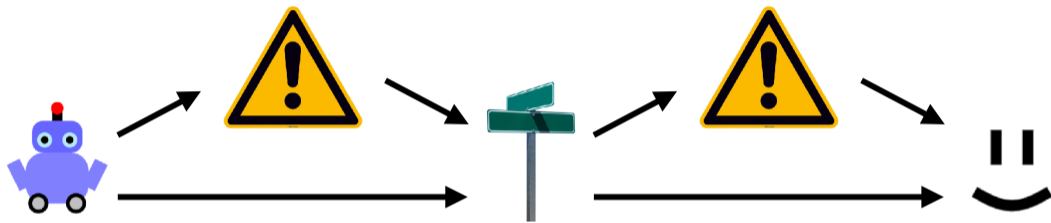


Shields to Guarantee Probabilistic Safety in MDPs

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

July 8, 2026

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

Shielding as a Game

- ▶ Shielding is a **two-player game** played on a Markov decision process
- ▶ The shield wants to prevent an agent from violating a formal specification

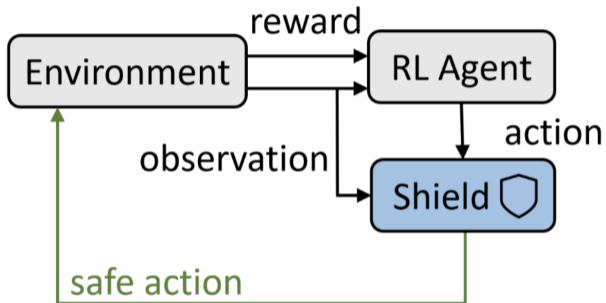
Shielding as a Game

- ▶ Shielding is a **two-player game** played on a Markov decision process
- ▶ The shield wants to prevent an agent from violating a formal specification
- ▶ At each step:
 - ▶ The shield decides which actions to allow
 - ▶ Then, the agent plays some strategy / policy

Shielding as a Game

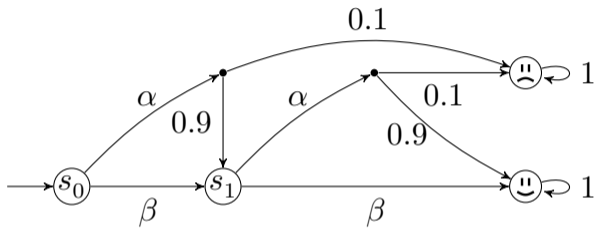
- ▶ Shielding is a **two-player game** played on a Markov decision process
- ▶ The shield wants to prevent an agent from violating a formal specification
- ▶ At each step:
 - ▶ The shield decides which actions to allow
 - ▶ Then, the agent plays some strategy / policy
- ▶ Goal: Whatever the agent does, the shielded agent should be safe

Shielding an Agent



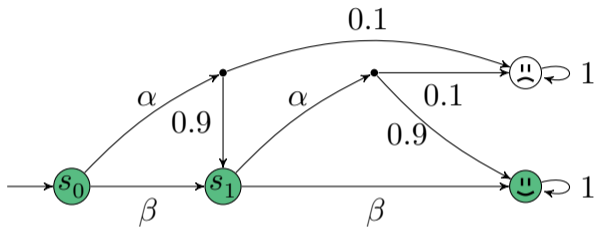
Picture: Könighofer et al., “Shields for Safe Reinforcement Learning”

Qualitative Shielding on an MDP



Ensure: Agent crashes **with probability zero**

Qualitative Shielding on an MDP

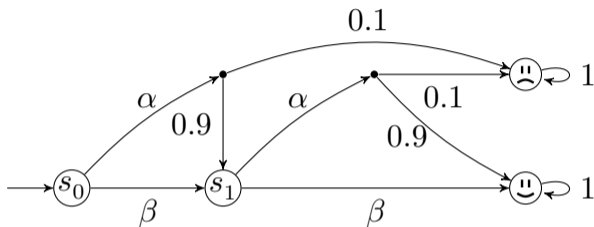


Ensure: Agent crashes **with probability zero**

\Rightarrow **there is a straightforward winning set.**

Shielding strategy: Precompute the winning set and ensure the agent stays in it.

