Dynamic Shielding for Reinforcement Learning in Black-Box Environments

Masaki Waga[1], Ezequiel Castellano[2], Sasinee Pruekprasert[3], Stefan Klikovits[2], Toru Takisaka[4], Ichiro Hasuo[2]

Kyoto University[1], National Institute of Informatics[2],
National Instutite of Advanced Industrial Science and Technology[3],
University of Electric Science and Technology of China[4]

Originally Presented at ATVA 2022 26th Oct. 2022

prevent unsafe exploration

Dynamic Shielding for Reinforcement Learning in Black-Box Environments

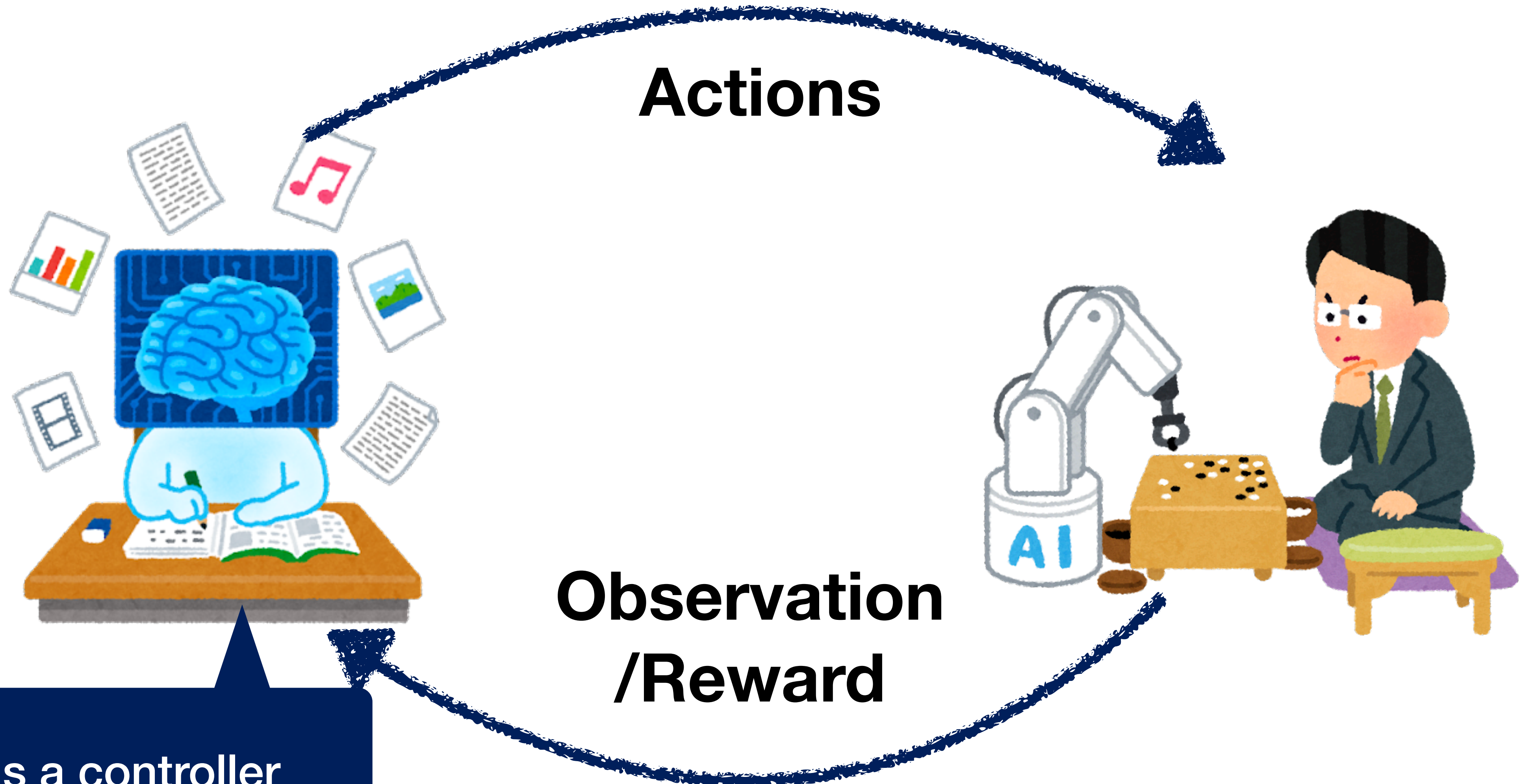
only for white-box env.
→ also for black-box env.

Masaki Waga[1], Ezequiel Castellano[2], Sasinee Pruekprasert[3], Stefan Klikovits[2], Toru Takisaka[4], Ichiro Hasuo[2]

Kyoto University[1], National Institute of Informatics[2],
National Instutite of Advanced Industrial Science and Technology[3],
University of Electric Science and Technology of China[4]

Originally Presented at ATVA 2022 26th Oct. 2022

Reinforcement Learning (RL)



Learns a controller
maximizing the reward
with trials-and-errors

Applications of RL

search.io
An Algolia Company

https://www.bbc.com/news/technology-35785875
https://chatbotslife.com/deep-learning-in-finance-learning-to-trade-with-q-rl-and-dqns-6c6cff4a1429

Applications of RL

search.io
An Algolia Company


BBC Sign in Home News Sport Reel Worklife

NEWS

Home | War in Ukraine | Coronavirus | Climate | Video | World | Asia | UK | Business | Tech | Science

Tech

Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol

12 March 2016

https://www.bbc.com/news/technology-35785875

https://chatbotslife.com/deep-learning-in-finance-learning-to-trade-with-q-rl-and-dqns-6c6cff4a1429

Applications of RL

search.io
An Algolia Company

BBC Sign in Home News Sport Reel Worklife

NEWS

Home | War in Ukraine | Coronavirus | Climate | Video | World | Asia | UK | Business | Tech | Science

Tech

Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol

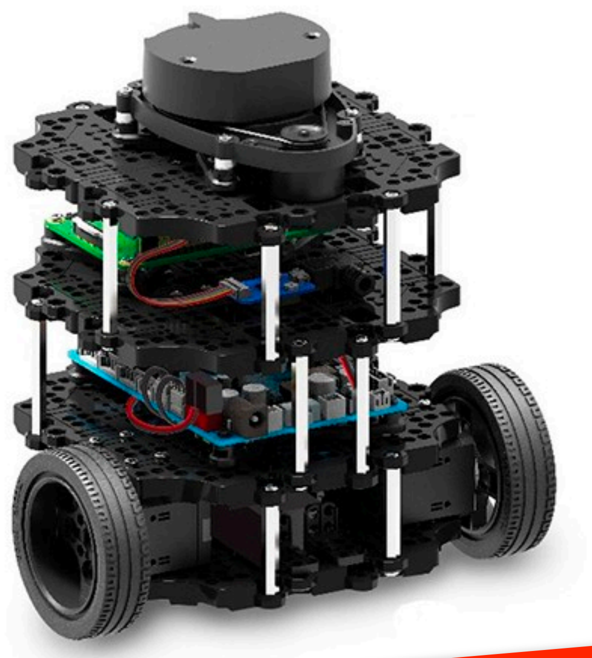
12 March 2016

https://www.bbc.com/news/technology-35785875

https://carla.org/2020/04/22/release-0.9.9/

https://chatbotslife.com/deep-learning-in-finance-learning-to-trade-with-q-rl-and-dqns-6c6cff4a1429

RL with Physical Env.



Undesirable actions may (eventually) break HW

https://www.roscomponents.com/1326-thickbox_default/turtlebot-3.jpg

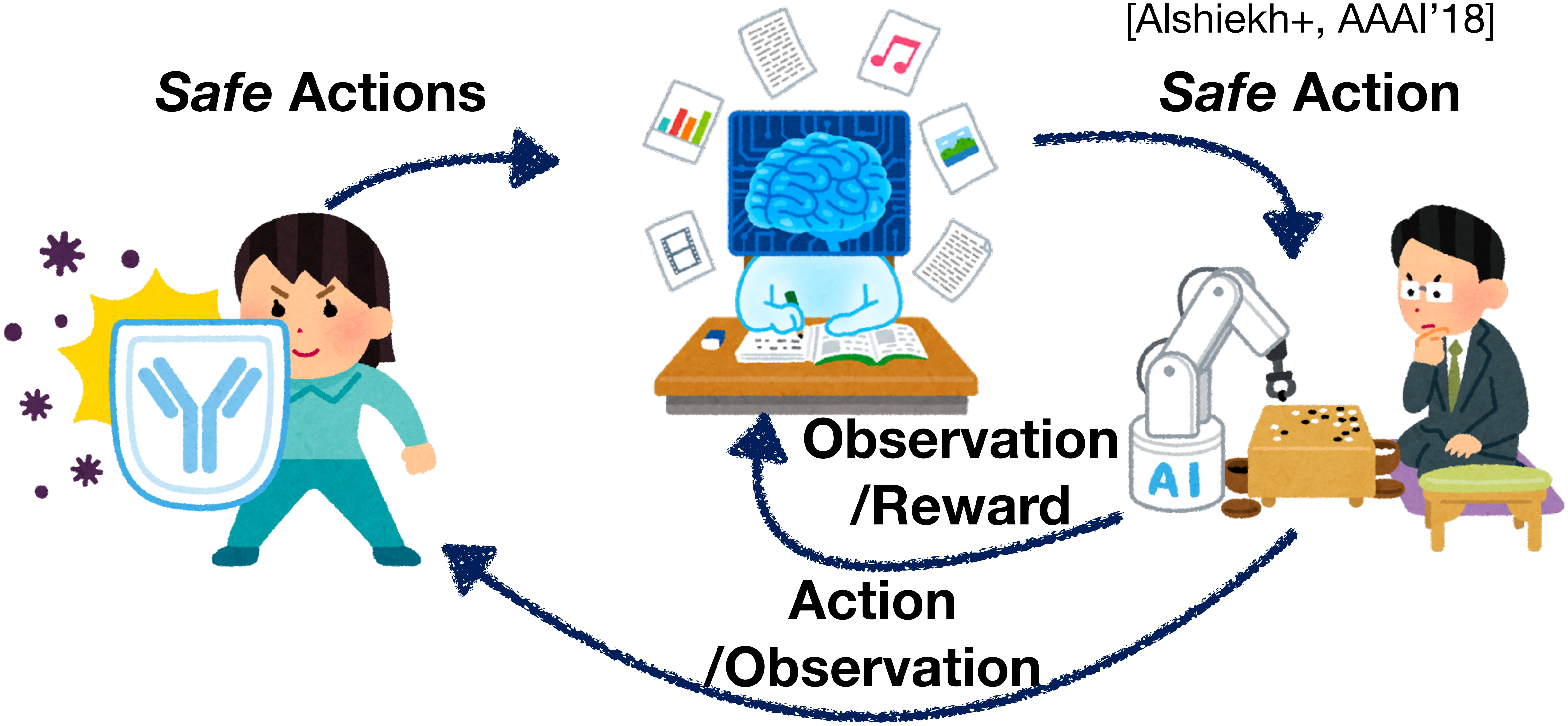
https://web.archive.org/web/20190417171518if_/http://emanual.robotis.com/assets/images/platform/turtlebot3/challenges/autorace_dankook_1.jpg

Q. Can we prevent undesired actions during training?

A. Yes if we have some prior knowledge of env.

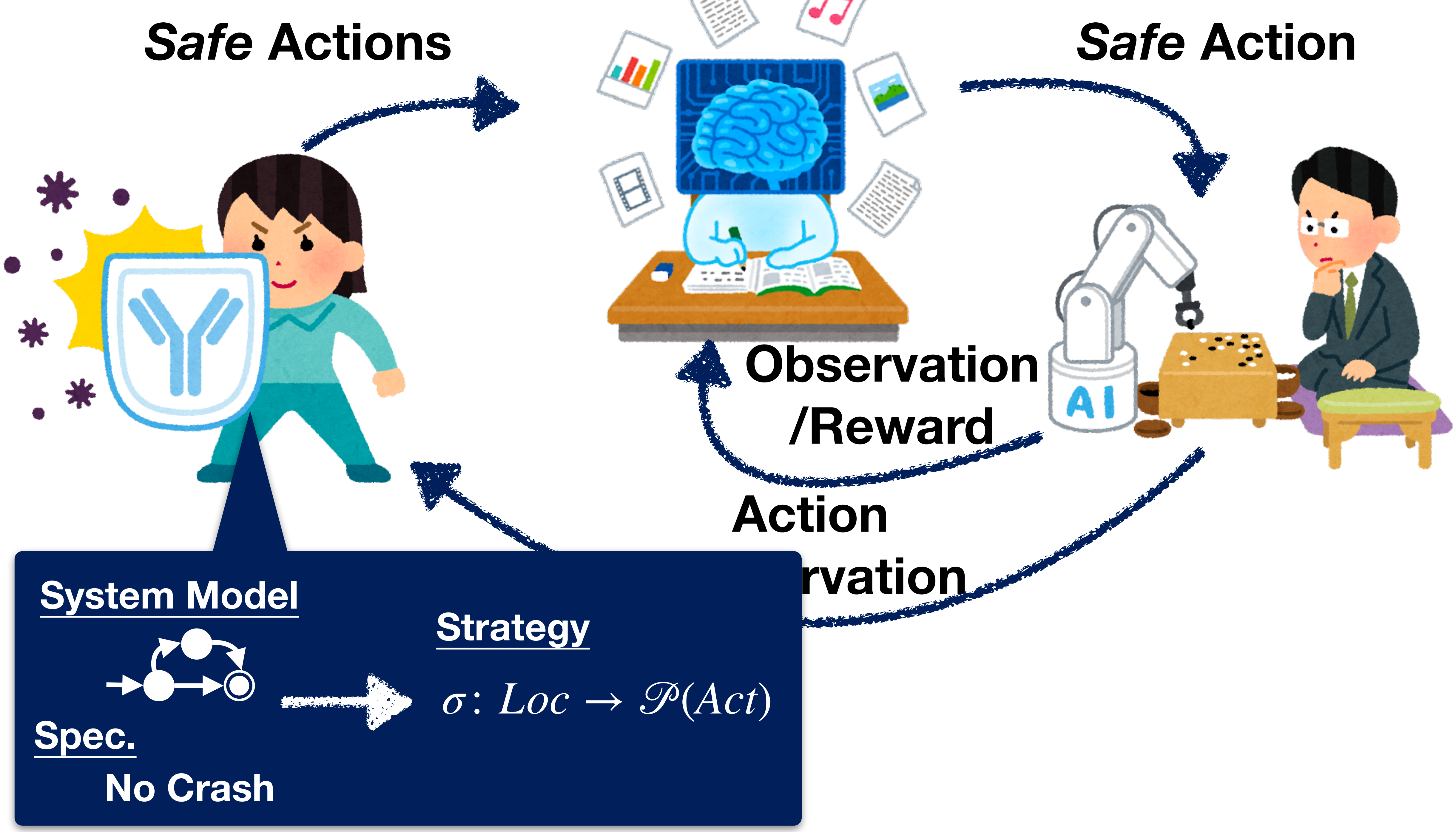
Safe RL with Shielding

[Alshiekh+, AAAI'18]



