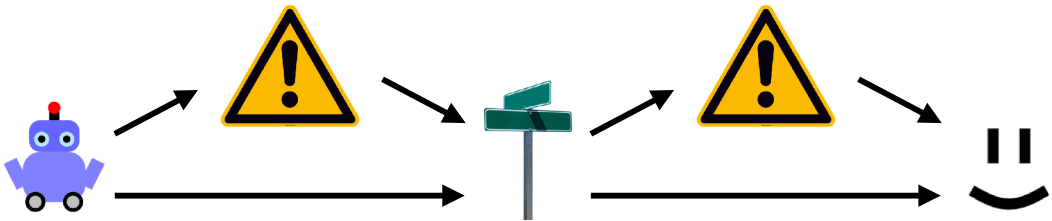
Shields to Guarantee Probabilistic Safety in MDPs

Linus Heck, Filip Macák, Roman Andriushchenko, Milan Češka, Sebastian Junges

CAV 2026

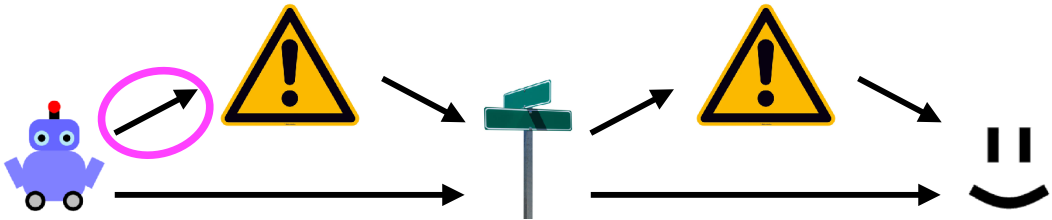
Radboud Universiteit

Classical Shielding: Example



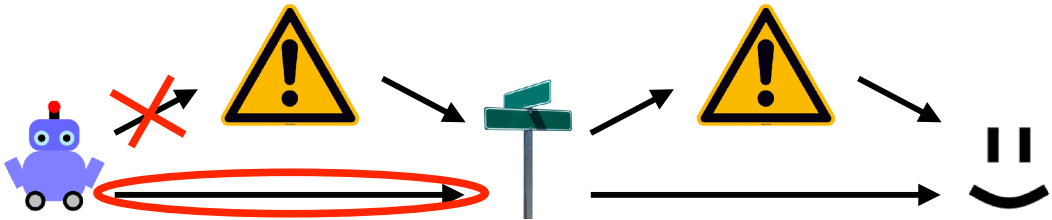
- Robot can choose going the dangerous or the safe path.
- Dangerous path has probability 0.1 of crashing.
- Specification: Robot should never crash.
- Shields are runtime enforcers that block unsafe actions.

Classical Shielding: Example



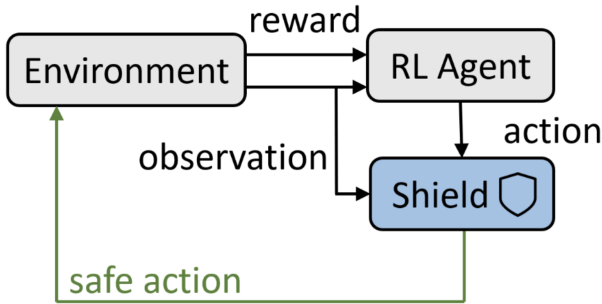
- Robot can choose going the dangerous or the safe path.
- Dangerous path has probability 0.1 of crashing.
- Specification: Robot should never crash.
- Shields are runtime enforcers that block unsafe actions.

Classical Shielding: Example



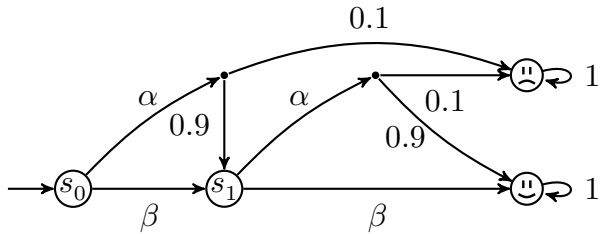
- Robot can choose going the dangerous or the safe path.
- Dangerous path has probability 0.1 of crashing.
- Specification: Robot should never crash.
- Shields are runtime enforcers that block unsafe actions.

