Do Androids Dream of Electric Fences?

Safe Reinforcement Learning with Imagination-Based Agents

Peter He, Borja Gonzalez Leon, Francesco Belardinelli

Peter He

Imperial College London
University College London
peter.he.21@ucl.ac.uk

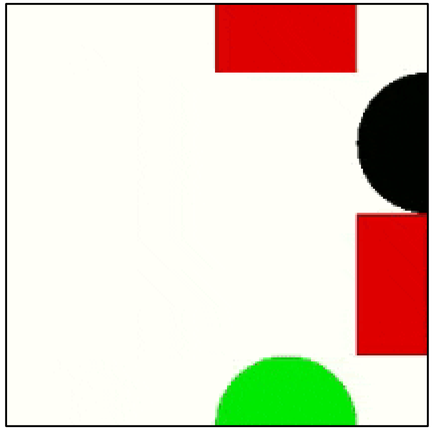
Borja Gonzalez Leon

Imperial College London

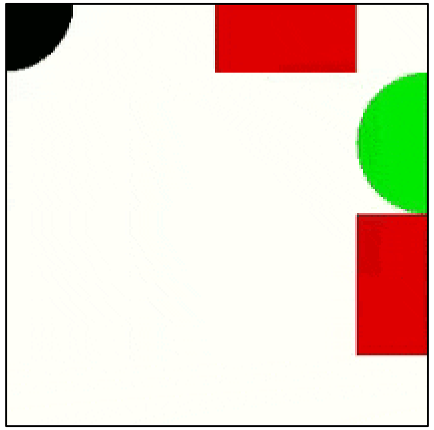
Francesco Belardinelli

Imperial College London

Coming Up In Today's Presentation…



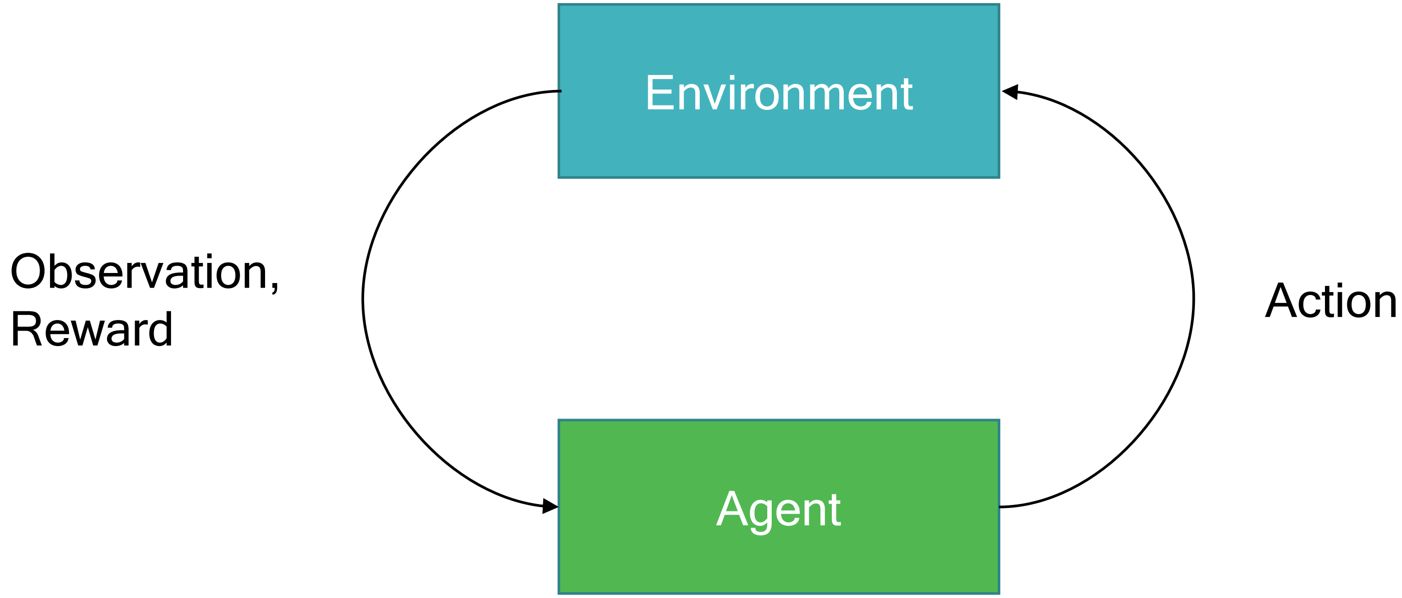
SOTA (Unsafe) RL Agent



Our Method

D. Hafner, et al. *Dream to Control: Learning Behaviors by Latent Imagination*. 2020.
D. Hafner, et al. *Mastering Atari with Discrete World Models*. 2021.