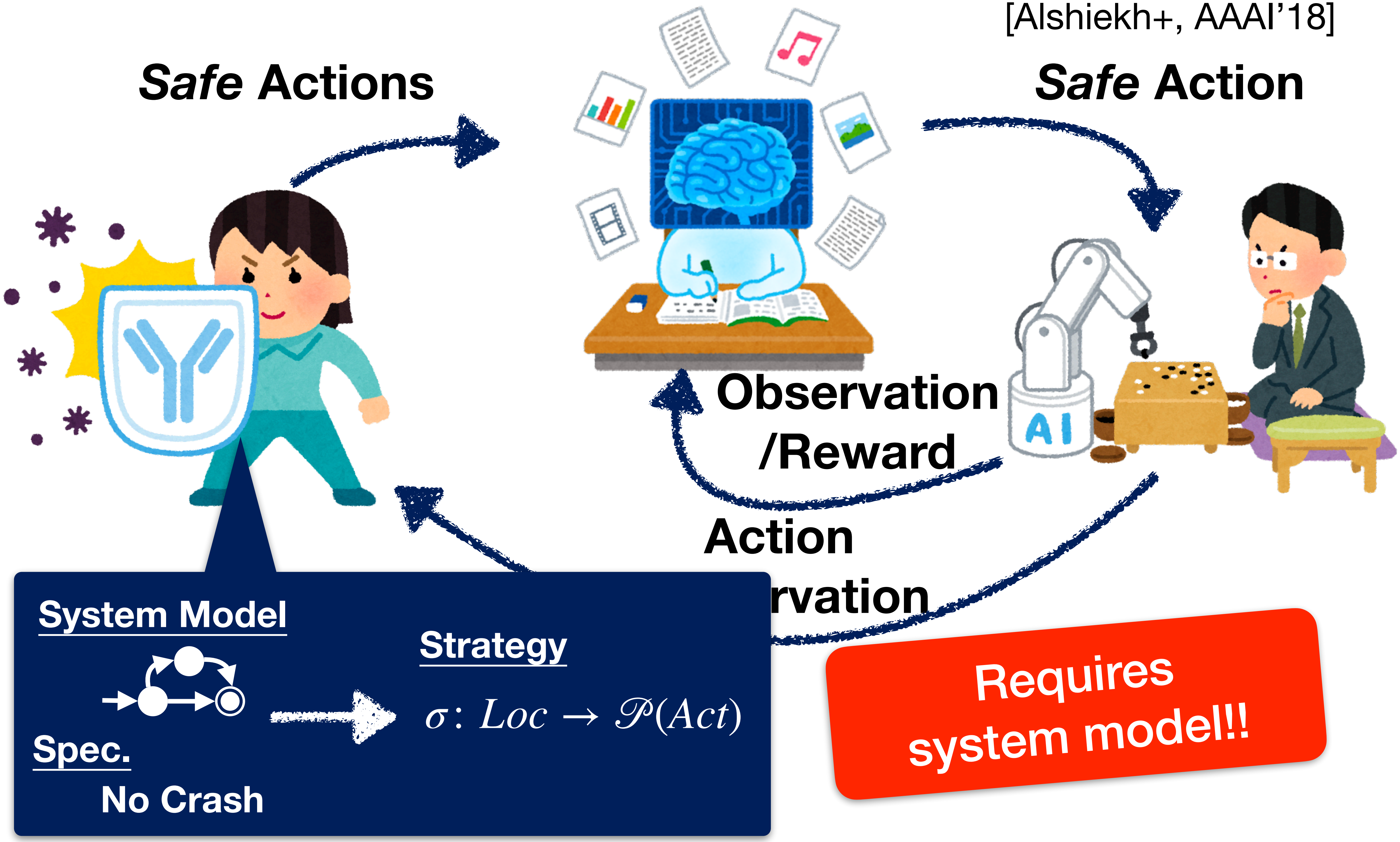
Safe RL with Shielding

[Alshiekh+, AAAI'18]



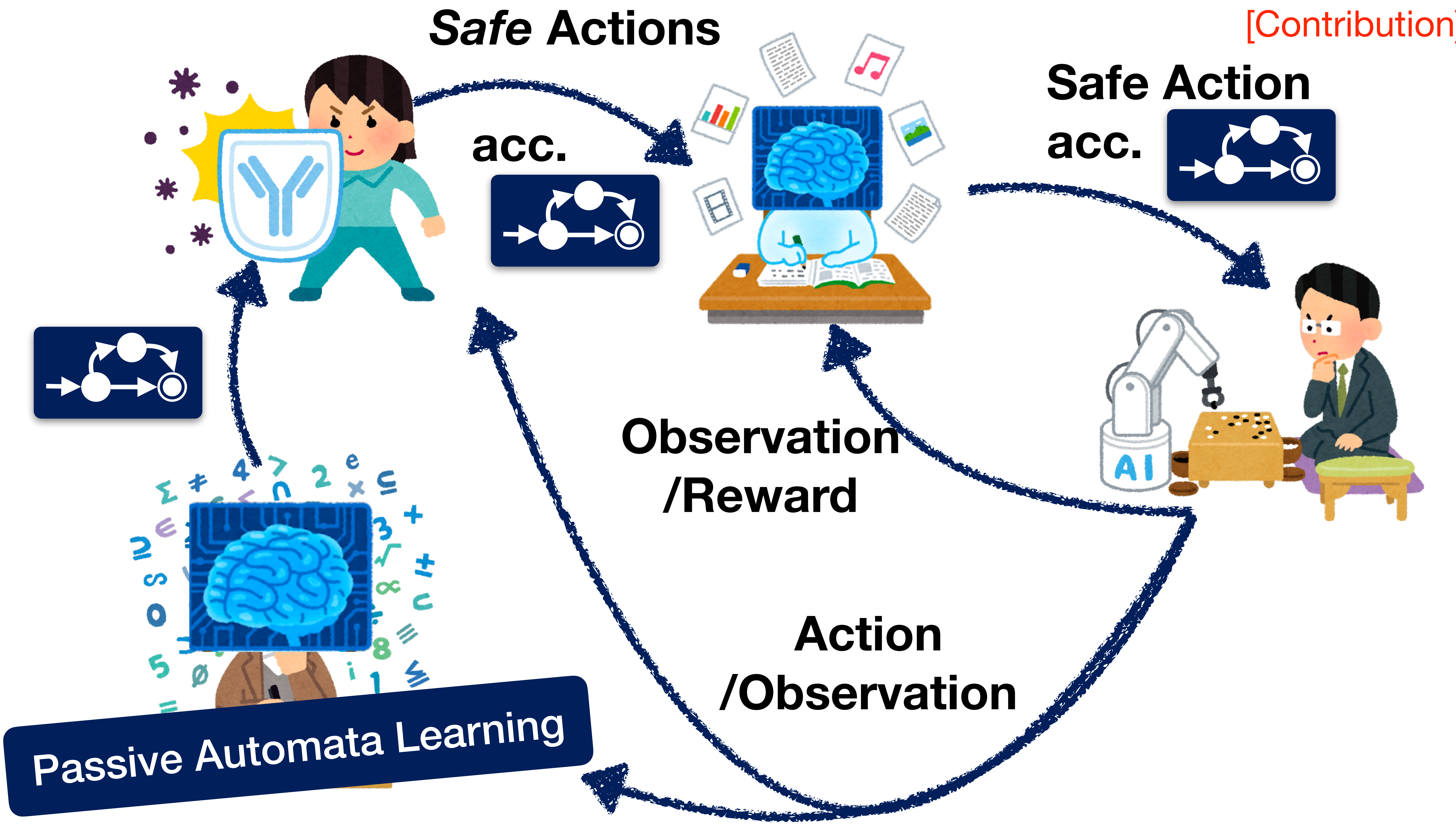
Safe RL with Shielding

[Alshiekh+, AAAI'18]



Q. Can we reduce undesired actions during training without prior system model?

Dynamic Shielding



Contributions

- Introduce the dynamic shielding scheme
  - Idea: passive automata learning + shielding
- Modified RPNI algorithm for passive autom. learning
  - to maintain necessary exploration
- Experiment results show that dynamic shielding reduces # of undesired actions during training

Outline

- Preliminaries
  - Static shielding
  - RPNI algorithm for passive automata learning
- Dynamic shielding + modification of RPNI algorithm
  - Idea 1: passive automata learning + shielding
  - Idea 2: additional requirements to deem two sequences are the same
- Experiments

(Preemptive) Shield

[Alshiekh+, 2018]

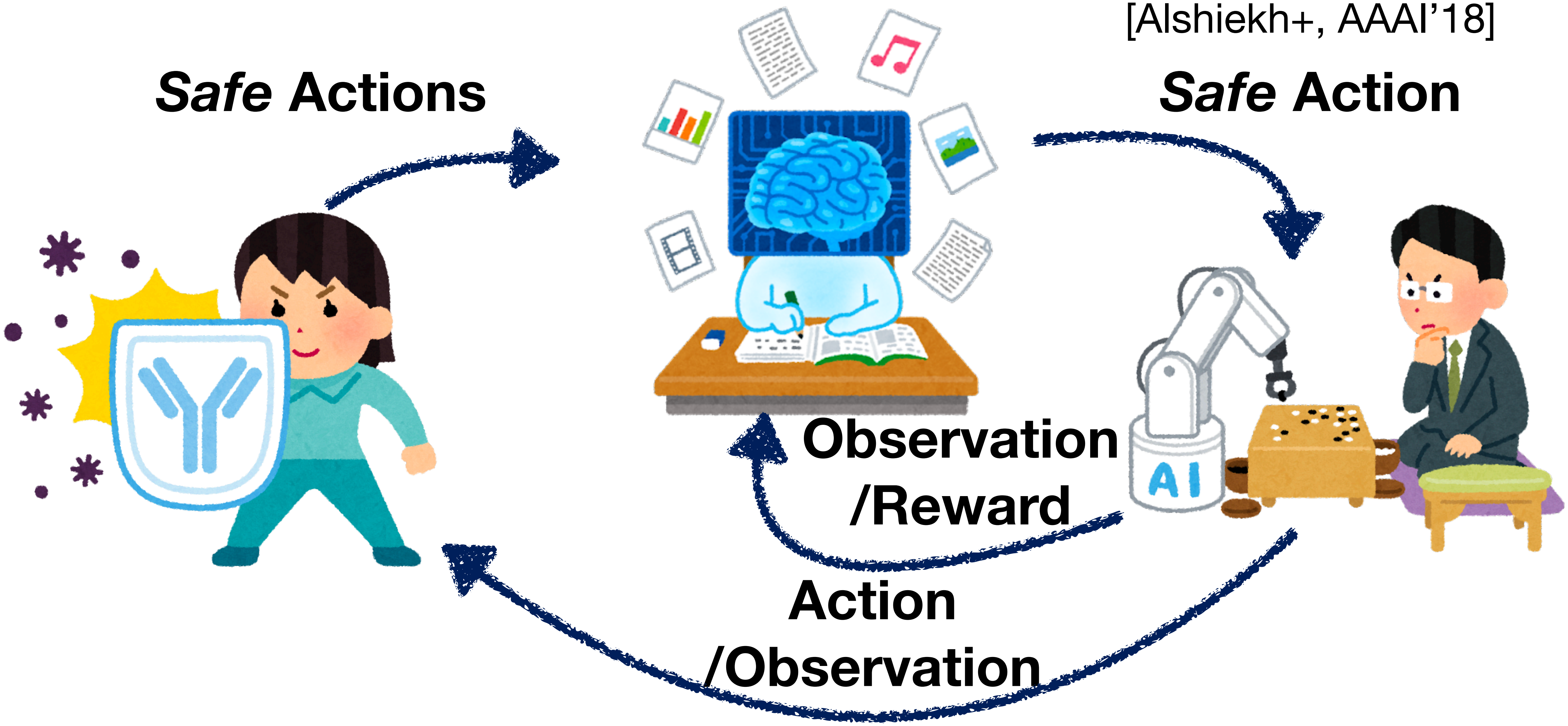
- Shield is stateful, i.e., Shield: $(Act \times Obs)^+ \to \mathscr{P}(Act)$
- We use a shield with finite state space
  → Mealy machine with input: $Act \times Obs$, output: $\mathscr{P}(Act)$

Shield Synthesis

1. Given: system model $\mathscr{M}$ and specification φ
   - $\mathscr{M}$: Mealy machine with 2 players
   - φ: safety LTL formula
2. Construct a safety game $\mathscr{G}$ by combining $\mathscr{M}$ and φ
3. Solve $\mathscr{G}$ to obtain the set of winning actions
   → Use it as the safe actions