Shielding an Agent



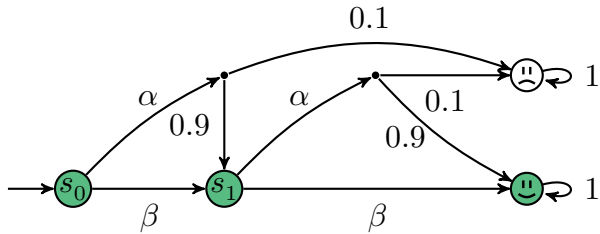
Picture: Könighofer et al., “Shields for Safe Reinforcement Learning”

Classical Shielding on an MDP



Ensure: Agent reaches ☹ with probability zero.

Classical Shielding on an MDP



Ensure: Agent reaches ☹ with probability zero.

Shielding strategy is known to give these guarantees:

- Safety: The shielded policy is safe (i.e., satisfies the specification)
- Permissiveness: If the policy is already safe, the shield does not interfere

Why Is Shielding With Probability Zero Not Enough?

Taking some risk is often required to achieve the intended goal. (Without taking some risk, you would not have been here at FLoC!)

Shielding with zero risk is very conservative.

Why Is Shielding With Probability Zero Not Enough?

Taking some risk is often required to achieve the intended goal. (Without taking some risk, you would not have been here at FLoC!)

Shielding with zero risk is very conservative.

$\Longrightarrow$ We want probabilistic shielding!

Natural notion of safety: Probability of reaching $\lightning$ is at most ν.

Overview

Shielding with Guarantees

Two desirable properties:

- Safety: The shielded policy is safe (i.e., satisfies the specification)
- Permissiveness: If the policy is already safe, the shield does not interfere

Shielding with Guarantees

Two desirable properties:

- Safety: The shielded policy is safe (i.e., satisfies the specification)
- Permissiveness: If the policy is already safe, the shield does not interfere

Many probabilistic shields in the literature do not give any guarantees.

Shielding with Guarantees

Two desirable properties:

- Safety: The shielded policy is safe (i.e., satisfies the specification)
- Permissiveness: If the policy is already safe, the shield does not interfere

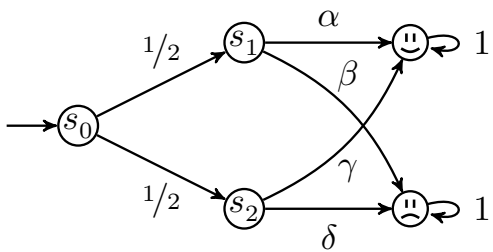
Many probabilistic shields in the literature do not give any guarantees.

We show: in general, a probabilistic shield cannot satisfy both guarantees.

No Safe and Permissive Probabilistic Shield Exists

Recall that shields are runtime enforcers.

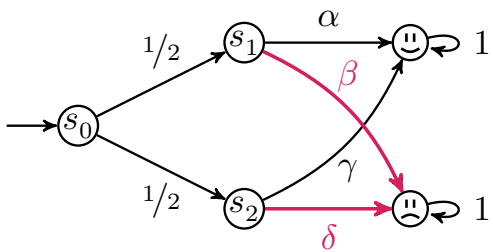
Assume: Some shield $\shield$ is safe and permissive reaching $\frownie$ with prob. at most $1/2$.



No Safe and Permissive Probabilistic Shield Exists

Recall that shields are runtime enforcers.

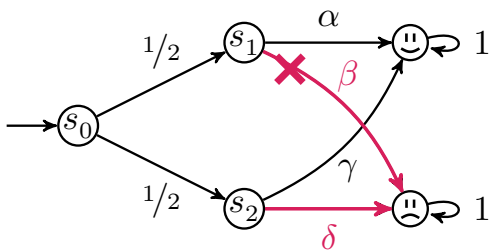
Assume: Some shield $\shield$ is safe and permissive reaching $\frownie$ with prob. at most $1/2$.



No Safe and Permissive Probabilistic Shield Exists

Recall that shields are runtime enforcers.