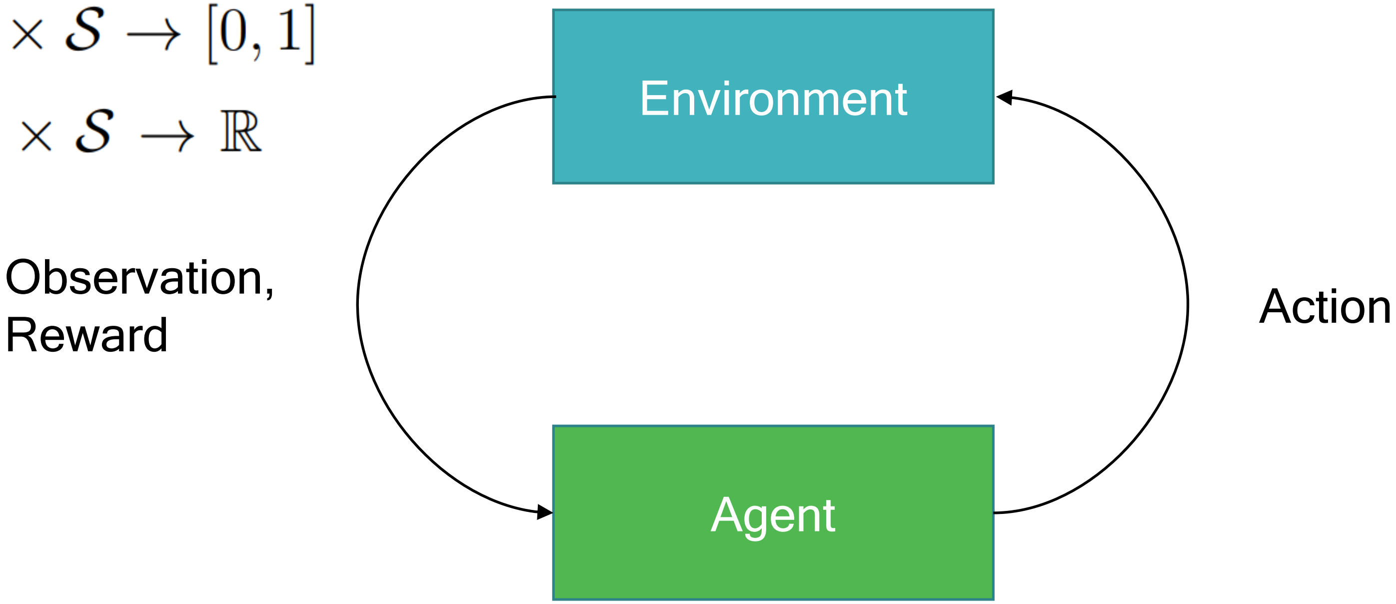
Reinforcement Learning



Reinforcement Learning

$$\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$$

$$\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$$

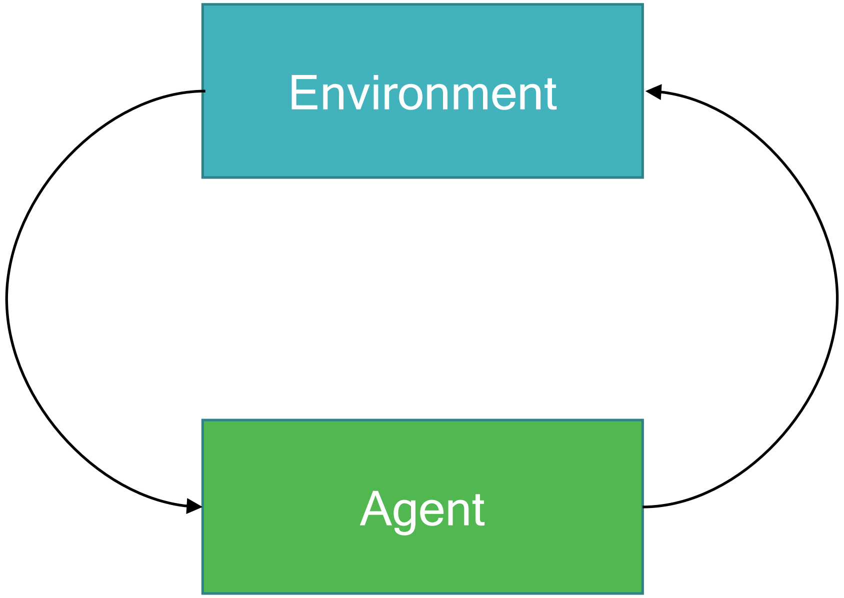


Reinforcement Learning

$$\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$$

$$\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$$

Observation,
Reward



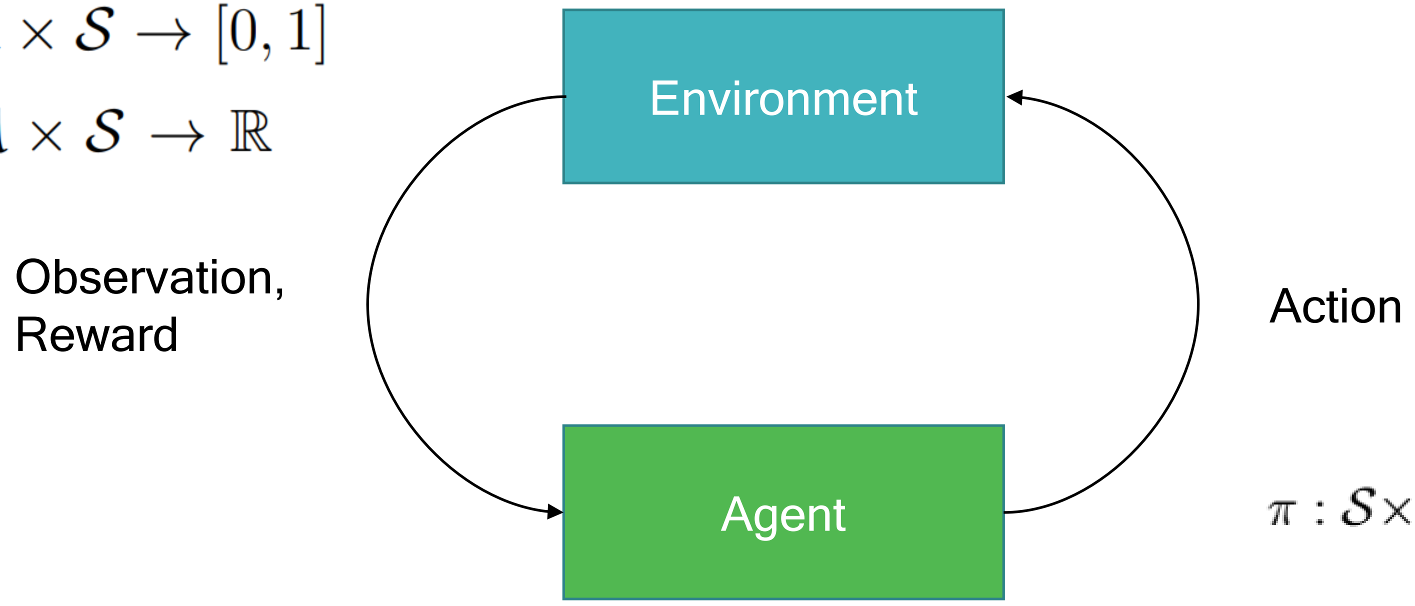
Action

$$\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$$

Reinforcement Learning

$$\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$$

$$\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$$



$$\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$$

Maximise $$\mathbb{E}\left[\sum_{t=0}^{H} \gamma^t \mathcal{R}(s_t, a_t, s_{t+1})\right]$$

Safety

“Bad things shouldn’t happen”

Safety

"Bad things shouldn't happen"

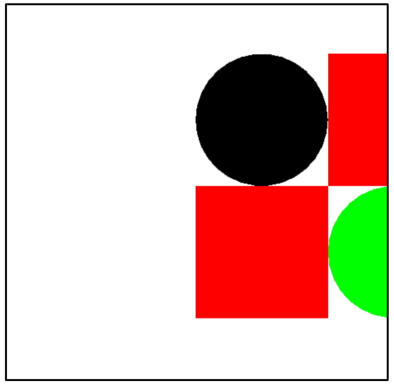
Safety

"Bad things shouldn't happen"

$$\phi, \phi' ::= true \mid p \mid \neg\phi \mid \phi \wedge \phi' \mid \bigcirc\phi \mid \phi \cup \phi'$$

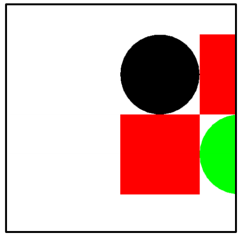
Safety

$$\phi, \phi' ::= true \mid p \mid \neg\phi \mid \phi \wedge \phi' \mid \bigcirc\phi \mid \phi \cup \phi'$$