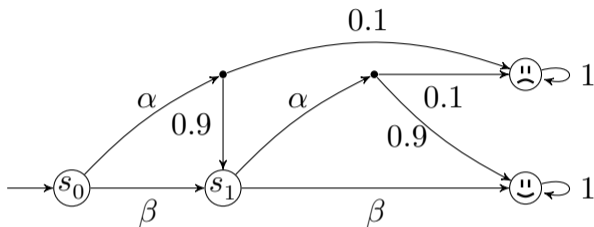
Probabilistic Shielding



Ensure: Agent crashes with probability at most 0.1 ($\Pr(s_0 \models \diamond \text{sad}) \leq 0.1$)

Q: Should we allow taking α at s_1 ?

Probabilistic Shielding



Ensure: Agent crashes with probability at most 0.1 ($\Pr(s_0 \models \diamond \text{sad face}) \leq 0.1$)

Q: Should we allow taking α at s_1 ?

A: It depends on what the policy played at s_0 !

\Rightarrow It's not that easy if we want to enforce a probabilistic guarantee.

Shielding with Guarantees

Two desirable properties:

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

Is a shield with these guarantees possible?

- ▶ $\varphi = \Pr(s_0 \models \diamond \neg \text{Bad}) \leq 0$: Yes! (Alshiekh et al., “Safe Reinforcement Learning via Shielding”)
- ▶ $\varphi = \Pr(s_0 \models \diamond \neg \text{Bad}) \leq \nu$ for $\nu > 0$: **No!** (Our result, in a few slides)

Shields: Definitions

A **history** is a sequence $h := s_0 d_1 \alpha_1 s_1 d_2 \alpha_2 \cdots s_t$ such that $s_0 \alpha_1 s_1 \alpha_2 \cdots s_t$ is a path, and for all $0 \leq k \leq t - 1$: $d_{k+1} \in \Delta(\mathbf{Act}(s_k))$.

Shields: Definitions

A **history** is a sequence $h := s_0 d_1 \alpha_1 s_1 d_2 \alpha_2 \cdots s_t$ such that $s_0 \alpha_1 s_1 \alpha_2 \cdots s_t$ is a path, and for all $0 \leq k \leq t - 1$: $d_{k+1} \in \Delta(\text{Act}(s_k))$.

A **shield** is a function $\square : \text{Hist}(\mathcal{M}) \times \Delta(\text{Act}) \rightarrow \Delta(\text{Act})$.

Shields: Definitions

A **history** is a sequence $h := s_0 d_1 \alpha_1 s_1 d_2 \alpha_2 \cdots s_t$ such that $s_0 \alpha_1 s_1 \alpha_2 \cdots s_t$ is a path, and for all $0 \leq k \leq t - 1$: $d_{k+1} \in \Delta(\text{Act}(s_k))$.

A **shield** is a function $\square : \text{Hist}(\mathcal{M}) \times \Delta(\text{Act}) \rightarrow \Delta(\text{Act})$.

Given a shield \square , its **policy transformer** is a function $\mathcal{T}_\square : \Pi_{\mathcal{M}} \rightarrow \Pi_{\mathcal{M}}$.

Shields: Definitions

A **history** is a sequence $h := s_0 d_1 \alpha_1 s_1 d_2 \alpha_2 \cdots s_t$ such that $s_0 \alpha_1 s_1 \alpha_2 \cdots s_t$ is a path, and for all $0 \leq k \leq t - 1$: $d_{k+1} \in \Delta(\text{Act}(s_k))$.

A **shield** is a function $\sqsupset : \text{Hist}(\mathcal{M}) \times \Delta(\text{Act}) \rightarrow \Delta(\text{Act})$.

Given a shield \sqsupset , its **policy transformer** is a function $\mathcal{T}_{\sqsupset} : \Pi_{\mathcal{M}} \rightarrow \Pi_{\mathcal{M}}$.

A shield \sqsupset **allows** a policy π if $\mathcal{T}_{\sqsupset}(\pi) = \pi$.

Shields: Definitions

A **history** is a sequence $h := s_0 d_1 \alpha_1 s_1 d_2 \alpha_2 \cdots s_t$ such that $s_0 \alpha_1 s_1 \alpha_2 \cdots s_t$ is a path, and for all $0 \leq k \leq t - 1$: $d_{k+1} \in \Delta(\text{Act}(s_k))$.

A **shield** is a function $\square : \text{Hist}(\mathcal{M}) \times \Delta(\text{Act}) \rightarrow \Delta(\text{Act})$.

Given a shield \square , its **policy transformer** is a function $\mathcal{T}_\square : \Pi_{\mathcal{M}} \rightarrow \Pi_{\mathcal{M}}$.

A shield \square **allows** a policy π if $\mathcal{T}_\square(\pi) = \pi$.