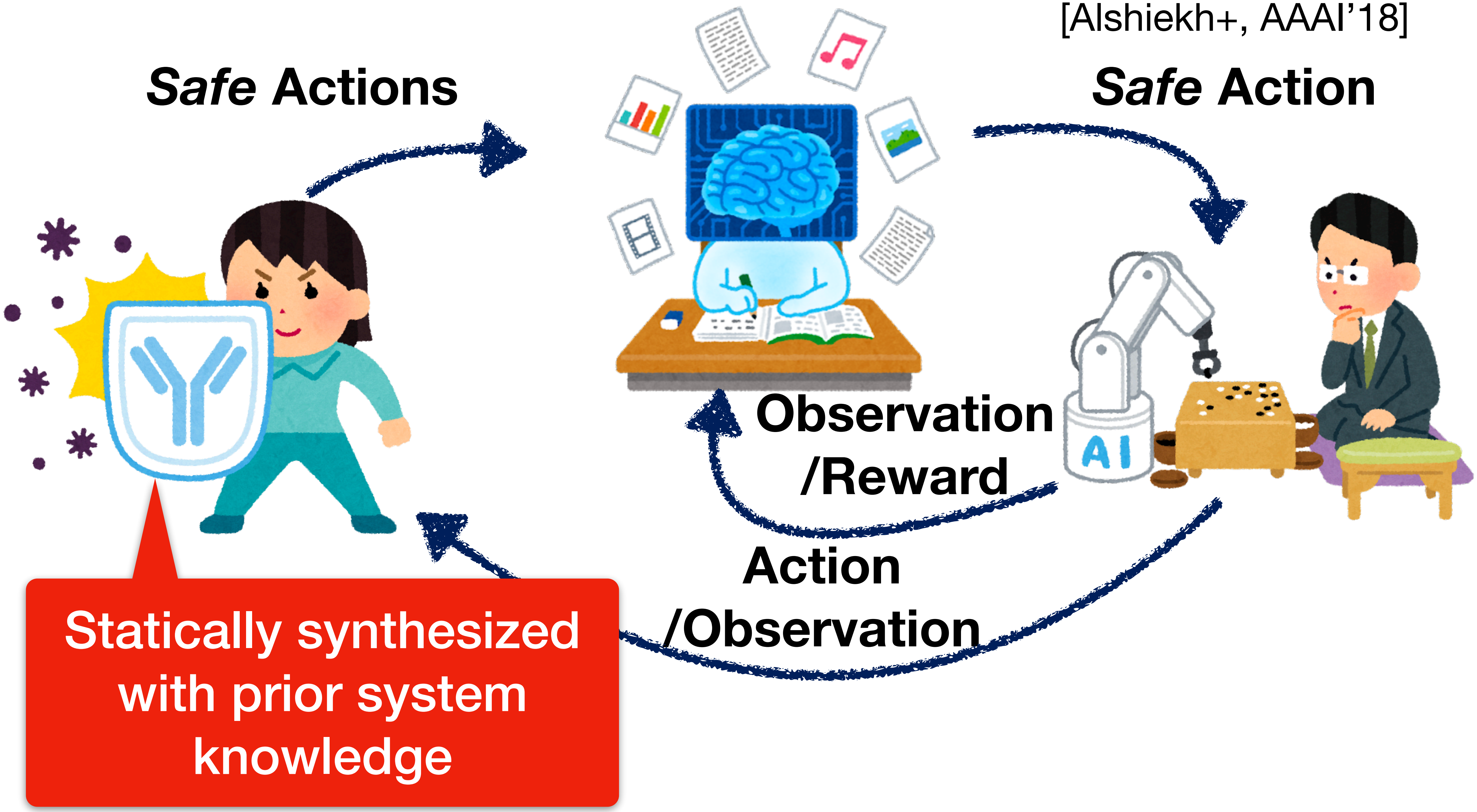
Safe RL with Shielding

[Alshiekh+, AAAI'18]



Safe RL with Shielding

[Alshiekh+, AAAI'18]



Outline

- Preliminaries
  - Static shielding
  - RPNI algorithm for passive automata learning
- Dynamic shielding + modification of RPNI algorithm
  - Idea 1: passive automata learning + shielding
  - Idea 2: additional requirements to deem two sequences are the same
- Experiments

Passive Automata Learning & RPNI-style Algorithm for Mealy machines

[Oncina & Garcia,1992]

Given: Set $T \subseteq Act^{+} \times Obs$ of words with labels (training data)

Learn: Mealy machine $\mathscr{M}$ compatible with T
i.e. $\forall (w, o) \in T \,.\, \mathscr{M}(w) = o$

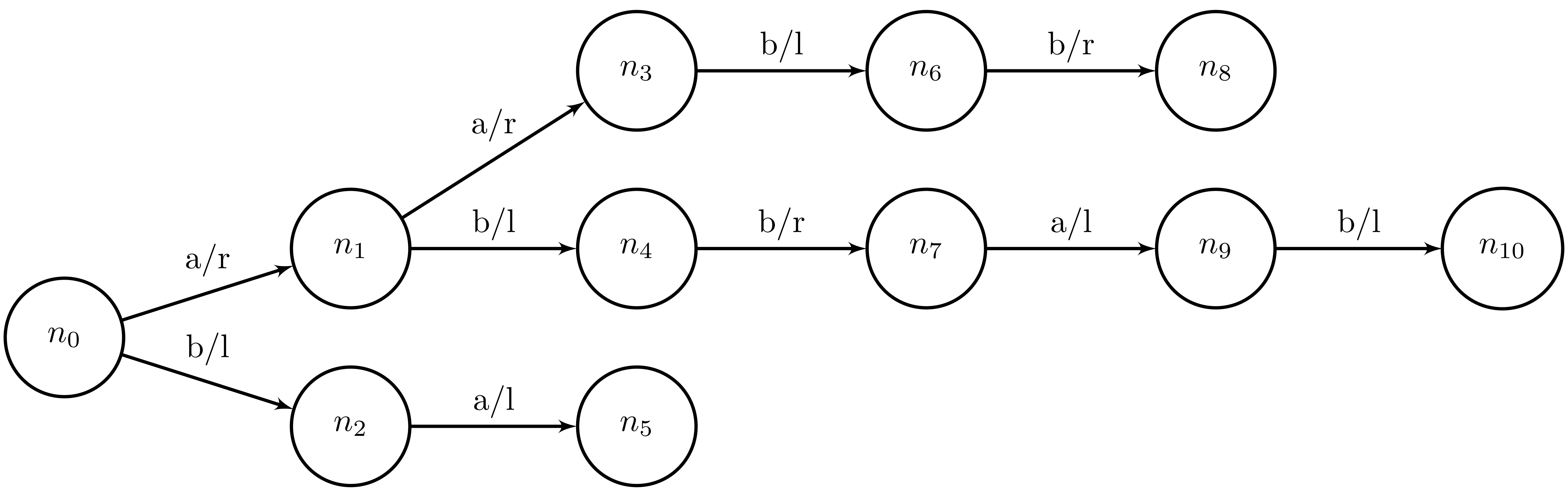
Idea:

1. Construct a prefix tree $\tilde{T}$ from T
2. Merge nodes of $\tilde{T}$ unless it makes nondeterministic branching

RPNI-style algorithm for Mealy machines

1. Construct a prefix tree $\tilde{T}$ from T

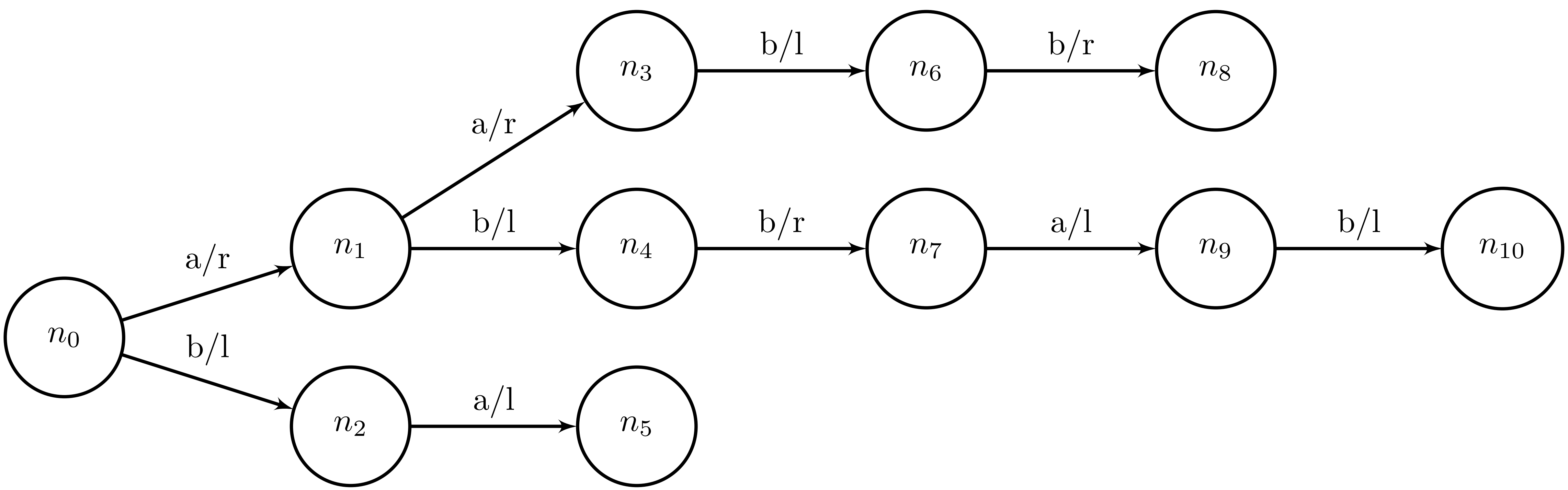
Initial prefix tree representing the training data



RPNI-style algorithm for Mealy machines

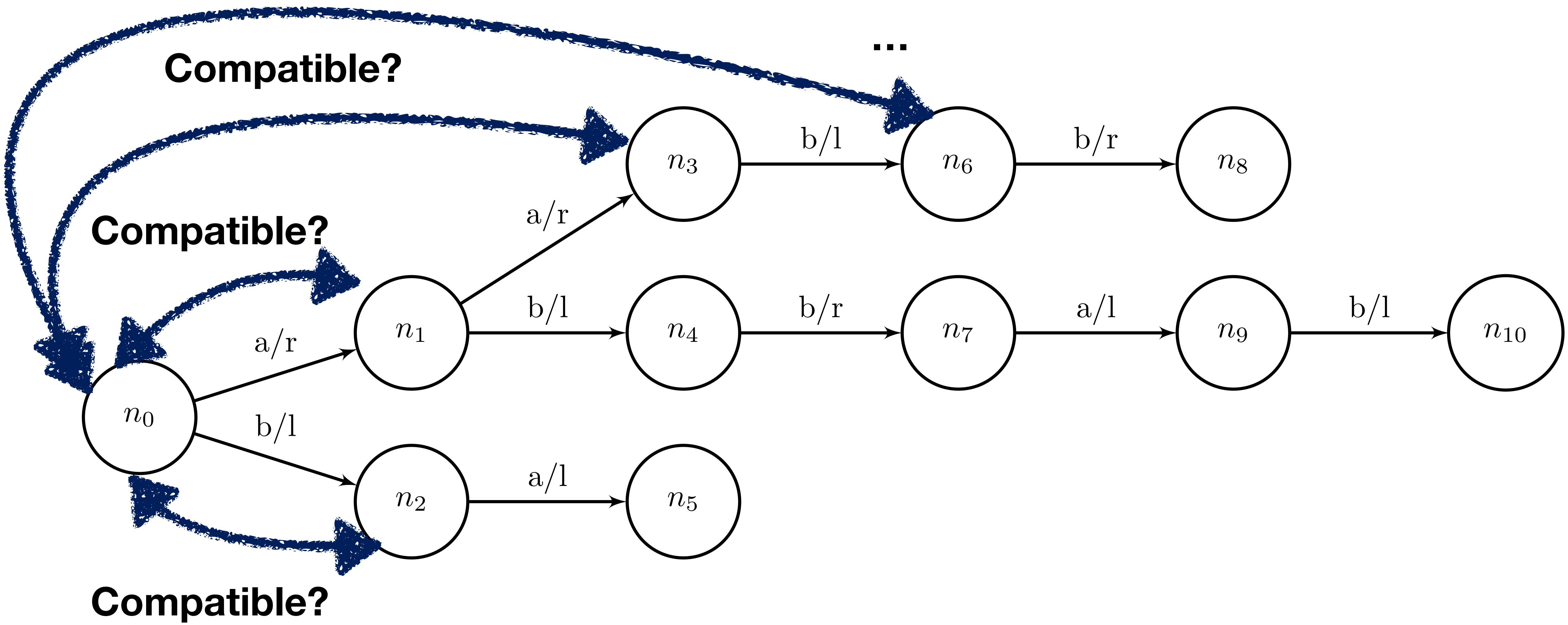
1. Construct a prefix tree $\tilde{T}$ from T