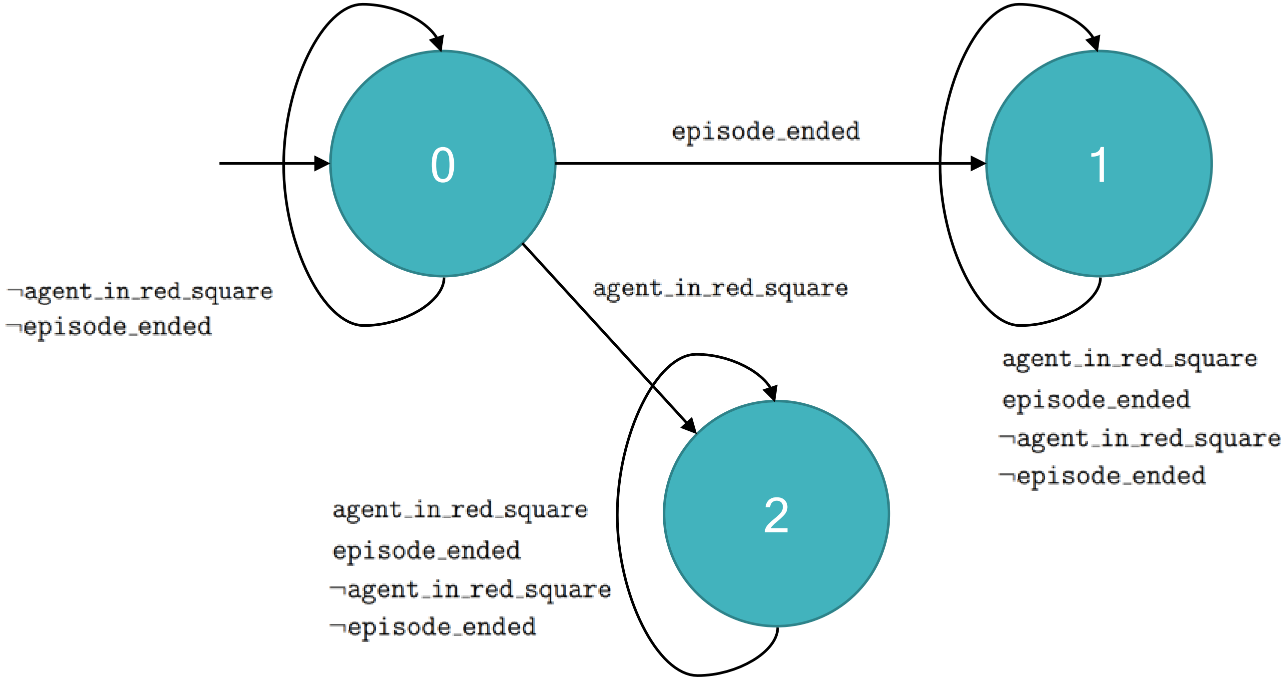


$\neg\texttt{agent\_in\_red\_square} \cup \texttt{episode\_ended}$

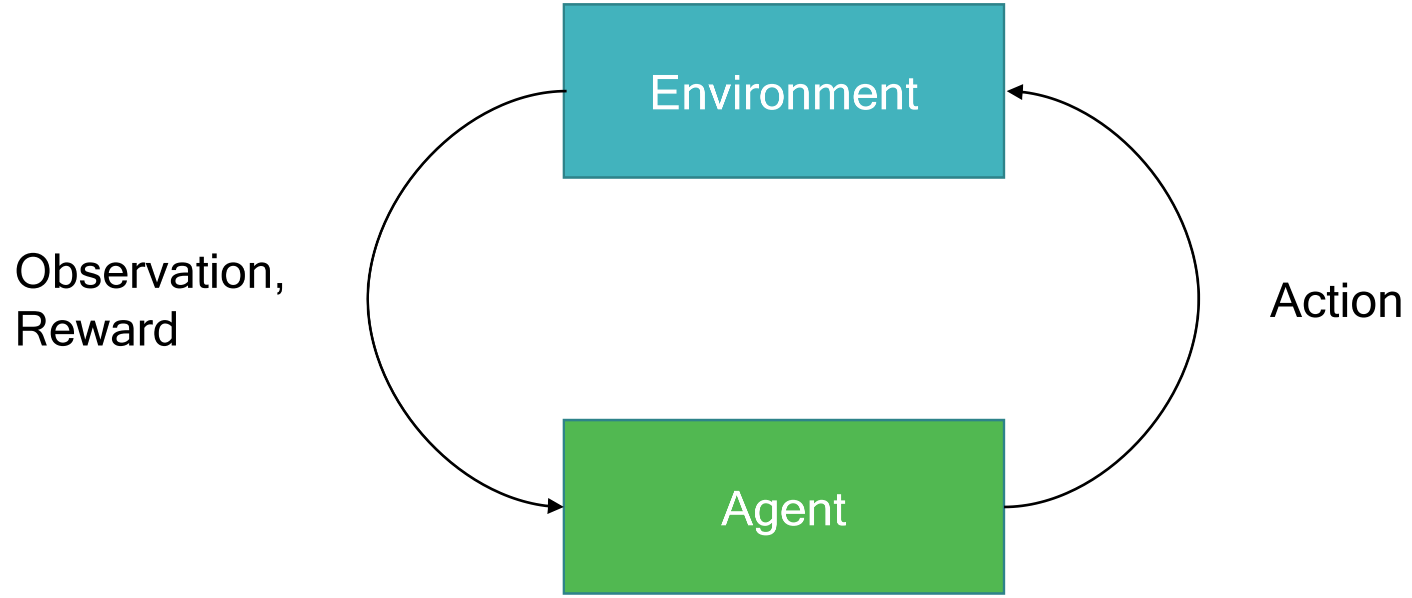
Safety



$\neg$agent_in_red_square $\cup$ episode_ended

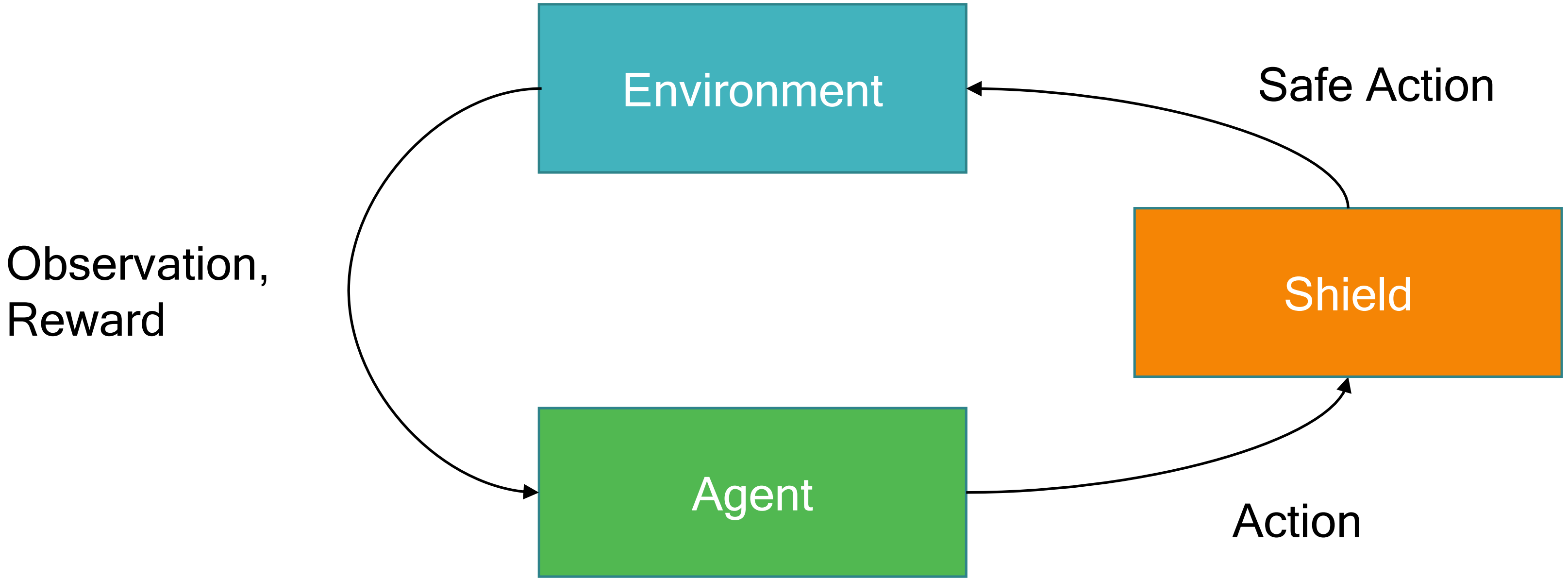


Shielding



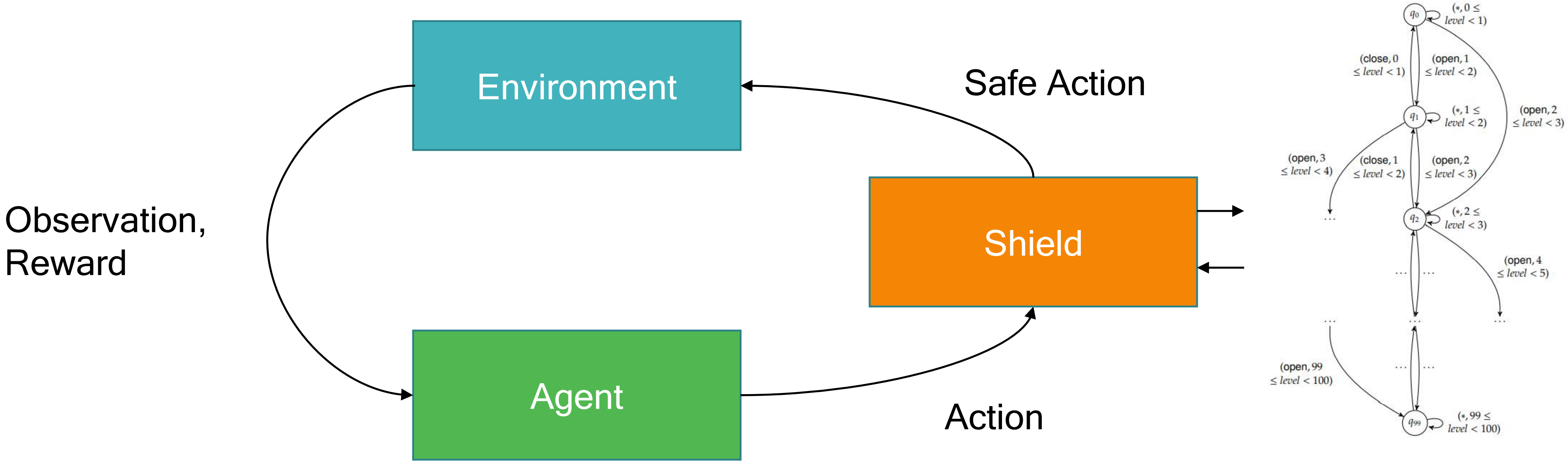
M. Alshiekh et al. *Safe Reinforcement Learning via Shielding*. 2018

Shielding



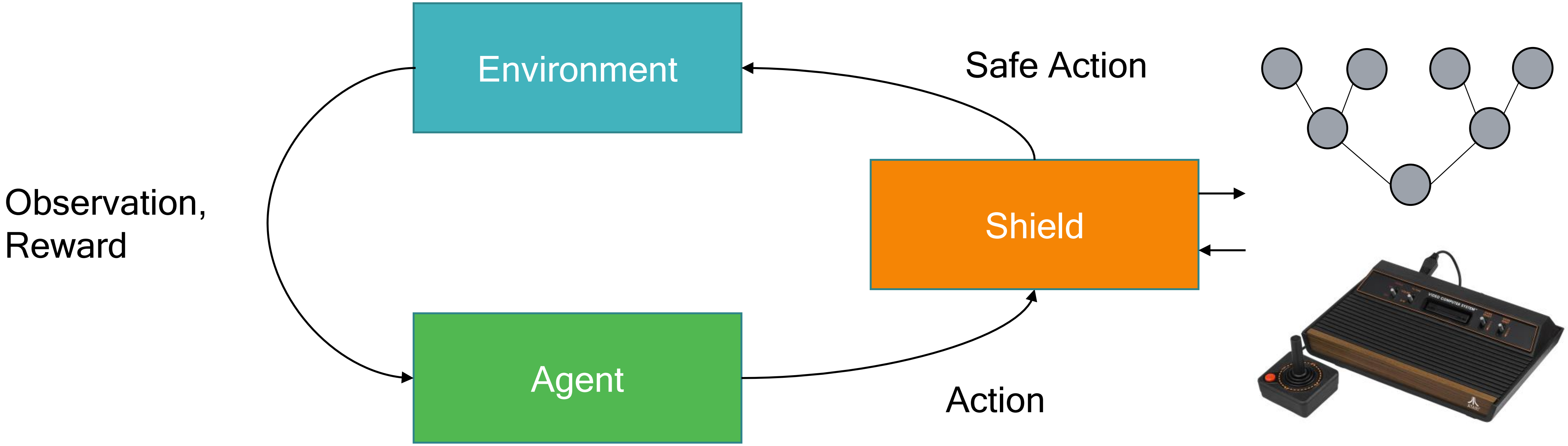
M. Alshiekh et al. *Safe Reinforcement Learning via Shielding*. 2018

Shielding



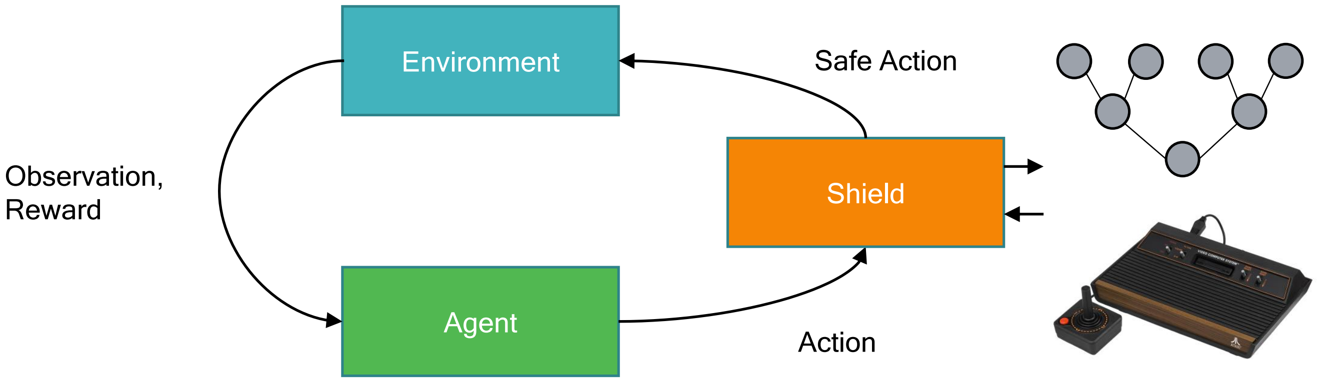
M. Alshiekh et al. *Safe Reinforcement Learning via Shielding*. 2018

Shielding



M. Giacobbe et al. *Shielding Atari Games with Bounded Prescience.* 2021.

Shielding



Given some finite trace $\rho = s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} \dots \xrightarrow{a_{n-1}} s_n$

And the set of all finite traces of length H from state s, $\varrho_H(s)$

A policy π is H-bounded safe iff.

