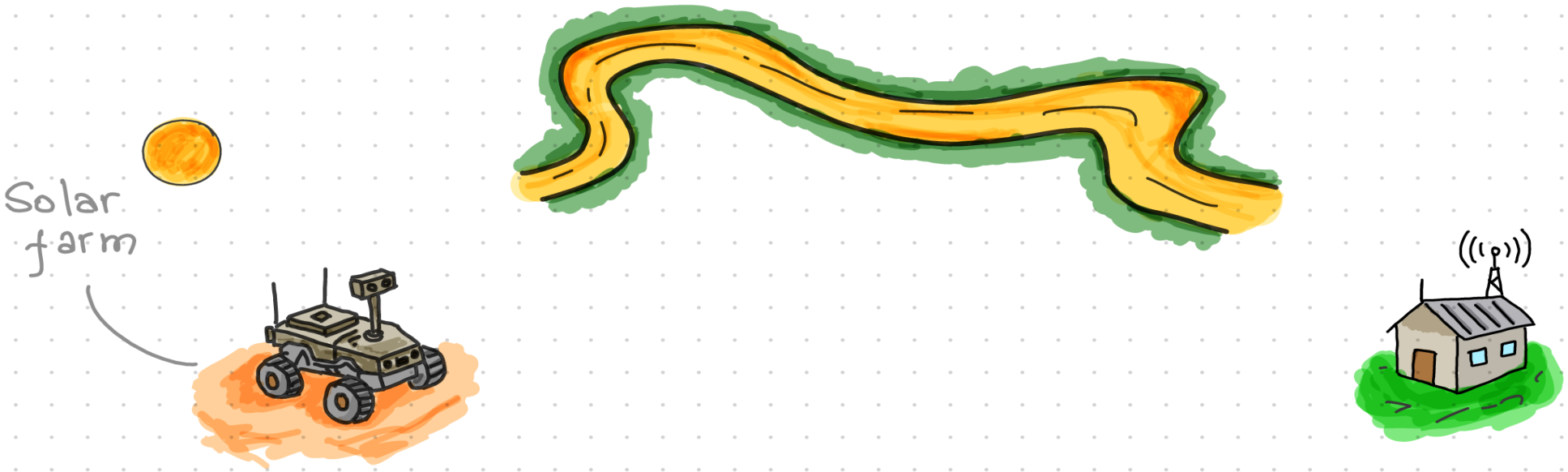



Testing is
sufficient

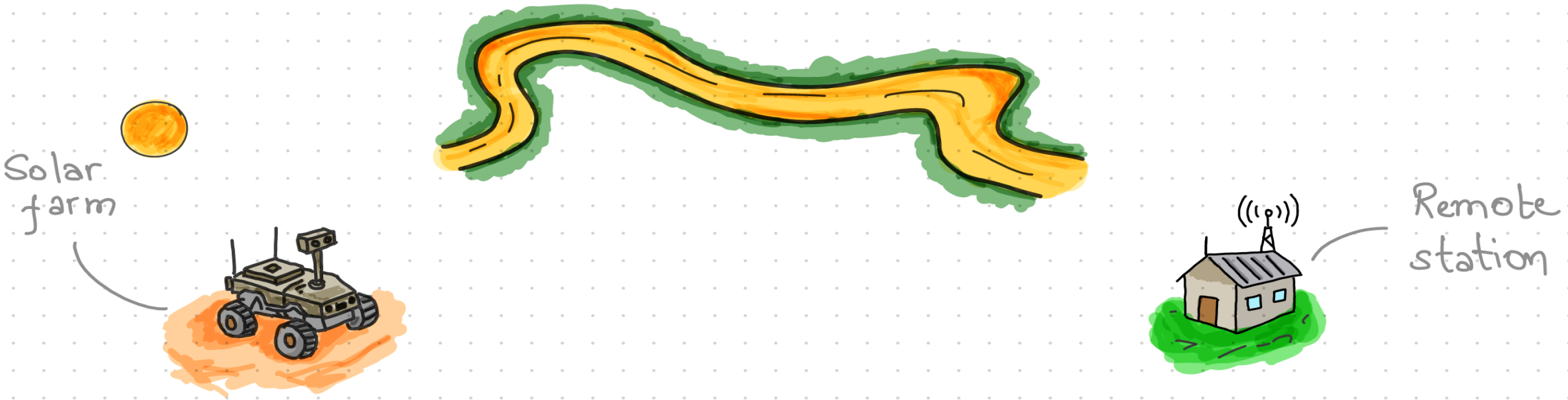

Testing is sufficient
Guarantees are warrented

A simple scenario

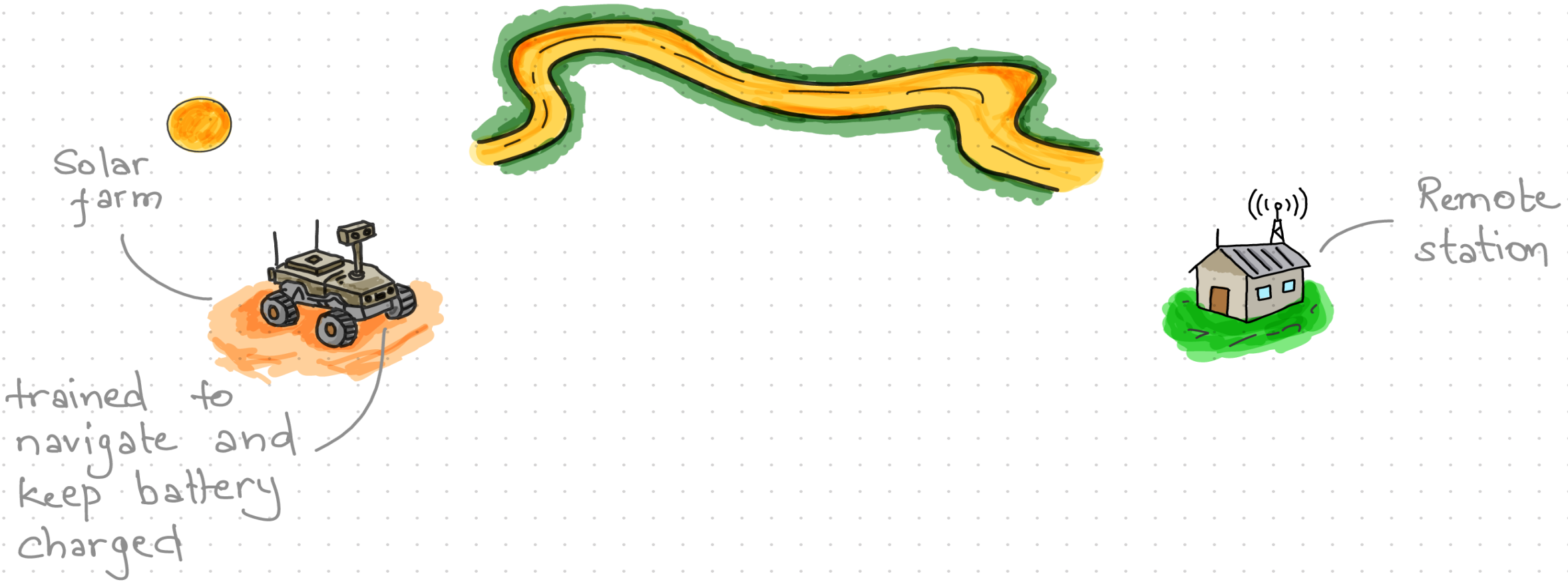
A simple scenario



A simple scenario

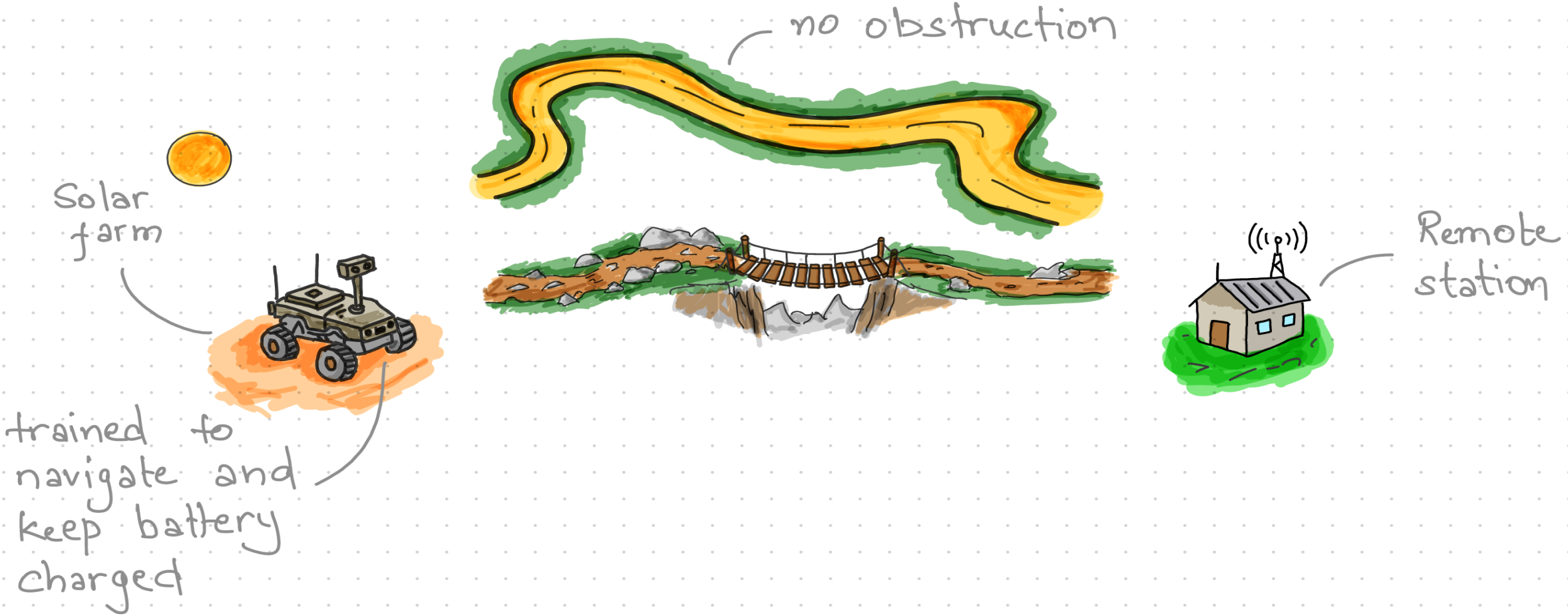


A simple scenario

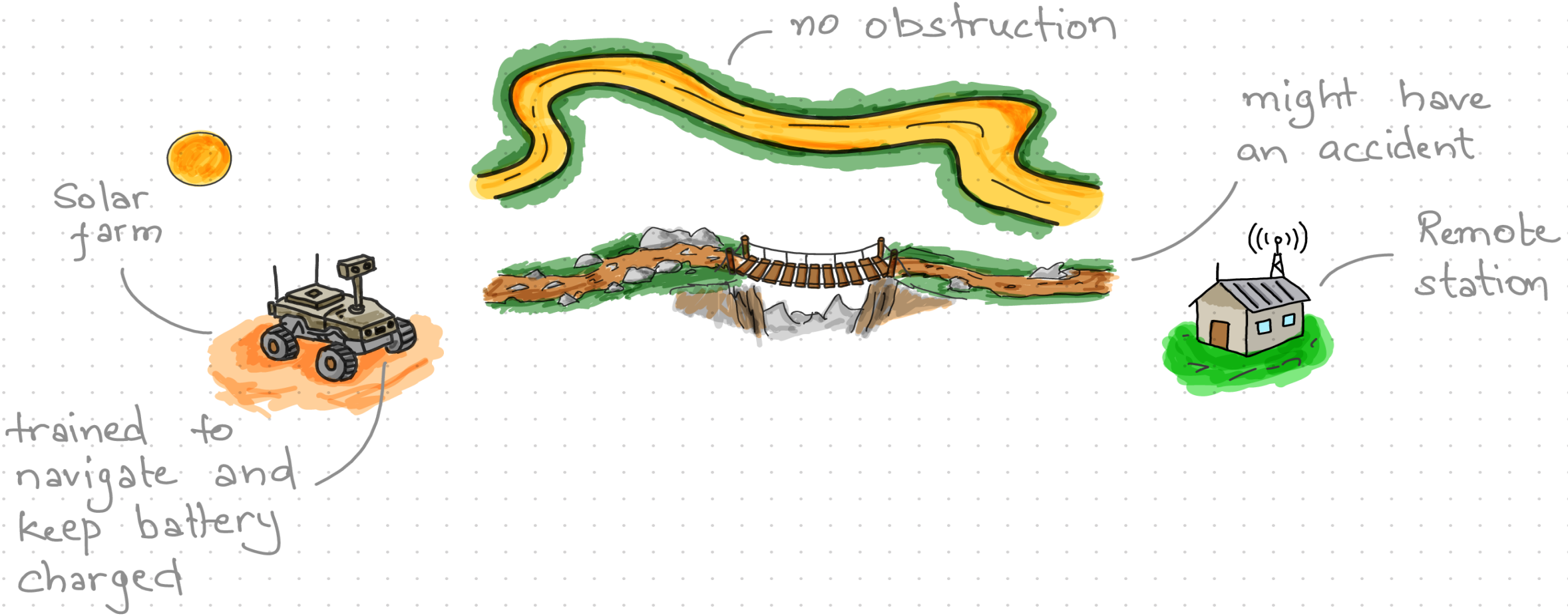


A simple scenario

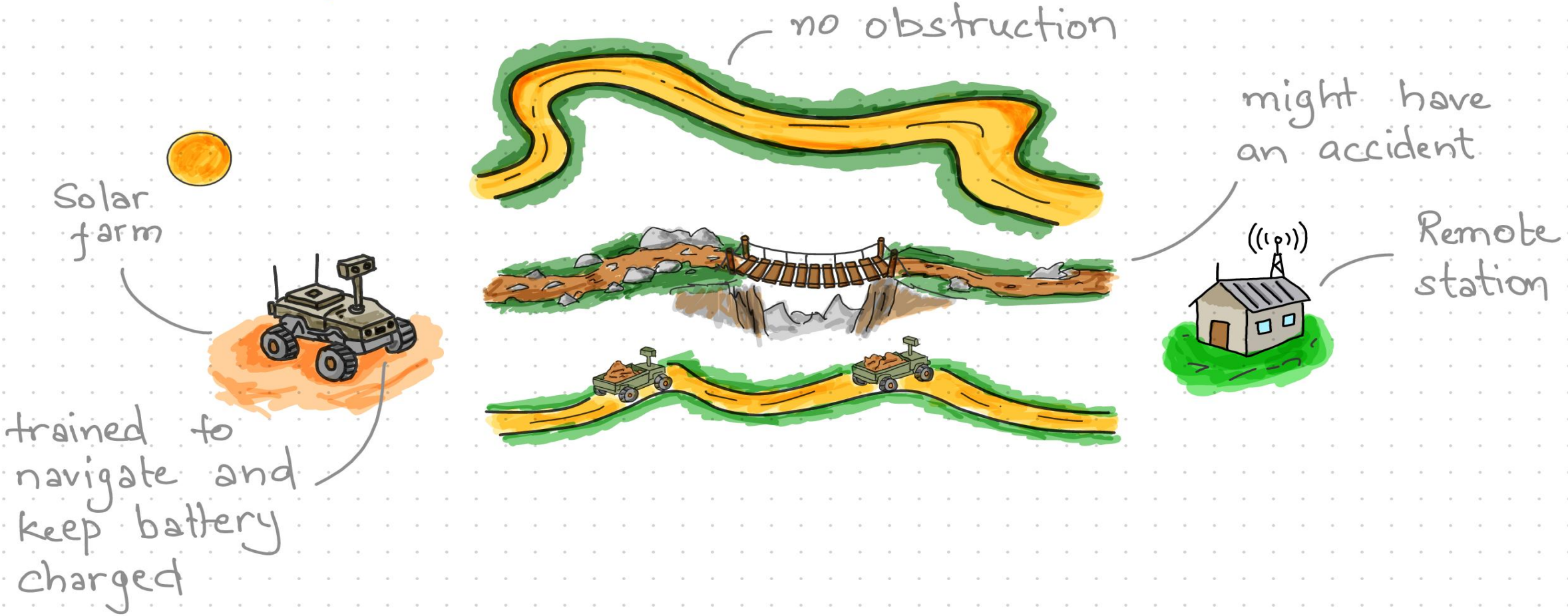
A simple scenario



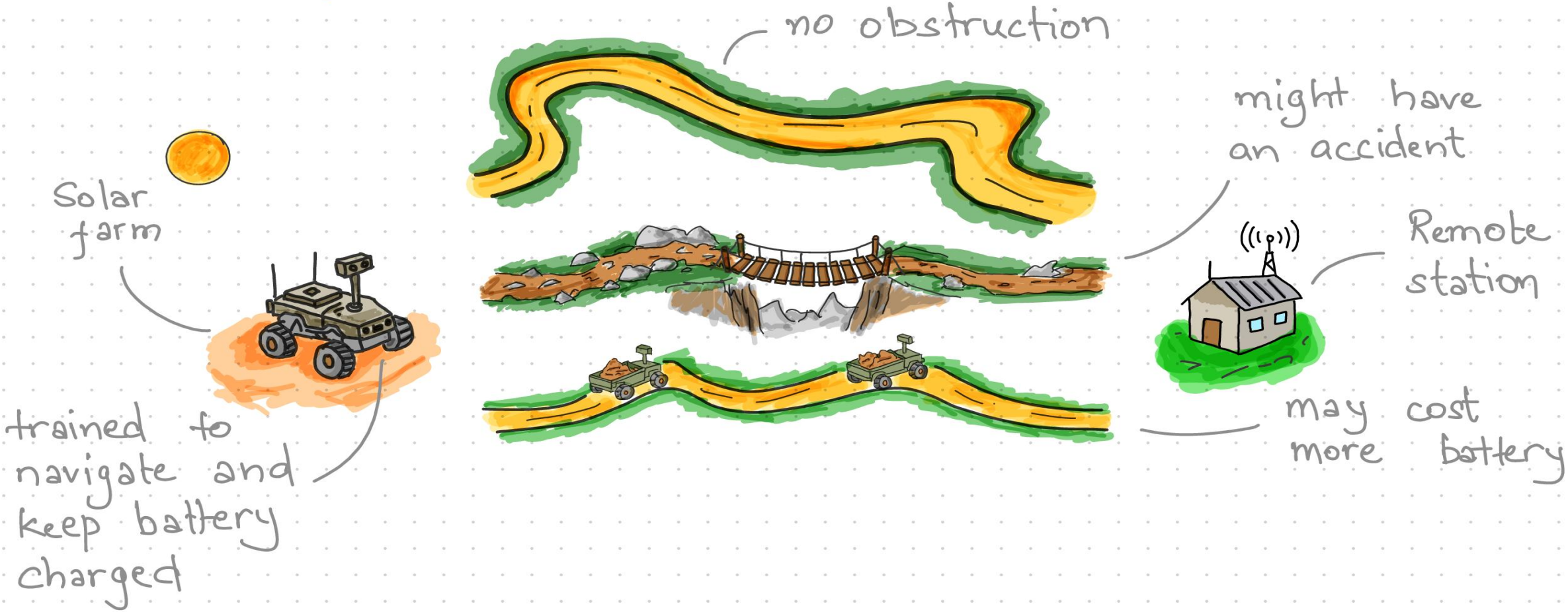
A simple scenario



A simple scenario



A simple scenario



A simple scenario

A simple scenario

A simple scenario

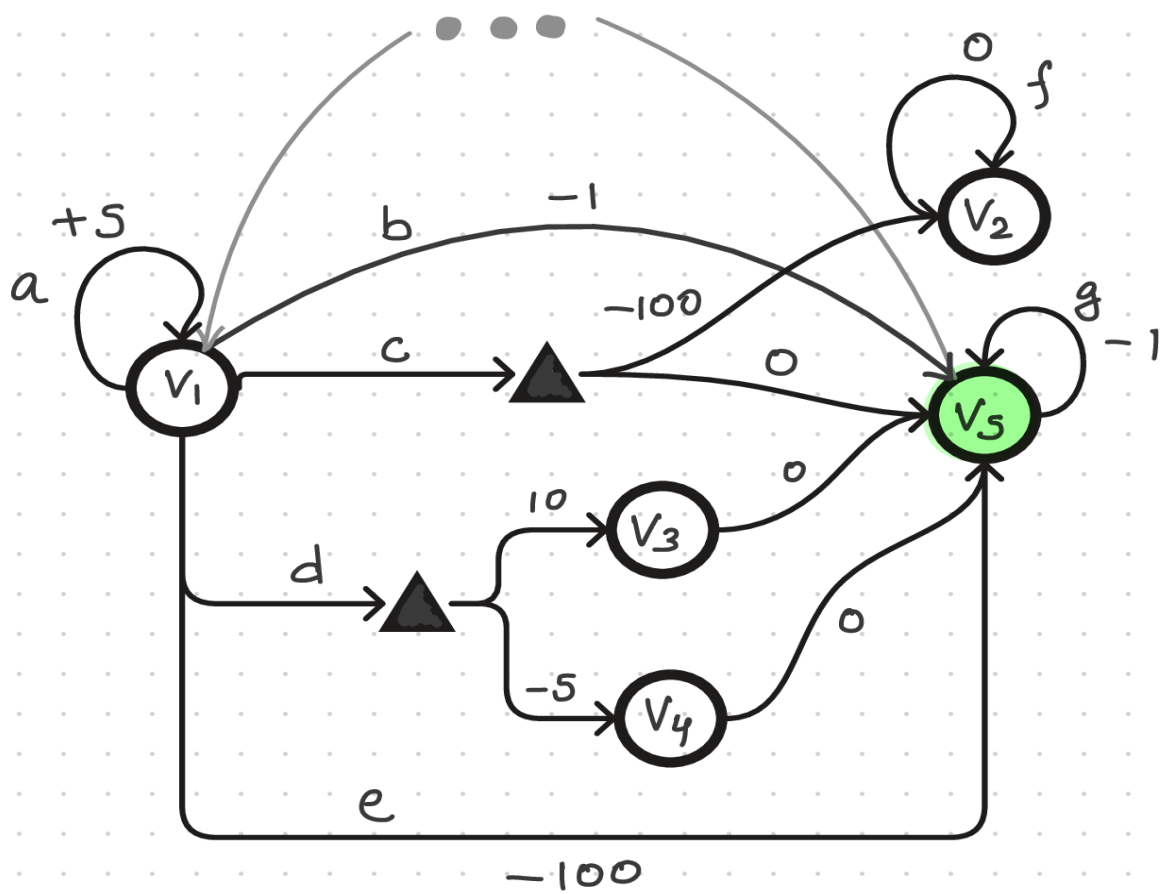
Can we force the rover to also keep going to the remote station?