We write $\text{Allow}(\square) := \{\pi \in \Pi_{\mathcal{M}} \mid \mathcal{T}_\square(\pi) = \pi\}$.

Strong Guarantees

Let \mathcal{M} be an MDP and \square a shield.

(S+): Strong Safety. $\mathcal{T}_{\square}(\Pi_{\mathcal{M}}) \subseteq \text{Safe}(\Pi_{\mathcal{M}})$.

(P+): Strong Permissiveness. $\text{Safe}(\Pi_{\mathcal{M}}) \subseteq \text{Allow}(\square)$.

Strong Guarantees

Let \mathcal{M} be an MDP and \square a shield.

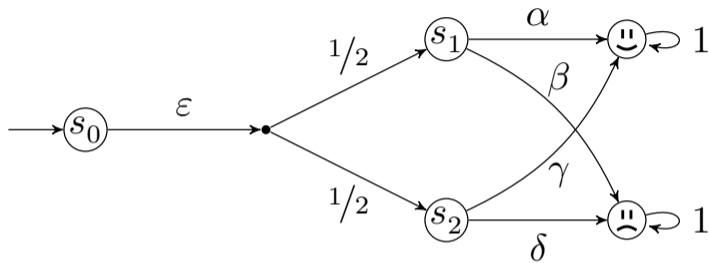
(S+): Strong Safety. $\mathcal{T}_{\square}(\Pi_{\mathcal{M}}) \subseteq \text{Safe}(\Pi_{\mathcal{M}})$.

(P+): Strong Permissiveness. $\text{Safe}(\Pi_{\mathcal{M}}) \subseteq \text{Allow}(\square)$.

Theorem. For safety threshold $\nu = 0$, the qualitative shield satisfies **(S+)** and **(P+)**.

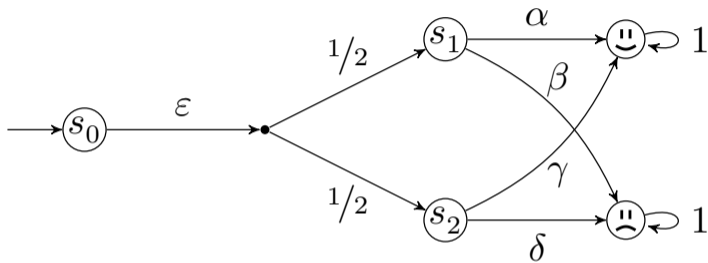
Theorem. For a safety threshold $0 < \nu < 1$, there is an (acyclic, 5-state) MDP such that no shield satisfies **(S+)** and **(P+)**.

Proof: No shield satisfies **(S+)** and **(P+)**



Assume: \square satisfies **(S+)** and **(P+)** for $\nu = 0.5$.

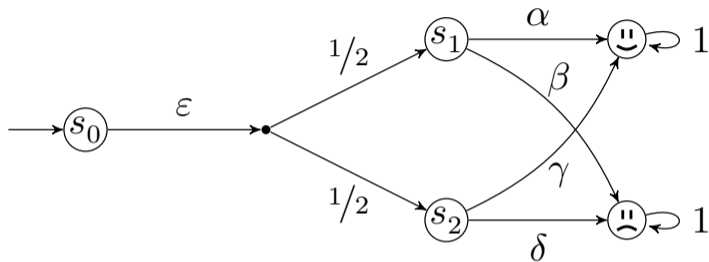
Proof: No shield satisfies **(S+)** and **(P+)**



Assume: \square satisfies **(S+)** and **(P+)** for $\nu = 0.5$.

Unsafe policy: $\pi := \{s_1 \mapsto \beta, s_2 \mapsto \delta\}$. \square cannot allow π ! (Otherwise: violate **(S+)**)

Proof: No shield satisfies **(S+)** and **(P+)**

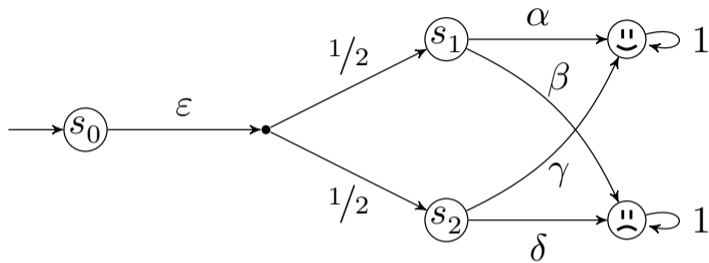


Assume: \square satisfies **(S+)** and **(P+)** for $\nu = 0.5$.