Initial prefix tree representing the training data



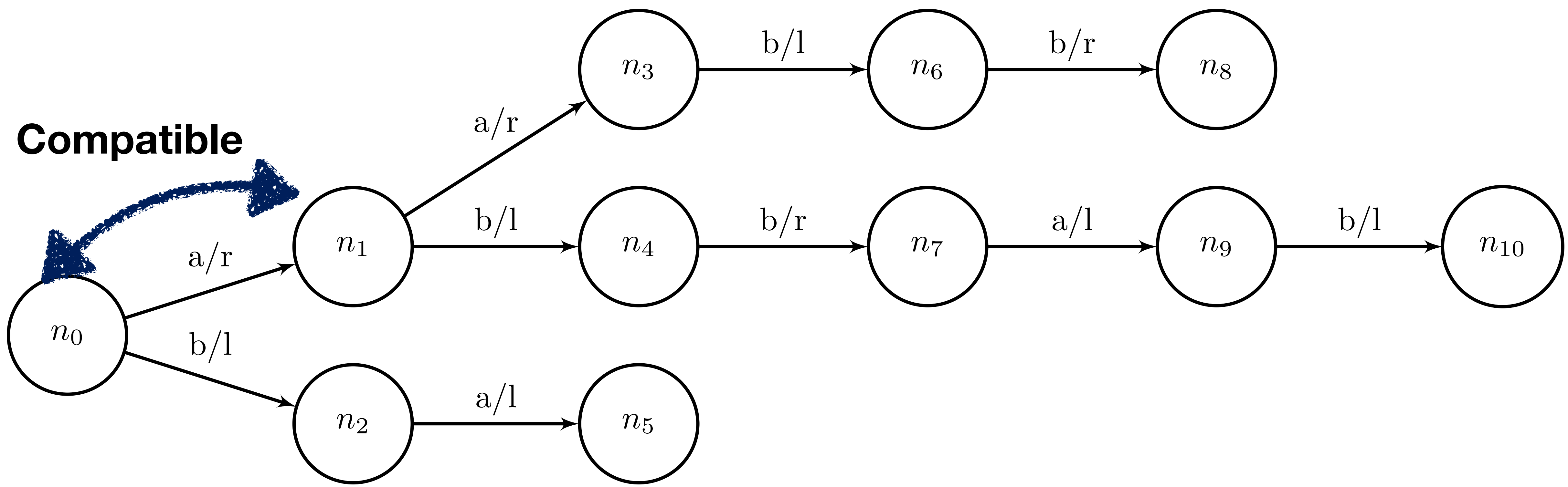
RPNI-style algorithm for Mealy machines

2. Merge nodes of $\tilde{T}$ unless it makes nondeterministic branching



RPNI-style algorithm for Mealy machines

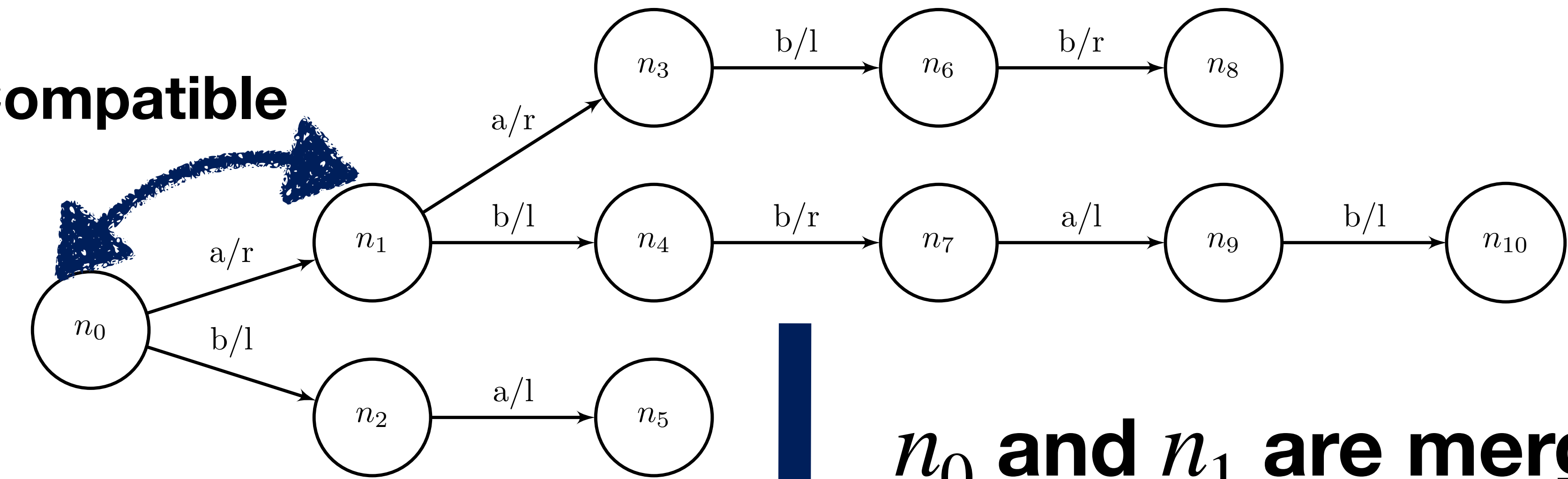
2. Merge nodes of $\tilde{T}$ unless it makes nondeterministic branching



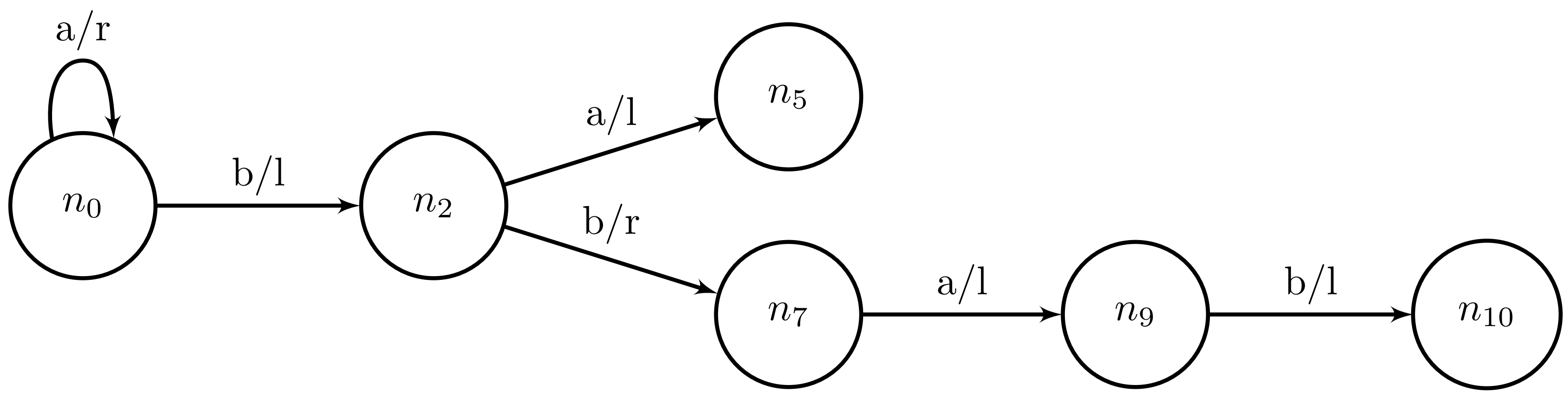
RPNI-style algorithm for Mealy machines

2. Merge nodes of $\tilde{T}$ unless it makes nondeterministic branching

Compatible

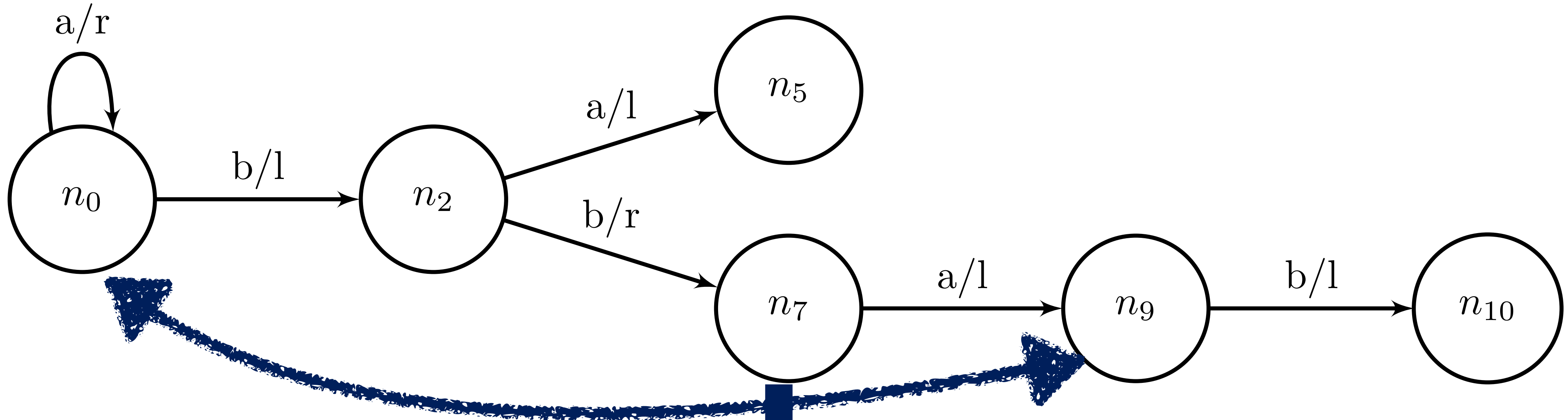


n_0 and n_1 are merged



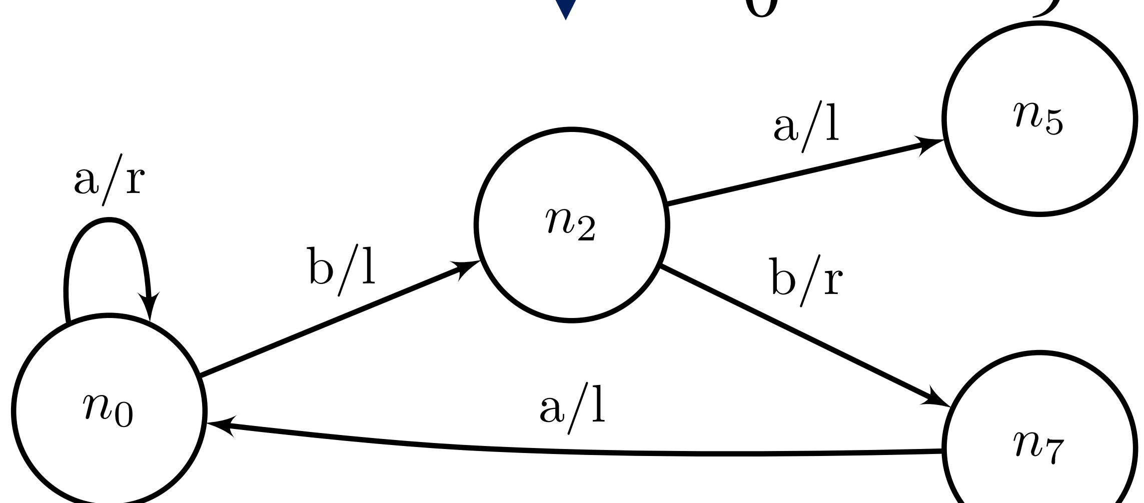
RPNI-style algorithm for Mealy machines

2. Merge nodes of $\tilde{T}$ unless it makes nondeterministic branching



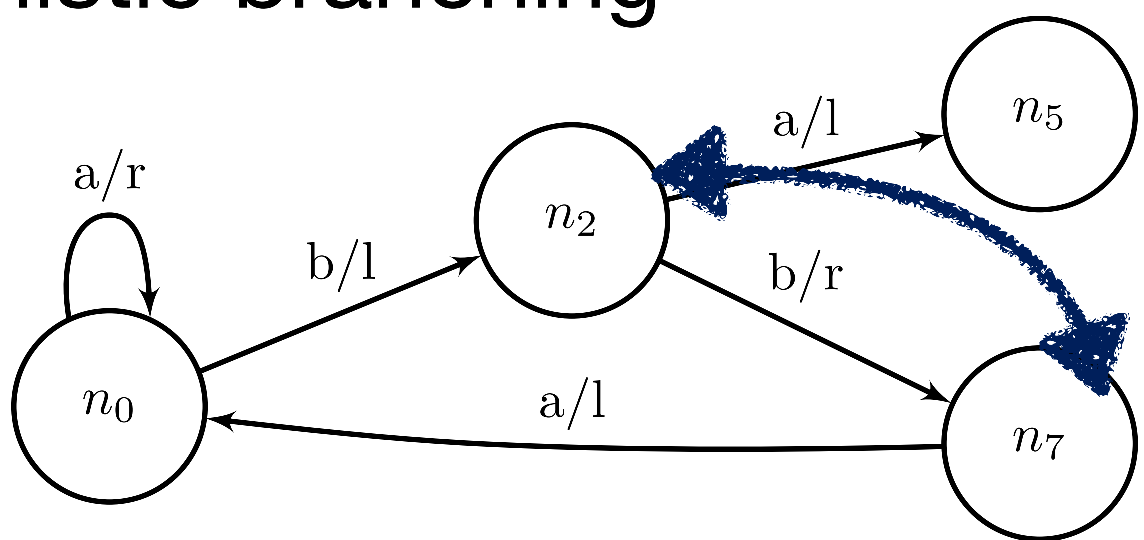
Compatible

n_0 **and** n_9 **are merged**



RPNI-style algorithm for Mealy machines

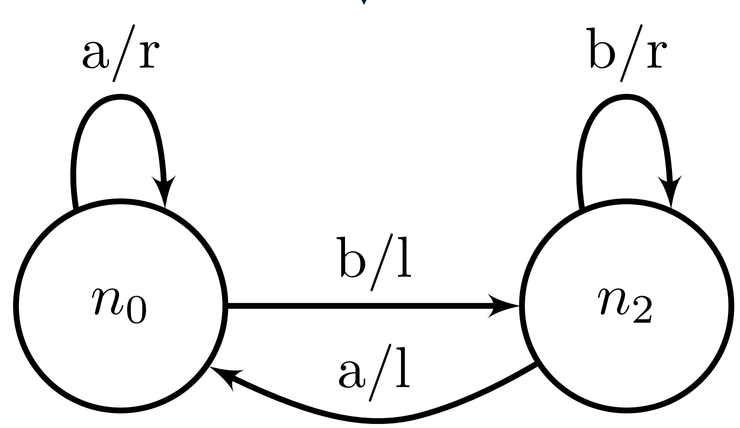
2. Merge nodes of $\tilde{T}$ unless it makes nondeterministic branching