Today

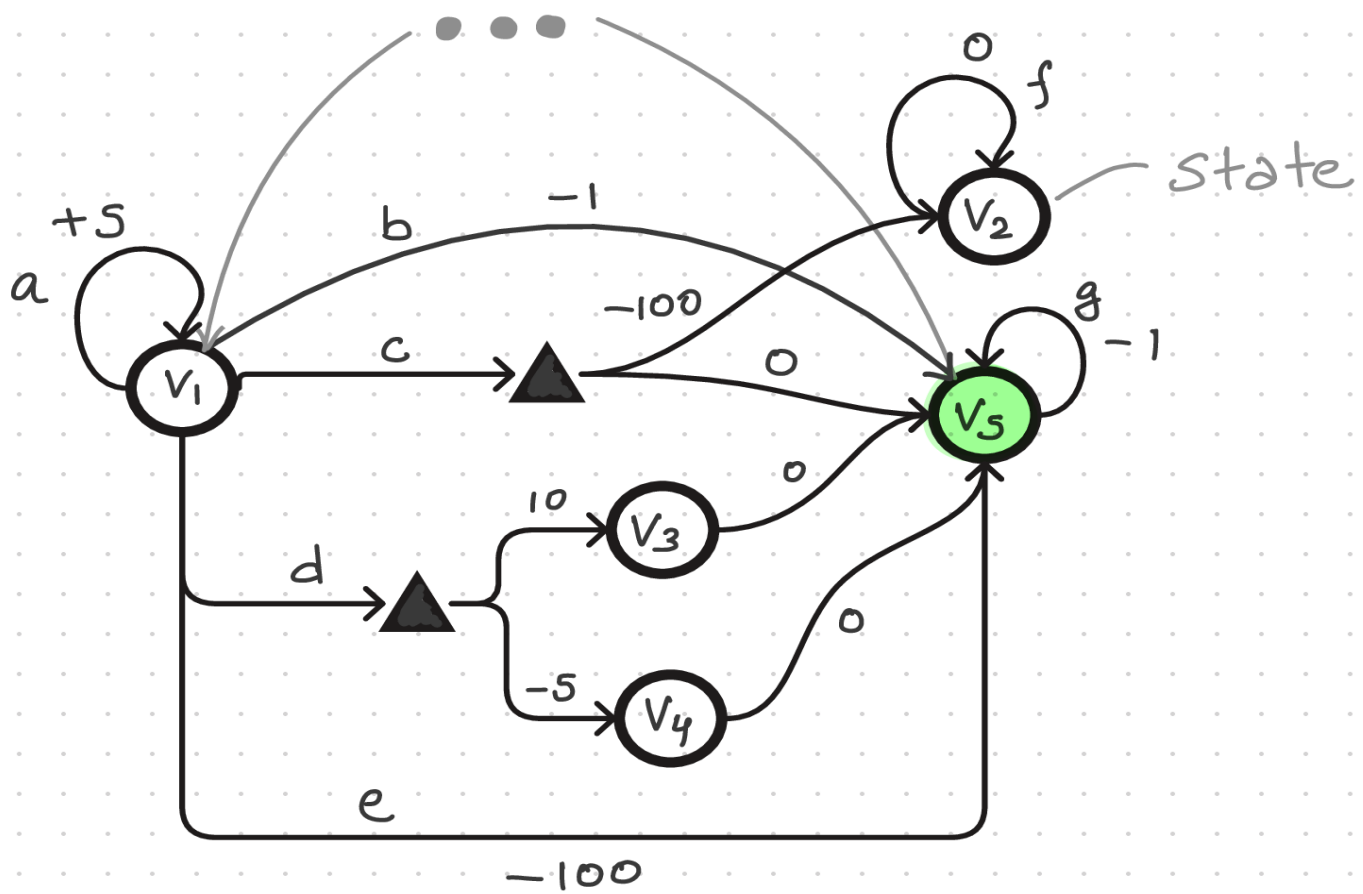
how we can enforce
additional ongoing
qualitative requirements
at runtime

[AA, SP Nayak, R Raha, AK Schmuck - AAMAS '26]

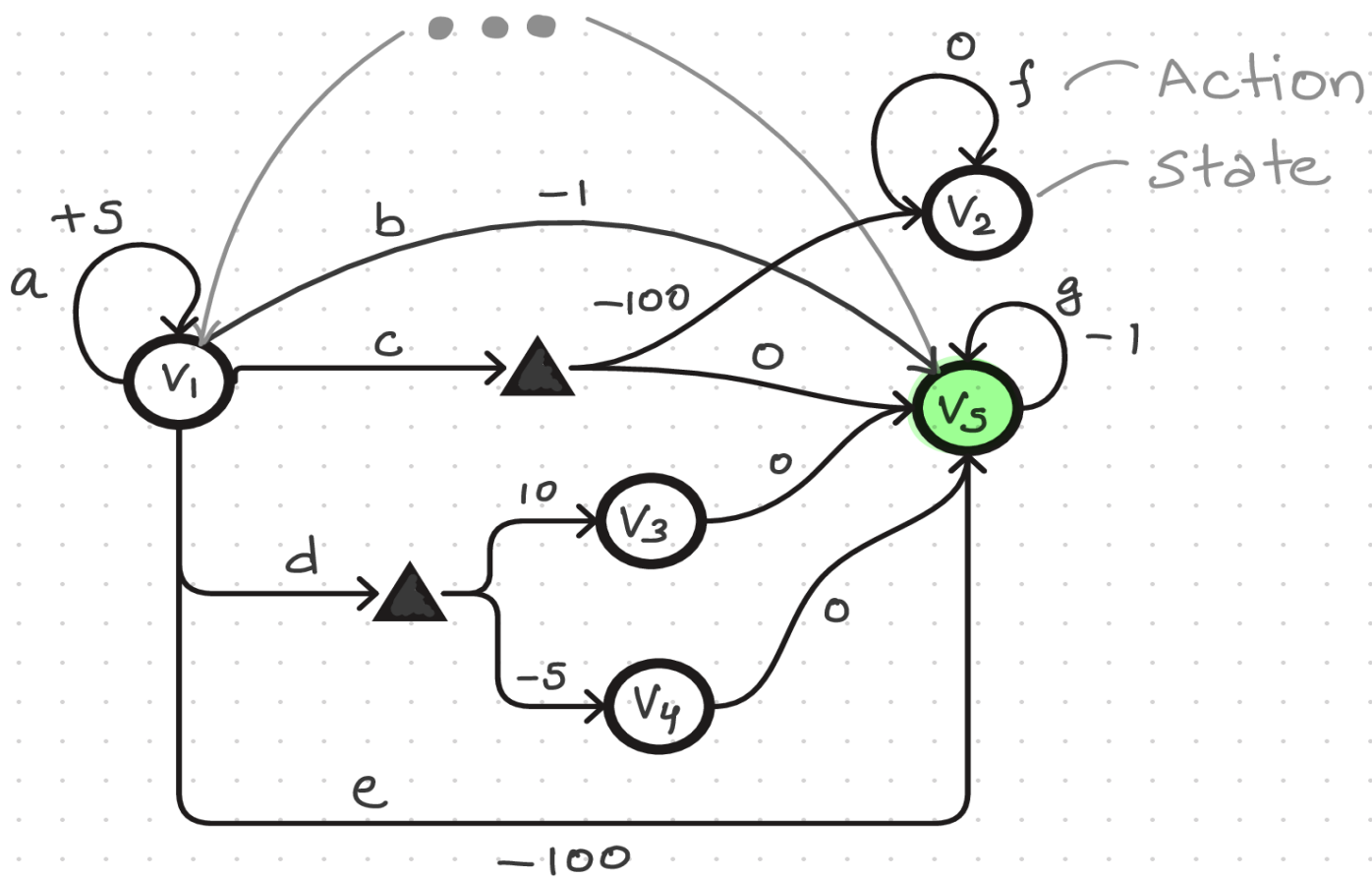
Markov Decision Processes



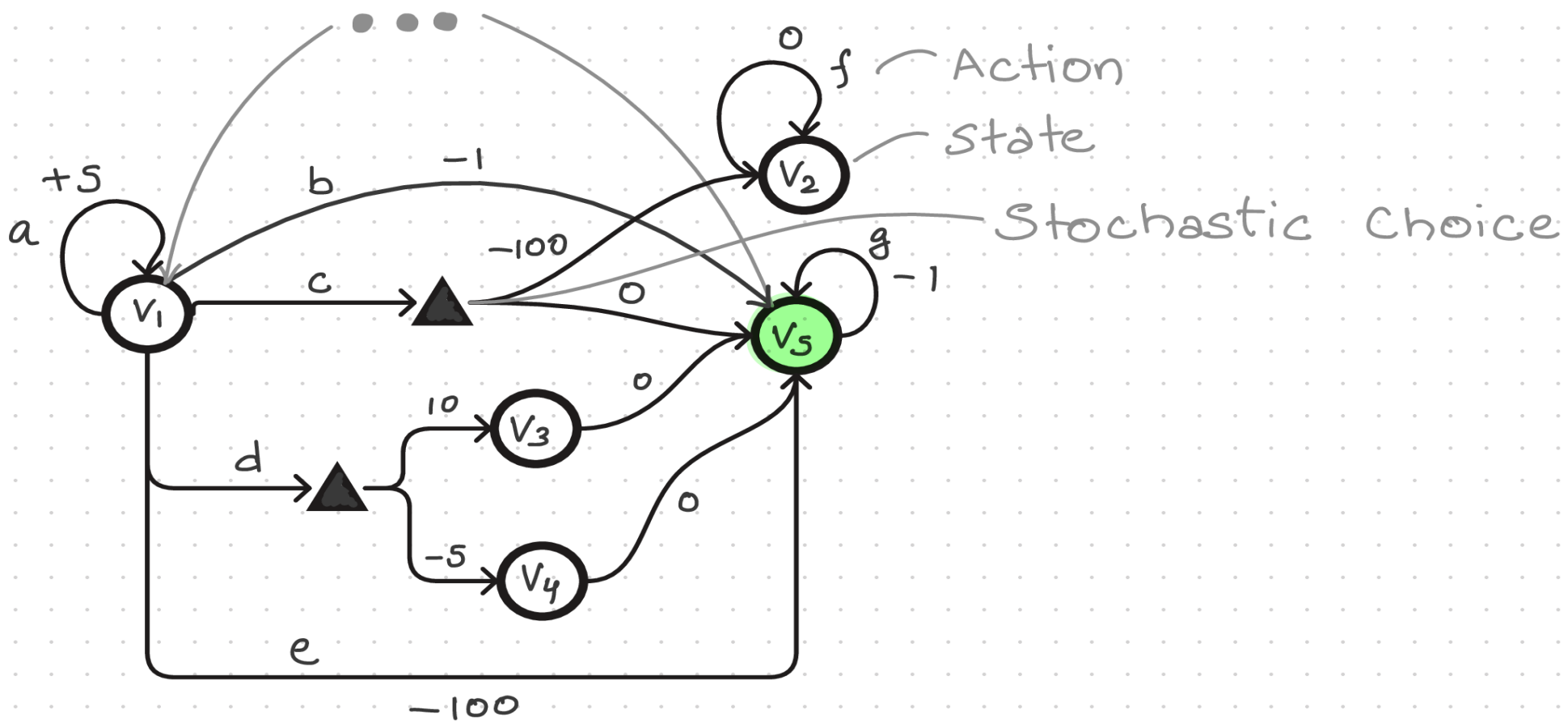
Markov Decision Processes



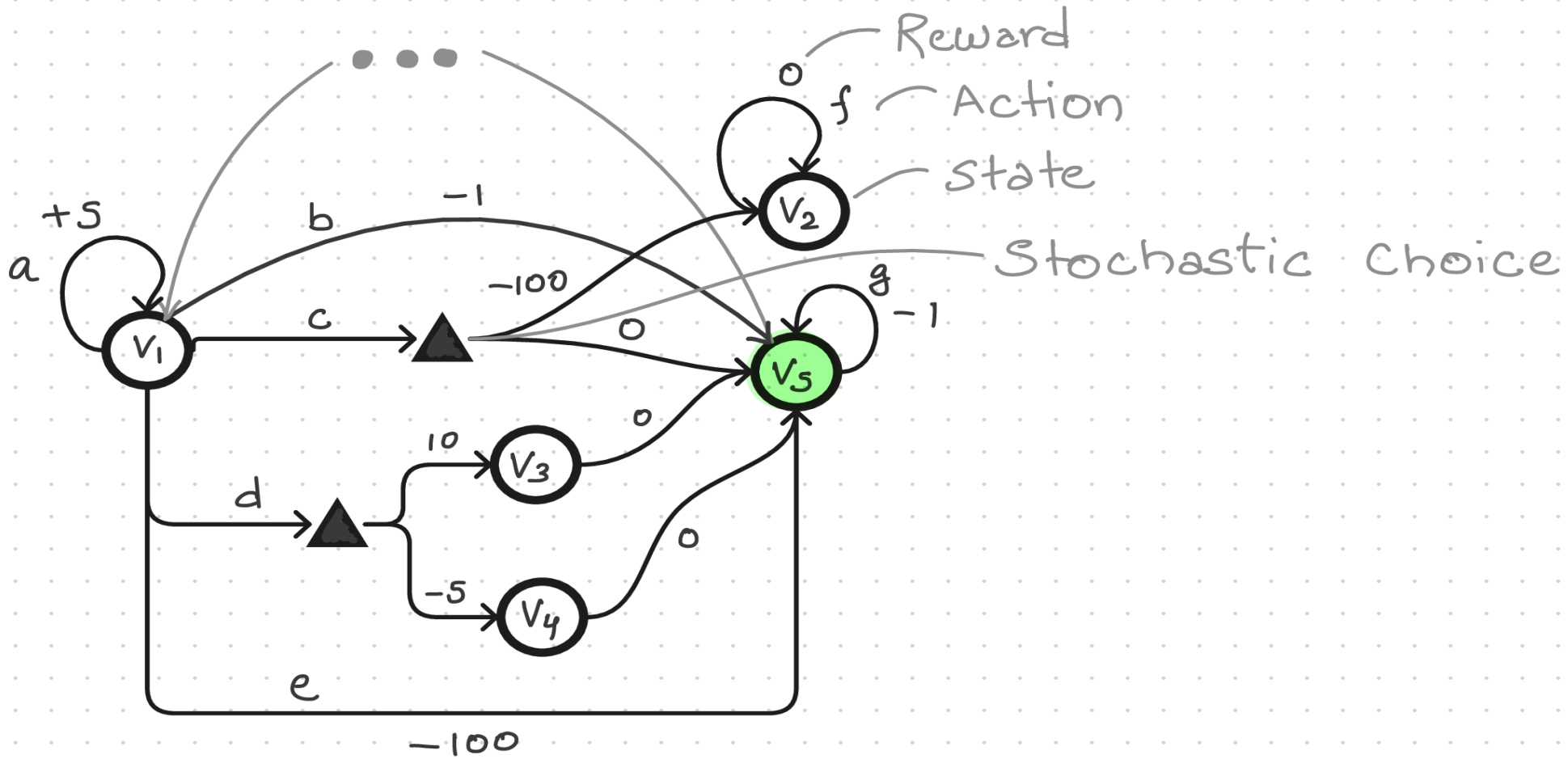
Markov Decision Processes



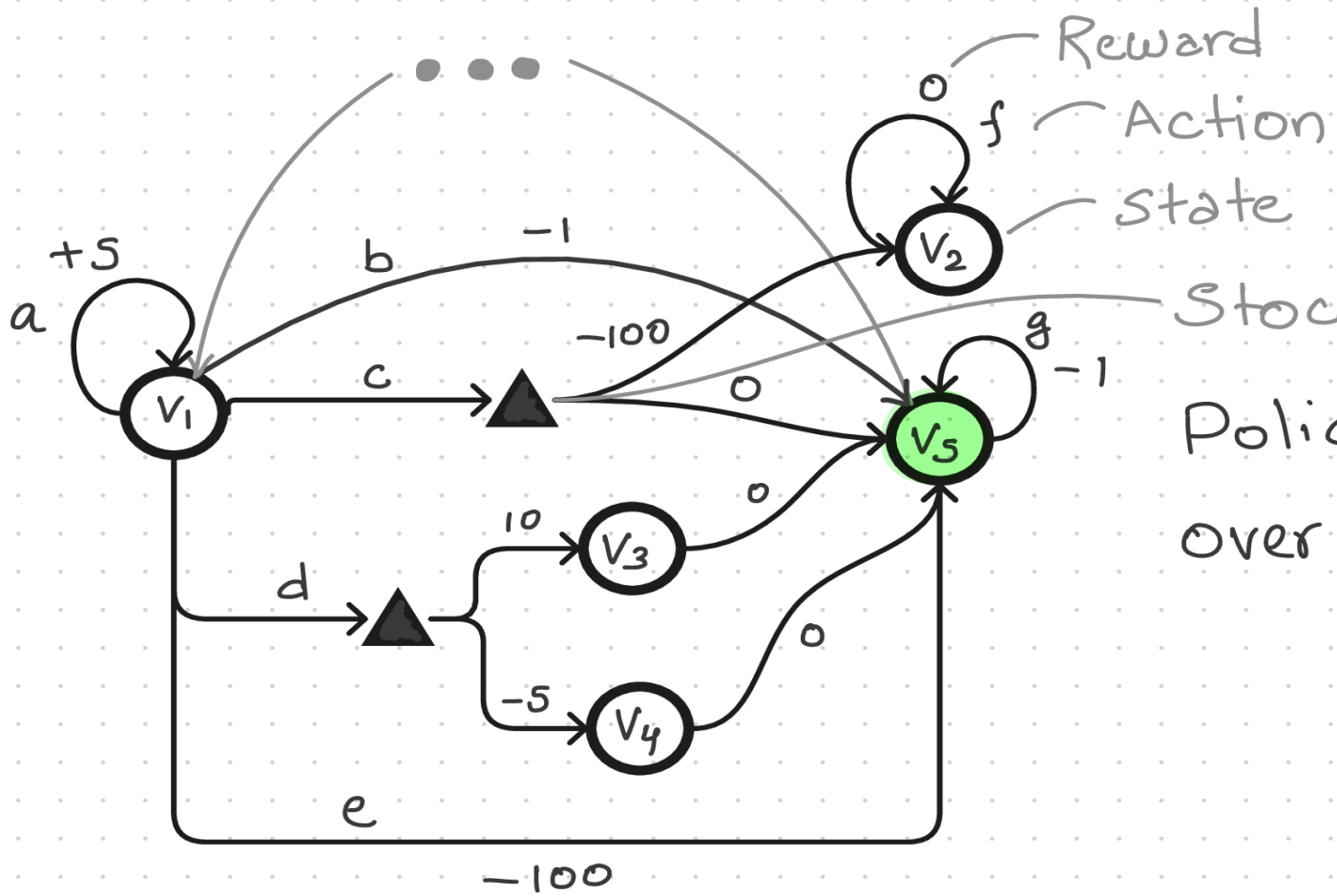
Markov Decision Processes



Markov Decision Processes

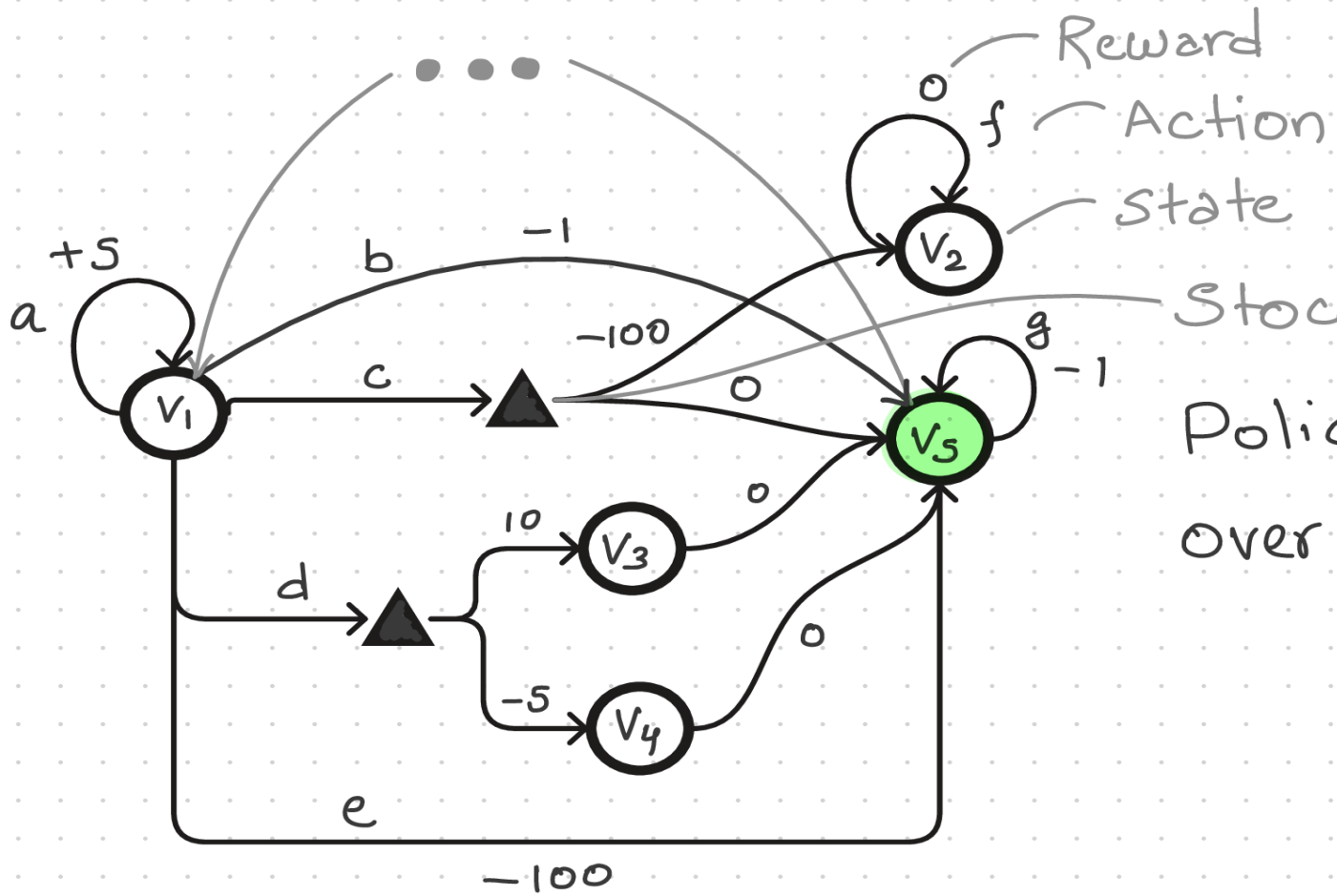


Markov Decision Processes



Policy is a distribution over actions

Markov Decision Processes



Policy is a distribution over actions

At v_1

a : 0.7

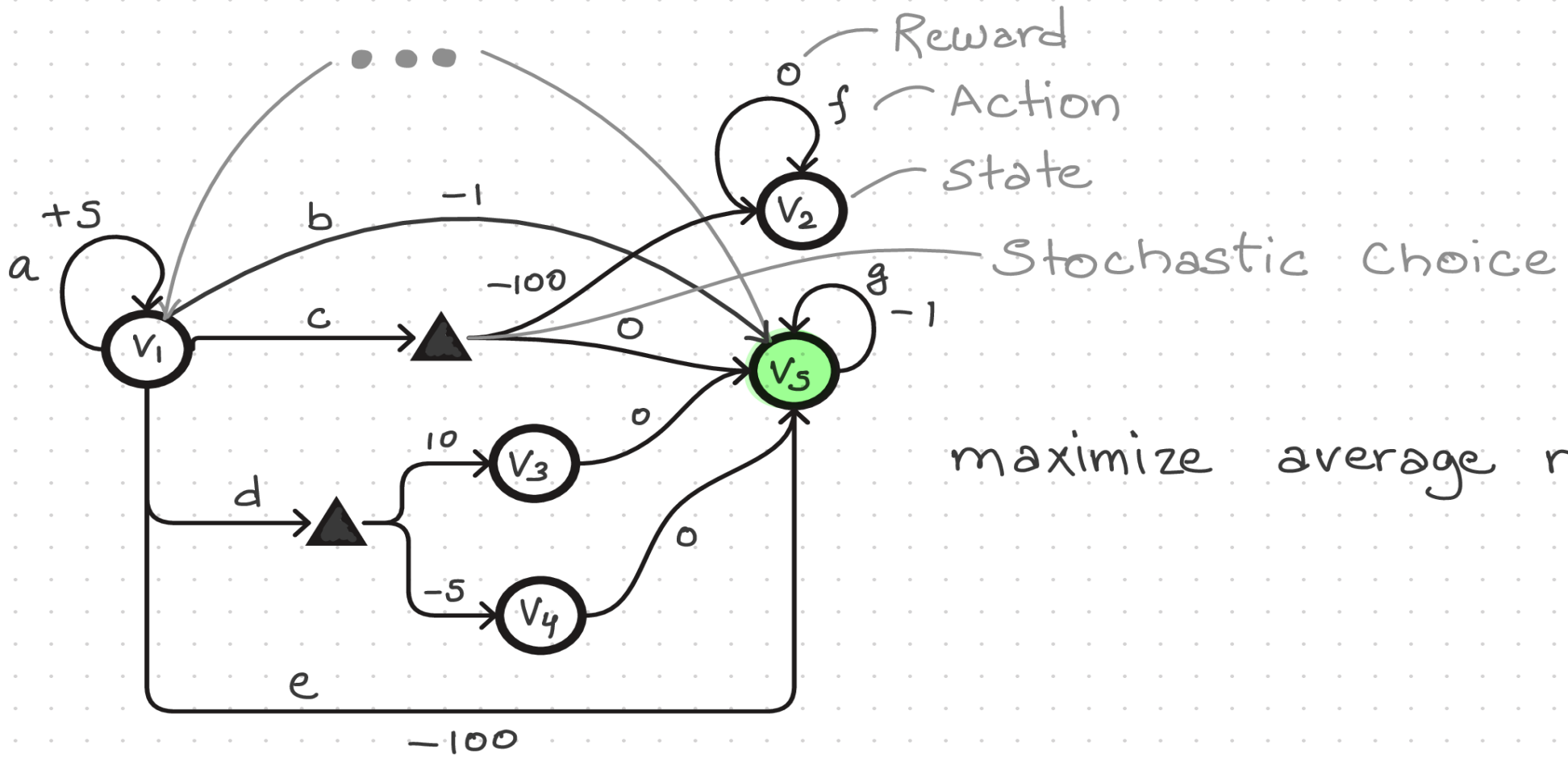
b : 0.1

c : 0.1

d : 0.1

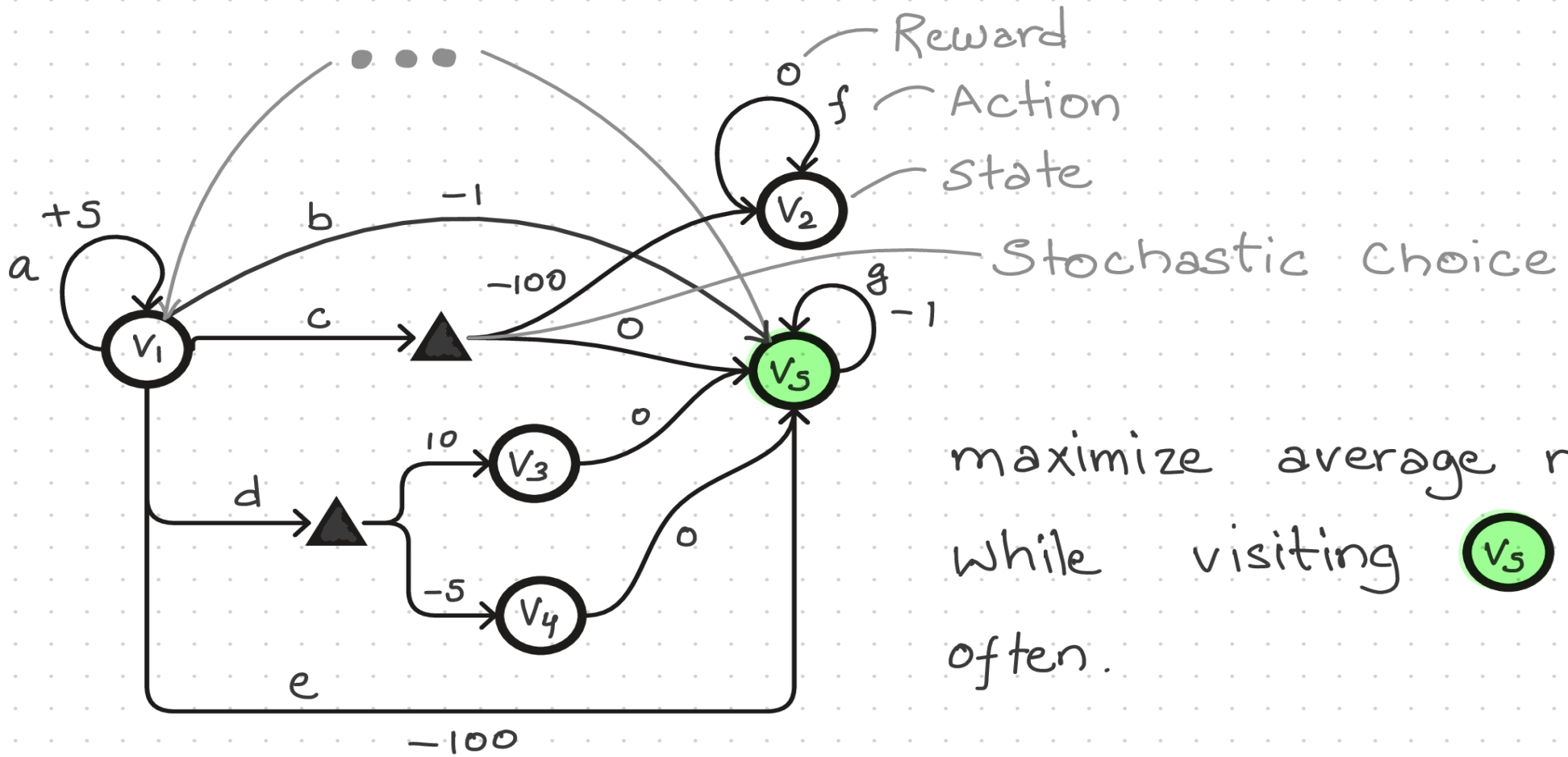
e : 0

Markov Decision Processes



maximize average reward

Markov Decision Processes



maximize average reward while visiting V_5 infinitely often.