Unsafe policy: $\pi := \{s_1 \mapsto \beta, s_2 \mapsto \delta\}$. \square cannot allow π ! (Otherwise: violate **(S+)**)

\square blocks $\beta \Rightarrow \square$ blocks safe policy $\pi := \{s_1 \mapsto \beta, s_2 \mapsto \gamma\} \Rightarrow \square$ violates **(P+)**.

Proof: No shield satisfies **(S+)** and **(P+)**



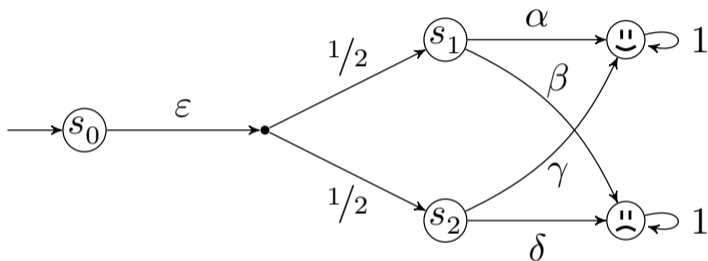
Assume: \square satisfies **(S+)** and **(P+)** for $\nu = 0.5$.

Unsafe policy: $\pi := \{s_1 \mapsto \beta, s_2 \mapsto \delta\}$. \square cannot allow π ! (Otherwise: violate **(S+)**)

\square blocks $\beta \Rightarrow \square$ blocks safe policy $\pi := \{s_1 \mapsto \beta, s_2 \mapsto \gamma\} \Rightarrow \square$ violates **(P+)**.

\square blocks $\delta \Rightarrow$ symmetric argument.

The Best Safe Shields on this MDP



Intuitively, “good safe shields” on this MDP for $\nu = 0.5$ block **either** β **or** δ .

We’ll formalize this notion.

Saturated Permissiveness

Guarantee (P*): Saturated Permissiveness. For all $\pi \in \Pi_{\mathcal{M}} \setminus \text{Allow}(\sqcup)$, there is no safe shield \sqcup' such that $\text{Allow}(\sqcup) \cup \{\pi\} \subseteq \text{Allow}(\sqcup')$.

Saturated Permissiveness

Guarantee (P*): Saturated Permissiveness. For all $\pi \in \Pi_{\mathcal{M}} \setminus \text{Allow}(\sqcup)$, there is no safe shield \sqcup' such that $\text{Allow}(\sqcup) \cup \{\pi\} \subseteq \text{Allow}(\sqcup')$.

Lemma. Allow sets of distinct saturated shields are incomparable.

Saturated Permissiveness

Guarantee (P*): Saturated Permissiveness. For all $\pi \in \Pi_{\mathcal{M}} \setminus \text{Allow}(\sqcup)$, there is no safe shield \sqcup' such that $\text{Allow}(\sqcup) \cup \{\pi\} \subseteq \text{Allow}(\sqcup')$.

Lemma. Allow sets of distinct saturated shields are incomparable.

Theorem. Zorn's lemma implies that for any MDP and any $\nu \in [0, 1]$, a saturated shield exists.

(On acyclic MDPs, this is true without Zorn's lemma.)

Lattice of Shields

Idea: Order all shields by permissiveness.

Definition (Lattice of Shields). The **lattice of shields** is $L_{\square} := (\text{CS Shields}, \sqsubseteq)$ such that $\square \sqsubseteq \square'$ iff $\text{Allow}(\square) \subseteq \text{Allow}(\square')$.

Lattice of Shields

Idea: Order all shields by permissiveness.

Definition (Lattice of Shields). The **lattice of shields** is $L_{\sqsubseteq} := (\text{CS Shields}, \sqsubseteq)$ such that $\sqsubseteq \sqsubseteq \sqsupset'$ iff $\text{Allow}(\sqsubseteq) \subseteq \text{Allow}(\sqsupset')$.