Compatible

n_2 and n_7 are merged

Final result with no compatible nodes



Observation of the RPNI

- Learns a small Mealy machine by merging nodes
  - Generalization in machine learning
- No data → can be anything
  - Result can be largely different from the ground truth if the training data is small

Outline

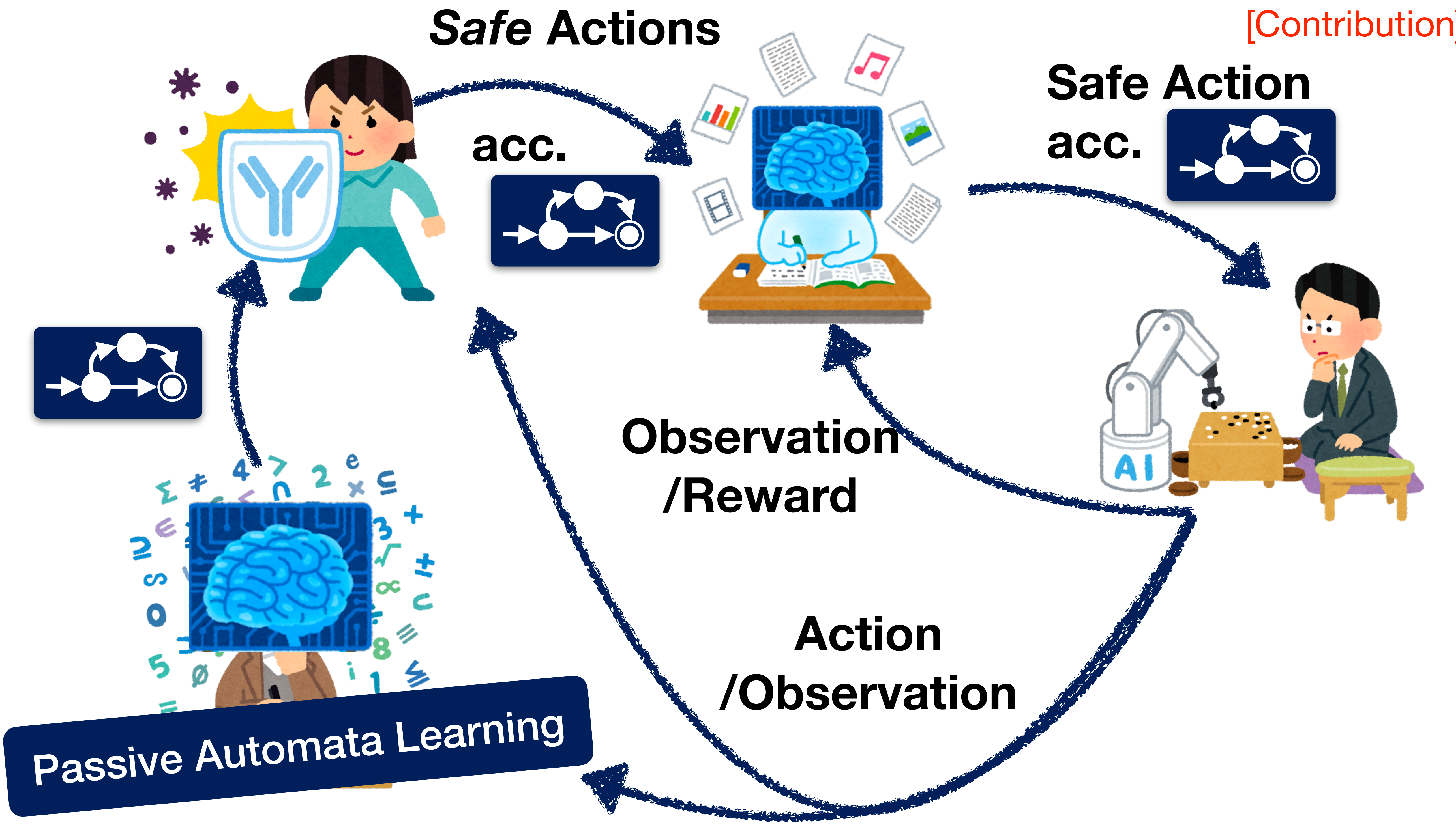
- Preliminaries
  - Static shielding
  - RPNI algorithm for passive automata learning
- Dynamic shielding + modification of RPNI algorithm
  - Idea 1: passive automata learning + shielding
  - Idea 2: additional requirements to deem two sequences are the same
- Experiments

Idea of Dynamic Shielding

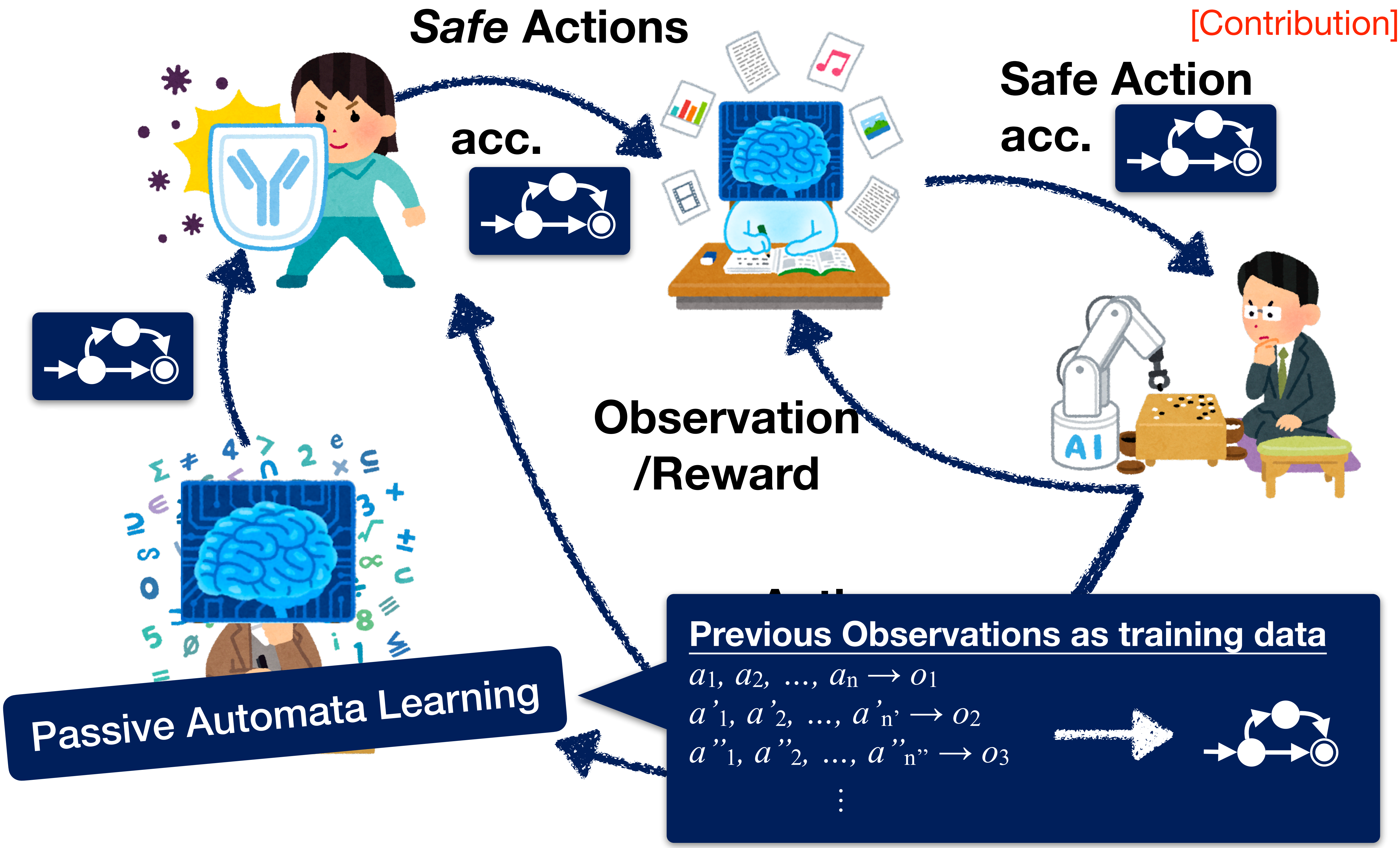
Explicitly learn the outcome of the actions
→ exploit it to avoid undesired behavior

- In the beginning, we know nothing
  → We cannot guarantee anything
- At a certain point, we know some of the unsafe actions
  → Use this information to prevent same mistake
  - By generalization, we also prevent similar mistakes

Dynamic Shielding



Dynamic Shielding



Difficulties in Dynamic Shielding

- Exploration is prevented if deemed to be unsafe
- At an early state, the training data is limited
  - Learned model is unreliable
- Learning algorithm should not merge nodes if the confidence of the similarity is low
  - Otherwise, necessary exploration may be prevented

RPNI algorithm with additional merging requirements [Contribution]

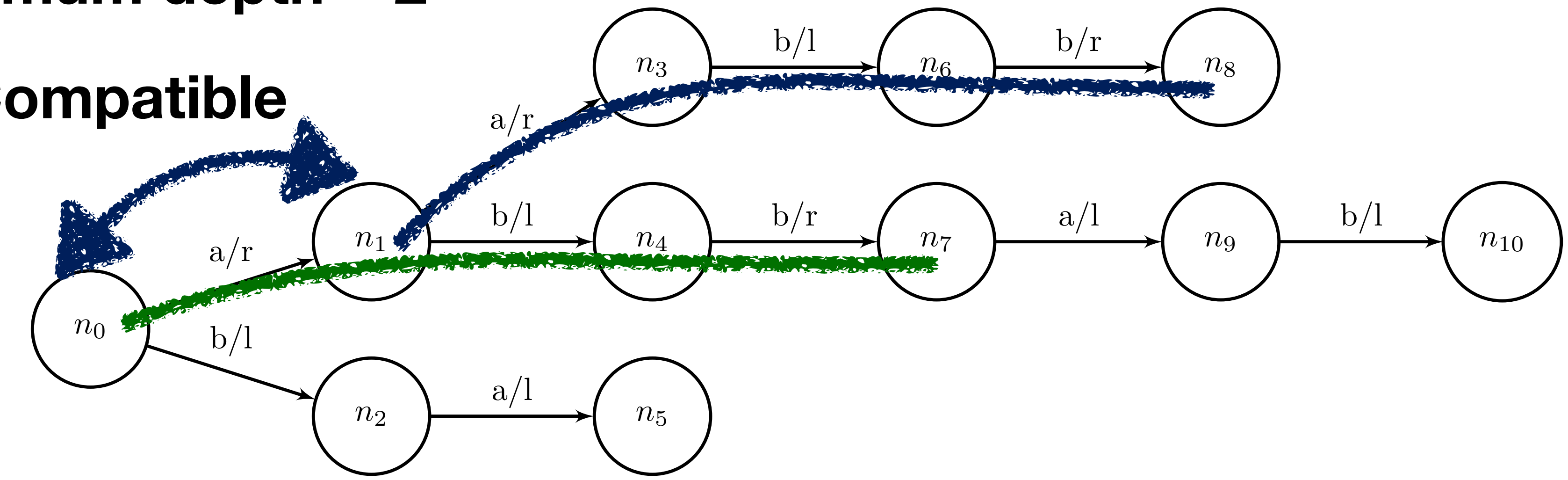
Idea: merge nodes only if we are confident enough

Evidence of the confidence:

- common children with enough depth

Example: minimum depth = 2

Compatible



Common children: “a/r, b/l, b/r” of depth 3
→ confident enough

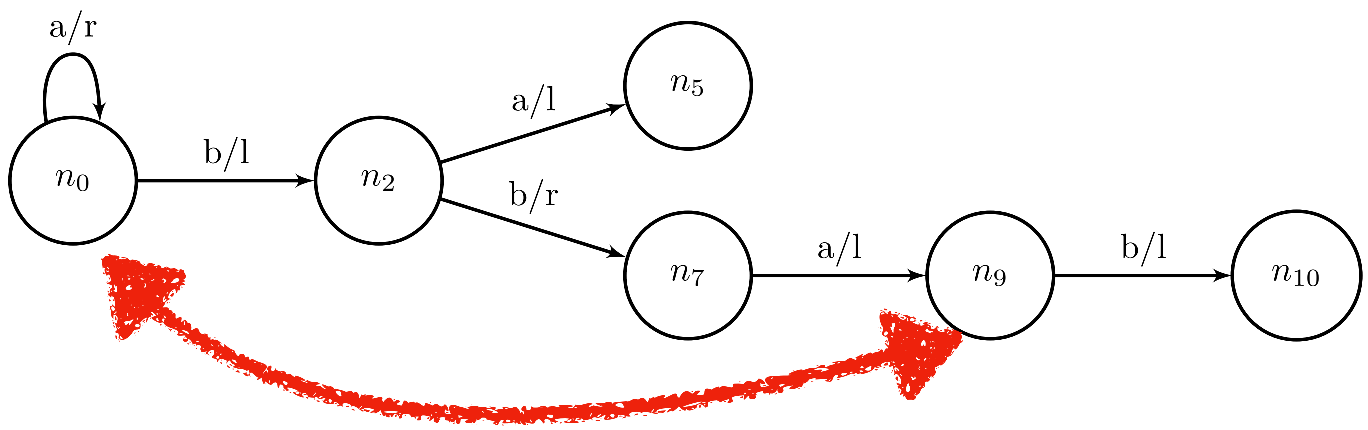
RPNI algorithm with additional merging requirements [Contribution]

Idea: merge nodes only if we are confident enough

Evidence of the confidence:

- common children with enough depth

Example: minimum depth = 2



Compatible but no common children with depth ≥ 2
→ we do not merge them!