Why not train?

Why not train?

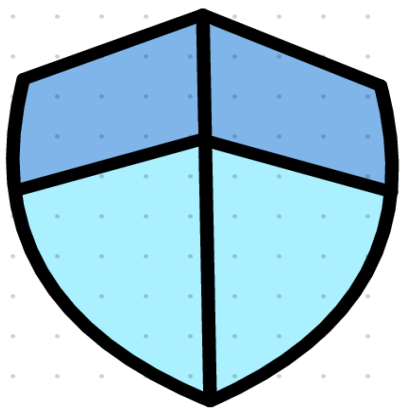
- Sparse reward on qualitative specification

Why not train?

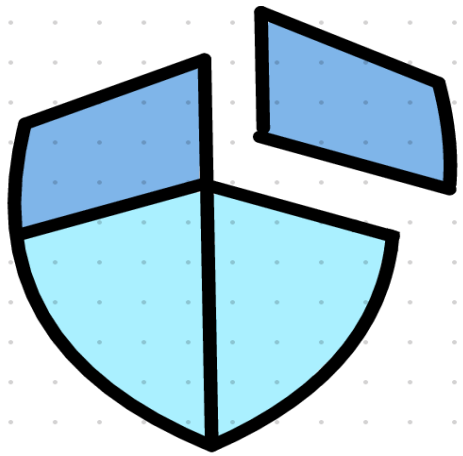
- Sparse reward on qualitative specification
- Difficult to get sure/almost sure guarantees

Why not train?

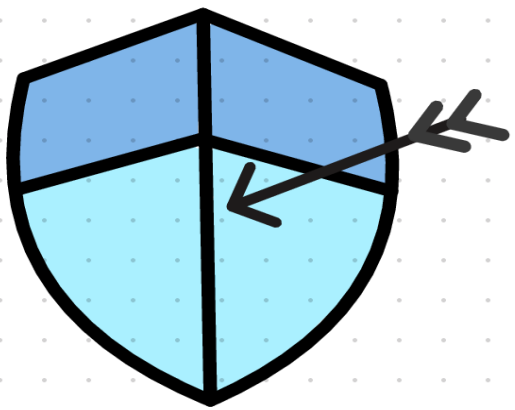
- Sparse reward on qualitative specification
- Difficult to get sure/almost sure guarantees
- Retraining is needed if specification changes



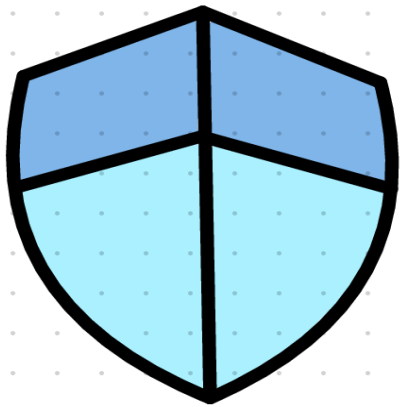
What is a shield?



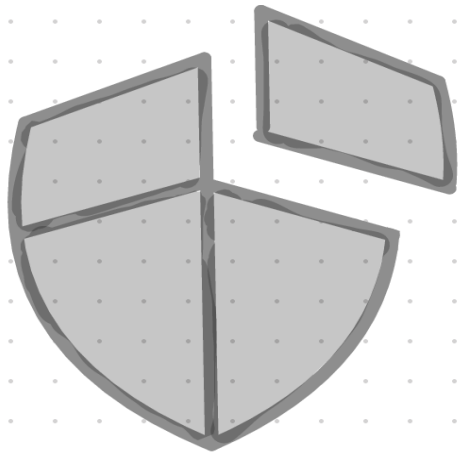
How to construct a shield?



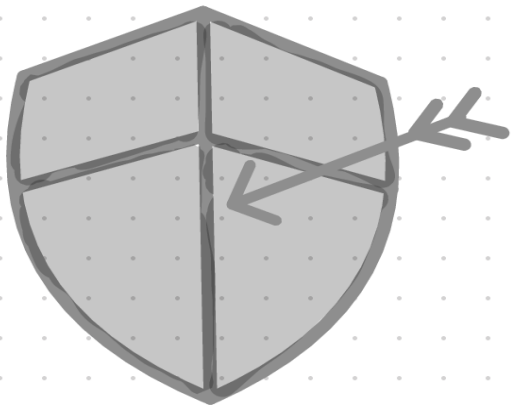
How to use a shield?



What is a shield?

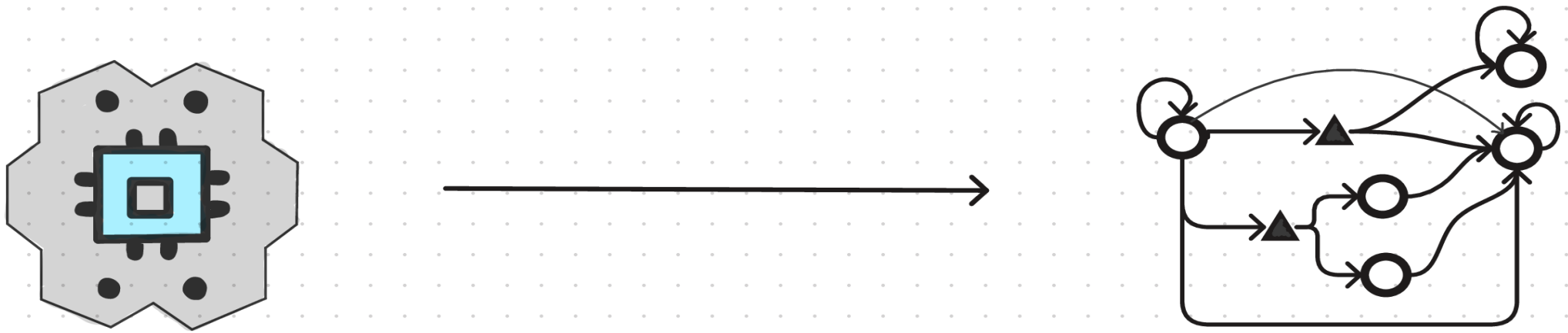


How to construct a shield?

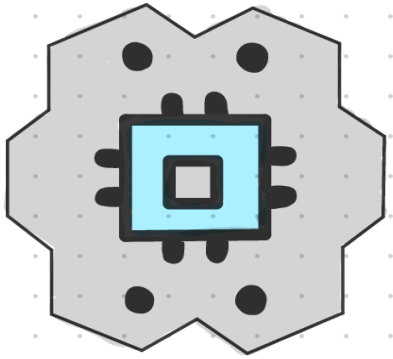


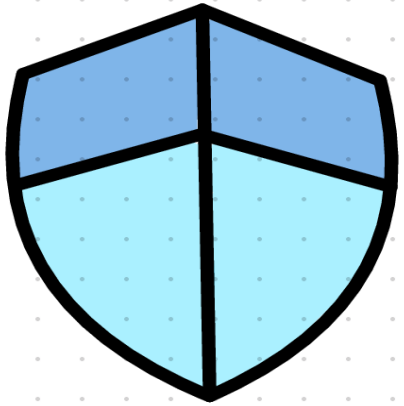
How to use a shield?