Theorem. $L_{\sqsubseteq} = (\text{CS Shields}, \sqsubseteq)$ is a complete lattice, where for $\sqsubseteq, \sqsupset' \in L$:

$$(\sqsubseteq \sqcup \sqsupset')(h, d) = d \text{ iff } \sqsubseteq(h, d) = d \text{ or } \sqsupset'(h, d) = d,$$

$$(\sqsubseteq \sqcap \sqsupset')(h, d) = d \text{ iff } \sqsubseteq(h, d) = d \text{ and } \sqsupset'(h, d) = d.$$

Existence of Saturated Shields

Proof sketch. A saturated shield is a maximal element within the set of safe shields. Need to show: **Every chain of safe shields has a safe upper bound.**

Existence of Saturated Shields

Proof sketch. A saturated shield is a maximal element within the set of safe shields. Need to show: **Every chain of safe shields has a safe upper bound.**

- ▶ Let X be a chain (totally ordered set) of safe shields.

Existence of Saturated Shields

Proof sketch. A saturated shield is a maximal element within the set of safe shields. Need to show: **Every chain of safe shields has a safe upper bound.**

- ▶ Let X be a chain (totally ordered set) of safe shields.
- ▶ We claim that the upper bound $\cup_X := \bigsqcup X$ is safe.

Existence of Saturated Shields

Proof sketch. A saturated shield is a maximal element within the set of safe shields. Need to show: **Every chain of safe shields has a safe upper bound.**

- ▶ Let X be a chain (totally ordered set) of safe shields.
- ▶ We claim that the upper bound $\sqcup_X := \bigsqcup X$ is safe.
- ▶ Suppose that \sqcup_X is unsafe. Then for some policy in $\pi \in \mathcal{T}_{\sqcup_X}(\Pi)$, the probability to reach \perp under π is **strictly greater** than ν .

Existence of Saturated Shields

Proof sketch. A saturated shield is a maximal element within the set of safe shields. Need to show: **Every chain of safe shields has a safe upper bound.**

- ▶ Let X be a chain (totally ordered set) of safe shields.
- ▶ We claim that the upper bound $\sqcup_X := \bigsqcup X$ is safe.
- ▶ Suppose that \sqcup_X is unsafe. Then for some policy in $\pi \in \mathcal{T}_{\sqcup_X}(\Pi)$, the probability to reach \perp under π is **strictly greater** than ν .
- ▶ Strict violation can be witnessed by a finite set of history-choice pairs.

Existence of Saturated Shields

Proof sketch. A saturated shield is a maximal element within the set of safe shields. Need to show: **Every chain of safe shields has a safe upper bound.**

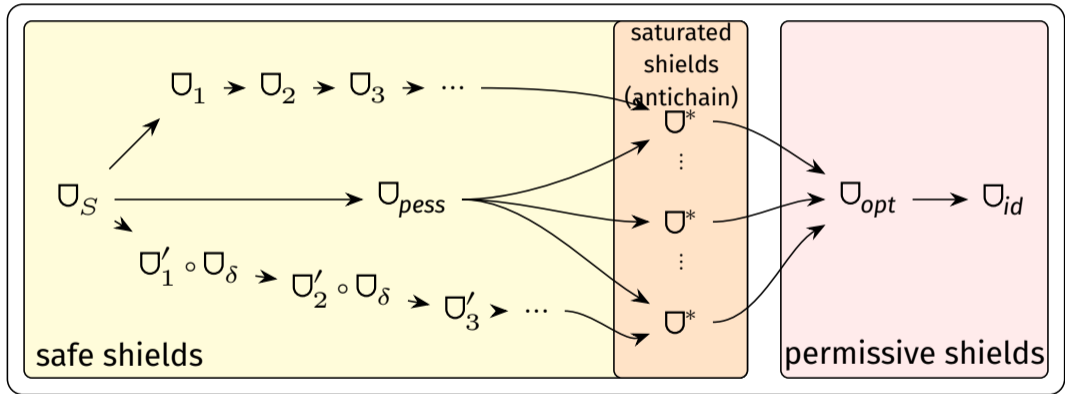
- ▶ Let X be a chain (totally ordered set) of safe shields.
- ▶ We claim that the upper bound $\sqcup_X := \bigsqcup X$ is safe.
- ▶ Suppose that \sqcup_X is unsafe. Then for some policy in $\pi \in \mathcal{T}_{\sqcup_X}(\Pi)$, the probability to reach \perp under π is **strictly greater** than ν .
- ▶ Strict violation can be witnessed by a finite set of history-choice pairs.
- ▶ Each of these history-choice pairs is allowed by some shield in X .