Heuristics to adaptively decide minimum depth

[Contribution]

- Merging should be less greedy in the beginning because:
  - the training data is small
  - we want varletry exploration
- Adaptively decide the minimum depth based on the episode length
- Concretely: $\left\lceil \frac{ep_{\max} - \sum_{i=0}^{N} |ep_i|/N}{\sum_{i=0}^{N} |ep_i|/N} \right\rceil$, with $ep_{\max}$: maximum episode length

  $\sum_{i=0}^{N} |ep_i|/N$: average episode length

Outline

- Preliminaries
  - Static shielding
  - RPNI algorithm for passive automata learning
- Dynamic shielding + modification of RPNI algorithm
  - Idea 1: passive automata learning + shielding
  - Idea 2: additional requirements to deem two sequences are the same
- Experiments

Setting of Experiments

- Implemented dynamic shielding with Python3 and Java
- Used 7 benchmarks mostly from the literature
  - discrete, continuous ($[-1,1]^4$), and image observation
- Baselines:
  - RL with no safety mechanism (Plain)
  - RL with safe padding (SafePadding)
    - A shielding-style method for black-box setting
    - No generalization by state merging
    - Different construction
- AMD EPYC 7702P, NVIDIA GeForce RTX 2080Ti, 125GiB RAM

RQ 1. Safety by Dynamic Shielding

Mean # of training episodes with undesired behaviors

	Plain	SafePadding	Dynamic Shielding (Ours)
WaterTank	1883.67	1892.4	**177.13**
GridWorld	6996.4	7322.23	**5623.43**
CliffWalk	1493.2	1528.67	**478.20**
Taxi	8723.13	2057.33	**37.77**
SelfDrivingCar	6403.07	6454.6	**5662.4**
SideWalk	373.6	427.93	**273.37**
CarRacing	180.13	141.17	**41.73**

RQ 1. Safety by Dynamic Shielding

Mean # of training episodes with undesired behaviors

	Plain	SafePadding	Dynamic Shielding (Ours)
WaterTank	1883.67	1892.4	**177.13**
GridWorld	6996.4	7322.23	**5623.43**
CliffWalk	1493.2	1528.67	**478.20**
Taxi	8723.13	2057.33	**37.77**
SelfDrivingCar	6403.07	6454.6	**5662.4**
SideWalk	373.6	427.93	**273.37**
CarRacing	180.13	141.17	**41.73**

RQ 1. Safety by Dynamic Shielding

Mean # of training episodes with undesired behaviors

	Plain	SafePadding	Dynamic Shielding (Ours)
WaterTank	1883.67	1892.4	**177.13**
GridWorld	6996.4	7322.23	**5623.43**
CliffWalk	1493.2	1[illegible]	**478.20**
Taxi	8723.13	2057.33	**37.77**
SelfDrivingCar	6403.07	6454.6	**5662.4**
SideWalk	373.6	[illegible]	**273.37**
CarRacing	180.13	141.17	**41.73**

≈ 0.4% of Plain

≈ 23% of Plain

RQ 2. Controller's Performance

Mean reward of the resulting controller in the testing phase

	Plain	SafePadding	Dynamic Shielding (Ours)
WaterTank	918.89	919.81	**921.81**
GridWorld	0.37	**0.46**	0.07
CliffWalk	-69.13	-66.00	**-65.93**
Taxi	-147.61	-139.62	**-92.93**
SelfDrivingCar	28.83	28.86	**29.81**
SideWalk	**0.93**	0.90	0.67
CarRacing	375.53	509.25	**622.07**

RQ 2. Controller's Performance

Mean reward of the resulting controller in the testing phase

	Plain	SafePadding	Dynamic Shielding (Ours)
WaterTank	918.89	919.81	**921.81**
GridWorld	0.37	**0.46**	0.07
CliffWalk	-69.13	-66.00	**-65.93**
Taxi	-147.61	-139.62	**-92.93**
SelfDrivingCar	28.83	28.86	**29.81**
SideWalk	**0.93**	0.90	0.67
CarRacing	375.53	509.25	**622.07**

≈ 166% of Plain

RQ 2. Controller’s Performance

Mean reward of the resulting controller in the testing phase

	Plain	SafePadding	Dynamic Shielding (Ours)
WaterTank	918.89	919.81	**921.81**
GridWorld	0.37	**0.46**	0.07
CliffWalk	-69.13	[illegible]	**-65.93**
Taxi	-147.61	-139.62	**-92.93**
SelfDrivingCar	28.83	28.86	**29.81**
SideWalk	**0.93**	0.90	0.67
CarRacing	375.53	509.25	**622.07**

Significantly worse

≈ 166% of Plain

RQ 3. Overhead of Dynamic Shielding

Mean exec. time [min] of the whole RL process

	Plain	SafePadding	Dynamic Shielding (Ours)
WaterTank	**31.01**	32.45	101.35
GridWorld	**2.95**	24.79	75.81
CliffWalk	**5.92**	6.09	13.98
Taxi	**5.601**	5.83	10.2
SelfDrivingCar	**14.43**	81.99	168.12
SideWalk	**12.71**	28.91	106.6
CarRacing	**127.5**	278.24	208.87

RQ 3. Overhead of Dynamic Shielding

Mean exec. time [min] of the whole RL process

	Plain	SafePadding	Dynamic Shielding (Ours)
WaterTank	**31.01**	32.45	101.35
GridWorld	**2.95**	24.79	75.81
CliffWalk	**5.92**	6.09	13.98
Taxi	**5.601**	5.83	10.2
SelfDrivingCar	**14.43**	81.99	168.12
SideWalk	**12.71**	28.91	106.6
CarRacing	**127.5**	278.24	208.87

RQ 3. Overhead of Dynamic Shielding

Significantly slower (≈ + 1-2 hours)

Mean exec. time [min] of the whole RL process

	Plain	SafePadding	Dynamic Shielding (Ours)
WaterTank	**31.01**	32.45	101.35
GridWorld	**2.95**	24.79	75.81
CliffWalk	**5.92**	6.09	13.98
Taxi	**5.601**	5.83	10.2
SelfDrivingCar	**14.43**	81.99	168.12
SideWalk	**12.71**	28.91	106.6
CarRacing	**127.5**	278.24	208.87

RQ 3. Overhead of Dynamic Shielding

Significantly slower
(≈ + 1-2 hours)

Mean exec. time [min] of the whole RL process

	Plain	SafePadding	Dynamic Shielding (Ours)
WaterTank	**31.01**	32.45	101.35
GridWorld	**2.95**	[illegible]	75.81
CliffWalk	**5.92**	6.09	13.98
Taxi	**5.601**	[illegible]	10.2
SelfDrivingCar	**14.43**	[illegible]	168.12
SideWalk	**12.71**	[illegible]	106.6
CarRacing	**127.5**	278.24	208.87

≈ +70 min. of Plain

≈ +94 min. of Plain

≈ +81 min. of Plain

Conclusions & Future works

- Improve the safety of exploration in RL with black-box env.
  - Idea: passive automata learning + shielding
- Undesired behaviors were significantly prevented
  - Note: $\gg 0$ but (hopefully) still useful for some usage
- Current limitation: Leaned system model is deterministic
  → Future work: Extension for stochastic models

Appendix

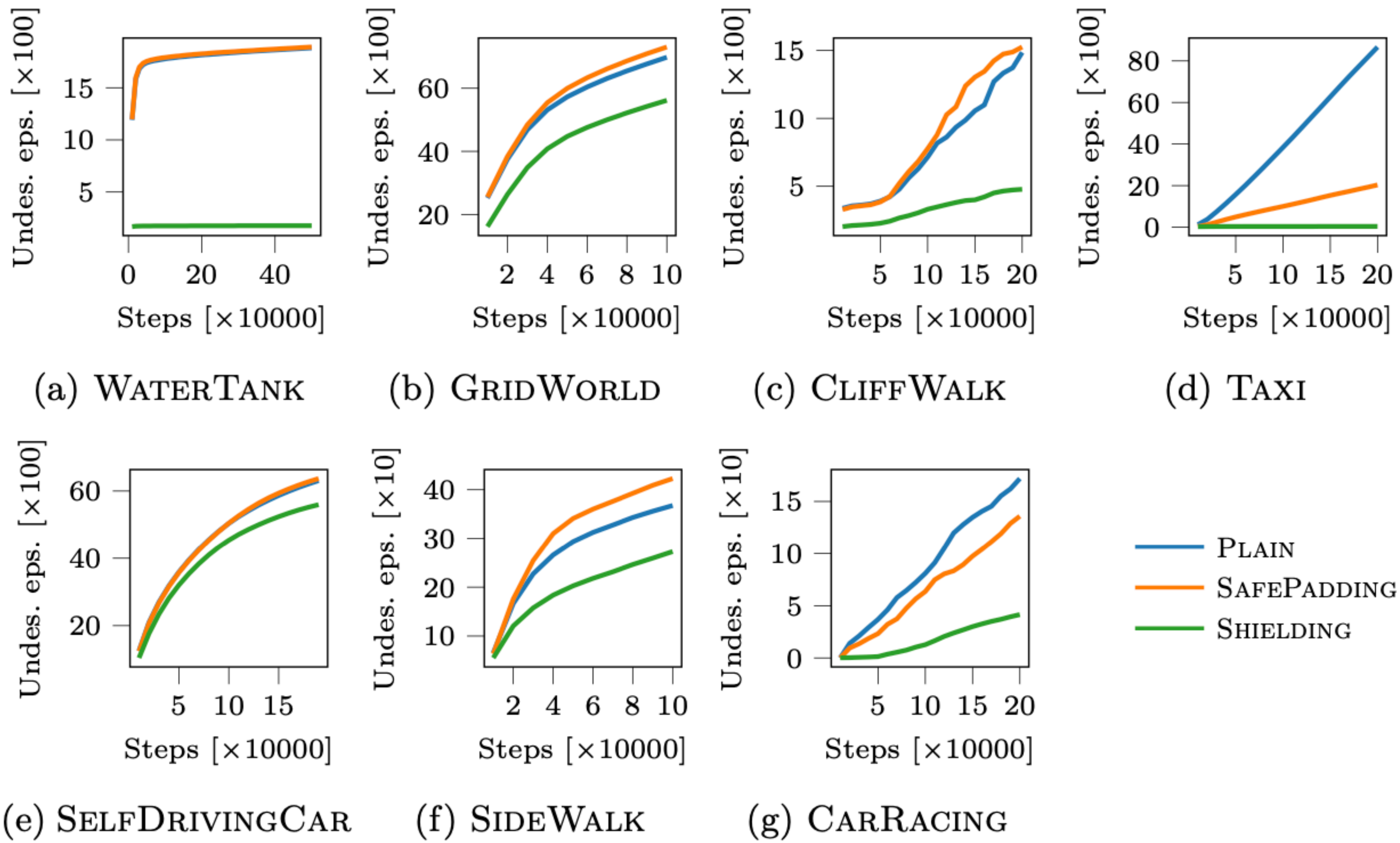
Detail of our Implementation

- Implemented in Python3 and Java
- Used libraries (only major ones):
  - Stable Baselines 3 or Keras-RL (in Python3): for RL
  - LearnLib (in Java): for the RPNI algorithm
    - Our modification of the RPNI algorithm is also in Java
  - Bridging between Python3 and Java: py4j
- Available at: https://doi.org/10.5281/zenodo.6906673