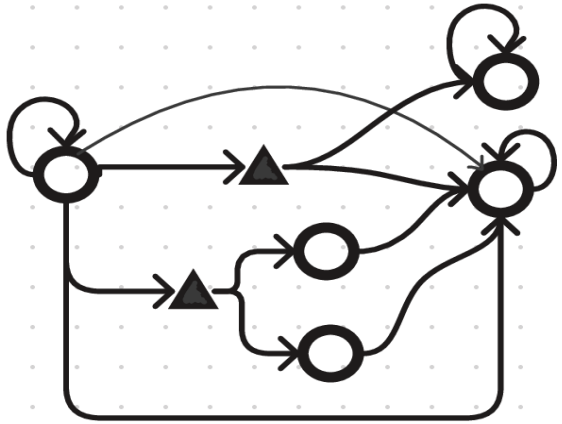
Shield



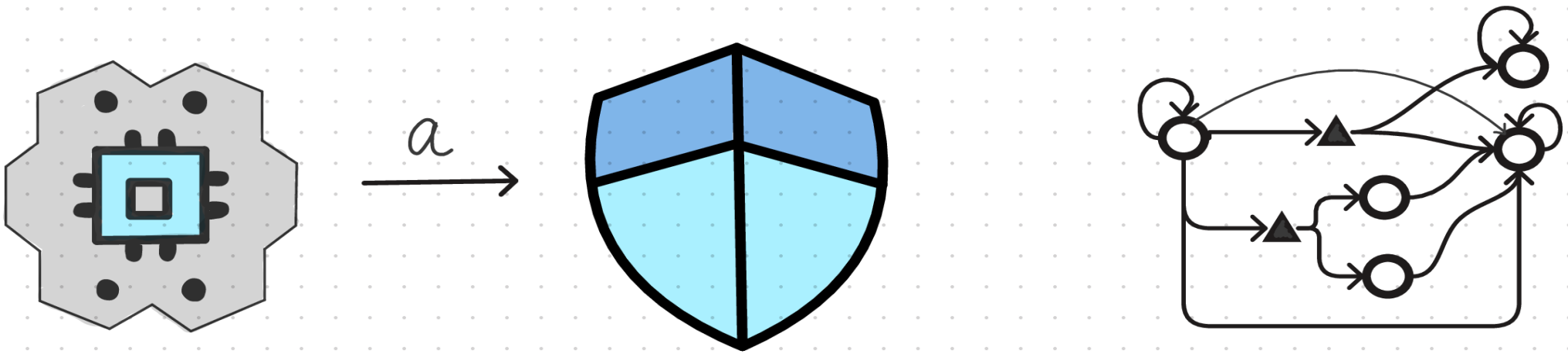
Shield



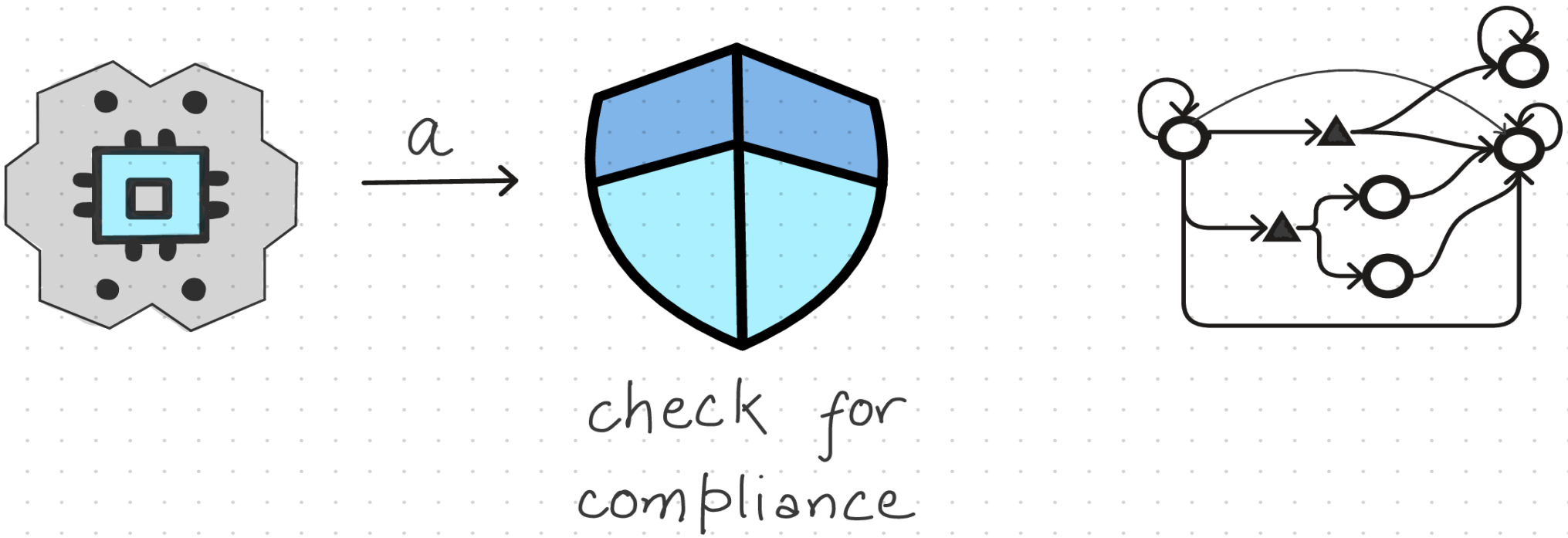




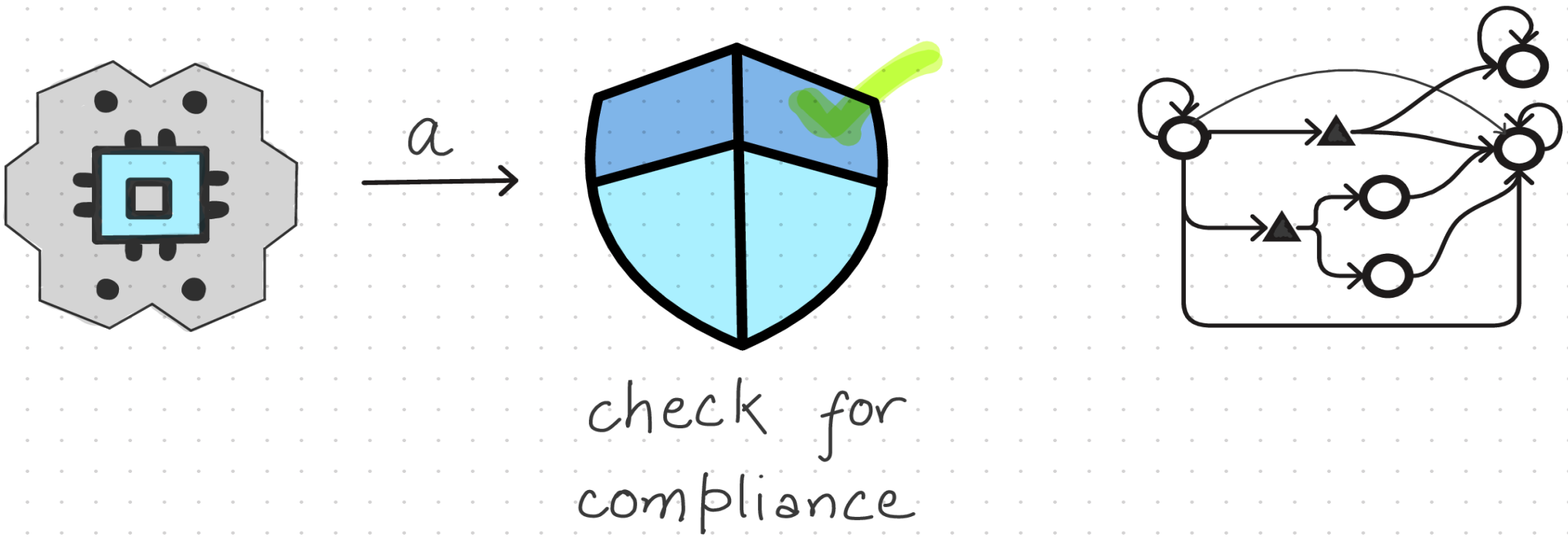
Shield



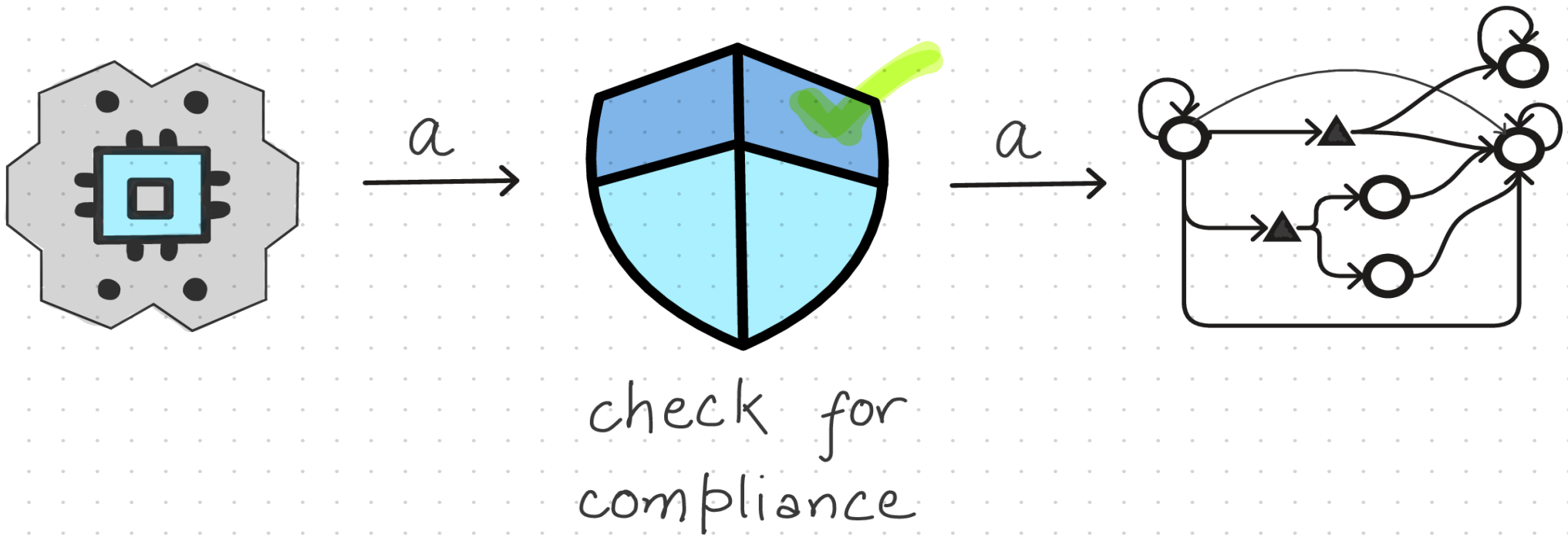
Shield



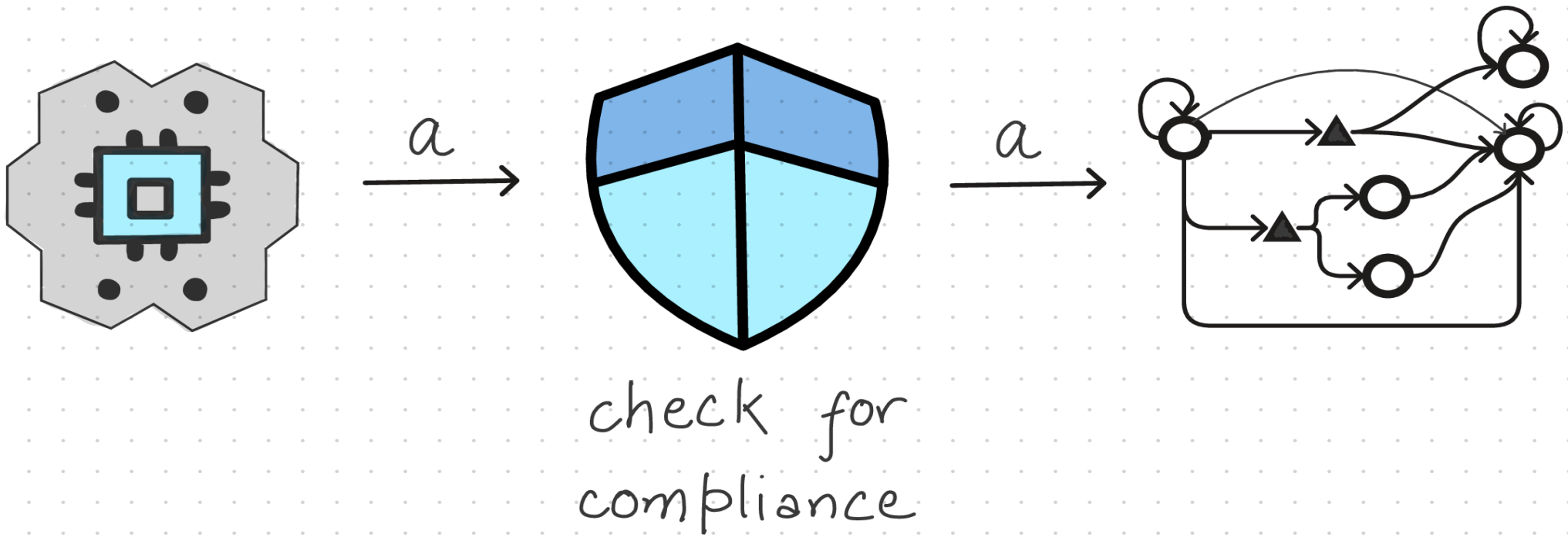
Shield



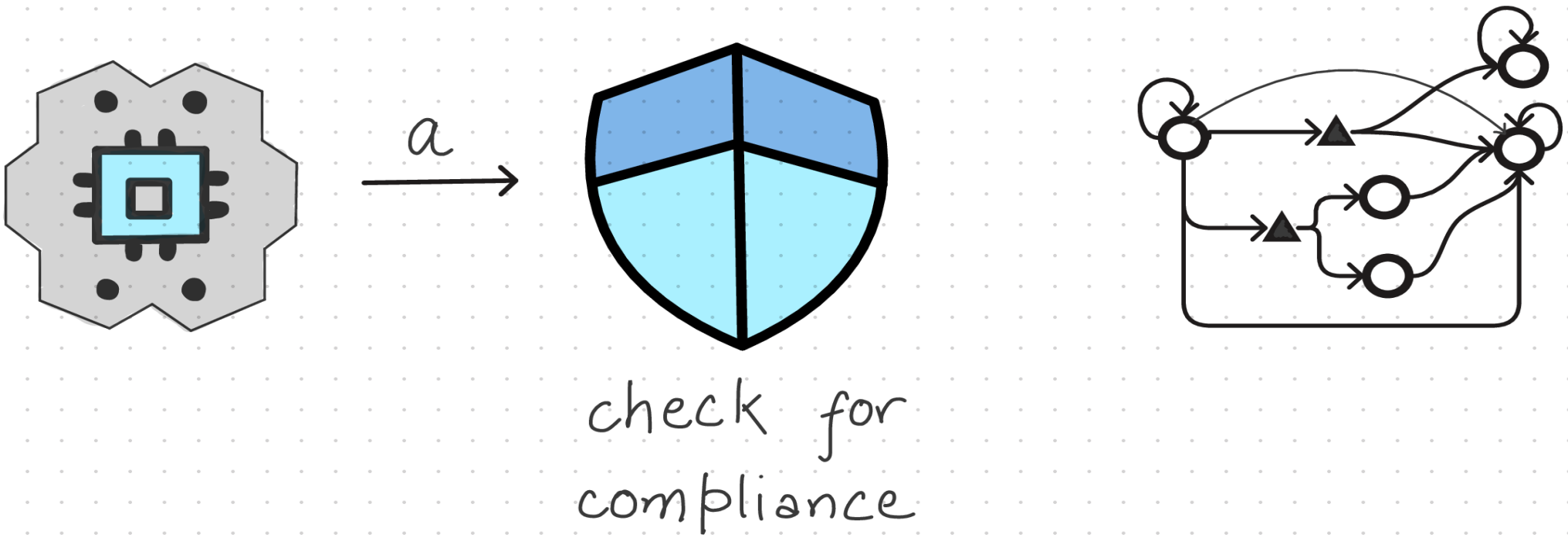
Shield



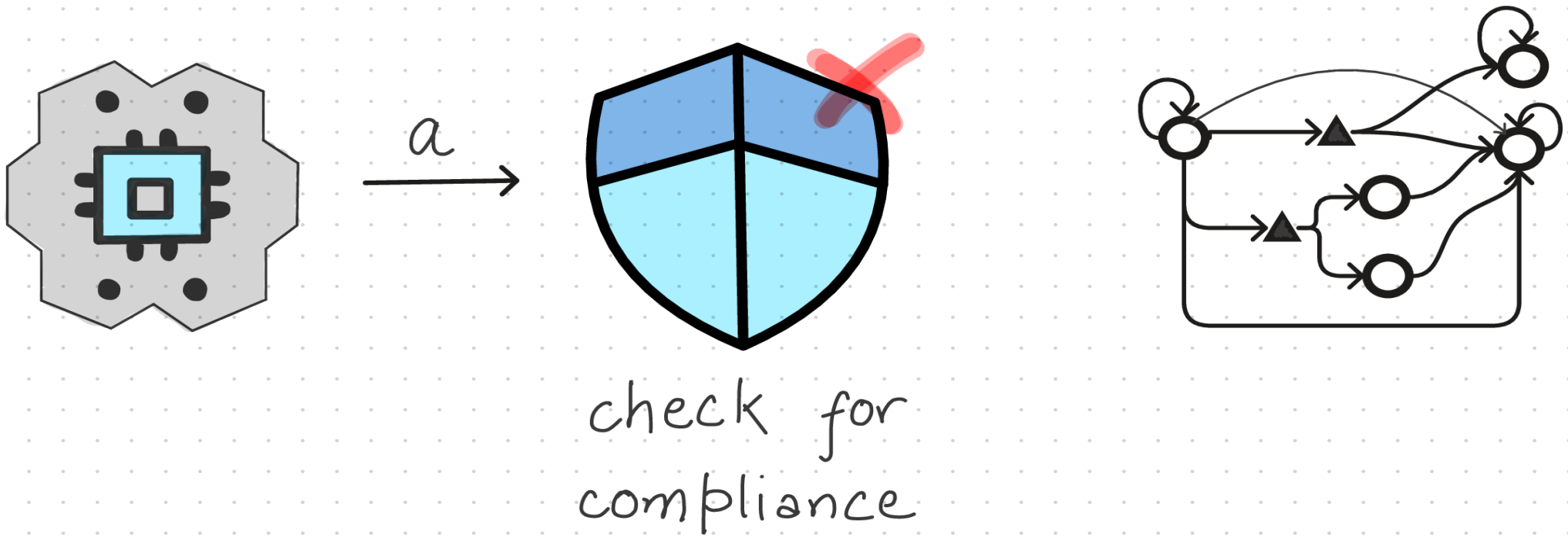
Shield



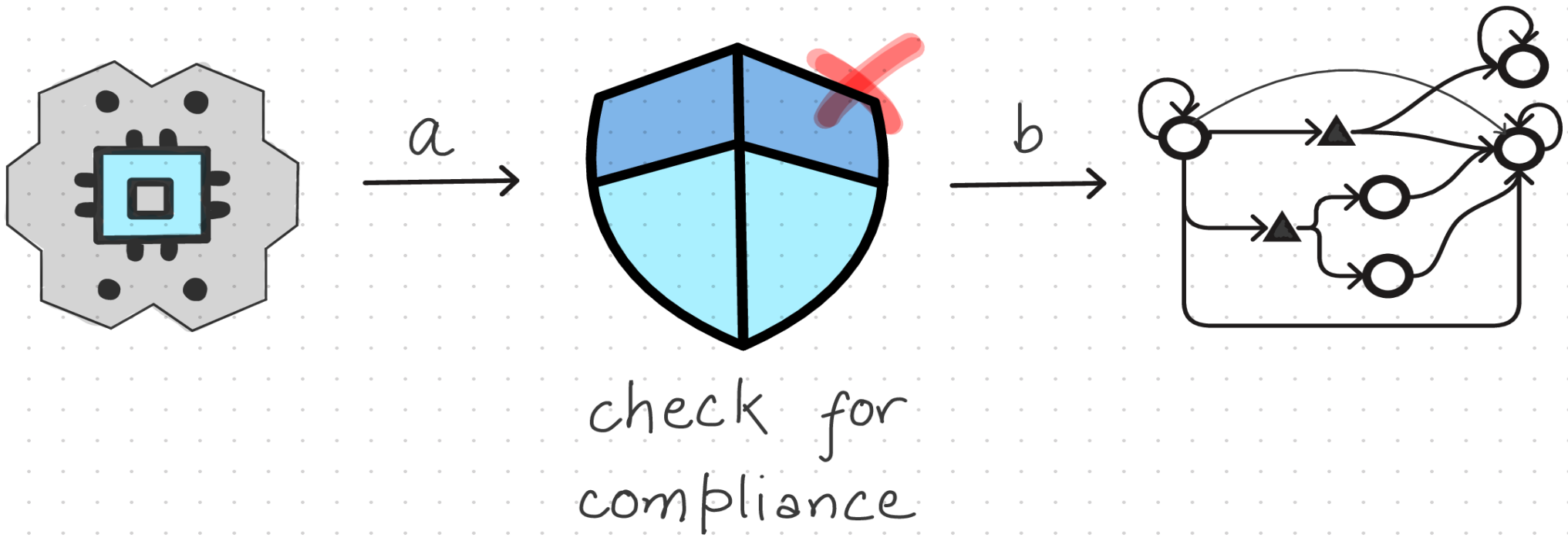
Shield



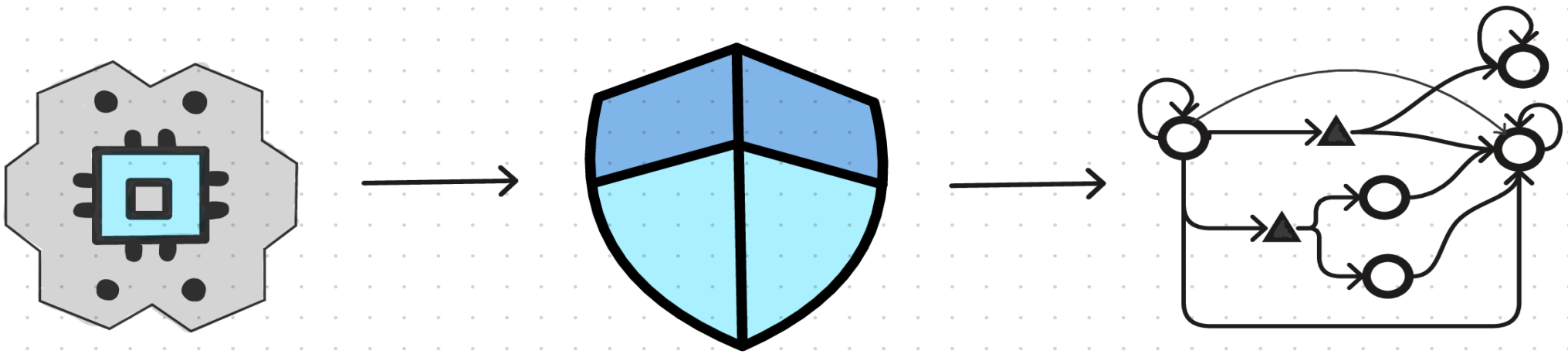
Shield



Shield



Shield



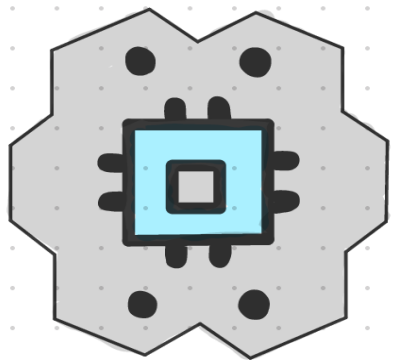
Shield

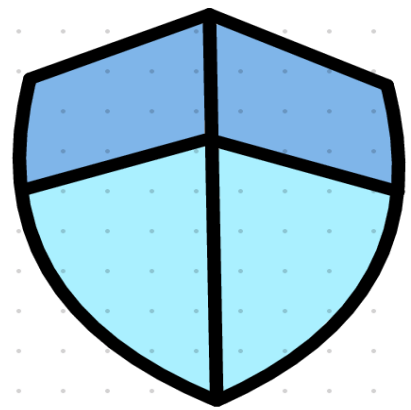
another
specification

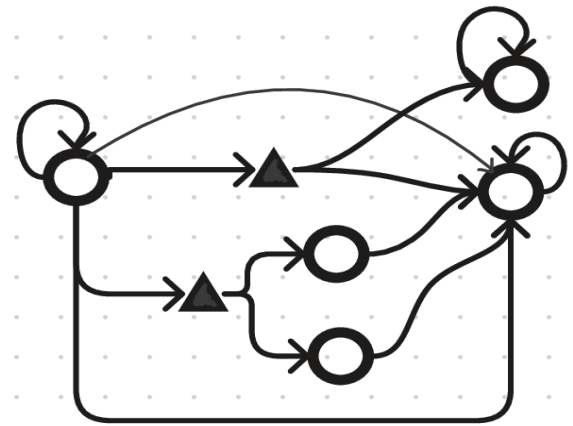
Shield

Allows decoupling of specification satisfaction with reward optimization

Shielding is difficult

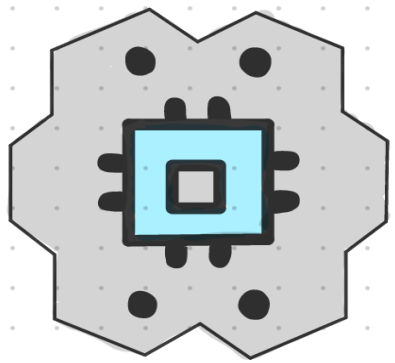


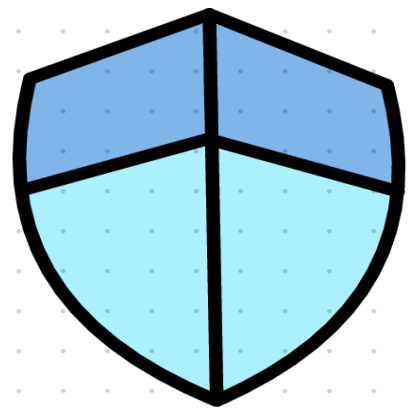


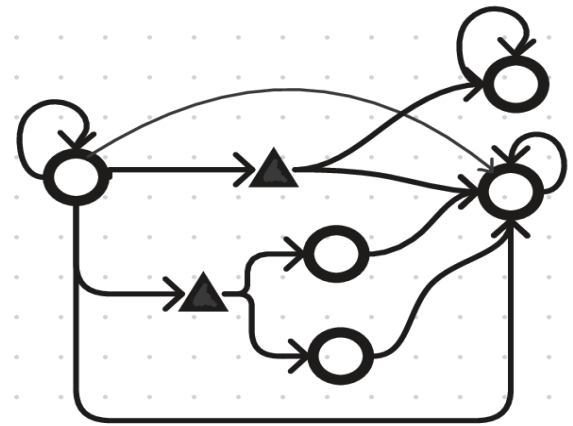


- Safety is easy ; liveness is not

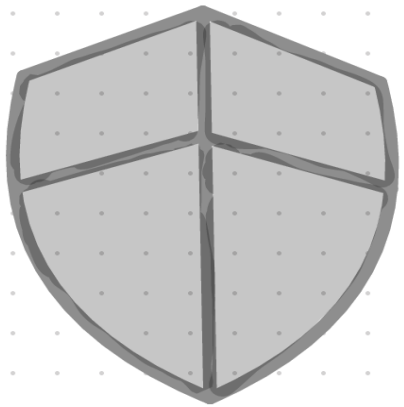
Shielding is difficult



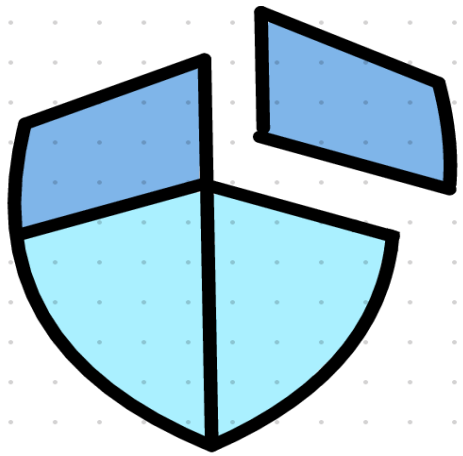




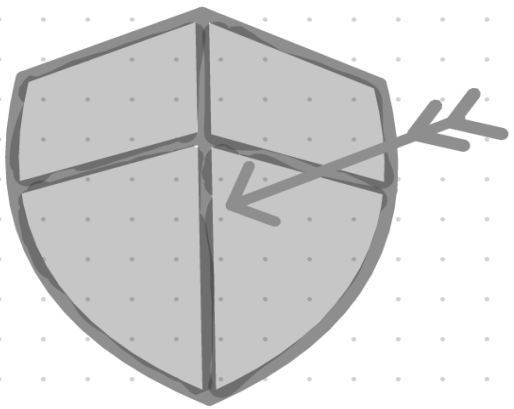
- Safety is easy ; liveness is not
- Should not interfere too much



What is a shield?

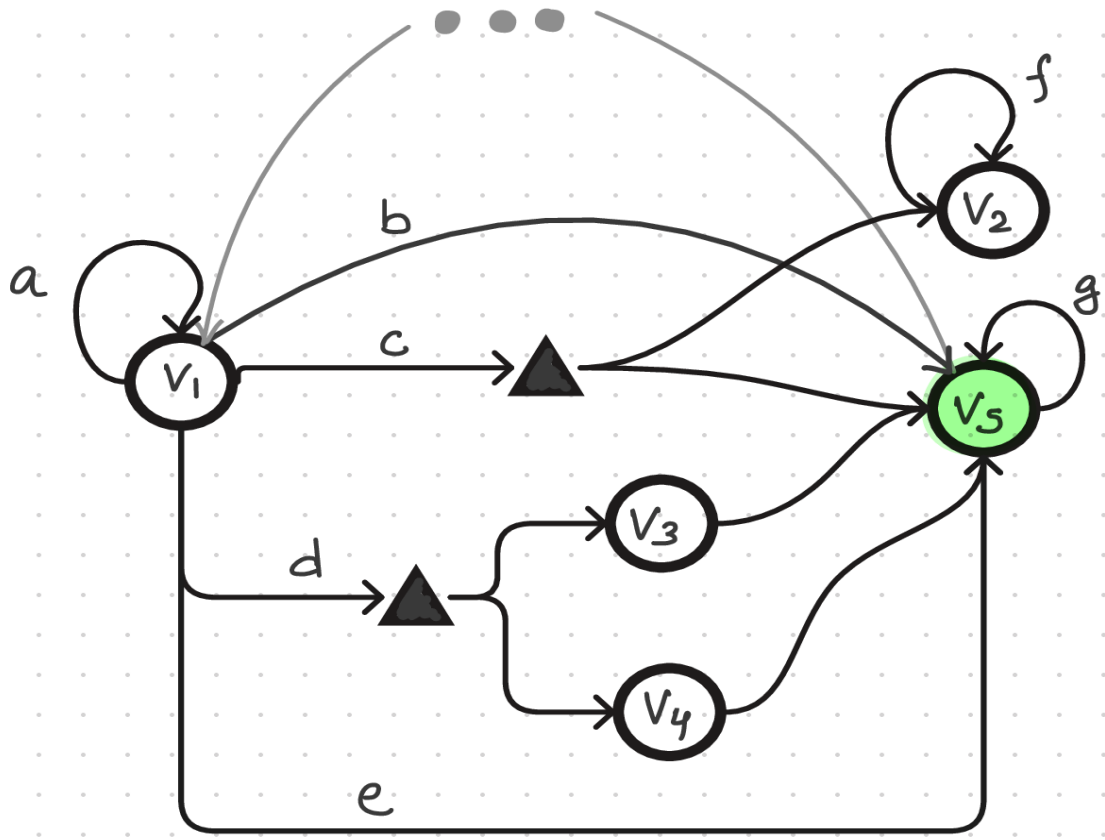


How to construct a shield?



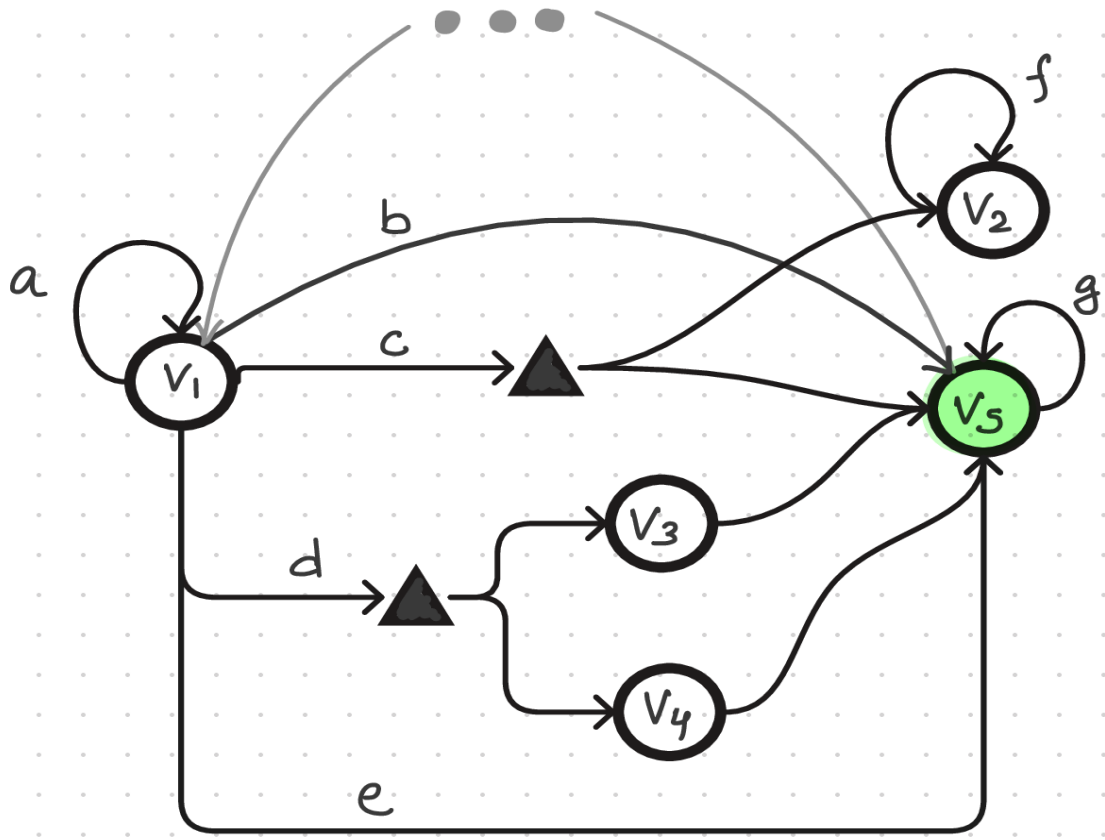
How to use a shield?

Games on graphs



2 players

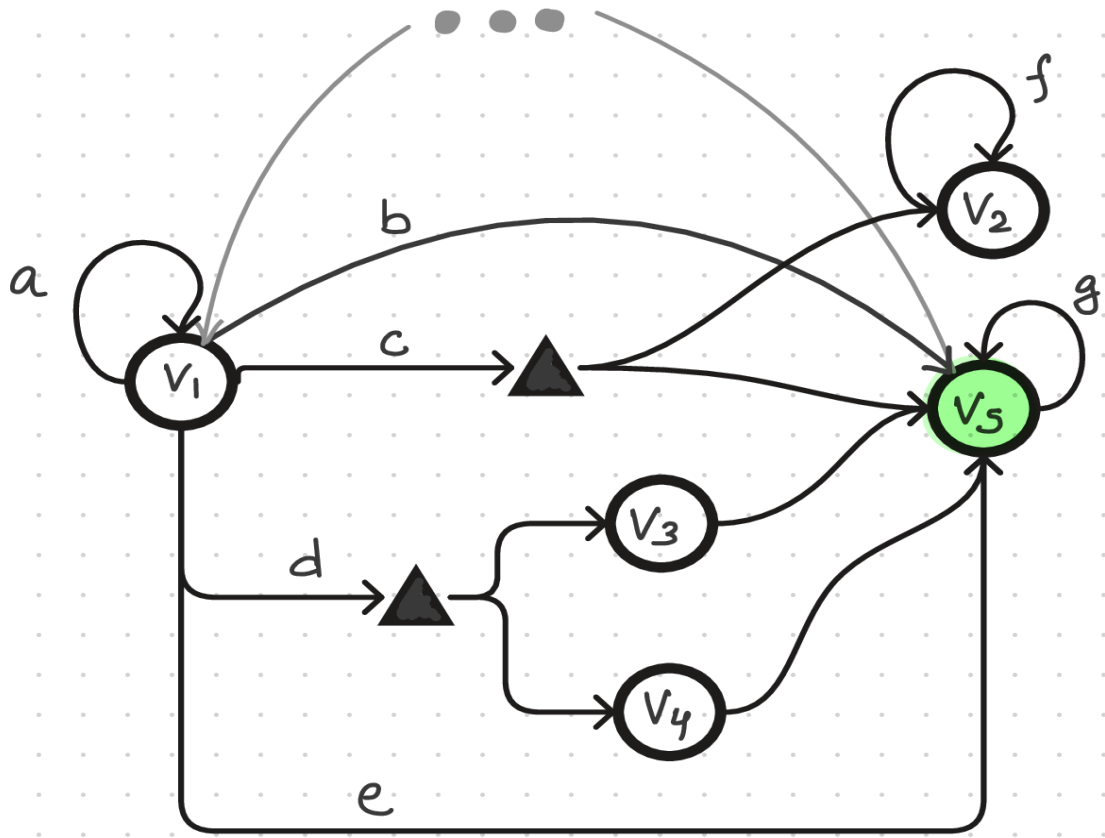
Games on graphs



2 players

○ wins if its specification is satisfied

Games on graphs

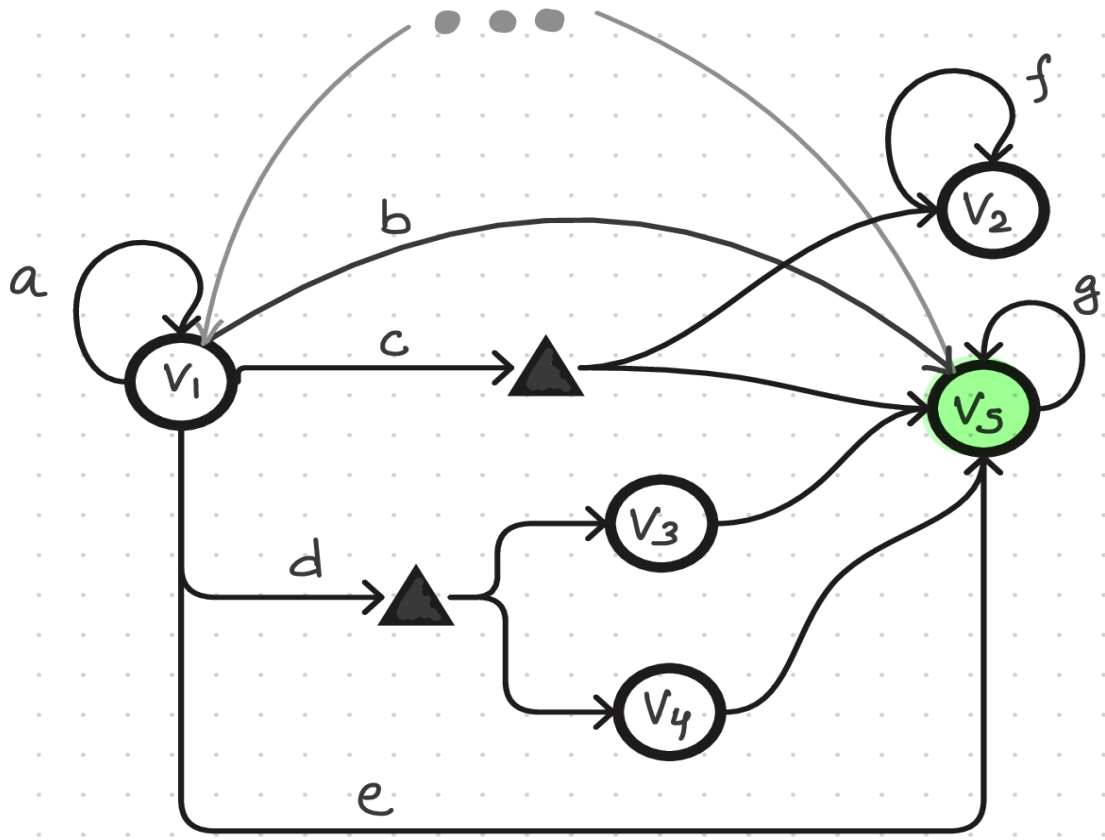


2 players

○ wins if its specification is satisfied

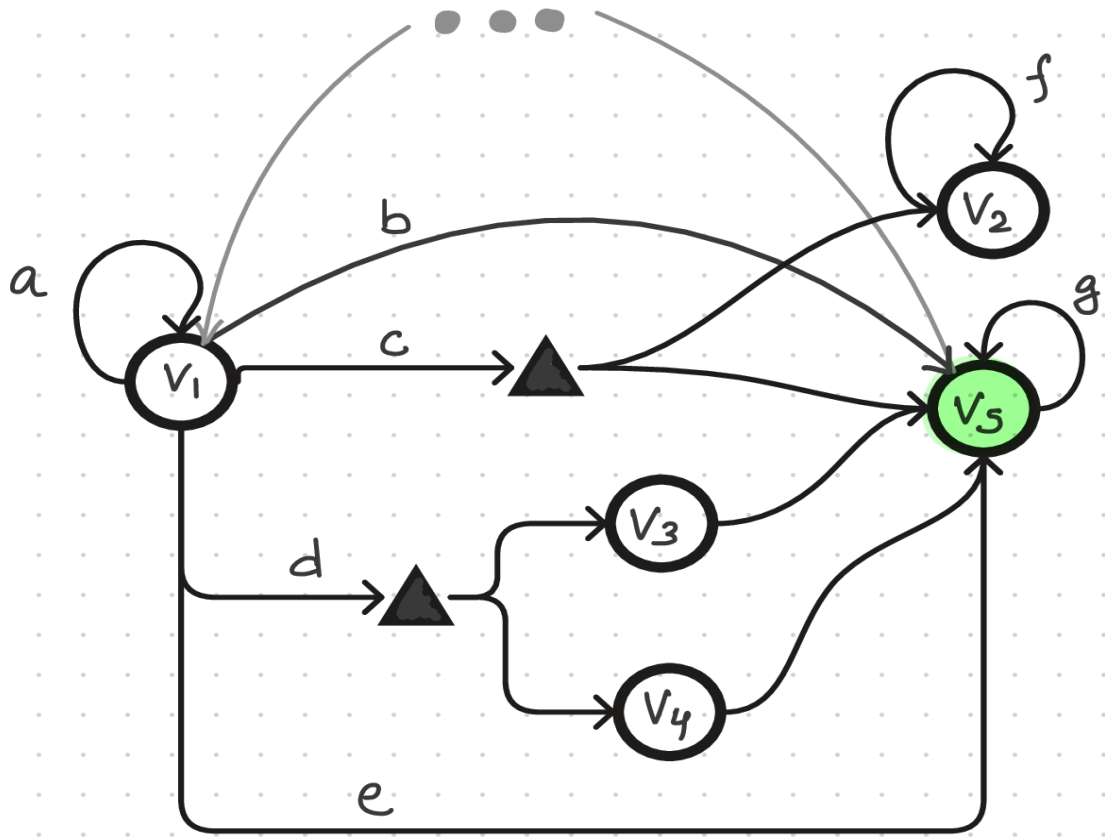
▲ wins if ○ fails

Games on graphs



find a strategy to visit v_5 infinitely often.