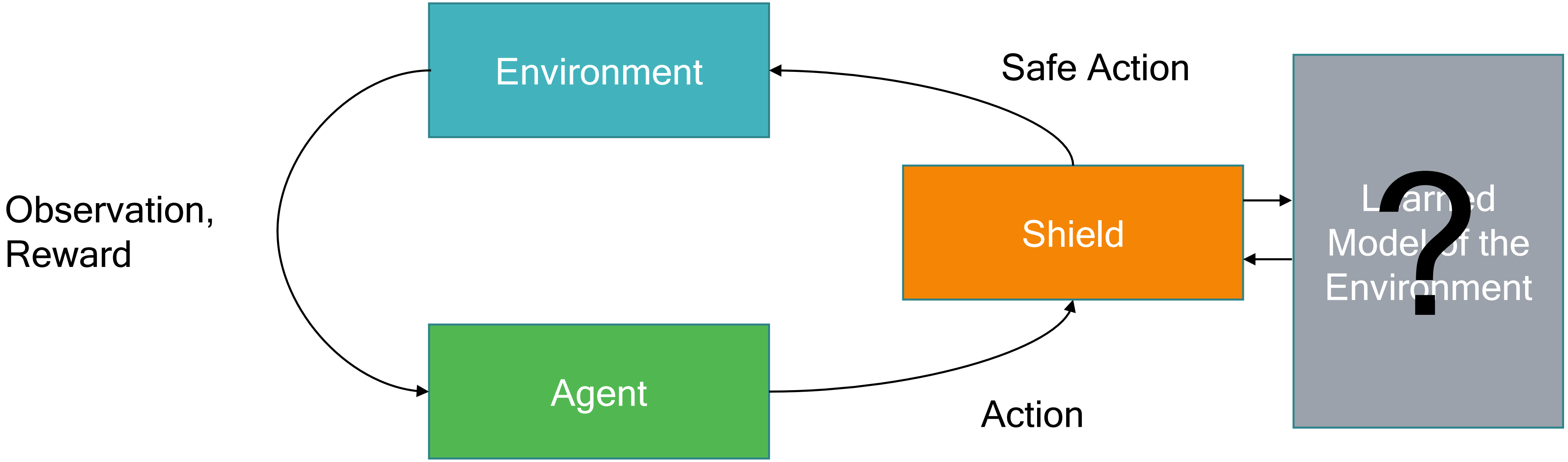
There are no safe traces

$$\forall s \in \mathcal{S}.(\exists \rho = (s,a) \in \varrho_H(s) S(\rho, \phi) \land \pi(s_0) = a_0) \lor \forall \rho \in \varrho_H(s). \neg S(\rho, \phi))$$

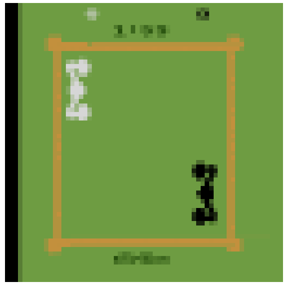
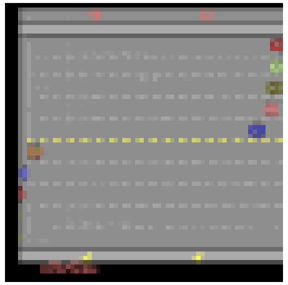
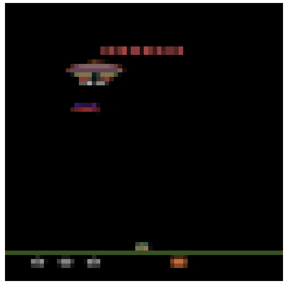
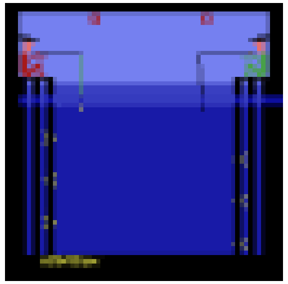
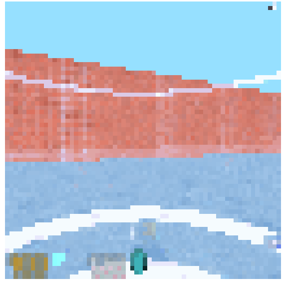
If there is a safe trace of length H, we take it

M. Giacobbe et al. *Shielding Atari Games with Bounded Prescience.* 2021.

Our Approach



Model-Based RL to the Rescue!

1. Learn a model of the environment
2. Learn a policy inside the model of the environment
3. Collect data in the real environment using the learned policy
4. Repeat until convergence

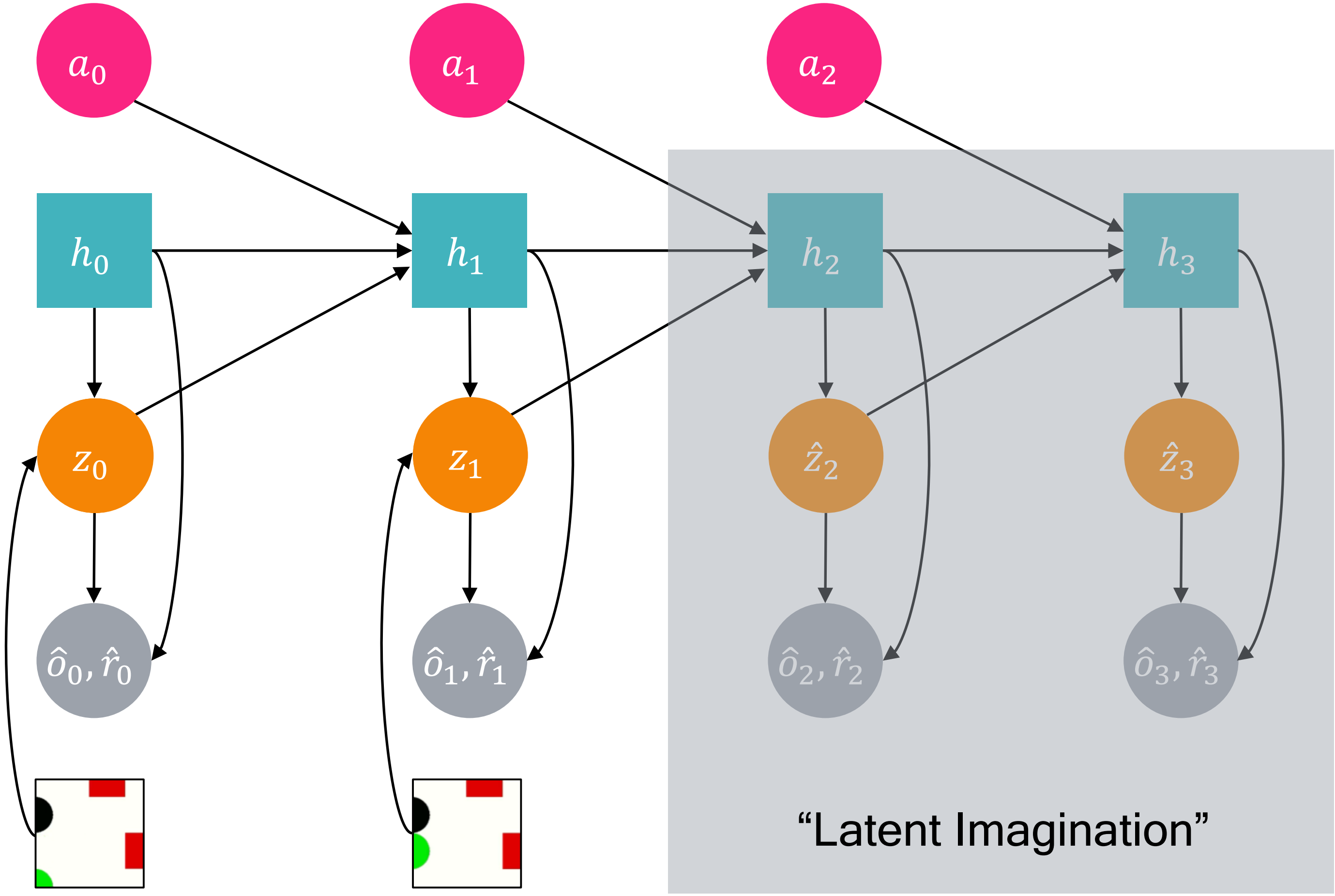
D. Hafner, et al. *Dream to Control: Learning Behaviors by Latent Imagination*. 2020.
D. Hafner, et al. *Mastering Atari with Discrete World Models*. 2021.