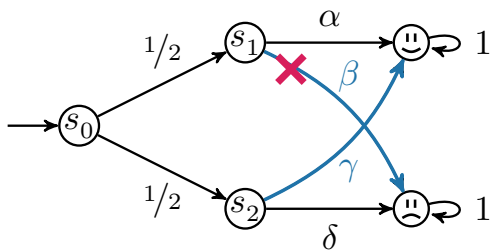
Assume: Some shield $\shield$ is safe and permissive reaching $\frownie$ with prob. at most $1/2$.



No Safe and Permissive Probabilistic Shield Exists

Recall that shields are runtime enforcers.

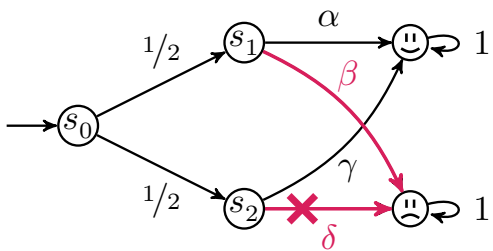
Assume: Some shield $\cup$ is safe and permissive reaching $\frowny$ with prob. at most $1/2$.



No Safe and Permissive Probabilistic Shield Exists

Recall that shields are runtime enforcers.

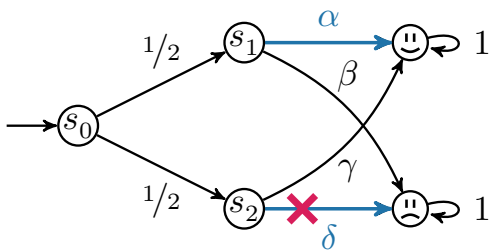
Assume: Some shield $\shield$ is safe and permissive reaching $\frownie$ with prob. at most $1/2$.



No Safe and Permissive Probabilistic Shield Exists

Recall that shields are runtime enforcers.

Assume: Some shield $\shield$ is safe and permissive reaching $\frownie$ with prob. at most $1/2$.

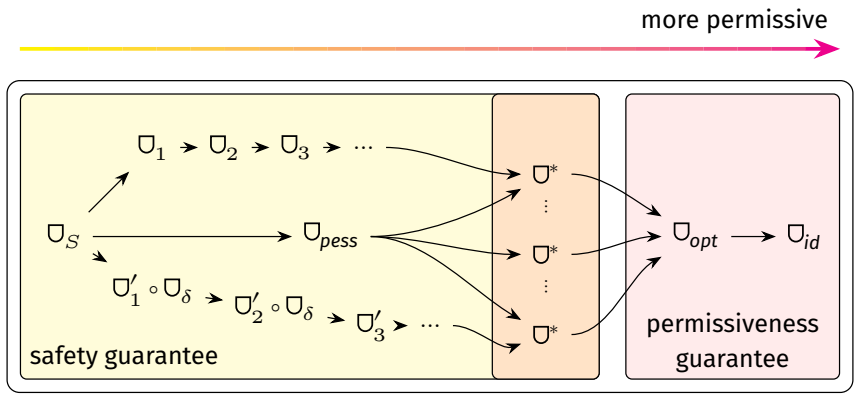


Contradiction: The shield is either not safe or not permissive. (Theorem 3)

Overview

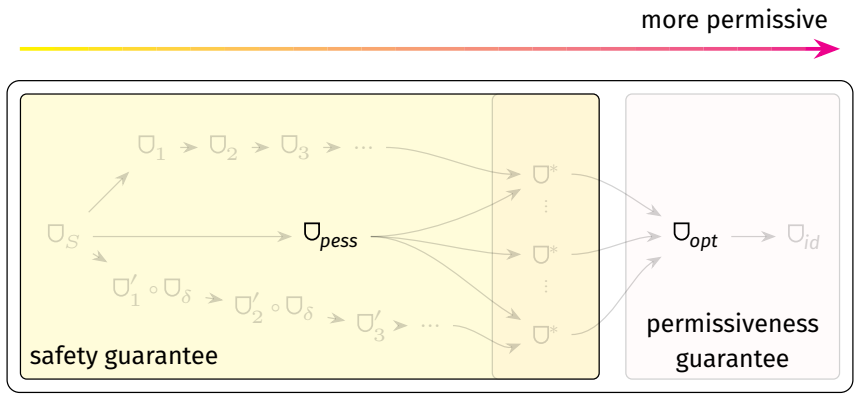
The Space of Probabilistic Shields

We want safe shields that are as permissive as possible. We introduce a lattice of shields and populate it with many new shields:



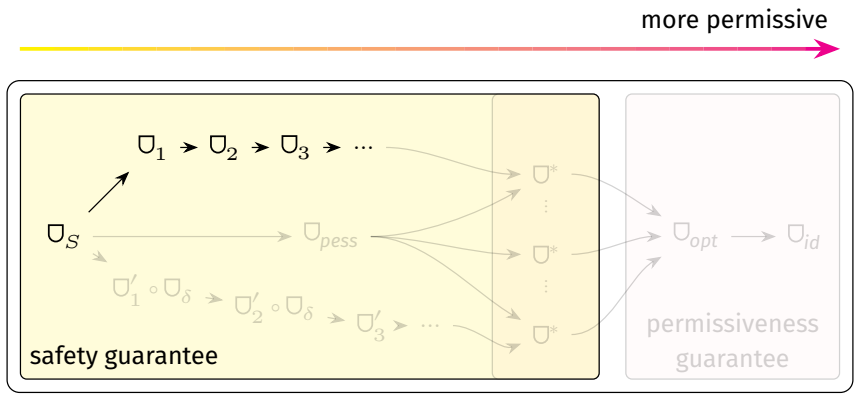
The Space of Probabilistic Shields

We want safe shields that are as permissive as possible. We introduce a lattice of shields and populate it with many new shields:



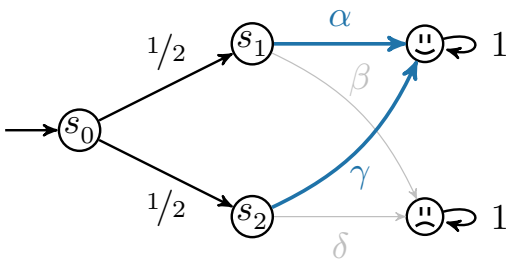
The Space of Probabilistic Shields

We want safe shields that are as permissive as possible. We introduce a lattice of shields and populate it with many new shields:

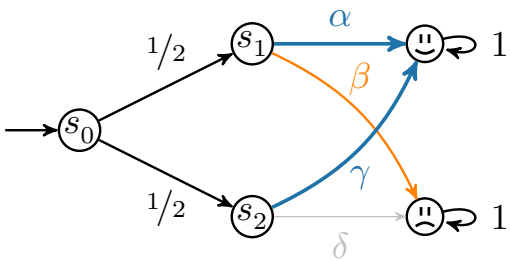


Constructing Shields

1. Start with the classical shield

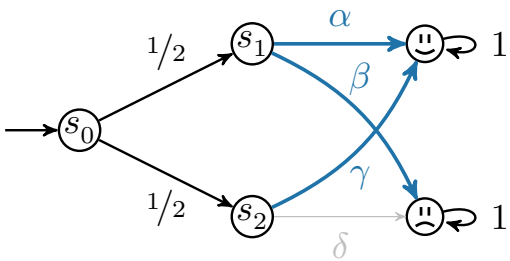


Constructing Shields



1. Start with the classical shield
2. Propose adding β

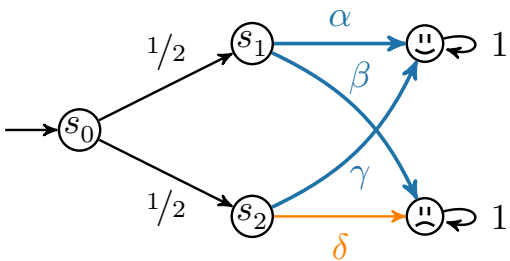
Constructing Shields



1. Start with the classical shield
2. Propose adding β
3. Adding β is safe $\Rightarrow$ move to $\shield_1$

$$\shield_S \longrightarrow \shield_1 \longrightarrow \cdots$$

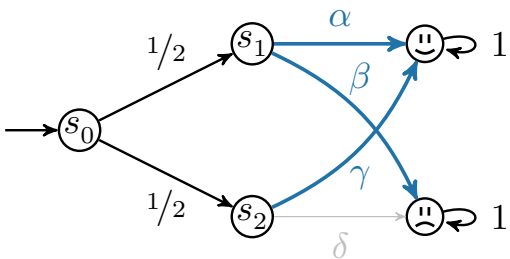
Constructing Shields



1. Start with the classical shield
2. Propose adding β
3. Adding β is safe $\Rightarrow$ move to $\shield_1$
4. Propose adding δ

$$\shield_S \longrightarrow \shield_1 \longrightarrow \cdots$$

Constructing Shields



1. Start with the classical shield
2. Propose adding β
3. Adding β is safe $\Rightarrow$ move to $\shield_1$
4. Propose adding δ
5. Adding δ is unsafe $\Rightarrow$ stay at $\shield_1$
6. ...

$$\shield_S \longrightarrow \shield_1 \longrightarrow \cdots$$

Overview

Experiments

We ran benchmarks on:

- eight different shields,
- three different MDPs,
- three different RL agents,
- three different thresholds.

Safety and Permissiveness in Practice

		Safety threshold		
Shield	Guarantee	.010	.050	.200
Classical	Safety	.000	.000	.000
δ-Shield	None	.000	.000	.000
δ^+-Shield	None	.001	.001	.240
Constructed (ours)	Safety	.010	.050	.200

Probability of reaching $\lightning$ for different shields for drone-greedy.

Safety and Permissiveness in Practice

		Safety threshold		
Shield	Guarantee	.010	.050	.200
Classical	Safety	.000	.000	.000
δ-Shield	None	.000	.000	.000
δ^+-Shield	None	.001	.001	.240
Constructed (ours)	Safety	.010	.050	.200

Probability of reaching $\lightning$ for different shields for drone-greedy.

Safety and Permissiveness in Practice

		Safety threshold		
Shield	Guarantee	.010	.050	.200
Classical	Safety	.000	.000	.000
δ-Shield	None	.000	.000	.000
δ^+-Shield	None	.001	.001	.240
Constructed (ours)	Safety	.010	.050	.200

Probability of reaching $\lightning$ for different shields for drone-greedy.

Safety and Permissiveness in Practice

		Safety threshold		
Shield	Guarantee	.010	.050	.200
Classical	Safety	.000	.000	.000
δ-Shield	None	.000	.000	.000
δ^+-Shield	None	.001	.001	.240
Constructed (ours)	Safety	.010	.050	.200

Probability of reaching $\lightning$ for different shields for drone-greedy.

Safety and Permissiveness in Practice

		Safety threshold		
Shield	Guarantee	.010	.050	.200
Classical	Safety	.000	.000	.000
δ-Shield	None	.000	.000	.000
δ^+-Shield	None	.001	.001	.240
Constructed (ours)	Safety	.010	.050	.200

Probability of reaching $\lightning$ for different shields for drone-greedy.

Safety and Permissiveness in Practice

		Safety threshold		
Shield	Guarantee	.010	.050	.200
Classical	Safety	.000	.000	.000
δ-Shield	None	.000	.000	.000
δ^+-Shield	None	.001	.001	.240
Constructed (ours)	Safety	.010	.050	.200

Our shield allows over $3\times$ as many actions!

Probability of reaching ☹ for different shields for drone-greedy.

The additive δ-shield (from literature) is unsafe in 11/27 cases.

Takeaways

Probabilistic shielding

- is necessary,
- should be done with safety guarantees,
- is practically feasible.

Probabilistic shielding requires trading off safety and permissiveness.

This paper introduces a broad framework for probabilistic shielding and safe probabilistic shields.

David et al. (2015), "Uppaal Stratego".
Alshiekh et al. (2018), "Safe Reinforcement Learning via Shielding".
Jansen et al. (2020), "Safe Reinforcement Learning Using Probabilistic Shields".
Hamel-De le Court, Belardinelli, and Goodall (2025), "Probabilistic Shielding for Safe Reinforcement Learning".