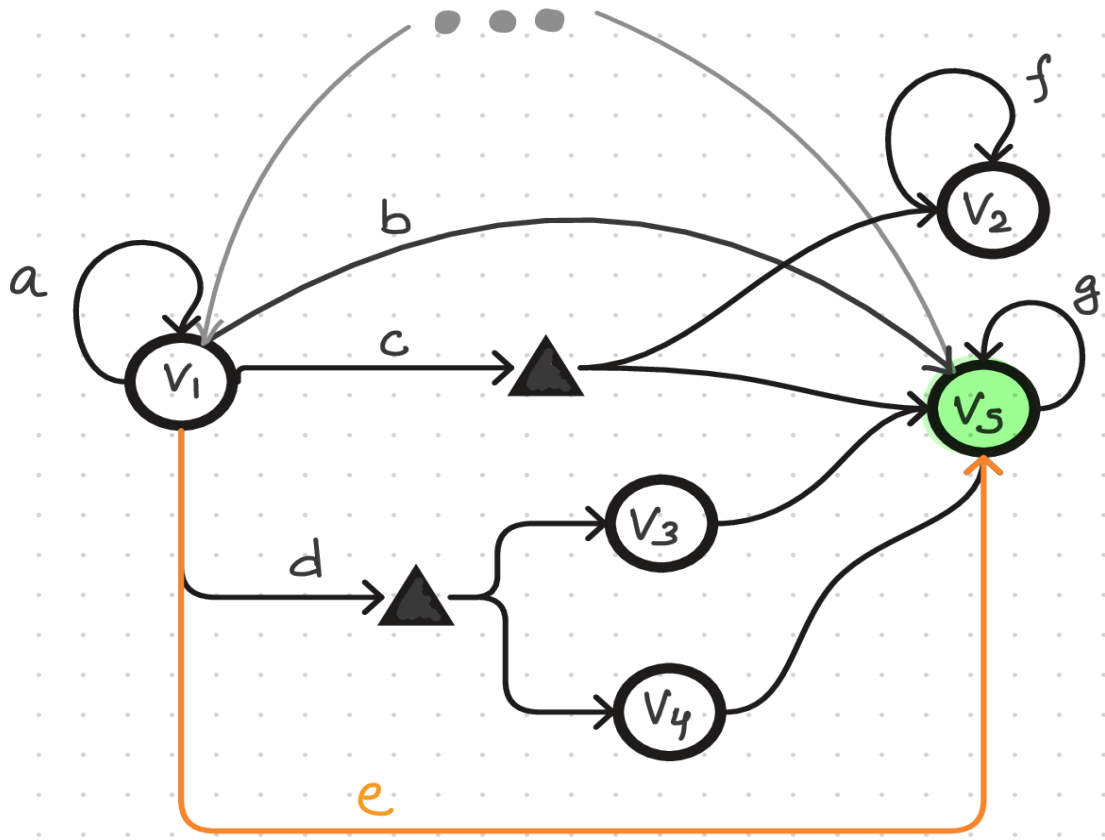
Games on graphs



find a strategy to visit v_5 infinitely often.

- one strategy might be incompatible with reward structure

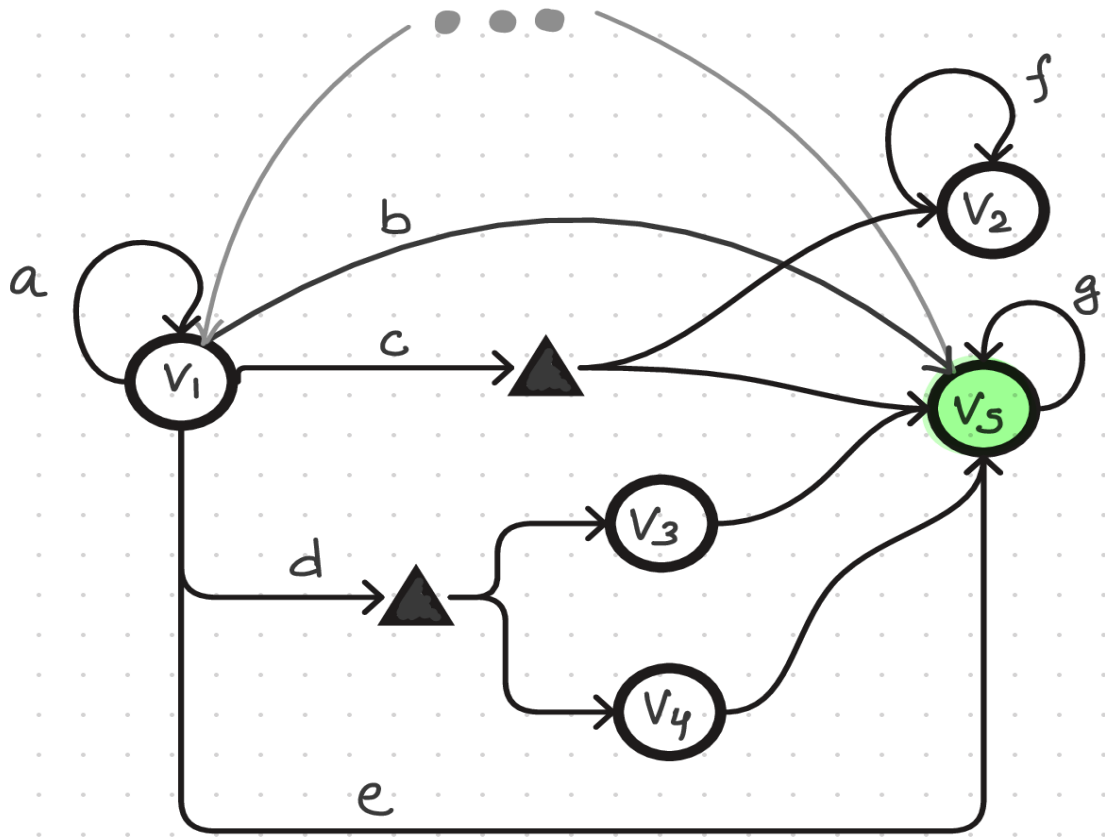
Games on graphs



find a strategy to visit v_5 infinitely often.

- one strategy might be incompatible with reward structure

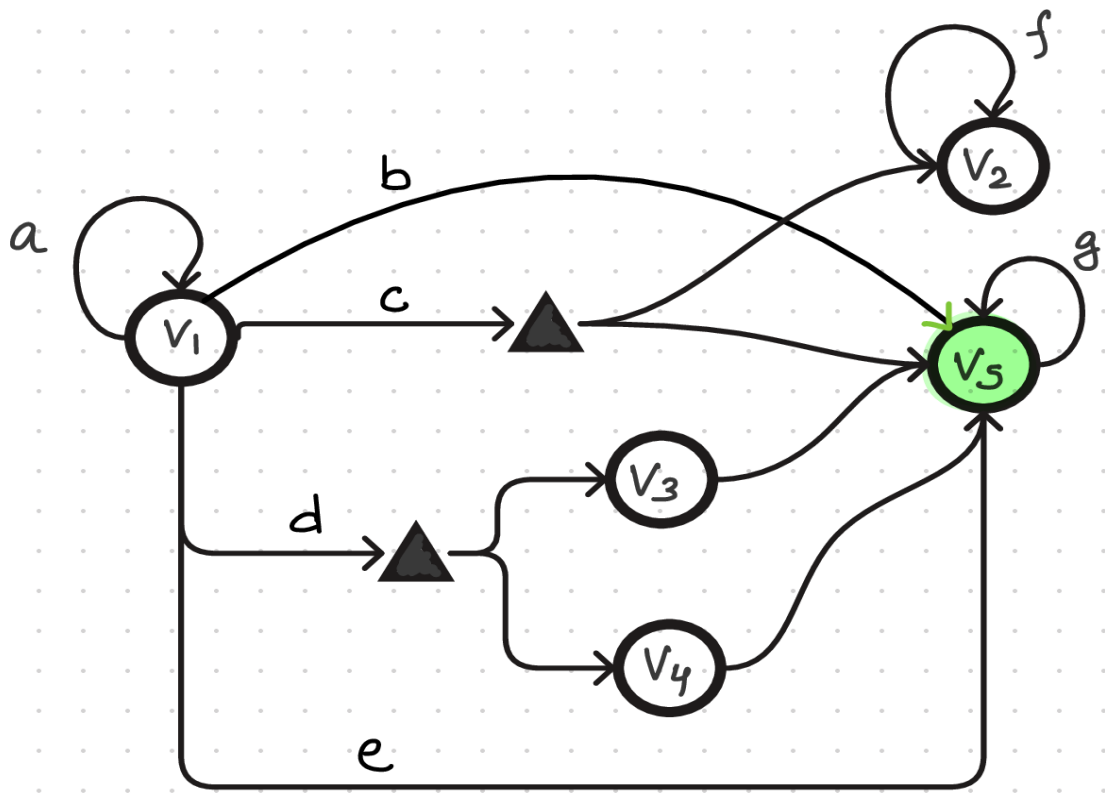
Games on graphs



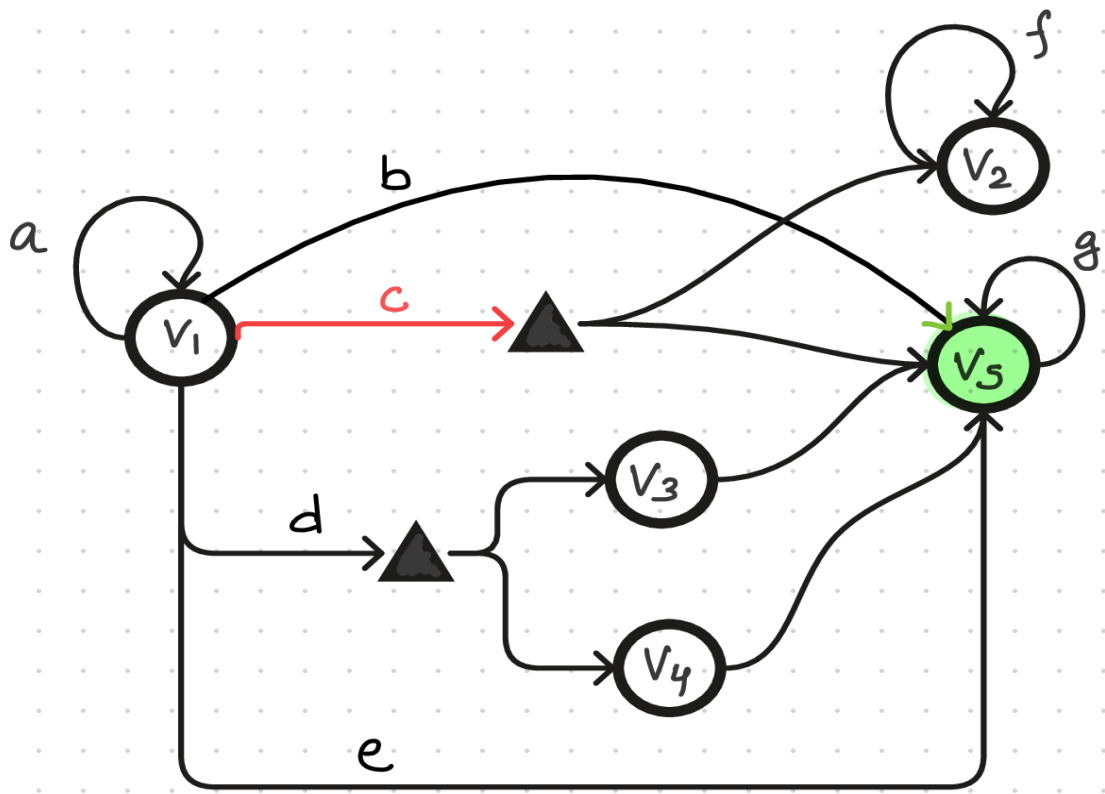
find a strategy to visit v_5 infinitely often.

- one strategy might be incompatible with reward structure
- there might be infinitely many strategies

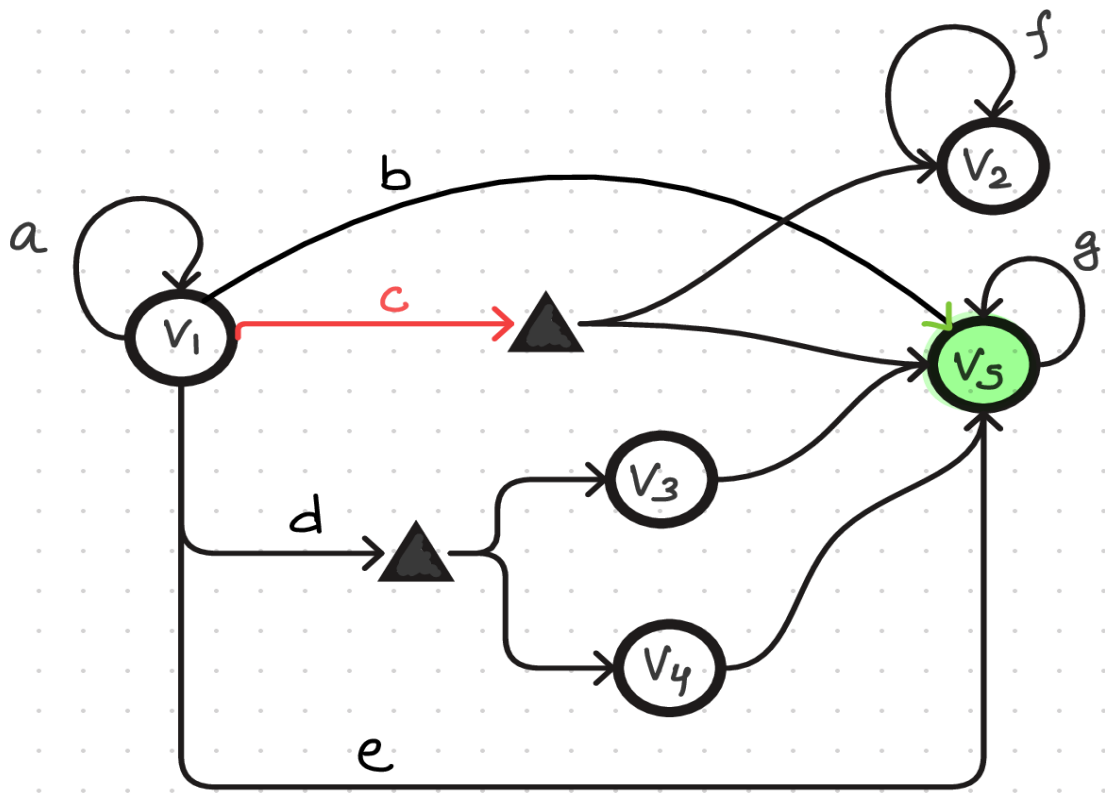
Strategy templates



Strategy templates



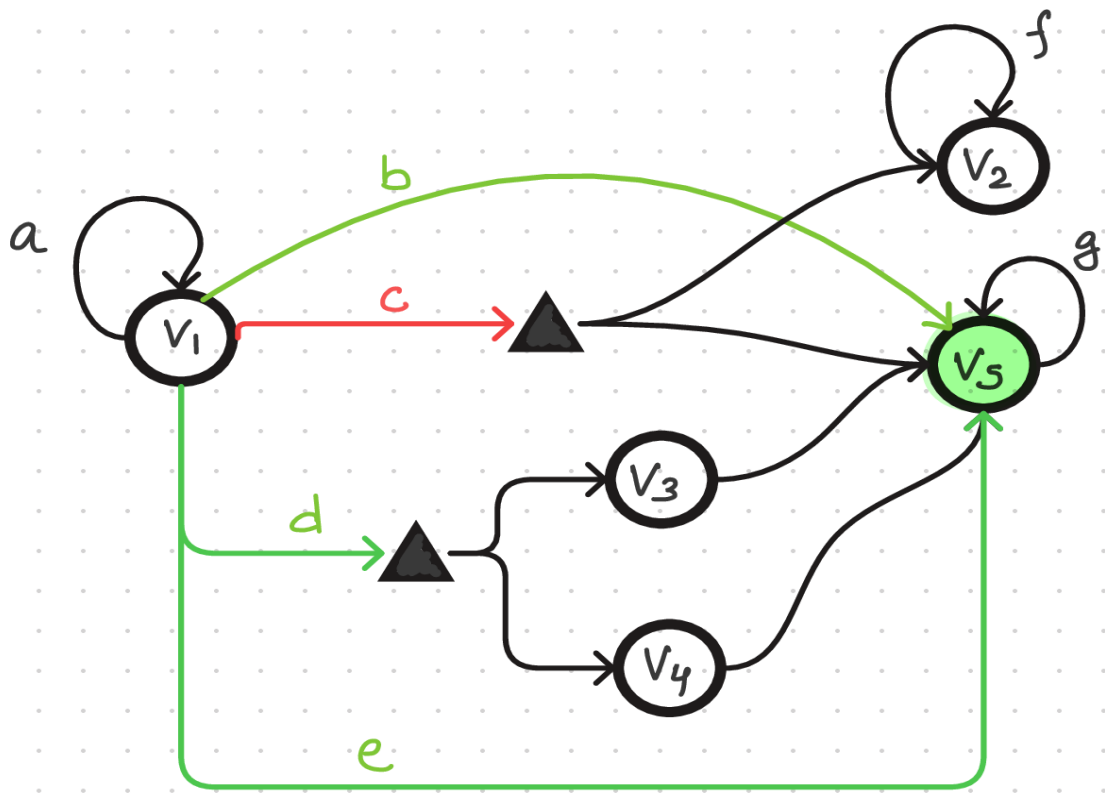
Strategy templates



At v_1,

- do not take action c

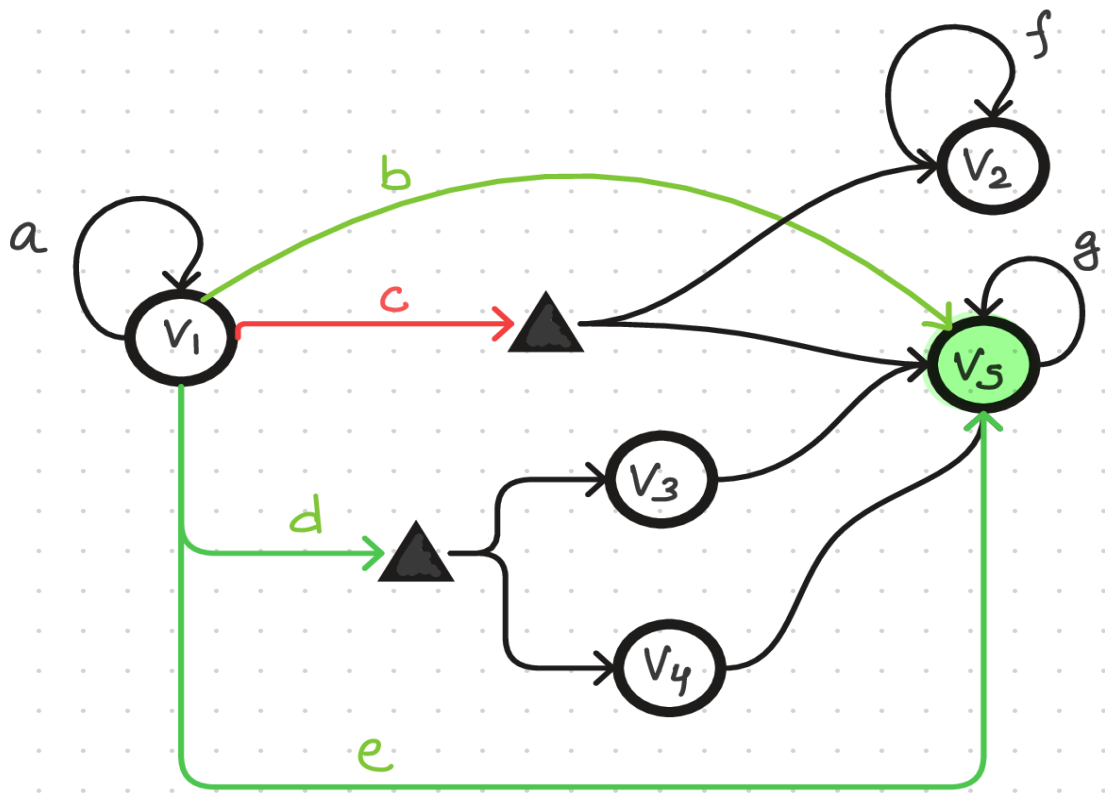
Strategy templates



At v_1,

- do not take action c

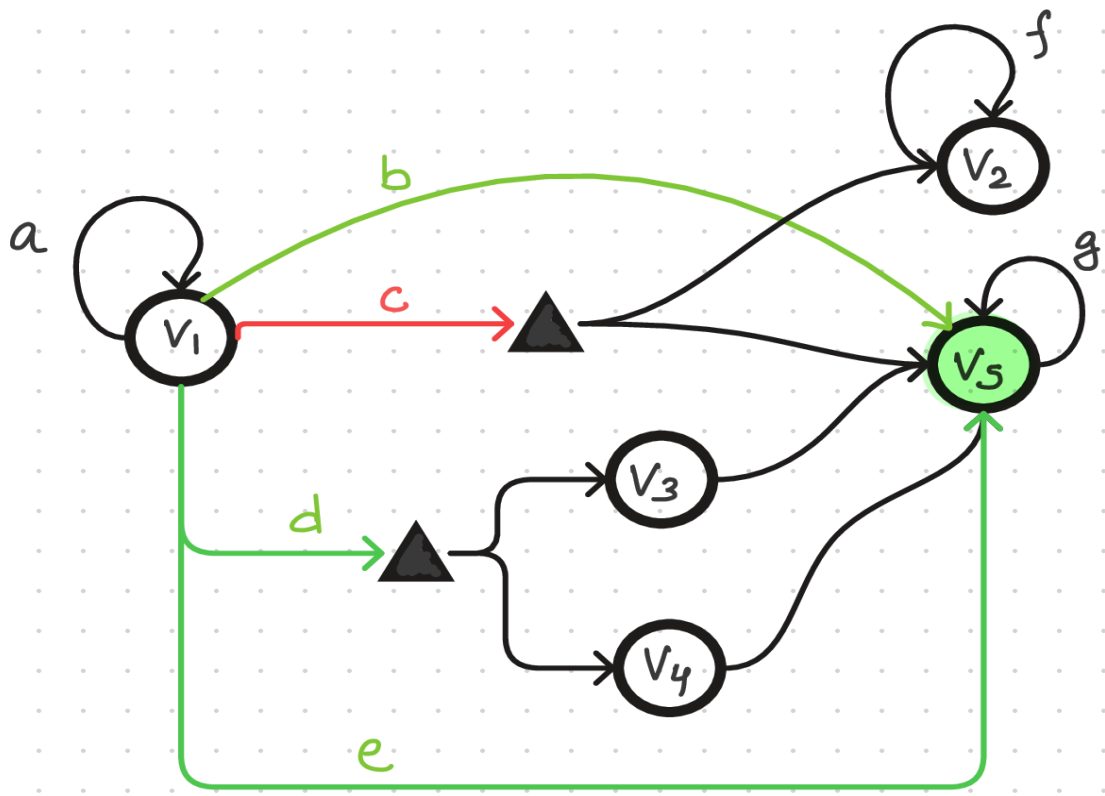
Strategy templates



At v_1,

- do not take action c
- infinitely often take one of action b, d, or e.

Strategy templates

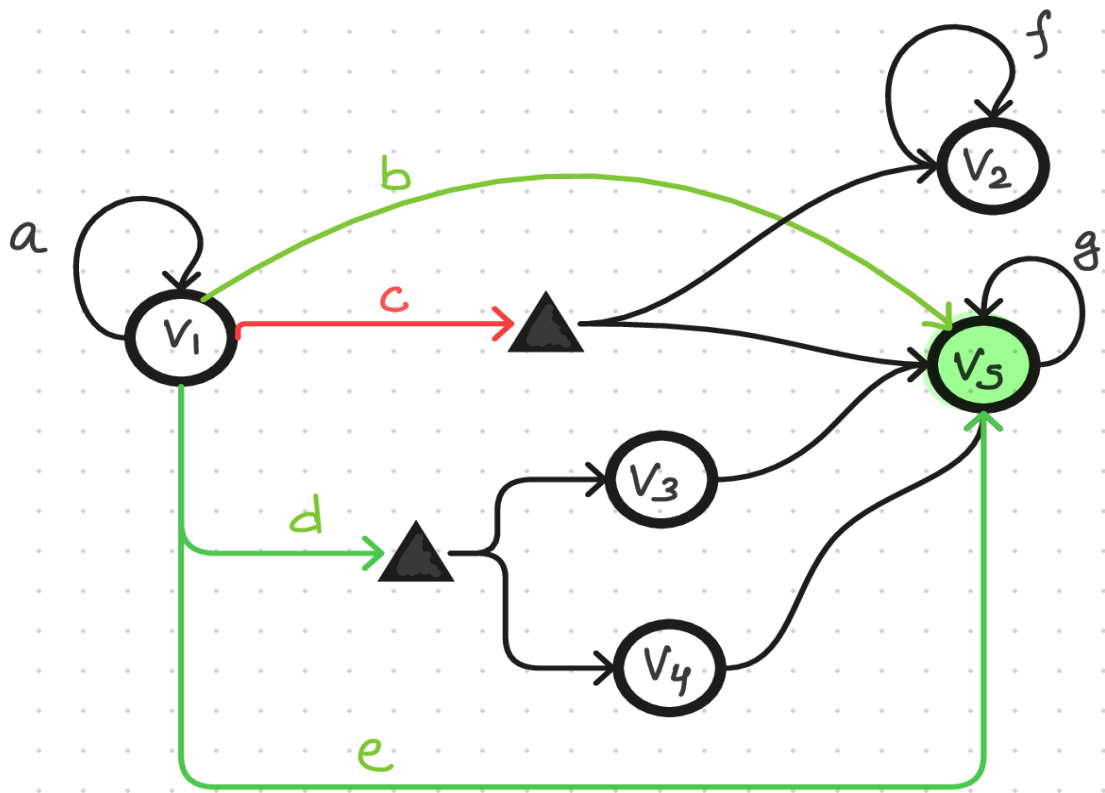


safety template

At v_1,

- do not take action c
- infinitely often take one of action b, d, or e.