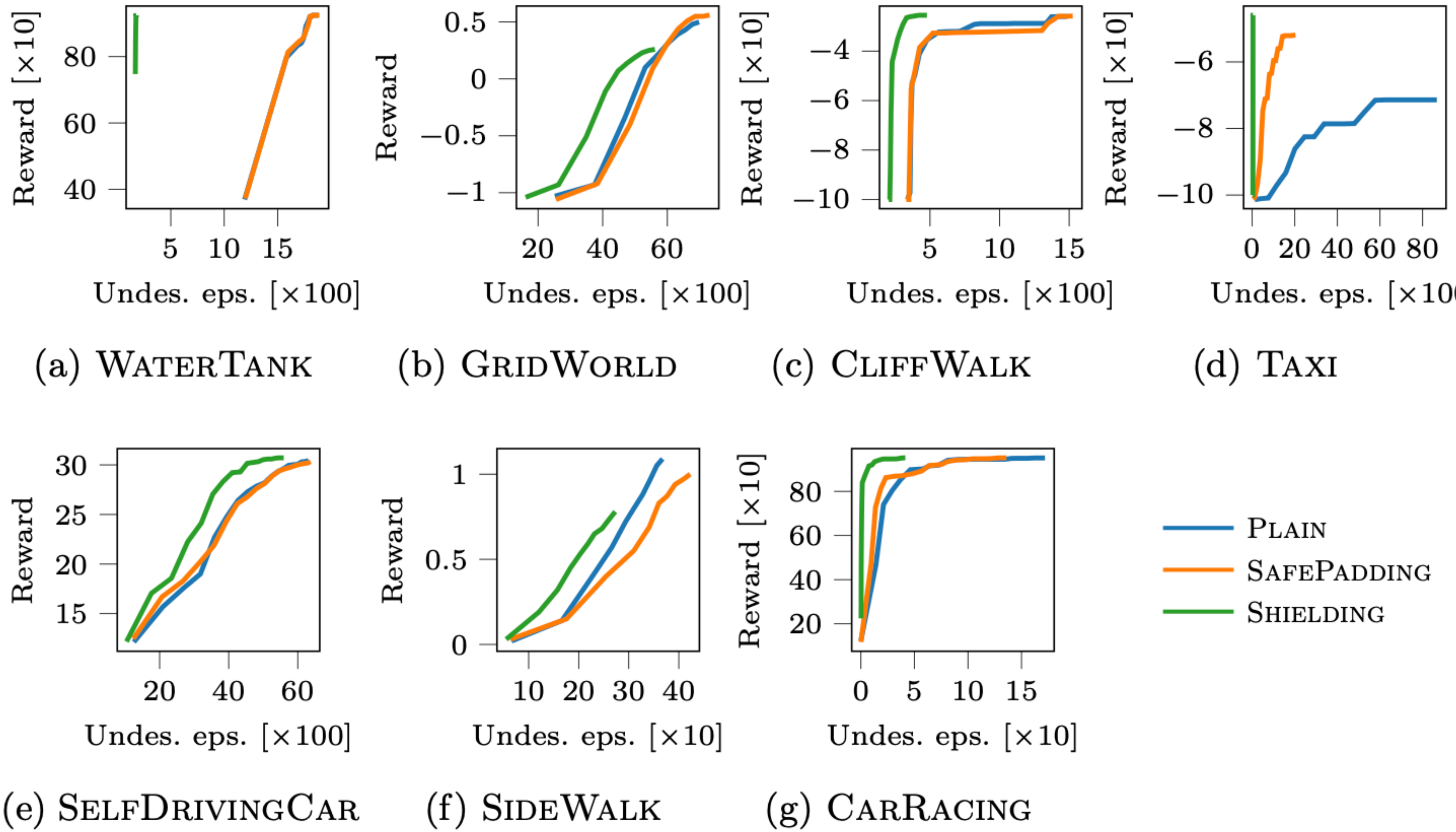
List of the Benchmarks

	Benchmark's origin	Observation space (size)	Network	Learning algorithm	# of steps
WaterTank	Alshiekh et al. [1]	Discrete (714)	MLP	PPO	500,000
GridWorld	Our original	Discrete (625)	MLP	PPO	100,000
Taxi	OpenAI Gym [7]	Discrete (500)	MLP	PPO	200,000
CliffWalk	OpenAI Gym [7]	Discrete (48)	MLP	PPO	200,000
SelfDrivingCar	Alshiekh et al. [1]	Continuous ($[-1, 1]^4$)	MLP	DQN	200,000
SideWalk	MiniWorld [9]	Image ($80 \times 60 \times 3 \times 256$)	CNN	PPO	100,000
CarRacing	OpenAI Gym [7]	Image ($96 \times 96 \times 3 \times 256$)	CNN	PPO	200,000

Other Experiment Results (Safety)



Other Experiment Results (Performance)



(a) WATERTANK (b) GRIDWORLD (c) CLIFFWALK (d) TAXI

(e) SELFDRIVINGCAR (f) SIDEWALK (g) CARRACING

Example: Limited Exploration due to Wrong Merging

Simple Grid World (A: agent; G: goal; X: Wall, should not hit)

```
XXXGXXX
XX   XX
XX X XX
   A
XXXXXXX
```

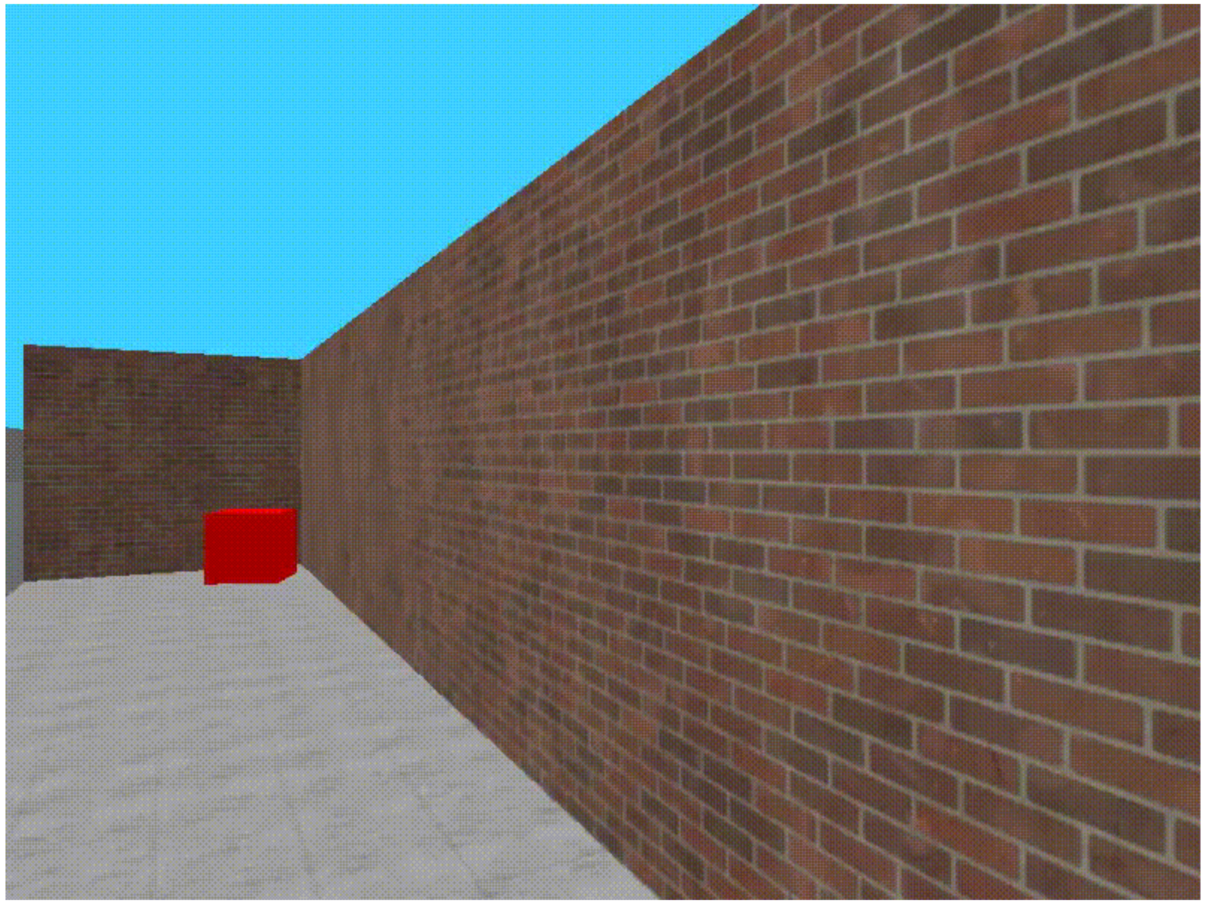
Training Data

```
→ ↑ ← (crash)      ← ↑ ← (crash)
→ ↑ → (crash)      ← ↑ → (crash)
→ ↑ ↑ ↑ (crash)    ← ↑ ↑ ↑ (crash)
→ ↑ ↑ → (crash)
                   ← ↑ ↑ ← (crash)
```

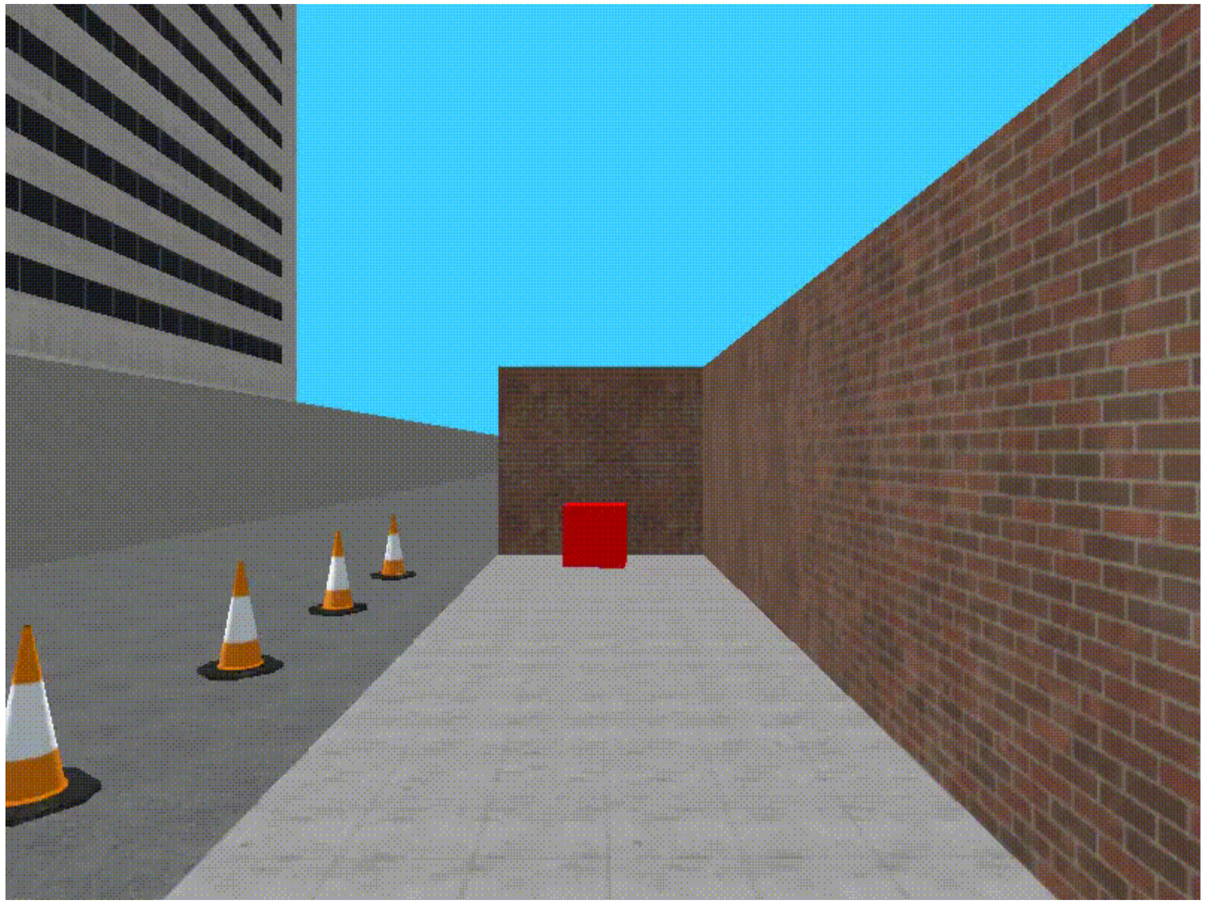
Outcome of “→ ↑” and “← ↑” seems the same from the training data
“→ ↑ ↑ ←” is deemed to be unsafe