Model-Based RL to the Rescue!

1. Learn a model of the environment
2. Learn the policy using the model of the environment
3. Collect data in the real environment using the learned policy
4. Also use the model of the environment to keep the agent safe
5. Repeat until convergence

New!

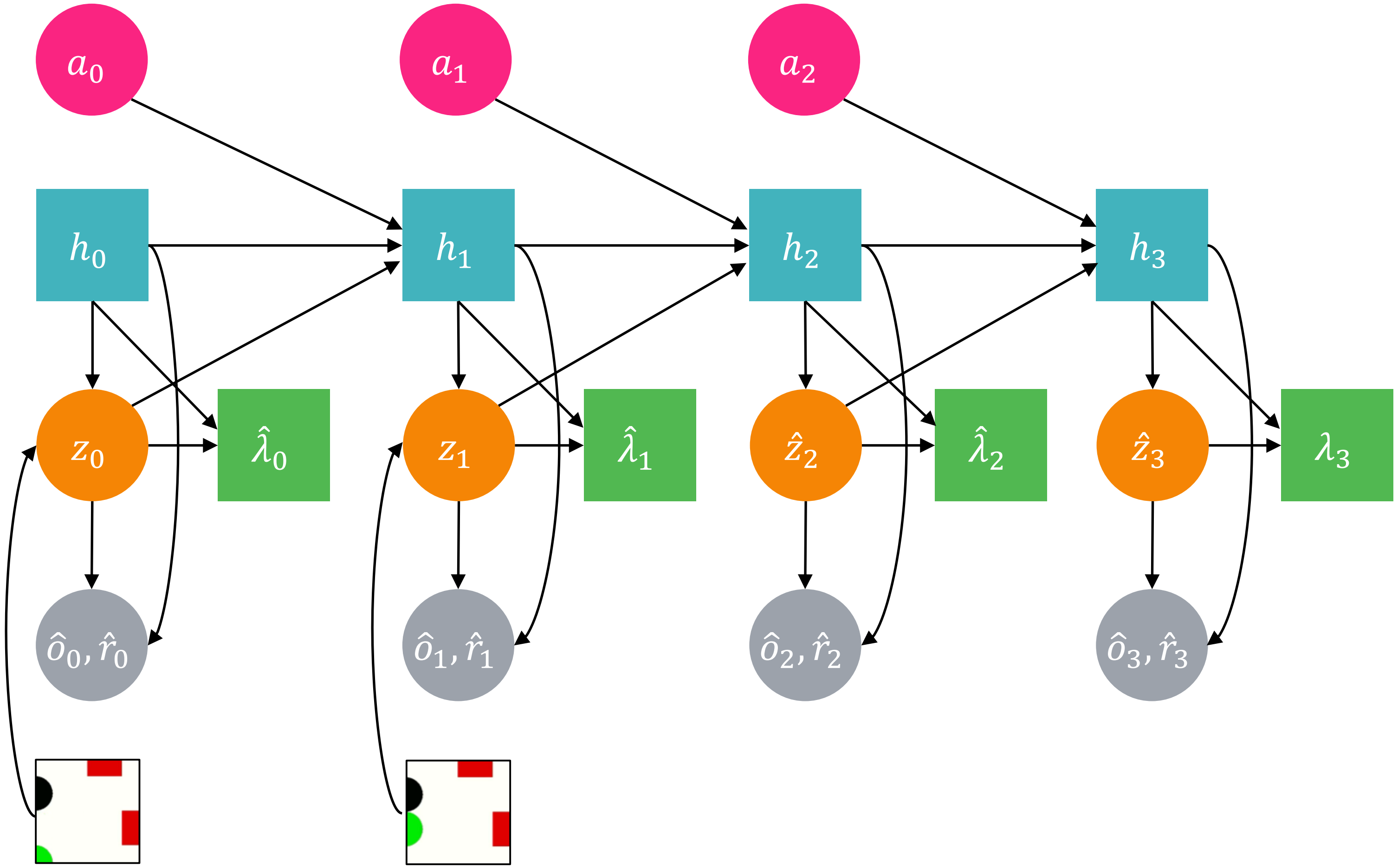
D. Hafner, et al. *Dream to Control: Learning Behaviors by Latent Imagination*. 2020.
D. Hafner, et al. *Mastering Atari with Discrete World Models*. 2021.



Recurrent State Space Model

$$h_t = f(h_{t-1}, z_{t-1}, a_{t-1})$$
$$z_t \sim q(z_t | h_t, o_t)$$
$$\hat{z}_t \sim p(\hat{z}_t | h_t)$$
$$\hat{o}_t \sim p(\hat{o}_t | h_t, z_t)$$
$$\hat{r}_t \sim p(\hat{r}_t | h_t, z_t)$$

D. Hafner et al. Learning Latent Dynamics for Planning from Pixels. 2019.

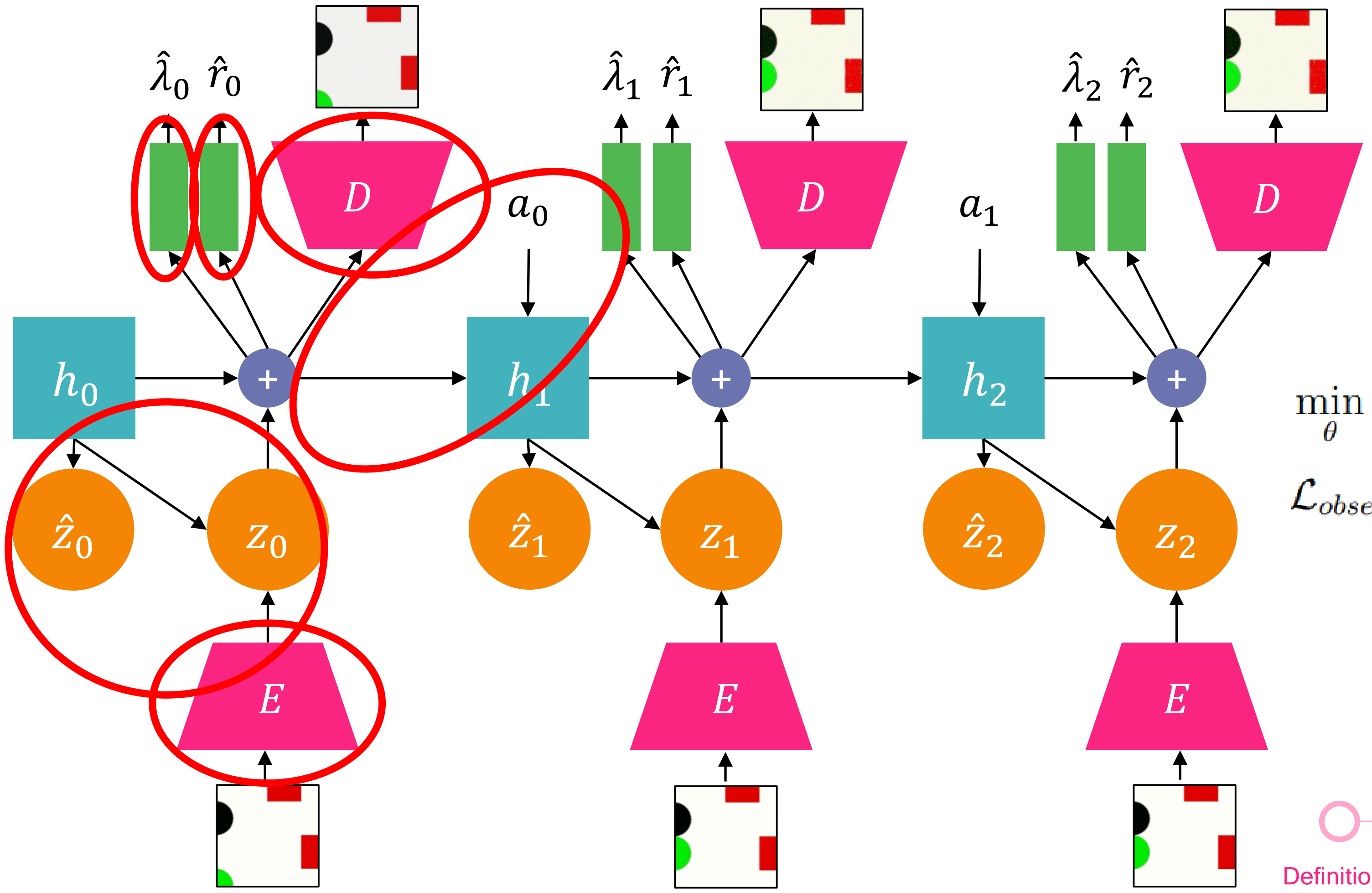


New!

Safety Recurrent State Space Model

New!