Existence of Saturated Shields

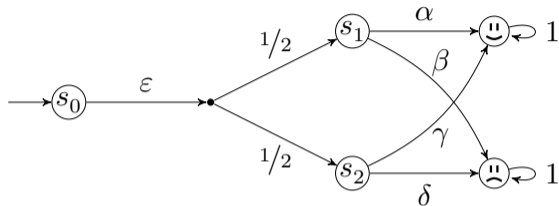
Proof sketch. A saturated shield is a maximal element within the set of safe shields. Need to show: **Every chain of safe shields has a safe upper bound.**

- ▶ Let X be a chain (totally ordered set) of safe shields.
- ▶ We claim that the upper bound $\sqcup_X := \bigsqcup X$ is safe.
- ▶ Suppose that \sqcup_X is unsafe. Then for some policy in $\pi \in \mathcal{T}_{\sqcup_X}(\Pi)$, the probability to reach \perp under π is **strictly greater** than ν .
- ▶ Strict violation can be witnessed by a finite set of history-choice pairs.
- ▶ Each of these history-choice pairs is allowed by some shield in X .
- ▶ As X is totally ordered, some shield in X allows **all** of these history-choice pairs \implies Some shield in X is unsafe. Contradiction.

Lattice of Shields

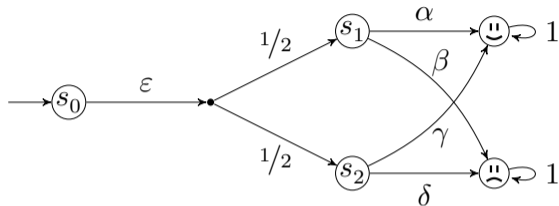


Constructed Shields: Intuition



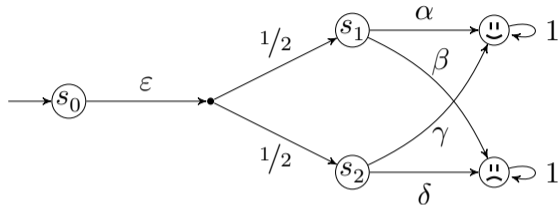
- Ensure $\varphi = \Pr(s_0 \models \diamond \text{sad}) \leq 0.5$

Constructed Shields: Intuition



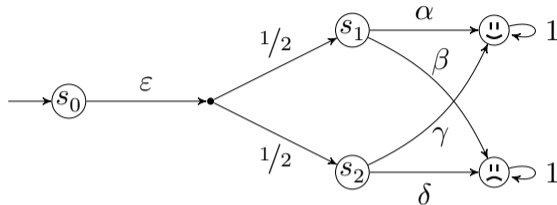
- ▶ Ensure $\varphi = \Pr(s_0 \models \diamond \text{smiley}) \leq 0.5$
- ▶ Start from shield \square_S only allowing α and γ

Constructed Shields: Intuition



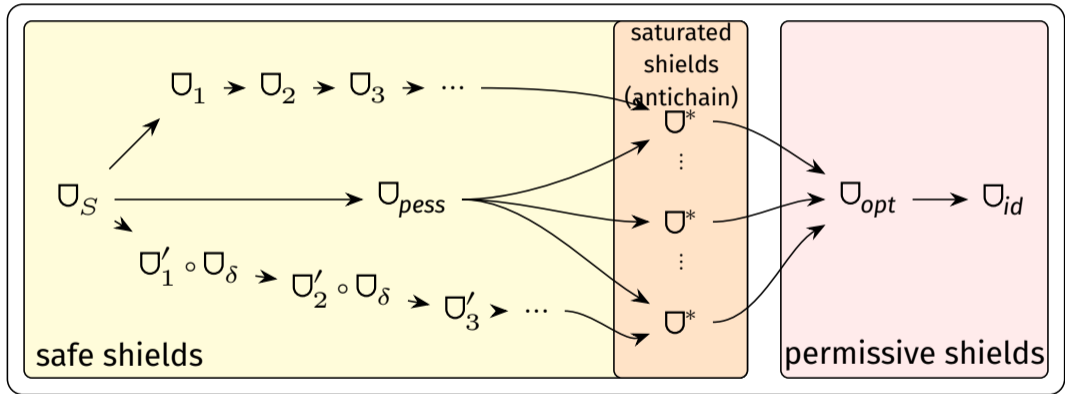
- ▶ Ensure $\varphi = \Pr(s_0 \models \diamond \text{sad}) \leq 0.5$
- ▶ Start from shield \square_S only allowing α and γ
- ▶ Policy tries to play $\beta \implies$ check for safety of additionally allowing β
- ▶ Allowing β is safe \implies permanently allow β