Benchmark: Sidewalk

Success

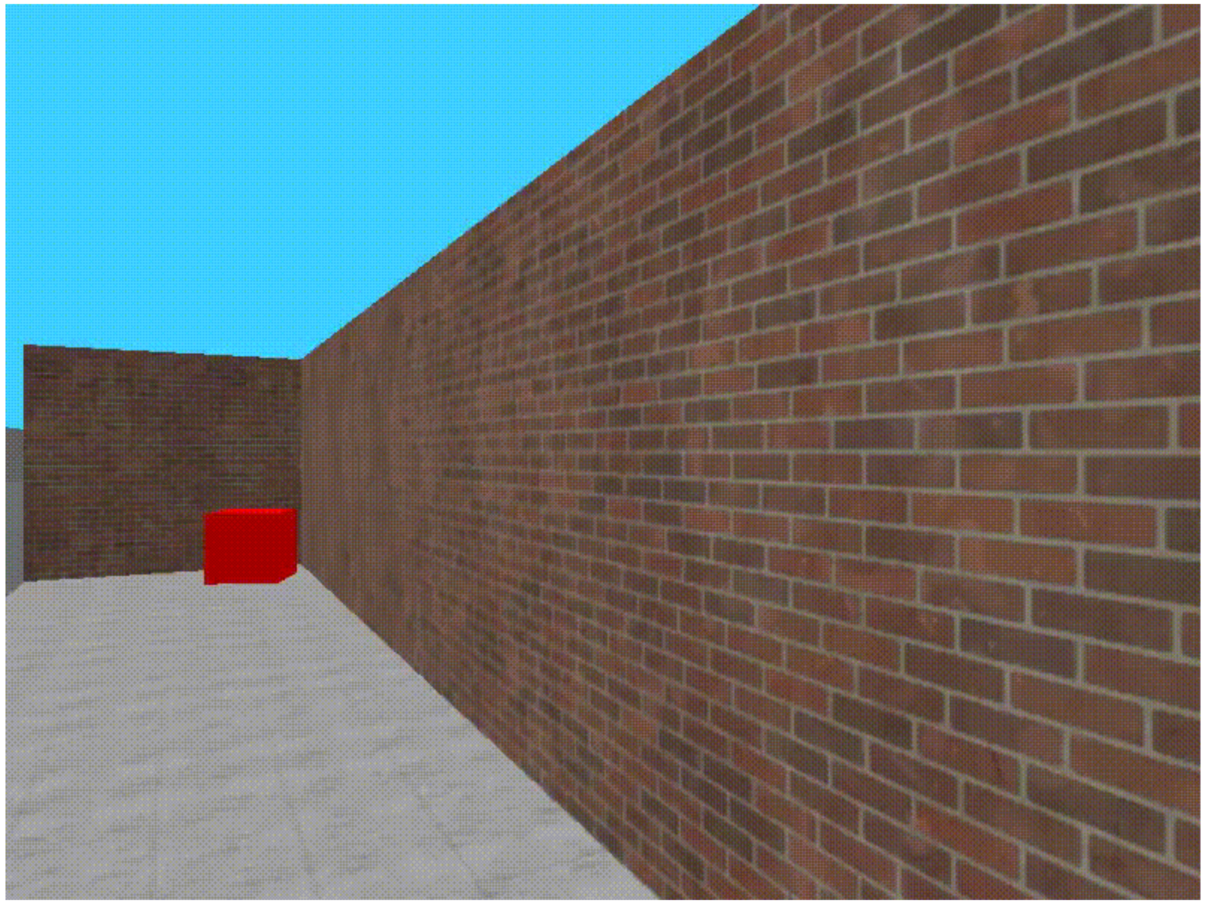


Unsafe

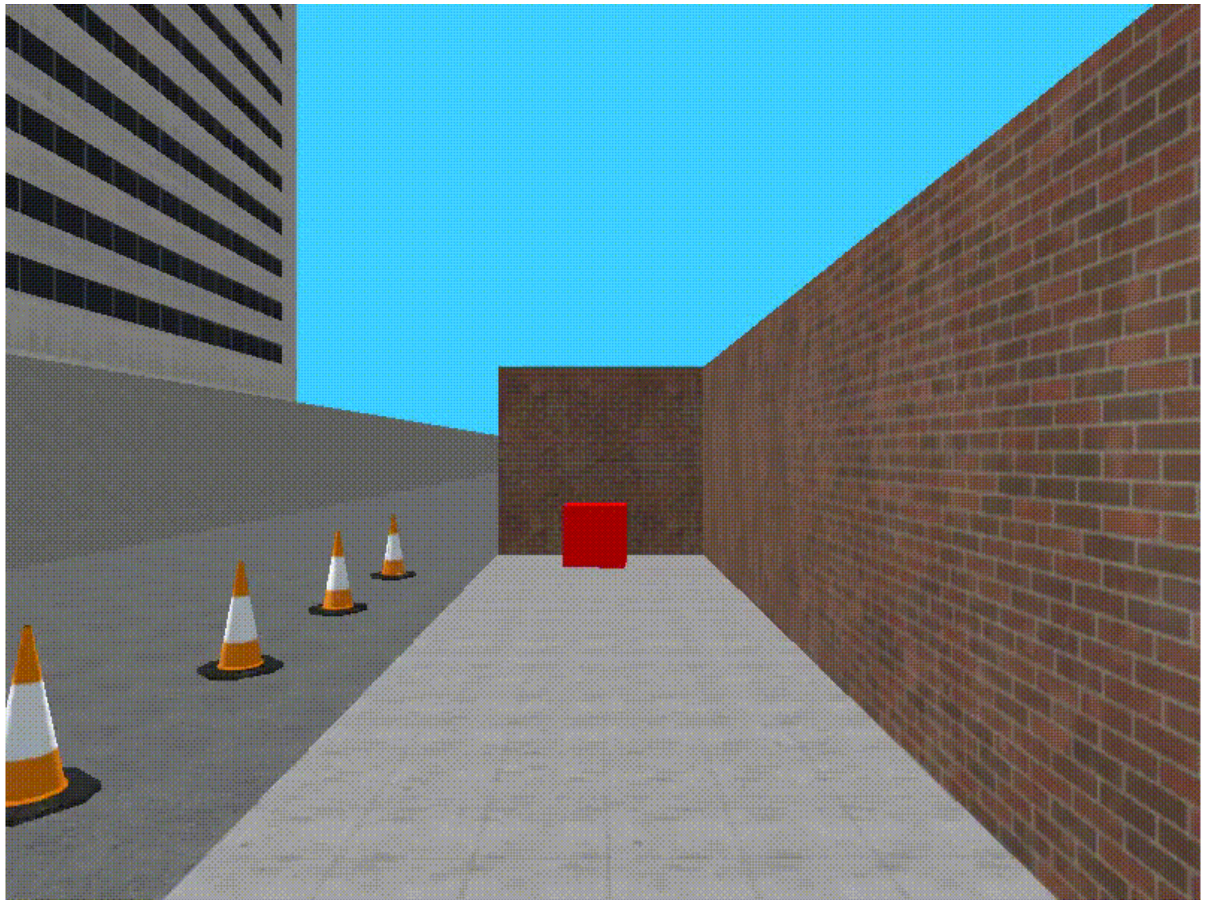


Benchmark: Sidewalk

Success

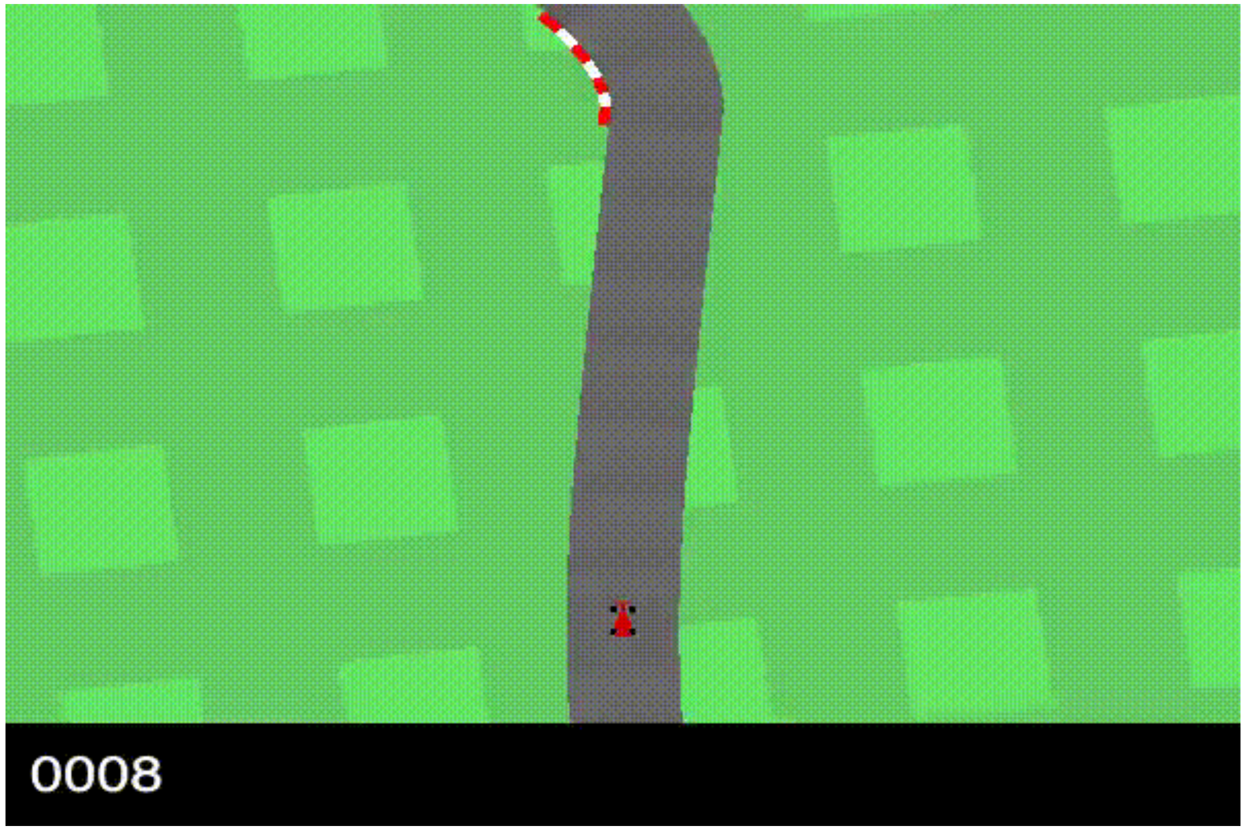


Unsafe

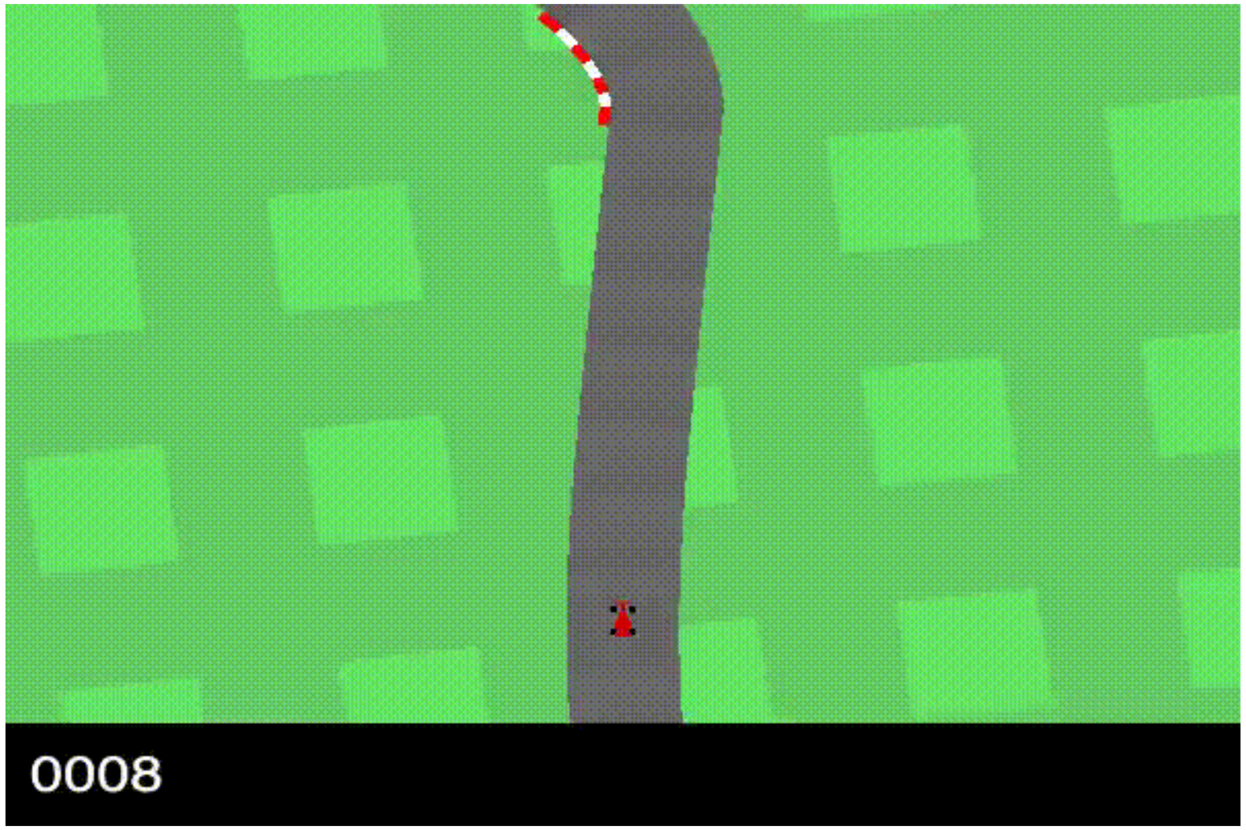


Benchmark: CarRacing

Success

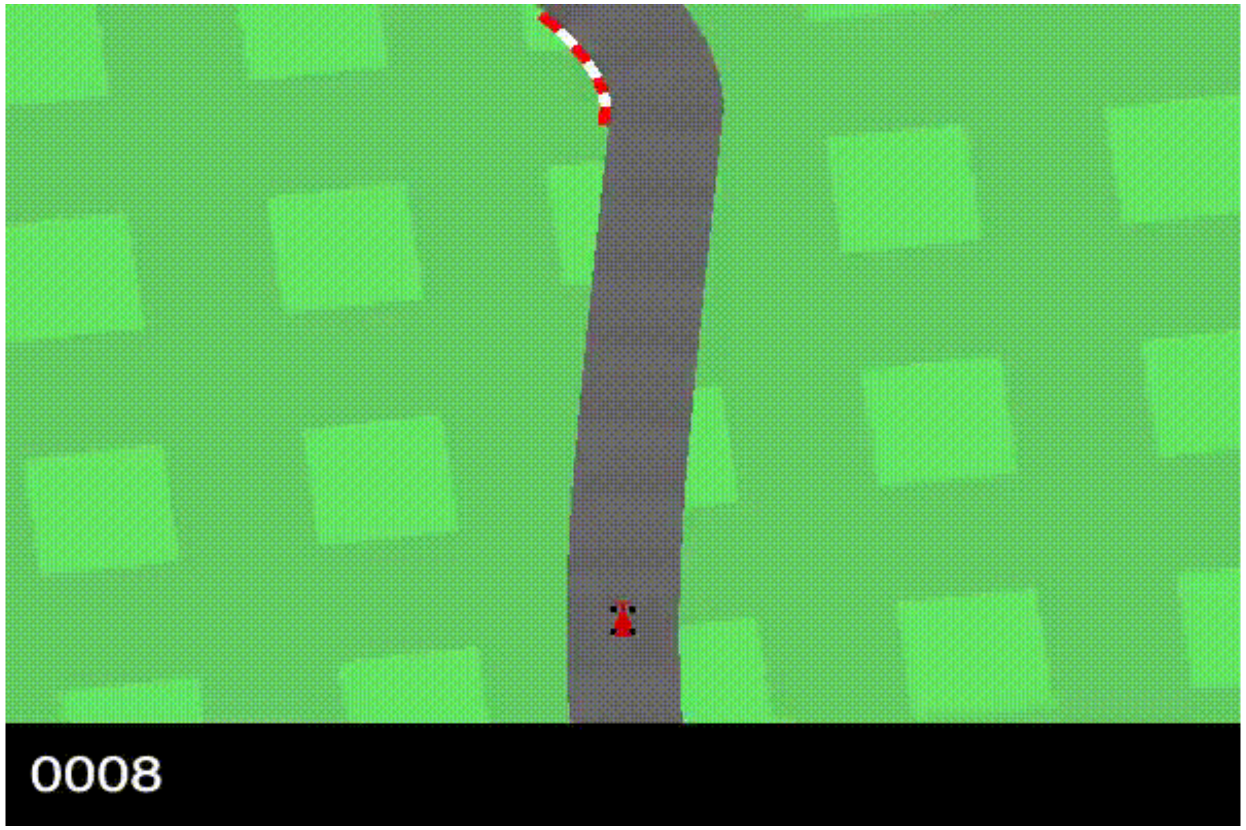


Unsafe



Benchmark: CarRacing

Success



Unsafe