Strategy templates



safety template

At v_1,

- do not take action c
- infinitely often take one of action b, d, or e.

live group template

Strategy templates

Theorem [AA, SPNayak, AK Schmuck - CAV '23]:

If you can satisfy the templates, you will win the game

Strategy templates

Theorem [AA, SPNayak, AK Schmuck - CAV'23]:

If you can satisfy the templates, you will win the game

Templates can be computed 'efficiently' for any ω-regular specification

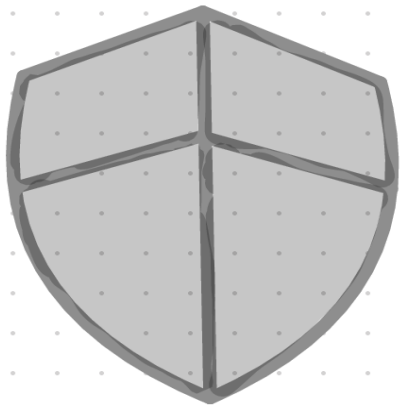
Strategy templates

Theorem [AA, SPNayak, AK Schmuck - CAV'23]:

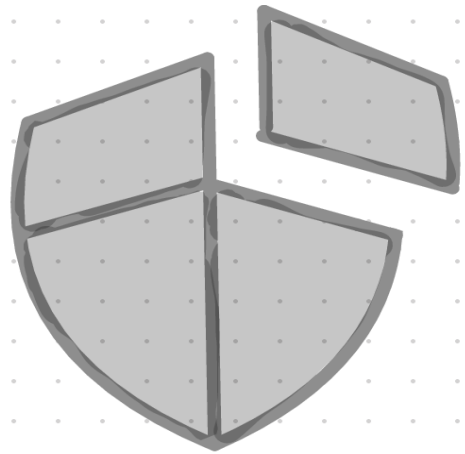
If you can satisfy the templates, you will win the game

Templates can be computed 'efficiently' for any ω-regular specification

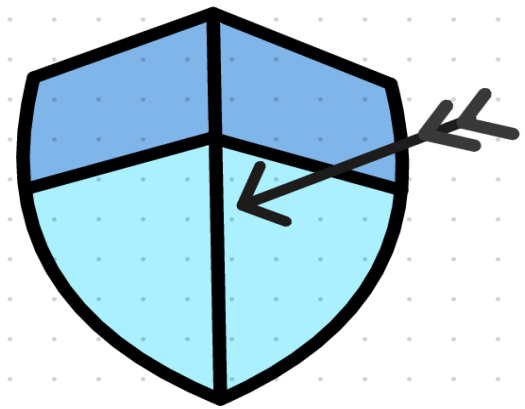
Decomposes a global specification to local decisions



What is a shield?

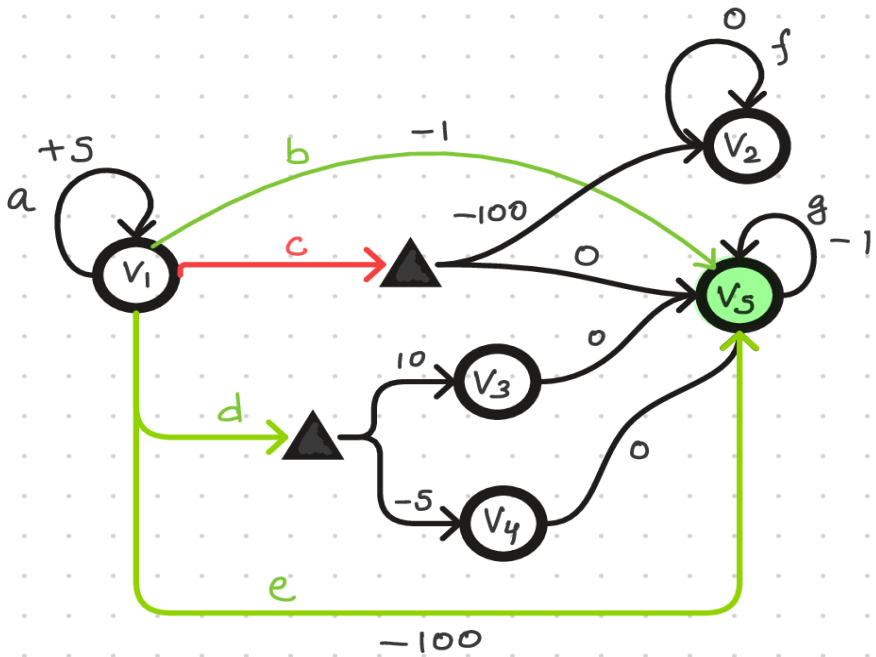


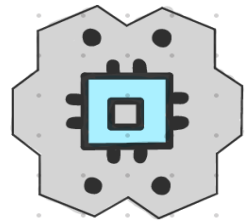
How to construct a shield?



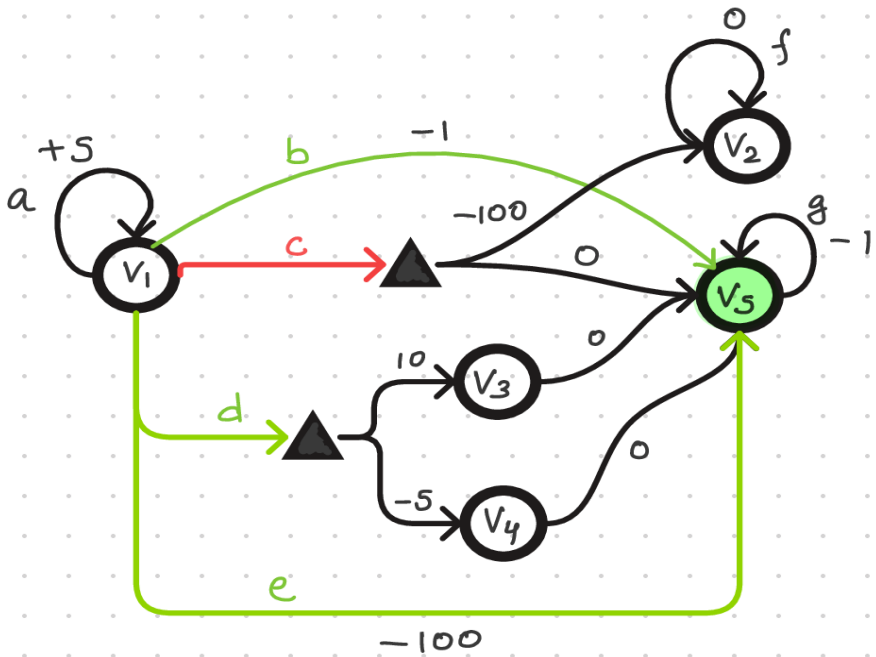
How to use a shield?

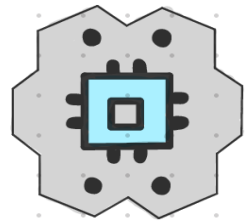
Shielding policies





Shielding policies





At v_1

a : 0.7

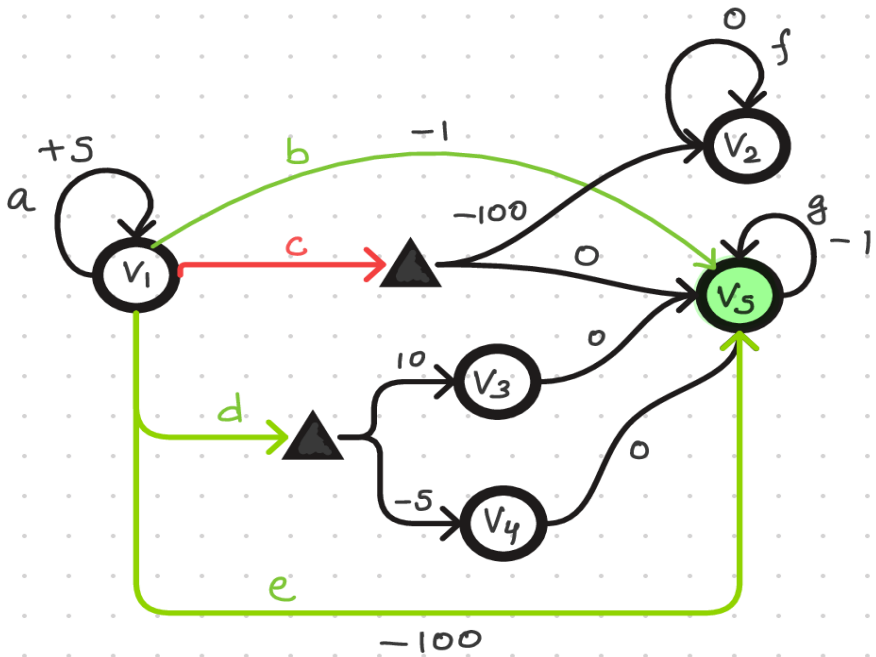
b : 0.1

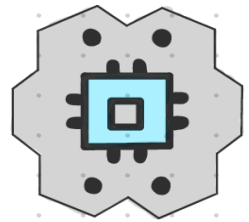
c : 0.1

d : 0.1

e : 0

Shielding policies





At v_1

a : 0.7

b : 0.1

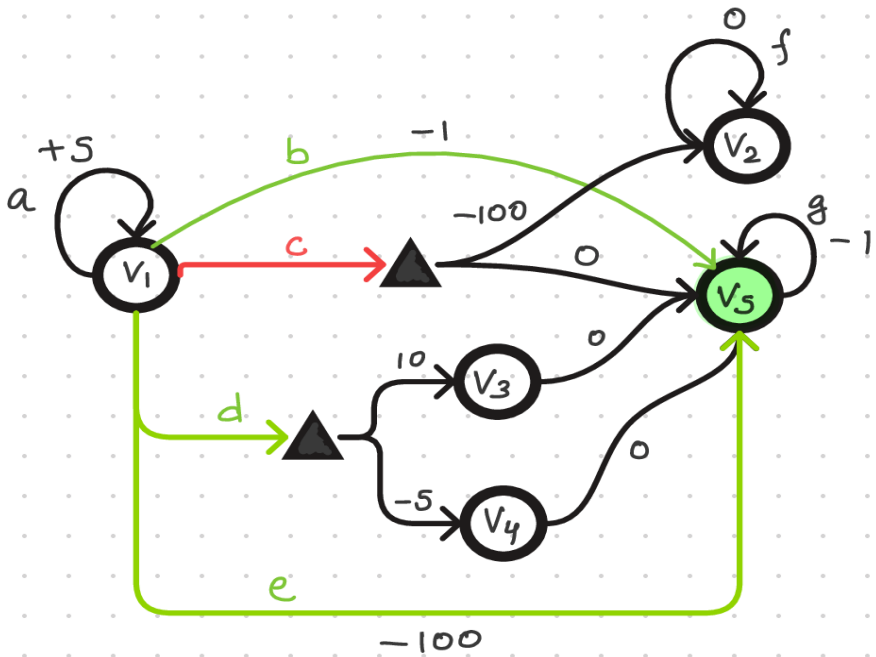
c : 0.1

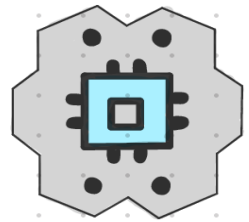
d : 0.1

e : 0

Pick a small α

Shielding policies





At v_1

a : 0.7

b : 0.1

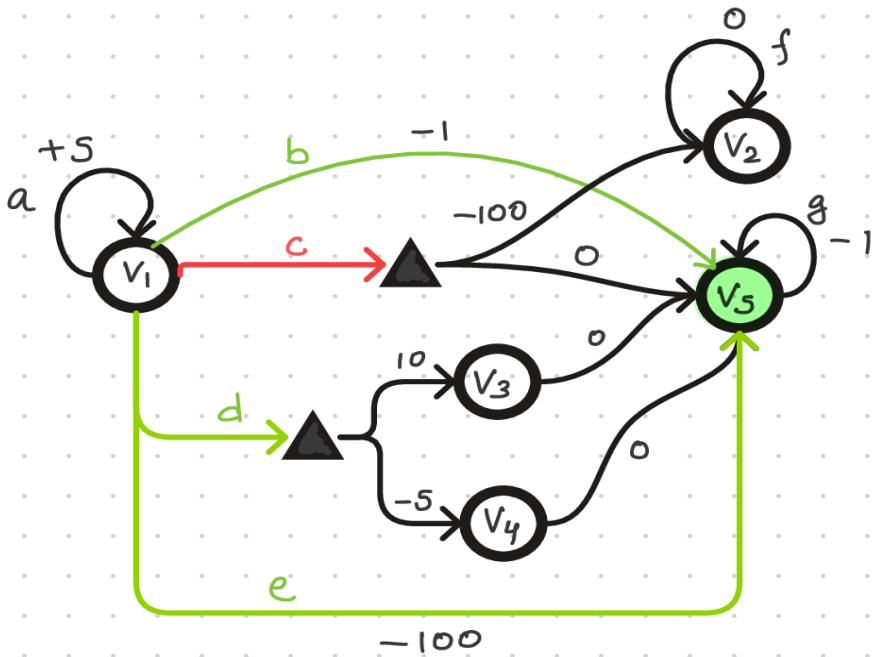
c : 0.1

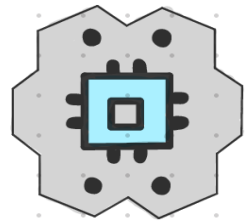
d : 0.1

e : 0

Pick a small $\alpha = 0.05$

Shielding policies





At v_1

a : 0.7

b : 0.1

c : 0.1

d : 0.1

e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7

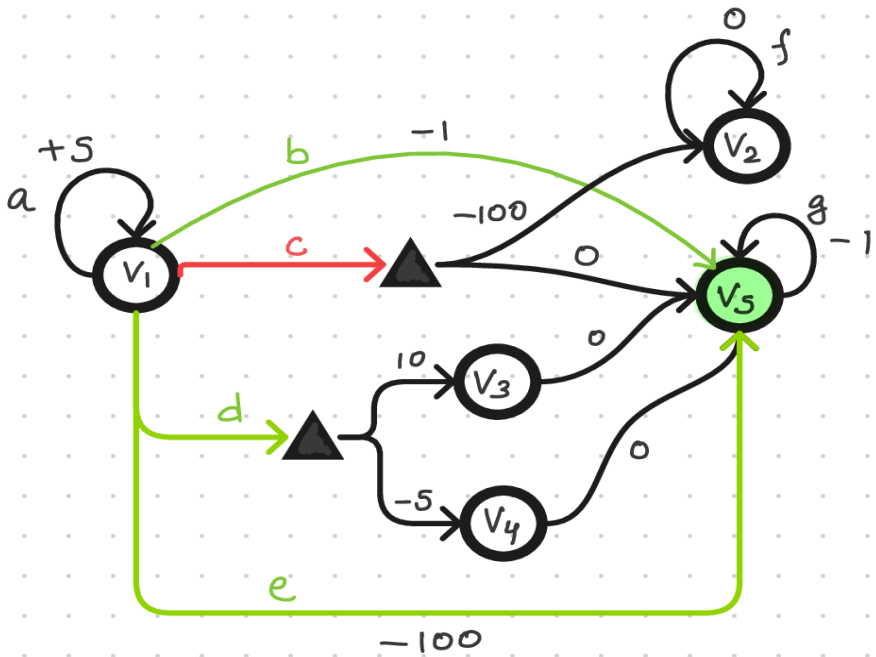
b : 0.1

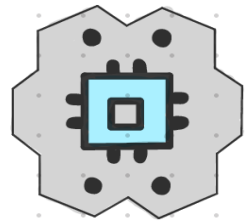
c : 0.1

d : 0.1

e : 0

Shielding policies





At v_1

a : 0.7

b : 0.1

c : 0.1

d : 0.1

e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7

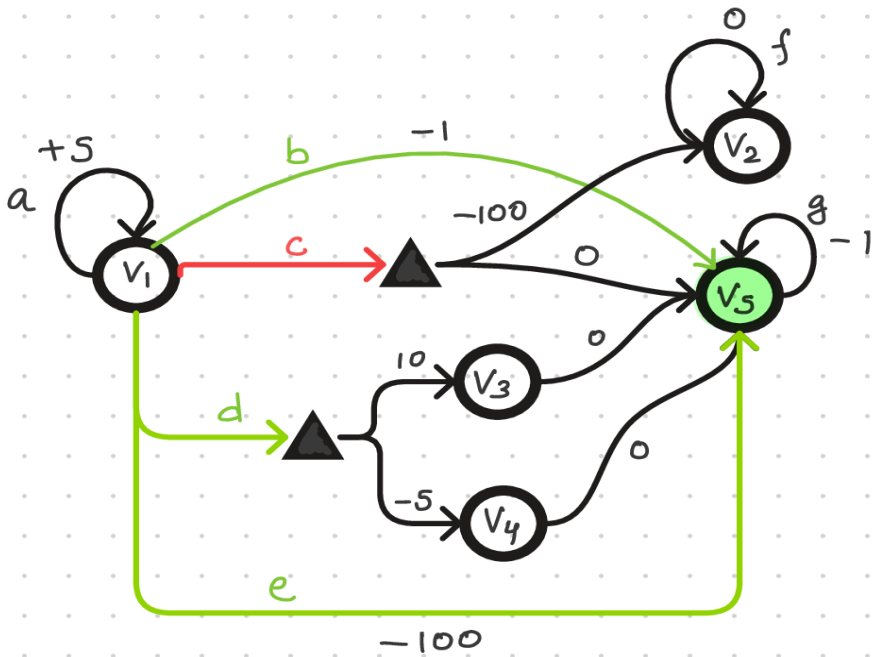
b : 0.1

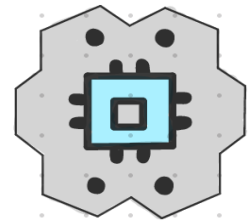
c : 0

d : 0.1

e : 0

Shielding policies





At v_1

a : 0.7

b : 0.1

c : 0.1

d : 0.1

e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7

b : 0.1

c : 0

d : 0.1

e : 0

norm. →

At v_1

a : 0.77

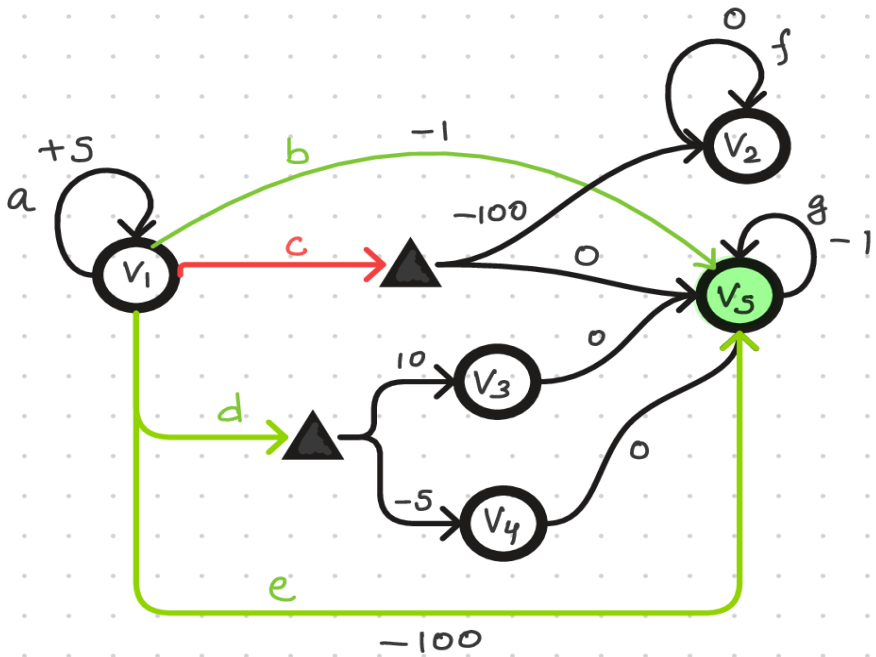
b : 0.11

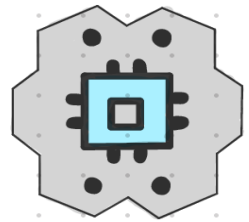
c : 0

d : 0.11

e : 0

Shielding policies





At v_1

a : 0.7

b : 0.1

c : 0.1

d : 0.1

e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7

b : 0.1

c : 0

d : 0.1

e : 0

norm. →

At v_1

a : 0.77

b : 0

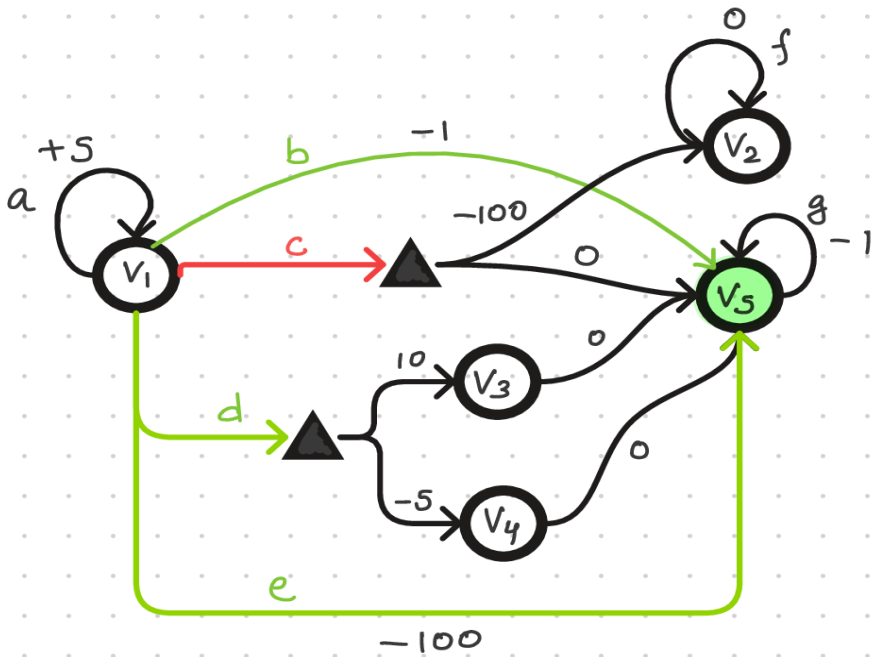
c : 0.11

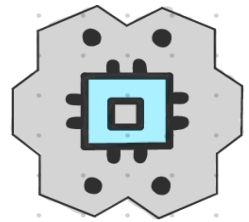
d : 0.11

e : 0

a

Shielding policies





At v_1

a : 0.7

b : 0.1

c : 0.1

d : 0.1

e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7

b : 0.1 + 0.05

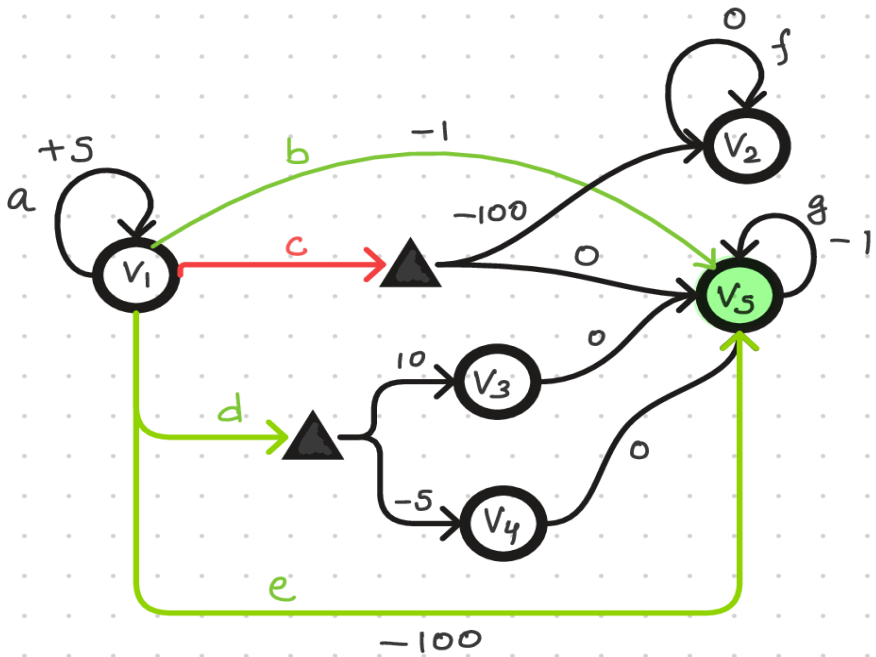
c : 0

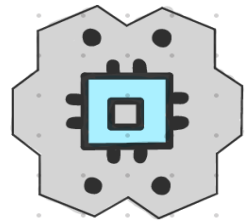
d : 0.1 + 0.05

e : 0 + 0.05

a

Shielding policies





At v_1

a : 0.7

b : 0.1

c : 0.1

d : 0.1

e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7

b : 0.15

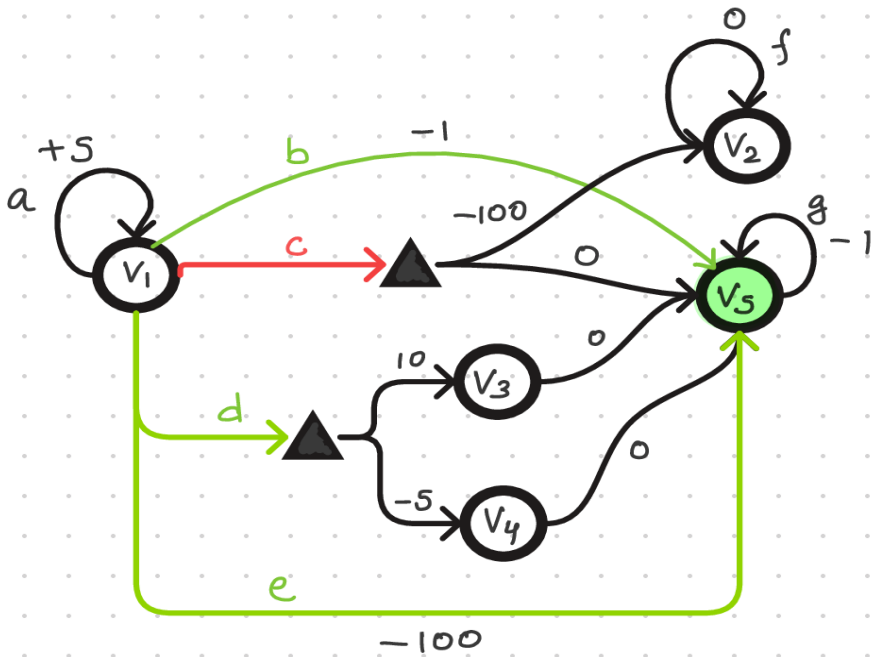
c : 0

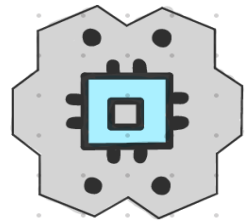
d : 0.15

e : 0.05

a

Shielding policies





At v_1

a : 0.7
b : 0.1
c : 0.1
d : 0.1
e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7
b : 0.15
c : 0
d : 0.15
e : 0.05