Constructed Shields: Intuition



- ▶ Ensure $\varphi = \Pr(s_0 \models \diamond \text{sad}) \leq 0.5$
- ▶ Start from shield \square_S only allowing α and γ
- ▶ Policy tries to play $\beta \implies$ check for safety of additionally allowing β
- ▶ Allowing β is safe \implies permanently allow β
- ▶ Policy tries to play $\delta \implies$ for safety of additionally allowing δ
- ▶ Allowing δ is unsafe \implies do not allow δ

Lattice of Shields



Safely Extending Shields

Definition (Safe Extension). Let $\square, \square' \in L_{\square}$. Then the **safe extension** of \square by \square' is the following (non-commutative and non-associative!) operation:

$$\square \oplus_{\text{safe}} \square' := \begin{cases} \square \sqcup \square' & \text{if } \square \sqcup \square' \text{ is safe,} \\ \square & \text{otherwise.} \end{cases}$$

Safely Extending Shields

Definition (Safe Extension). Let $\square, \square' \in L_{\square}$. Then the **safe extension** of \square by \square' is the following (non-commutative and non-associative!) operation:

$$\square \oplus_{\text{safe}} \square' := \begin{cases} \square \sqcup \square' & \text{if } \square \sqcup \square' \text{ is safe,} \\ \square & \text{otherwise.} \end{cases}$$

Safety can be checked in polynomial time.

Constructing Shields

Definition (Constructing Shields). Given a sequence of history-choice pairs $H = ((h_0, d_0), \dots, (h_t, d_t))$,

$$\sqcup_0 := \sqcup_S, \quad \sqcup_{i+1} := \sqcup_i \oplus_{\text{safe}} \sqcup_{(h_i, d_i)}.$$

- ▶ Constructed shields agree with some saturated shield on H .
- ▶ Constructed shields converge towards saturated permissiveness.

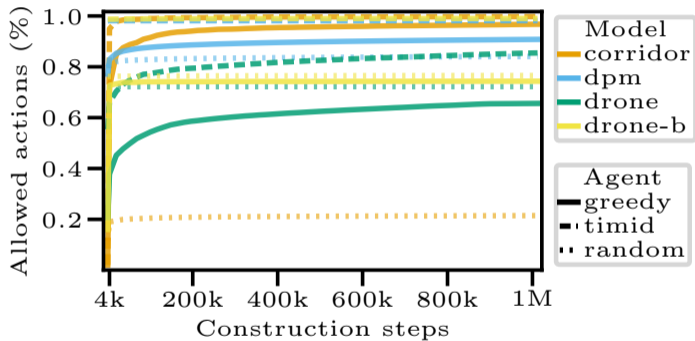
Online Shielding

- ▶ Online Shielding: Construct a shield **while** shielding
- ▶ When should we swap our shield with a more permissive one?

Online Shielding

- ▶ Online Shielding: Construct a shield **while** shielding
- ▶ When should we swap our shield with a more permissive one?
- ▶ One must be careful: Switching from one safe shield to another safe shield during an execution can break probabilistic guarantees!
- ▶ Solution: only modify shields between executions

Experiments: Convergence of Constructed Shields



Experiments: Safety and Permissiveness of Shields

Model	Agent (value)	ν	\square_{Safe}	\square_{δ^+}	\square_{onl}	\square_{off}	\square_{ML}					
			SAFE	UNSAFE	SAFE	SAFE	SAFE					
drone	greedy (.240)	0.01	.000	.207	.001	.211	.009	.214	.010	.218	.000	.208
		0.05	.000	.207	.001	.211	.049	.254	.050	.259	.000	.208
		0.2	.000	.207	.240	=1	.160	.538	.200	.656	.054	.239
	timid (.014)	0.01	.000	.554	.002	.728	.009	.808	.010	.830	.000	.554
		0.05	.000	.554	.002	.728	.008	.811	.011	.855	.000	.554
		0.2	.000	.554	.014	=1	.009	.810	.011	.855	.001	.640
	random (.962)	0.01	.000	.722	.052	.839	.000	.721	.000	.722	.000	.722
		0.05	.000	.722	.052	.839	.000	.723	.000	.722	.006	.749
		0.2	.000	.722	.052	.839	.000	.722	.000	.722	.027	.812

Experiments: Shield Queries per Second

Model	$ M $	\cup_S	\cup_{δ^+}	\cup_{δ}	\cup_{opt}	\cup_{pes}	\cup_{onl}	\cup_{off}	\cup_{nom}
corridor	15	2.5k	2.0k	2.0k	1.9k	1.9k	1.7k	2.0k	2.0k
dpm	797	2.8k	2.8k	2.8k	2.4k	2.6k	1.4k	2.2k	2.2k
drone	1859	2.5k	1.8k	1.8k	1.8k	2.3k	0.6k	1.6k	1.6k
drone-b	87k	1.9k	1.3k	1.3k	1.3k	1.5k	0.4k	1.3k	1.3k

Takeaways

- ▶ Shielding is a way to enforce safety for agents in MDPs
- ▶ It's harder with probabilistic guarantees
- ▶ Still, constructed shields can give you good shields in practice!