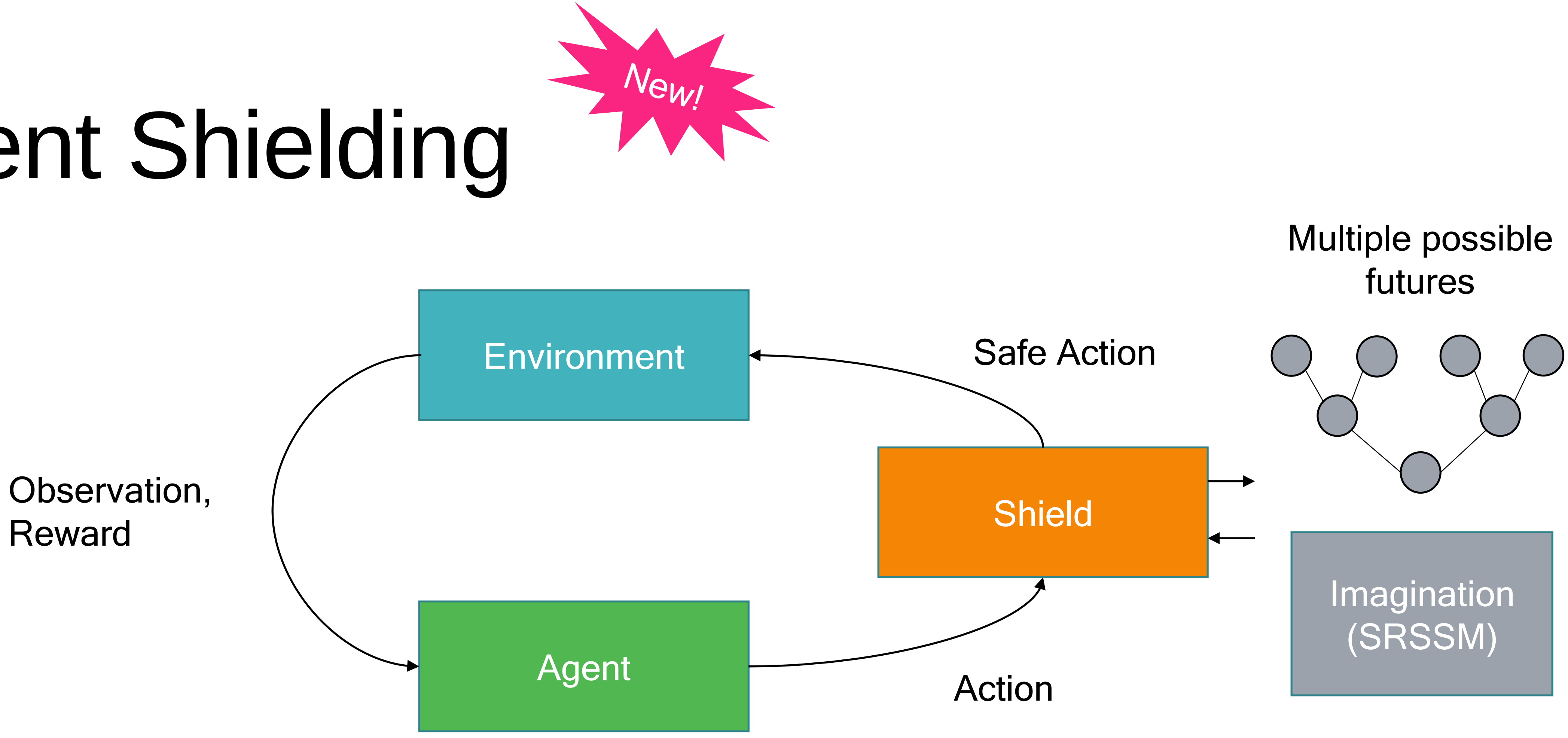
SRSSM with Neural Networks



$$\min_{\theta} \mathcal{L}_{model} = $$

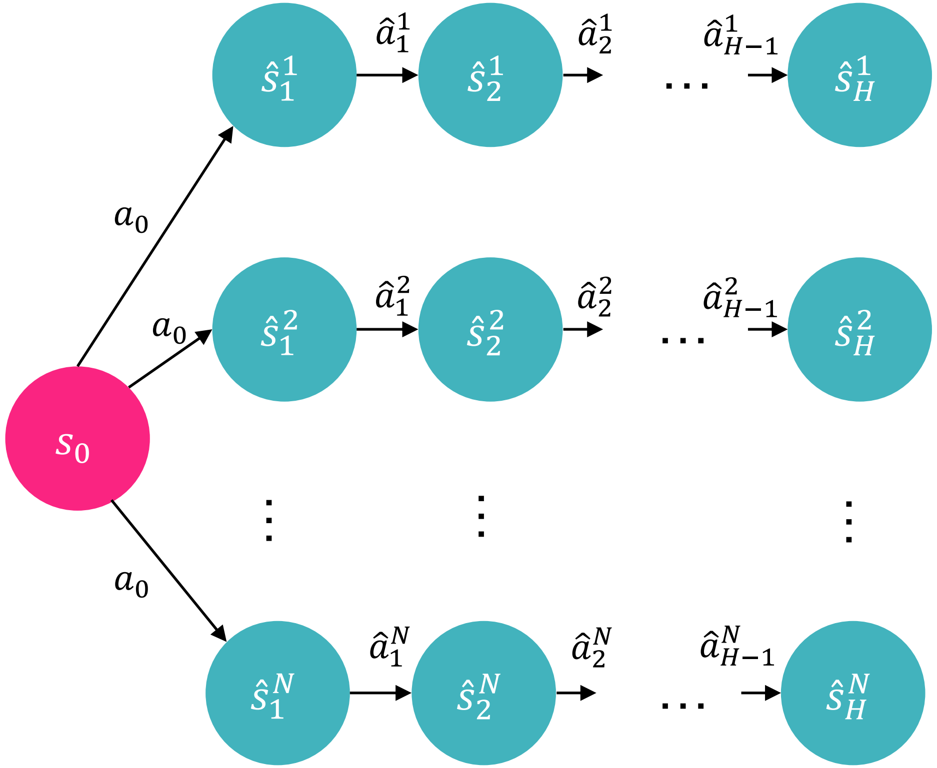
$$\mathcal{L}_{observation} + \mathcal{L}_{reward} + \mathcal{L}_{KL} + \mathcal{L}_{violation}$$

Latent Shielding



New!

Approximate Bounded Prescience

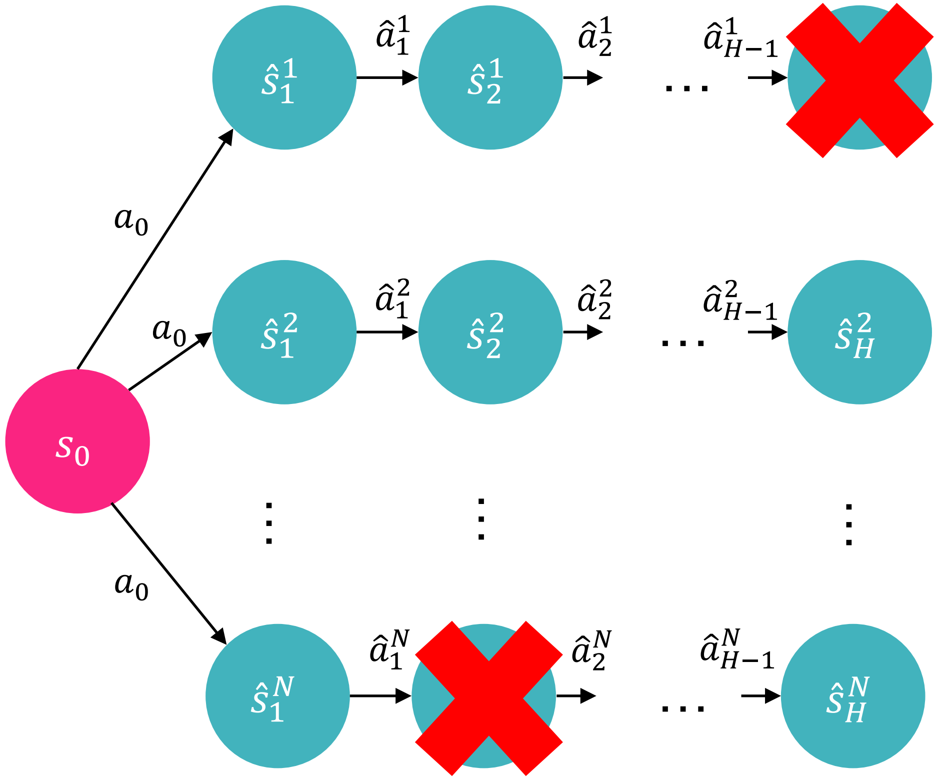


$$\hat{a}_h^n = \pi(\hat{s}_{h-1}^n) + \boxed{\eta}$$

Noise Term

New!

ABP Shielding for Latent Trajectories



$$\hat{a}_h^n = \pi(\hat{s}_{h-1}^n) + \underbrace{\boxed{\eta}}_{\text{Noise Term}}$$

New!