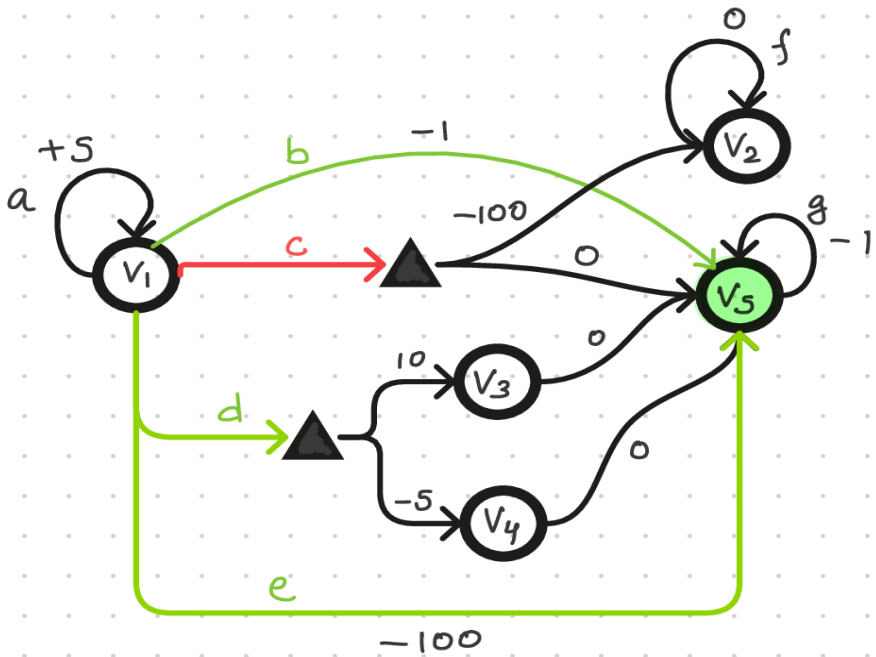
norm. →

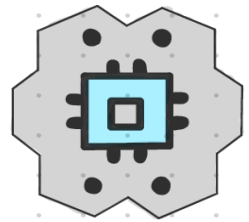
At v_1

a : 0.66
b : 0.14
c : 0
d : 0.14
e : 0.04

a

Shielding policies





At v_1

a : 0.7
b : 0.1
c : 0.1
d : 0.1
e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7
b : 0.15
c : 0
d : 0.15
e : 0.05

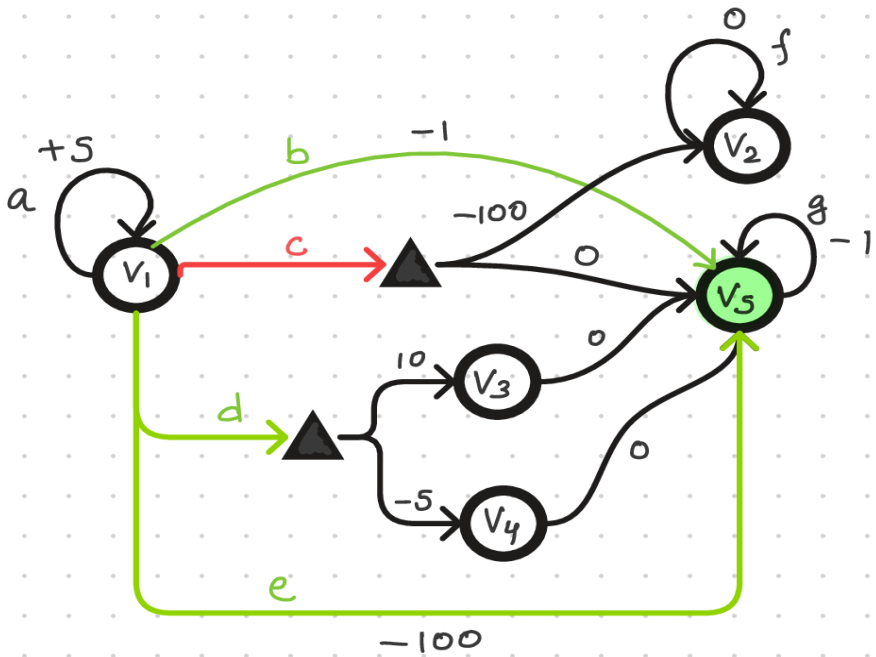
norm. →

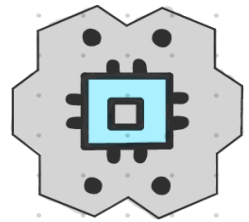
At v_1

a : 0.66
b : 0.14
c : 0
d : 0.14
e : 0.04

a a

Shielding policies





At v_1

a : 0.7

b : 0.1

c : 0.1

d : 0.1

e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7

b : 0.2

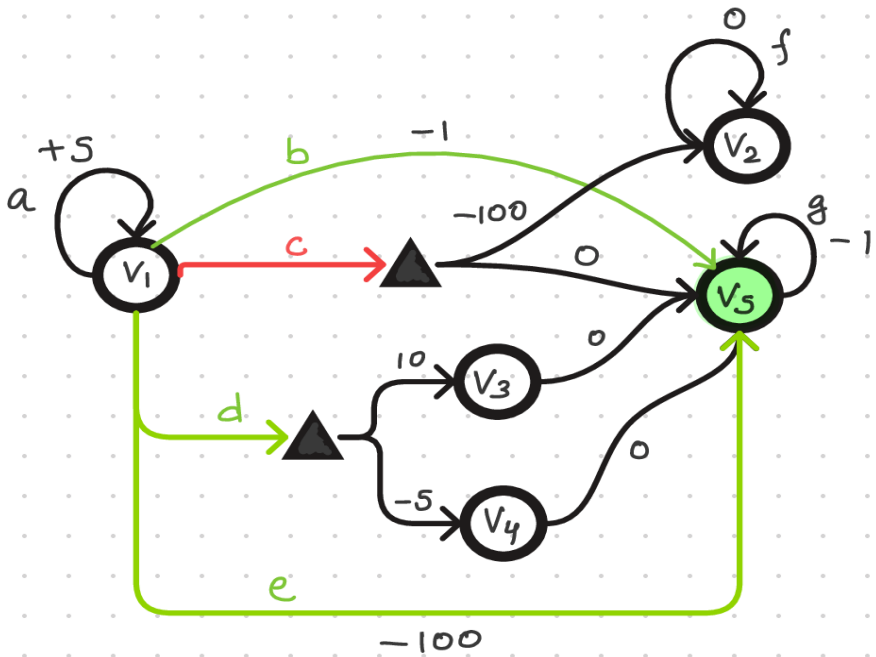
c : 0

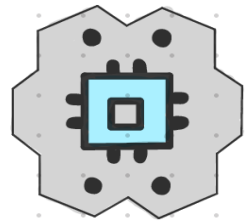
d : 0.2

e : 0.1

a a

Shielding policies





Pick a small $\alpha = 0.05$

At v_1

a : 0.7
b : 0.1
c : 0.1
d : 0.1
e : 0

At v_1

a : 0.7
b : 0.2
c : 0
d : 0.2
e : 0.1

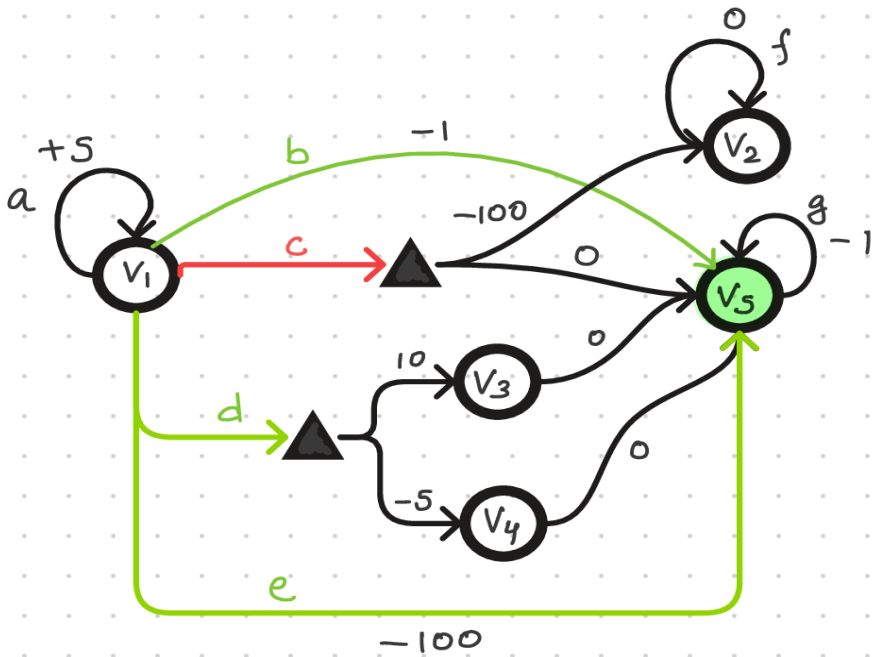
norm. →

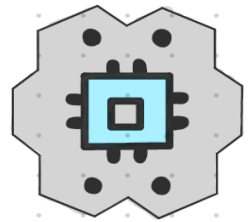
At v_1

a : 0.58
b : 0.16
c : 0
d : 0.16
e : 0.08

a a

Shielding policies





At v_1

a : 0.7
b : 0.1
c : 0.1
d : 0.1
e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7
b : 0.2
c : 0
d : 0.2
e : 0.1

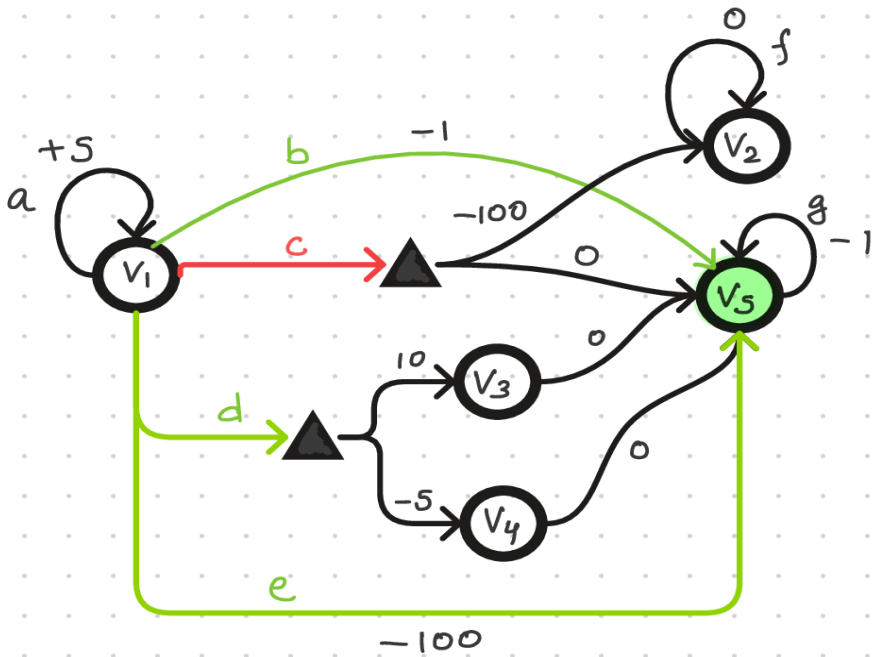
norm. →

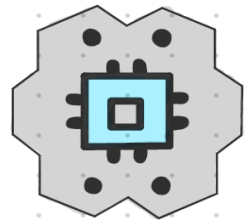
At v_1

a : 0.58
b : 0.16
c : 0
d : 0.16
e : 0.08

a a a a ...

Shielding policies





Pick a small $\alpha = 0.05$

At v_1

a : 0.7
b : 0.1
c : 0.1
d : 0.1
e : 0

At v_1

a : 0.7
b : 2
c : 0
d : 2
e : 1.9

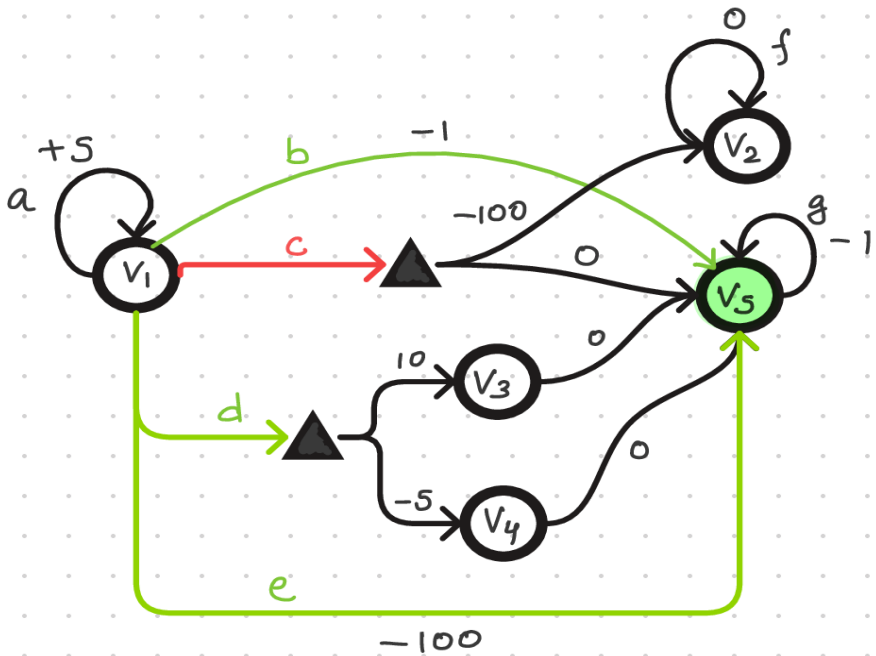
norm. →

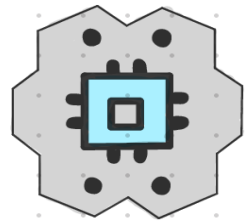
At v_1

a : 0.10
b : 0.30
c : 0
d : 0.30
e : 0.28

a a a a ...

Shielding policies





At v_1

a : 0.7
b : 0.1
c : 0.1
d : 0.1
e : 0

Pick a small $\alpha = 0.05$

At v_1

a : 0.7
b : 0.1
c : 0
d : 0.1
e : 0

norm. →

At v_1

a : 0.77
b : 0.11
c : 0
d : 0.11
e : 0

a a a a ... b

Shielding policies

Ensures specification satisfaction surely/almost surely

Shielding policies

Ensures specification satisfaction surely/almost surely

Interferes minimally

Shielding policies

Ensures specification satisfaction surely/almost surely

Interferes minimally

ε-close to original policy's expected reward

Shielding policies

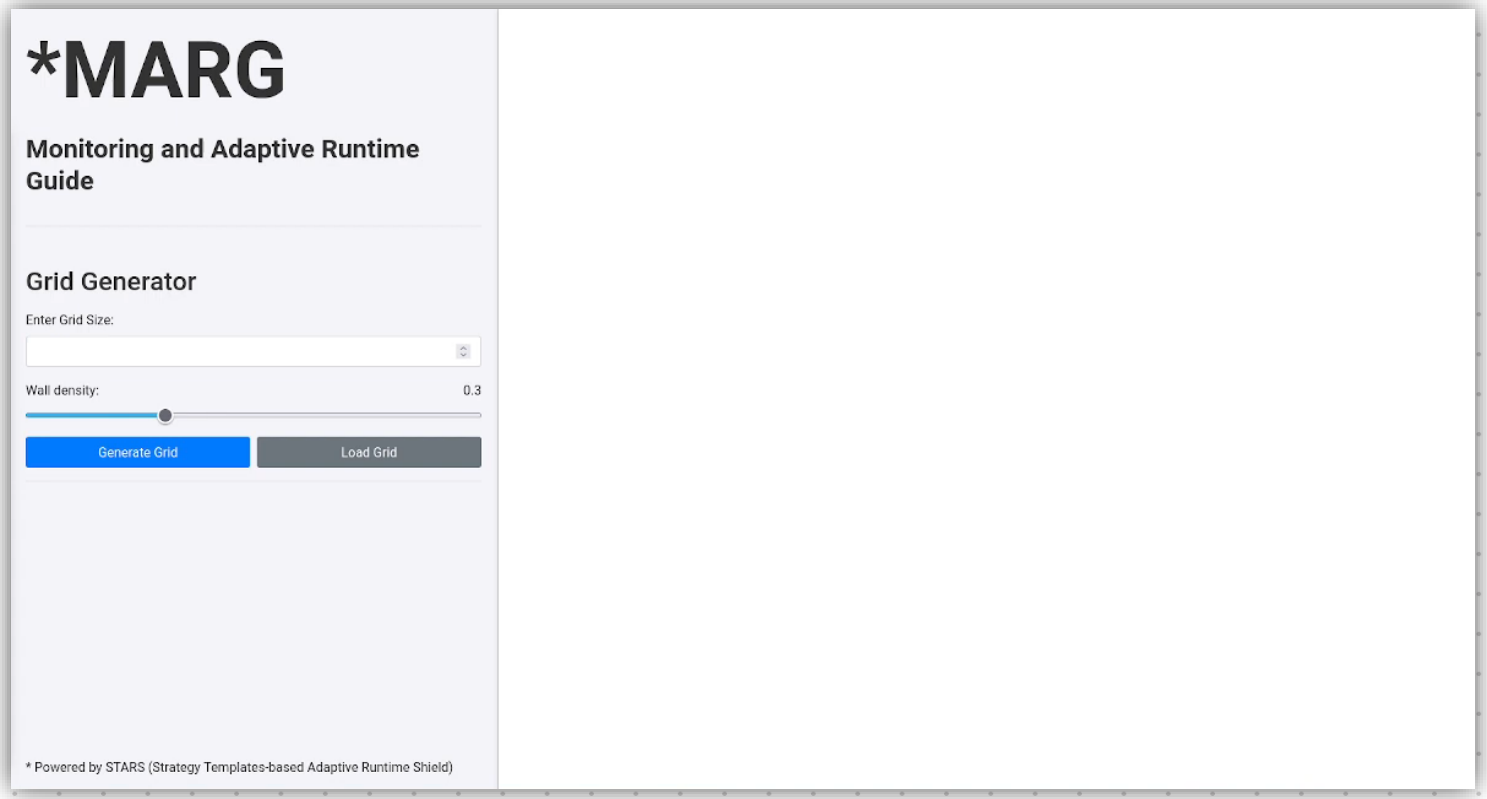
Ensures specification satisfaction surely/almost surely