Appendix

Value of a Shield

Definition (Value of Shield). Given a finite set of history-choice pairs $J \subseteq \text{Hist}(\mathcal{M}) \times \Delta(\text{Act})$, let $\square = \bigsqcup_{(h,d) \in J} \square_{(h,d)}$. We define **the value $V_{\square}(h)$ of \square for history $h \in \text{Hist}(\mathcal{M})$ as:**

$$V_{\square}(h) := \max \{C_{\square}(h, d) \mid (h, d) \in J'\} \cup \{V_{\min}(\text{last}(h))\}, \text{ where}$$
$$J' := J \cup \{(h, d \mid_{\text{Act}^{\text{Safe}}(\text{last}(h))}) \mid (h', d) \in J \text{ s.t. } h' \text{ suffix of } h\}, \text{ and}$$
$$C_{\square}(h, d) := \sum_{\alpha \in \text{Supp}(d)} d(\alpha) \sum_{s' \in S} \mathcal{P}(\text{last}(h), \alpha, s') \cdot V_{\square}(h \cdot d \cdot \alpha \cdot s').$$

Shields and their Allowed Policies

Given a set of policies $P \subseteq \Pi_{\mathcal{M}}$, the **history mixing** is the set $\text{mix}(P) \subseteq \Pi_{\mathcal{M}}$ such that $\pi \in \text{mix}(P)$ if:

$$\forall h \in \text{Hist}(\mathcal{M}). h \text{ is consistent with } \pi \implies (\exists \pi' \in P. h \text{ is consistent with } \pi').$$

Lemma. Given shield \square and policies $P \subseteq \text{Allow}(\square)$. Then $\text{mix}(P) \subseteq \text{Allow}(\square)$.

Lemma. Let P be a mix-closed set of policies that **covers all safe actions**. There is a unique **canonical** shield \square with $\text{Allow}(\square) = P$.

Q1: Safety and Permissiveness of Shields

Model	Agent (value)	ν	\mathbb{P}_{Safe} SAFE	\mathbb{P}_{δ^+} UNSAFE	\mathbb{P}_{onl} SAFE	\mathbb{P}_{off} SAFE	\mathbb{P}_{ML} SAFE					
corridor	greedy (.125)	0.05	.000	.020	.000	.020	.000	.020	.000	.020	.000	.020
		0.1	.000	.020	.000	.020	.100	.204	.100	.202	.000	.020
		0.2	.000	.020	.125	.735	.125	.936	.125	.968	.125	.735
	timid (.125)	0.05	.000	.020	.000	.020	.000	.020	.000	.020	.000	.020
		0.1	.000	.020	.000	.020	.100	.406	.100	.406	.000	.020
		0.2	.000	.020	.125	=1	.124	.993	.125	.997	.125	.925
	random (.733)	0.05	.000	.116	.000	.116	.049	.134	.050	.142	.000	.116
		0.1	.000	.116	.000	.116	.098	.153	.100	.148	.000	.116
		0.2	.000	.116	.000	.116	.199	.190	.200	.215	.000	.116
dpm	greedy (.479)	0.01	.000	.771	.479	=1	.010	.790	.010	.792	.009	.794
		0.05	.000	.771	.479	=1	.047	.839	.050	.848	.044	.844
		0.2	.000	.771	.479	=1	.062	.870	.097	.908	.155	.924
	timid (.063)	0.01	.000	.981	.063	=1	.000	.979	.000	.981	.000	.981
		0.05	.000	.981	.063	=1	.000	.979	.000	.981	.010	.984
		0.2	.000	.981	.063	=1	.000	.979	.000	.981	.063	=1
	random (.131)	0.01	.000	.798	.131	=1	.005	.831	.006	.841	.006	.839
		0.05	.000	.798	.131	=1	.005	.831	.006	.841	.037	.989
		0.2	.000	.798	.131	=1	.005	.831	.006	.841	.131	=1