ABP Shieldin rajectories

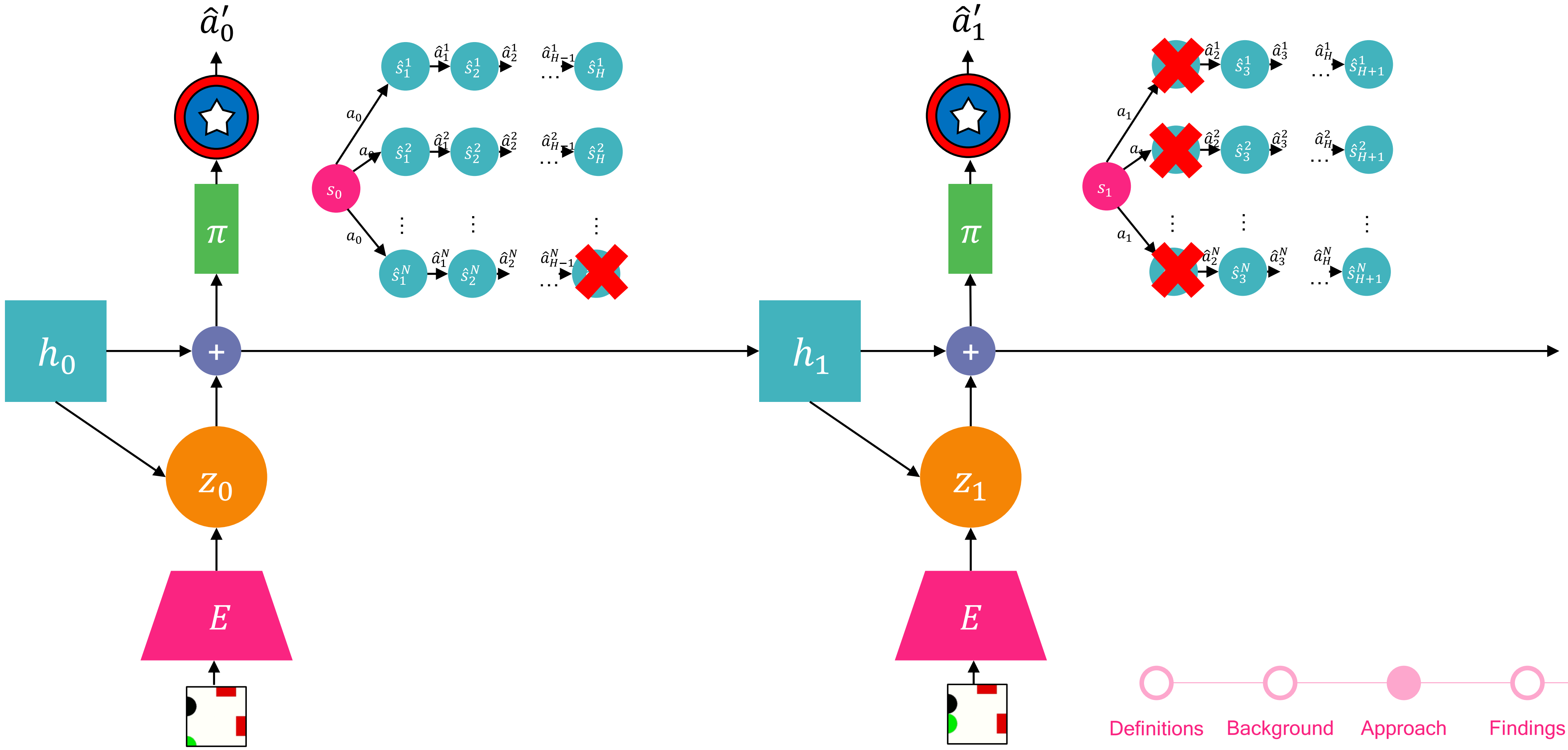
$\pi'(s_t) = \begin{cases} \pi(s_t \\ \varsigma(s_t \end{cases}$ $ation|\ \pi(s_t)) < \epsilon$

Safe Alternative Po

Safety Threshold

ABP Shielding for Latent Trajectories

New!



Training an Agent with Latent Shielding

New!

1. Learn a **SRSSM** model of the environment
2. Learn the policy using the model of the environment, **assigning a punishment to violation states**
3. Collect data in the real environment using the learned policy **with the shield**
4. Repeat until convergence

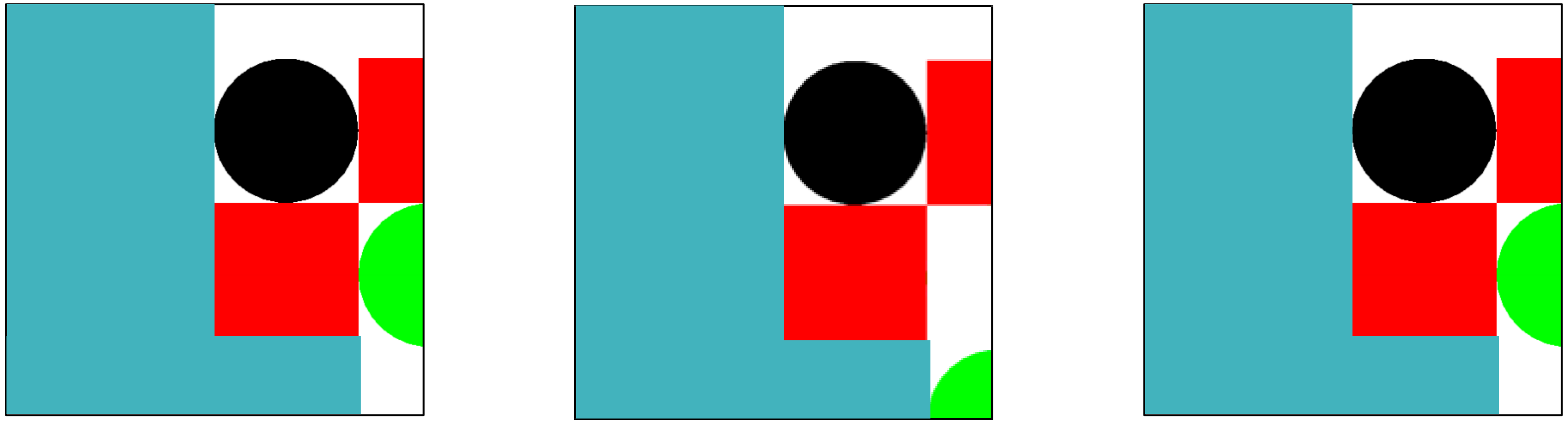
But It’s Not All Fun and Games…

New!

An inaccurate internal model of the environment can lead to the latent shield hindering exploration!

New!

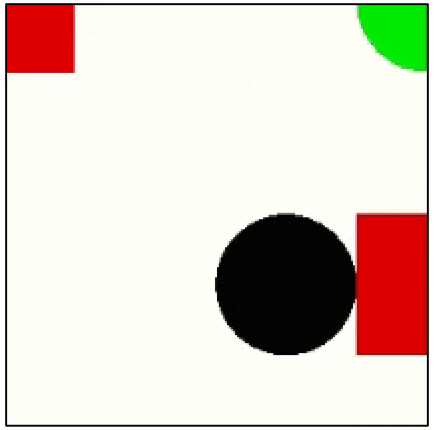
But It's Not All Fun and Games…



An inaccurate internal model of the environment can lead to the latent shield hindering exploration

But It's Not All Fun and Games…

New!



In fact, even bounded prescience shielding can hinder exploration

Shield Introduction Schedules

New!

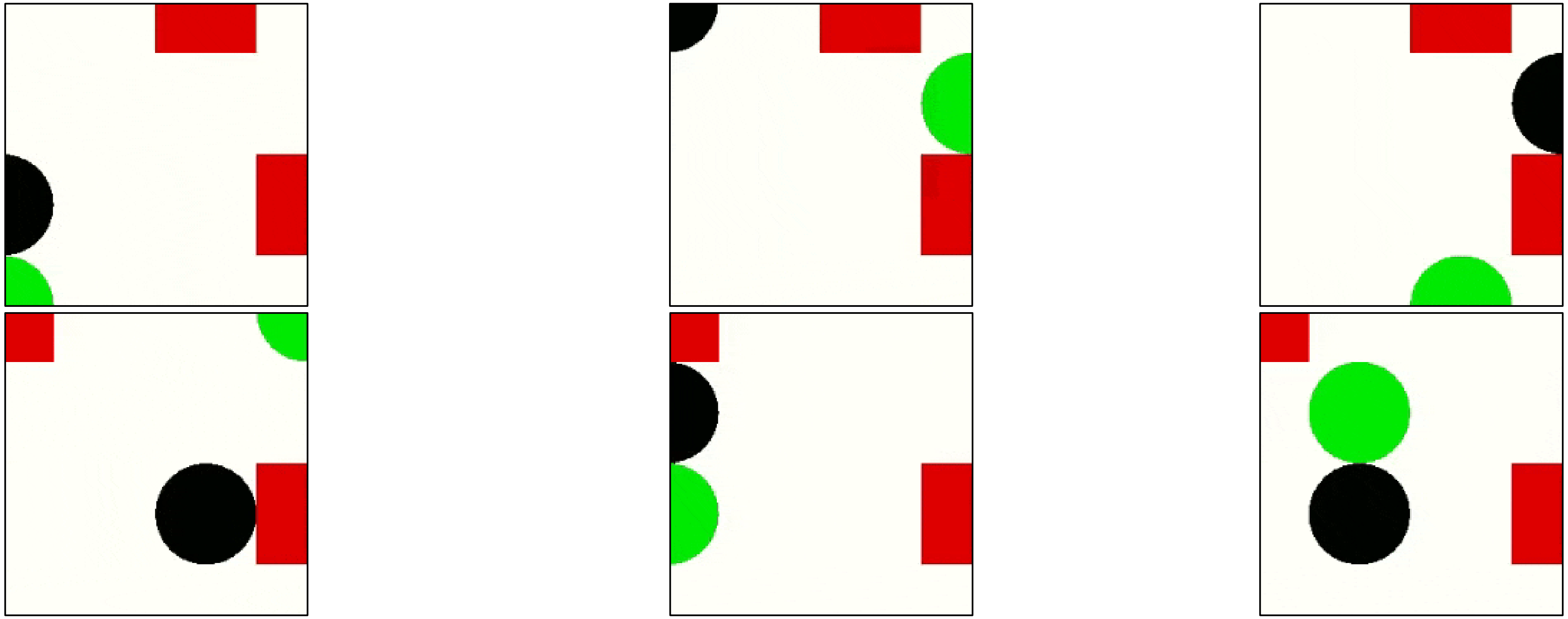
Potential implementations:

+ A gradually decaying probability of disabling the shield with respect to time
+ Enabling the shield once the change in dynamics model loss falls to below some threshold
+ Simply enabling shielding after a certain number of training episodes have been completed

Performance Evaluation

Seed 1

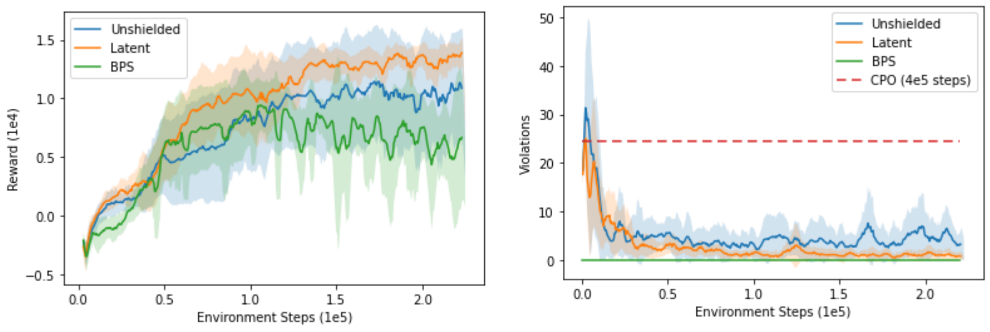
Seed 2



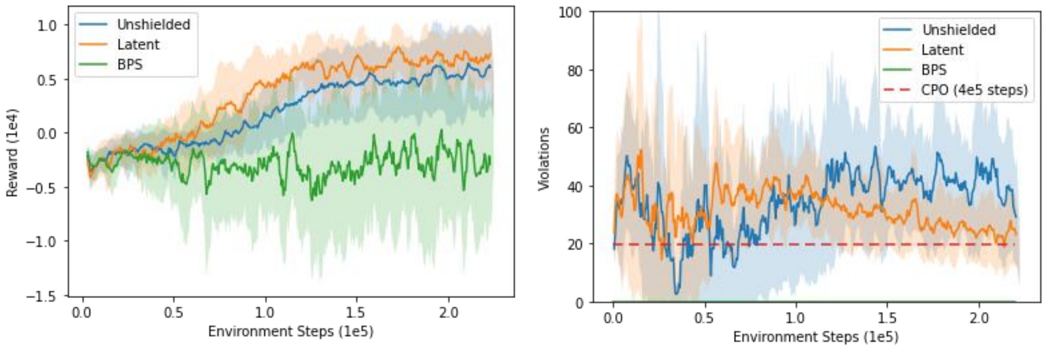
BPS (Giacobbe et al.) | Latent Shield (ours) | Baseline (Hafner et al.)

M. Giacobbe et al. *Shielding Atari Games with Bounded Prescience.* 2021.
D. Hafner, et al. *Dream to Control: Learning Behaviors by Latent Imagination*. 2020.
D. Hafner, et al. *Mastering Atari with Discrete World Models*. 2021.

Performance Evaluation



Static Gridworld



Procedurally Generated Gridworld

(see paper for MORE graphs)

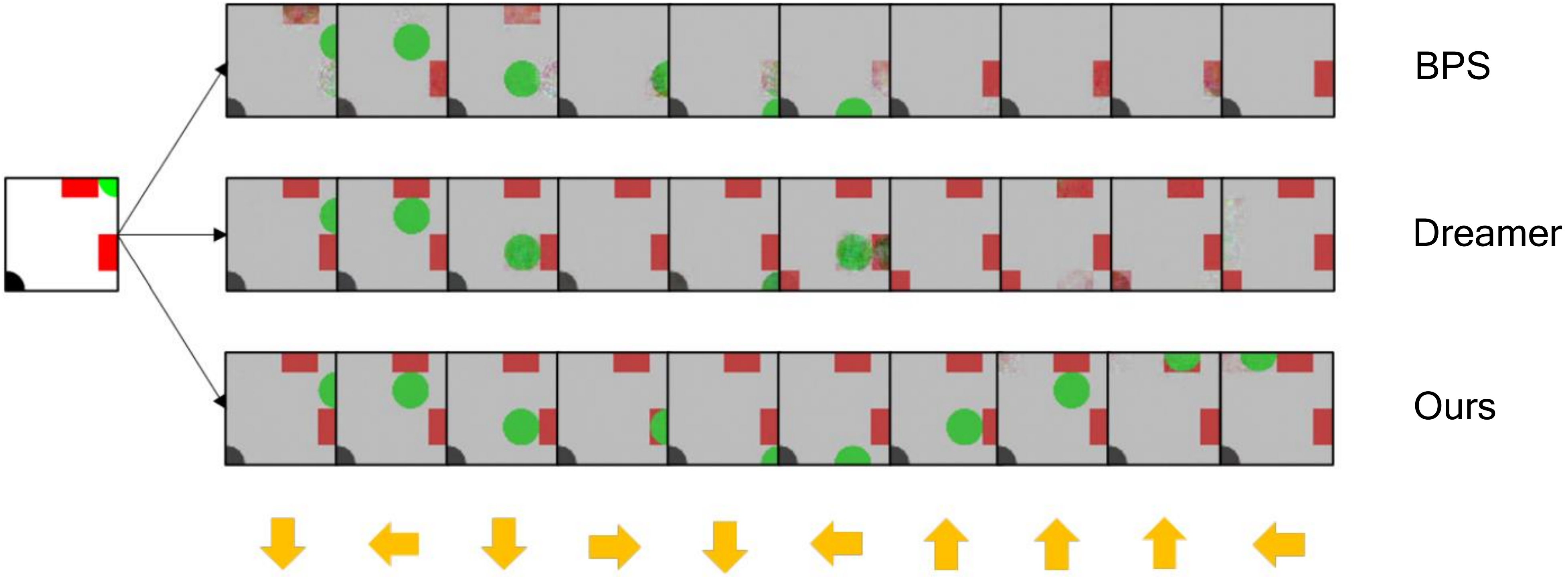
M. Giacobbe et al. *Shielding Atari Games with Bounded Prescience*. 2021.
D. Hafner, et al. *Dream to Control: Learning Behaviors by Latent Imagination*. 2020.
D. Hafner, et al. *Mastering Atari with Discrete World Models*. 2021.
J. Achiam, et al. Constrained Policy Optimization. 2017.

Performance Evaluation

	Flavour	Metric	Latent	Unshielded	BPS	CPO
Visual Grid World	Fixed	Testing Reward	**15067 (434)**	13148 (249)	12468 (620)	-2925 (1065)
		Testing Violations	0.30 (0.76)	2.25 (1.60)	**0 (0)**	13.43 (19.25)
		Training Violations	1262 (172)	2306 (833)	**0 (0)**	16455 (1435)
	Procedural	Testing Reward	**8084 (2221)**	6825 (1427)	1938 (3552)	-1588 (2051)
		Testing Violations	4.50 (3.59)	33.7 (16.28)	**0 (0)**	19.60 (13.83)
		Training Violations	14018 (1852)	15309 (4686)	**0 (0)**	18705 (3756)
Cliff Driver	$p_{stick} = 0.1$	Testing Reward	8.57 (2.96)	**10.76 (3.29)**	10.50 (3.28)	7.56 (2.86)
		Testing Violations	**0 (0)**	**0 (0)**	**0 (0)**	3.40 (1.91)
		Training Violations	58.2 (9.60)	90.0 (9.10)	**24.0 (13.02)**	973.0 (357.7)
	$p_{stick} = 0.5$	Testing Reward	**8.10 (4.99)**	6.63 (8.07)	7.10 (9.52)	6.44 (3.00)
		Testing Violations	**0.18 (0.84)**	0.54 (1.53)	0.22 (1.18)	0.48 (1.24)
		Training Violations	91.8 (16.85)	157.6 (18.4)	**80.4 (17.43)**	3126 (2823)

M. Giacobbe et al. *Shielding Atari Games with Bounded Prescience.* 2021.
D. Hafner, et al. *Dream to Control: Learning Behaviors by Latent Imagination*. 2020.
D. Hafner, et al. *Mastering Atari with Discrete World Models*. 2021.
J. Achiam, et al. Constrained Policy Optimization. 2017.

Examining Latent Dynamics



Open Questions

+ What’s the best Shield Introduction Schedule?
+ How might we leverage uncertainty?
+ How might we leverage offline pre-training?

Takeaways

+ **Latent shielding** lets you shield agents in high-dimensional environments without knowledge of the dynamics *a priori*.
+ It does this by learning the environment model rather than having it be handcrafted.
+ Shielding can harm model-based DRL algorithms - introduce the shield gently with a **Shield Introduction Schedule**.