Interferes minimally

ε-close to original policy's expected reward

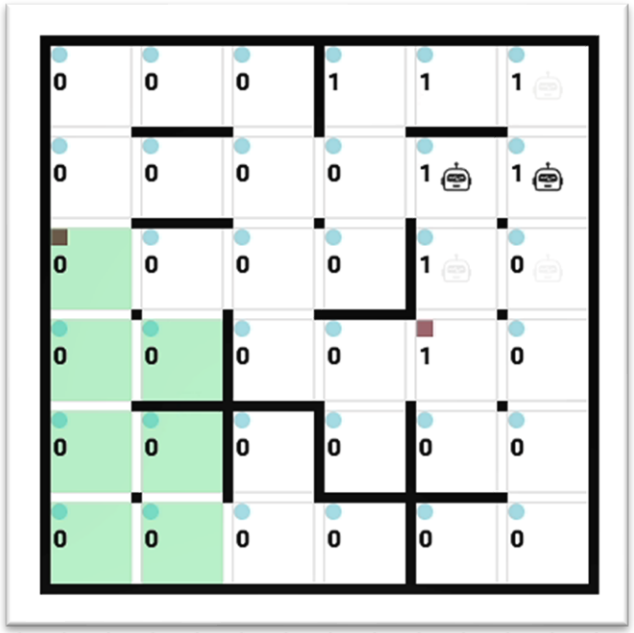
Compositional

Factory Bot

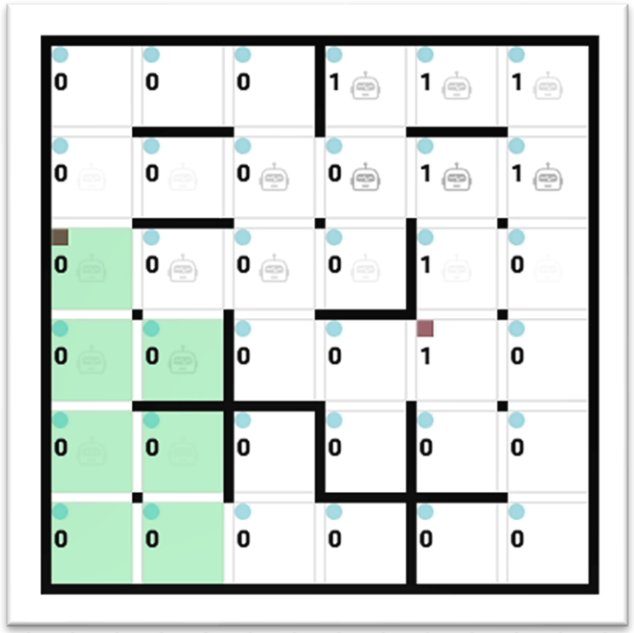


https://marg.mpi-sws.org

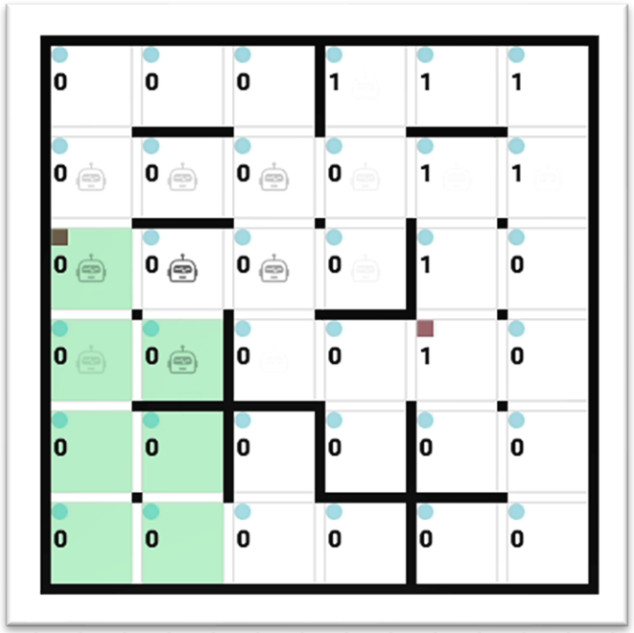
Factory Bot



No shield

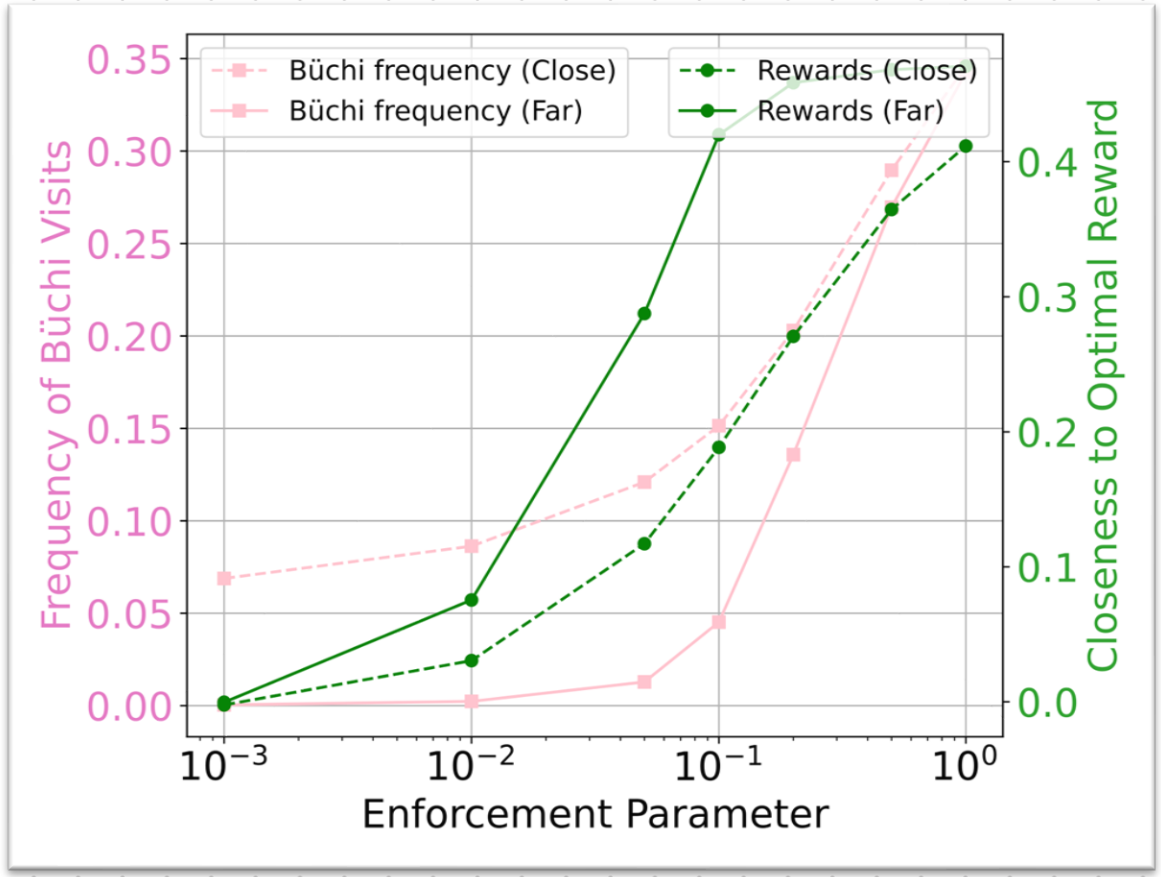


Low α



High α

Factory Bot



Enforcement parameter (α) $\uparrow$ $\Rightarrow$

Factory Bot

Enforcement parameter (α) ↑ $\Rightarrow$

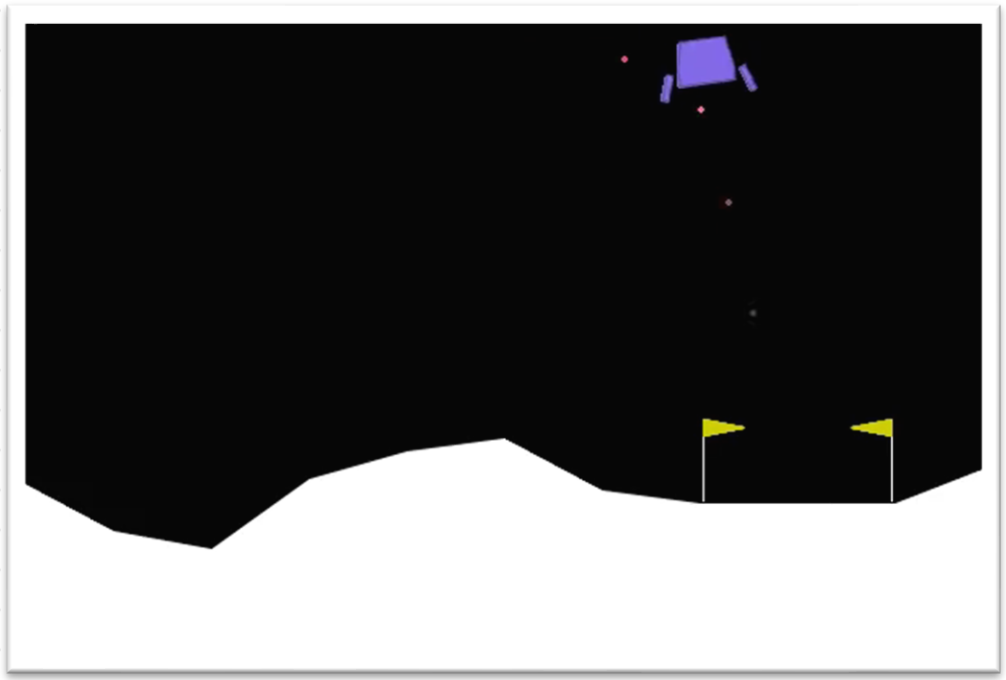
- Frequency of goal state visits ↑

Factory Bot

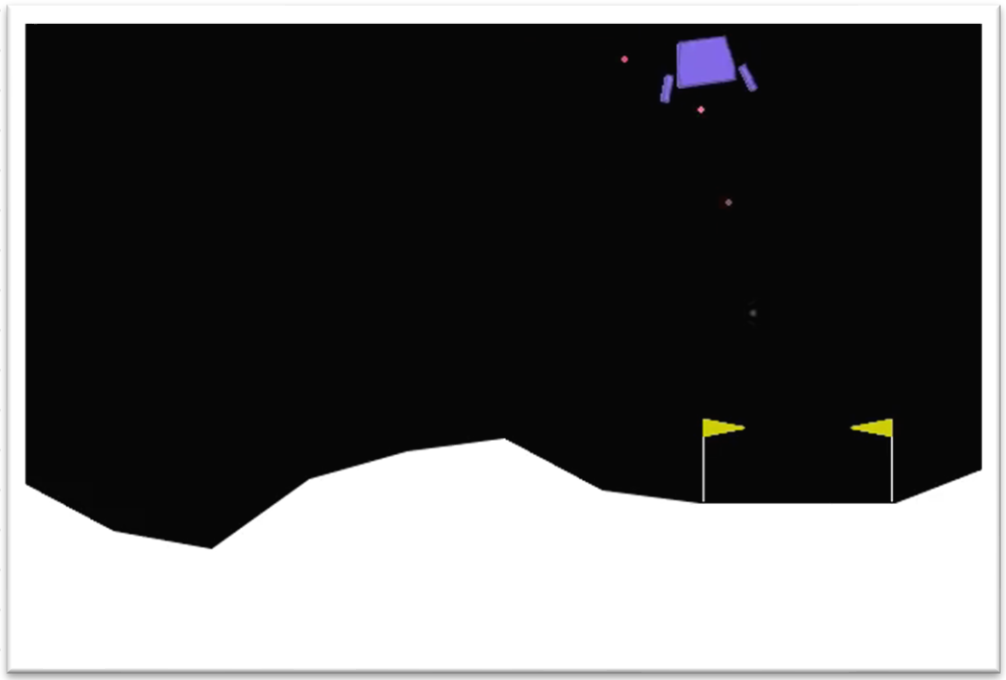
Enforcement parameter (α) ↑ ⇒

- Frequency of goal state visits ↑
- Distance from optimal reward ↑

Shielding under unknown MDPs

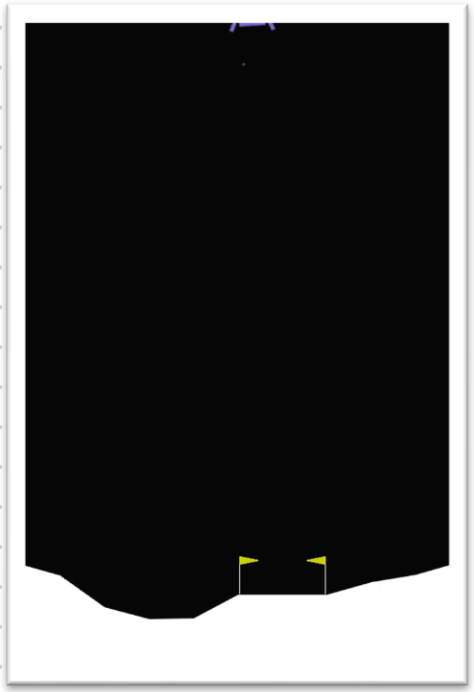


Shielding under unknown MDPs

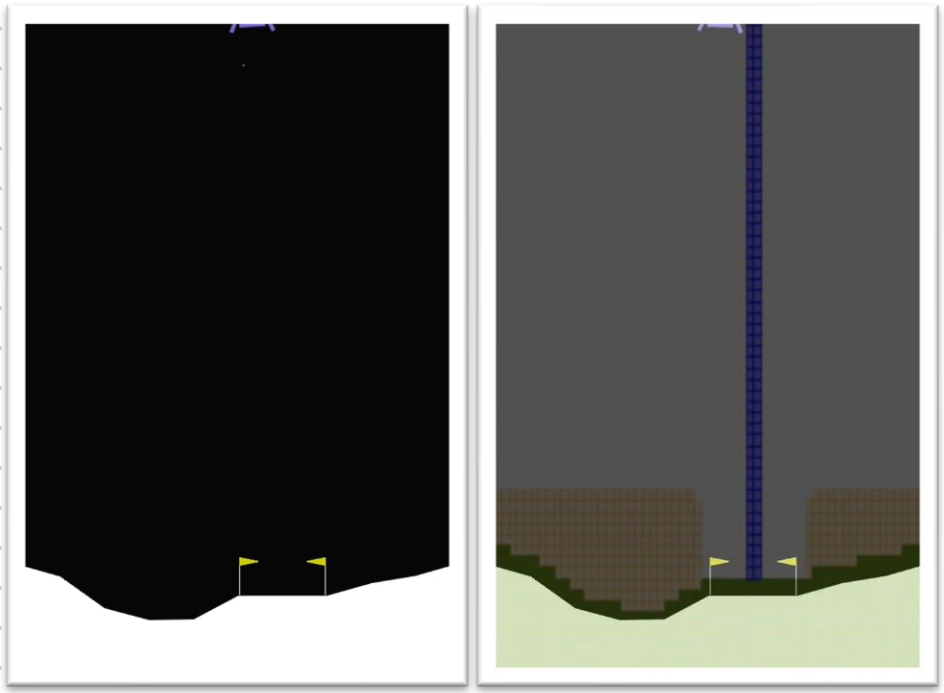


- Lunar Lander has 8-dim continuous space
- Underlying MDP is unknown

Shielding under unknown MDPs

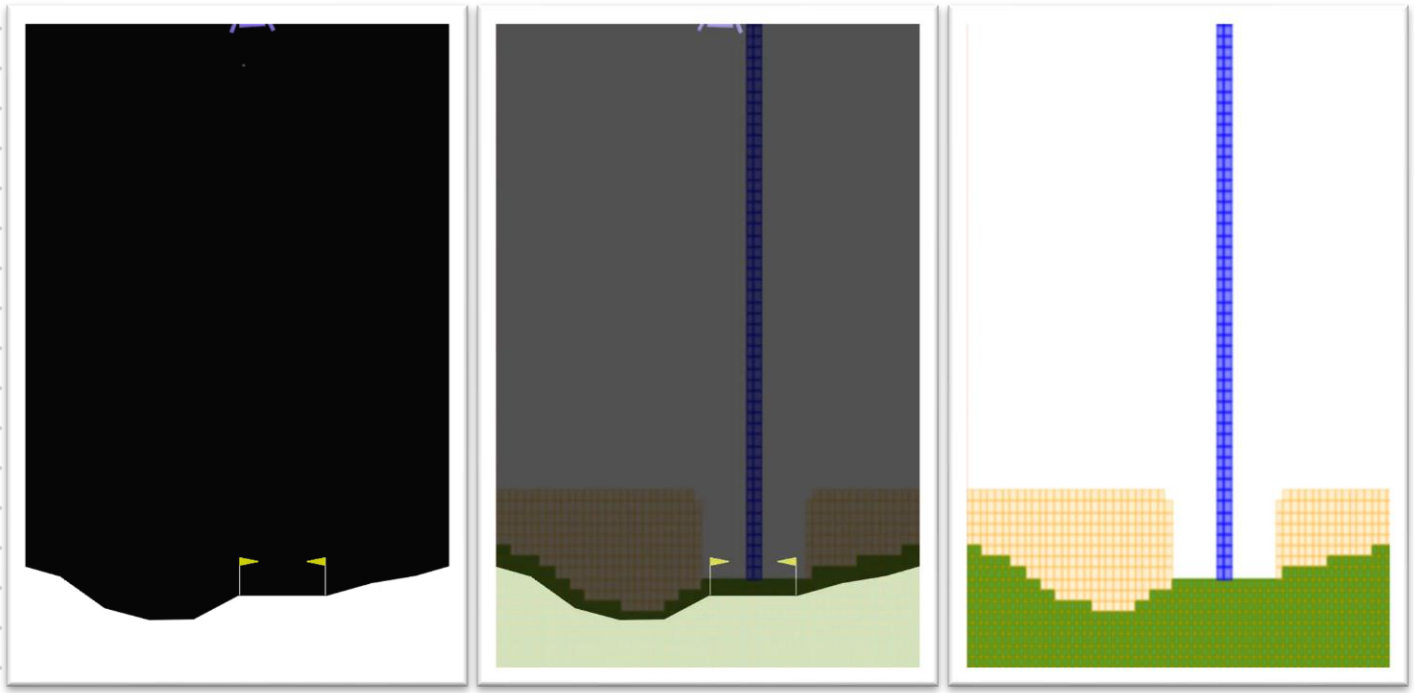


Shielding under unknown MDPs



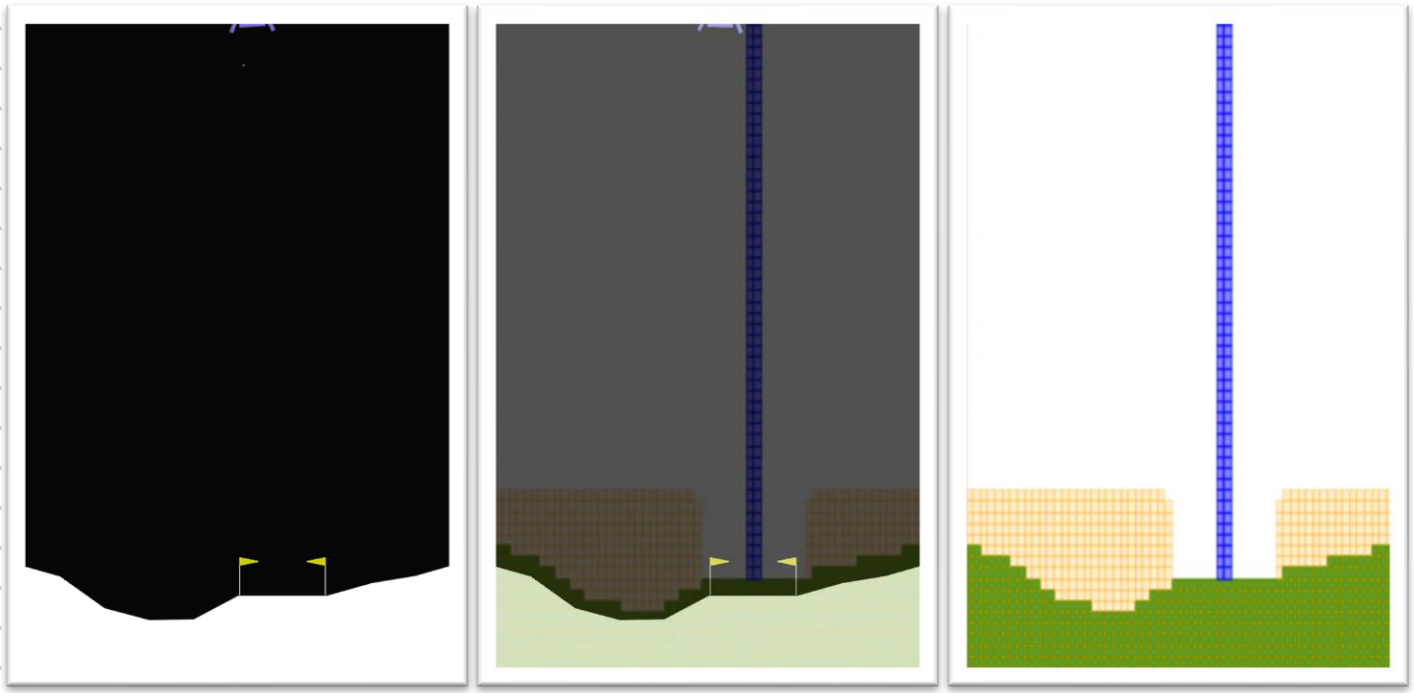
Discretization

Shielding under unknown MDPs



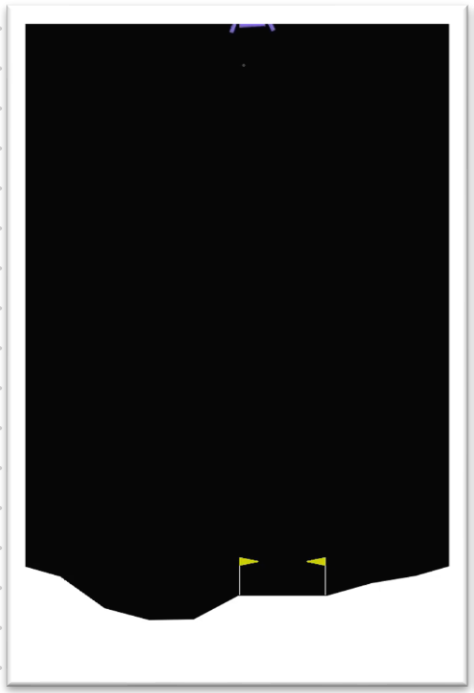
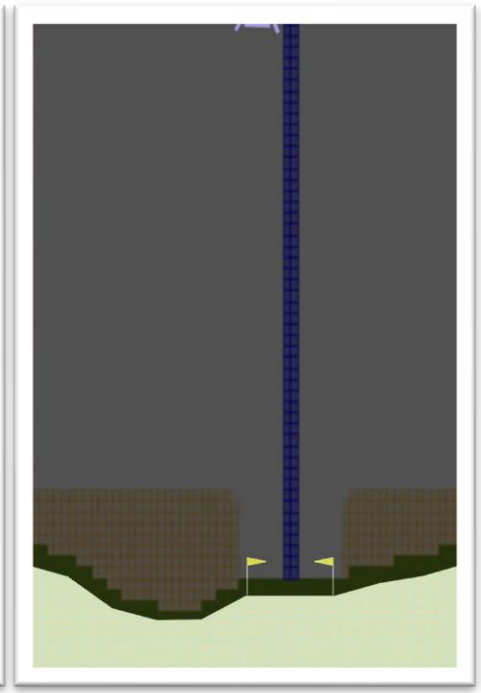
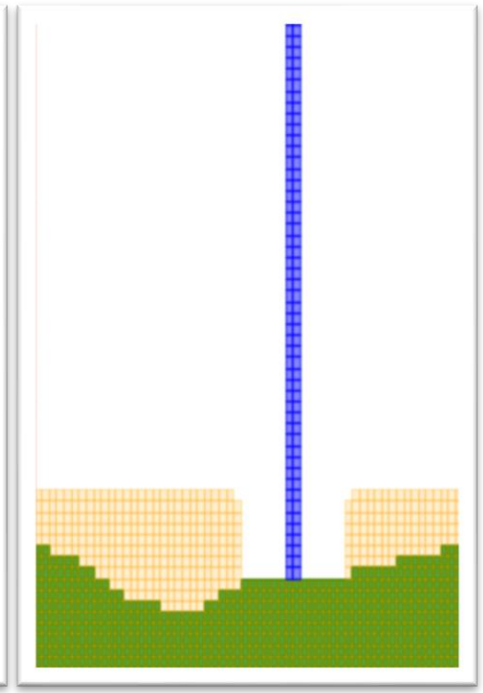
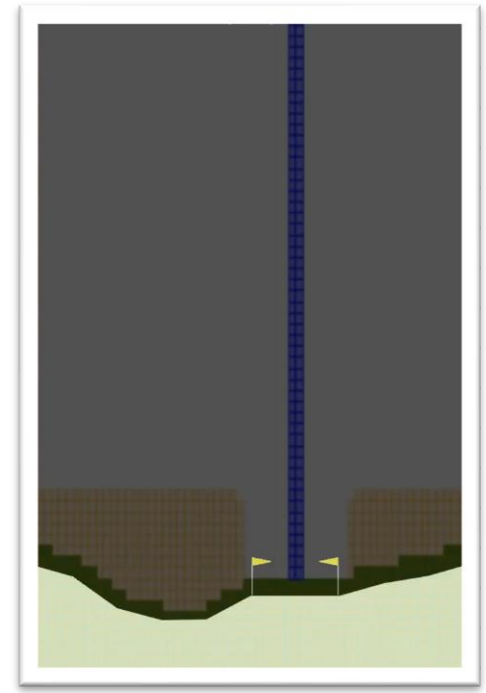
Discretization $\longrightarrow$ Abstraction

Shielding under unknown MDPs



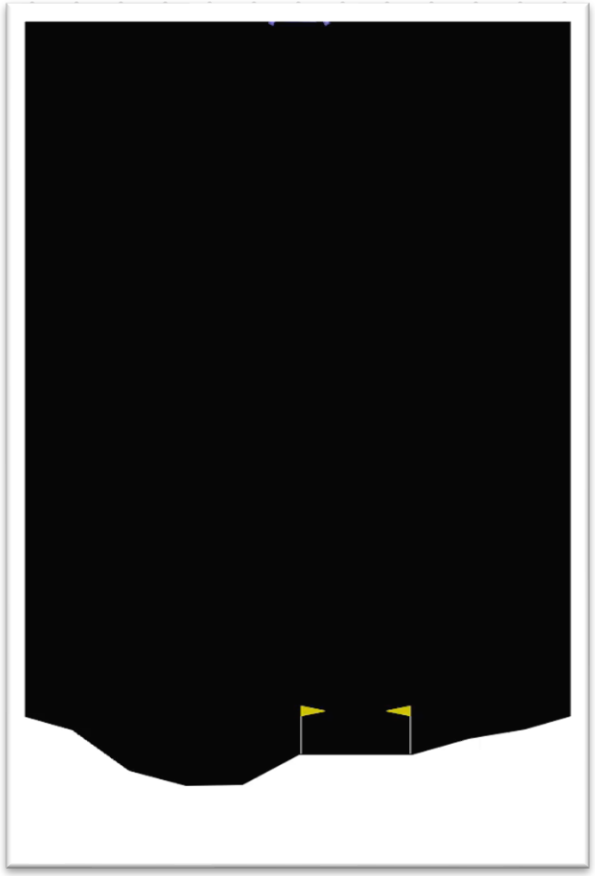
Discretization $\longrightarrow$ Abstraction $\longrightarrow$ Shield synthesis

Shielding under unknown MDPs

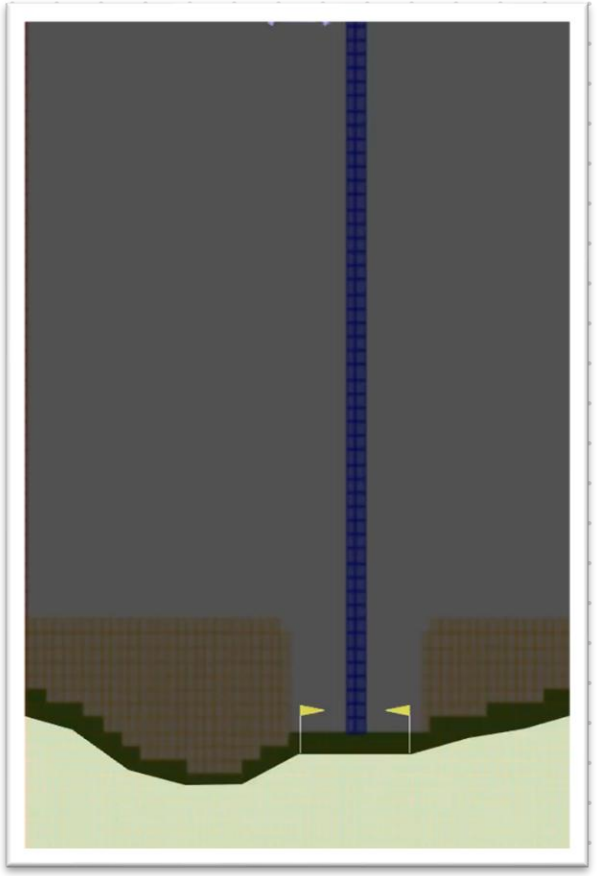





Discretization → Abstraction → Shield synthesis → Concretization

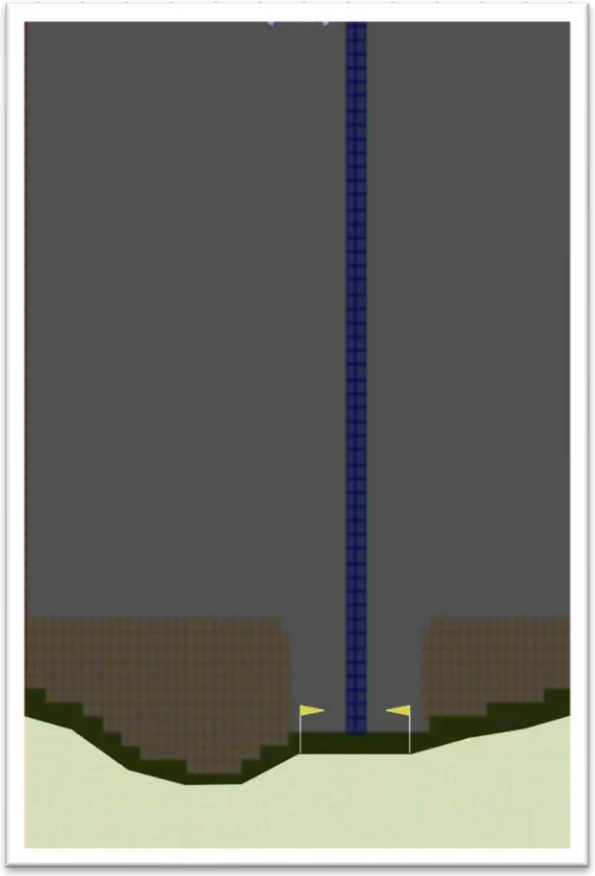
Lunar Lander



No shield

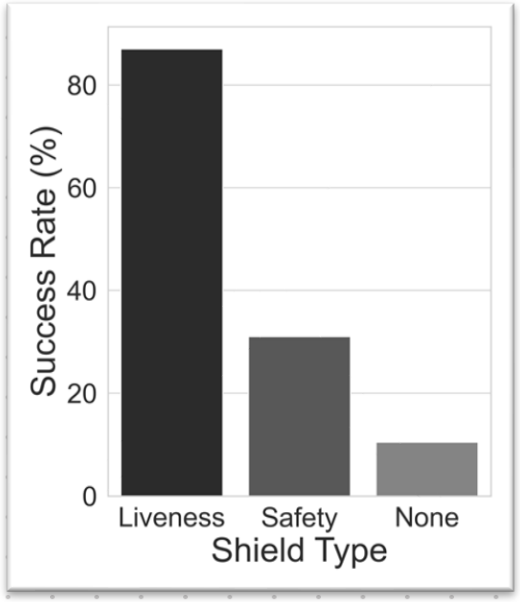


Low α

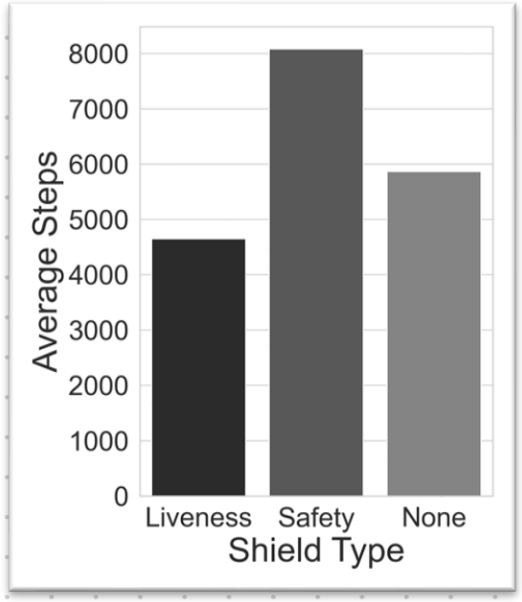


High α

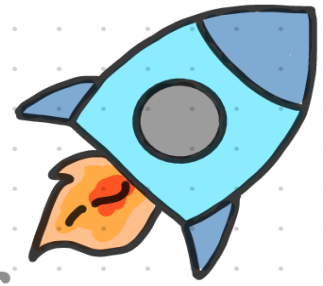
Lunar Lander



Reaches ~89% safe landing on helipad



Targetted nudging leads to faster landing



Thank You

Ashwani Anand, MPI-SWS

ashwani@mpi-sws.org

mpi-sws.org/~ashwani

Appendix : Minimal Interference

1. History based minimal interference

   For any bounded execution ρ, that can be extended to a run that satisfies the liveness template, the probability of observing ρ in shielded execution remains close to its probability under the nominal policy.

2. Minimal shielding cost

   Given a cost function capturing the interference as the difference between the action distributions of the original and the shielded policies based on the cost of shielding at the arbitrary finite execution ρ, the cost of a run is limit average of costs for finite prefixes.