Fear Field: Adaptive constraints for safe environment transitions in Shielded Reinforcement Learning

Odriozola-Olalde, H., Arana-Arexolaleiba, N., Zamalloa, M., Perez-Cerrolaza, J., & Arozamena-Rodríguez, J.

MONDRAGON UNIBERTSITATEA
GOI ESKOLA POLITEKNIKOA
FACULTY OF ENGINEERING

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
www.ikerlan.es

Index

Source: Wikimedia (CC BY-SA 3.0)

Problem statement

CONTROLLER RELATED TO THE TRAINING DATASET

AI-based controllers learn to solve the problem given a dataset.

TIME VARIANT ENVIRONMENTS

An autonomous system can find unexpected conditions.

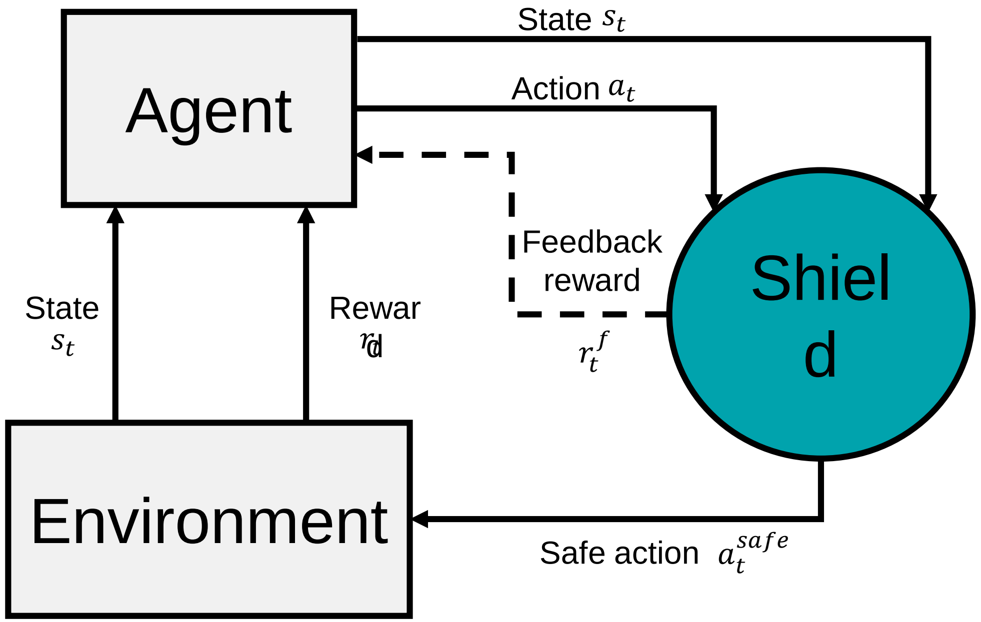
RISK ASSESSMENT

Safety specifications must be guaranteed even in unexpected conditions.

Shielded RL

Shielded Reinforcement Learning.

Shielded RL is a reactive method, i.e., it only corrects the agent's proposed action when it foresees that the action will lead the environment to violate the safety specifications (unsafe states).

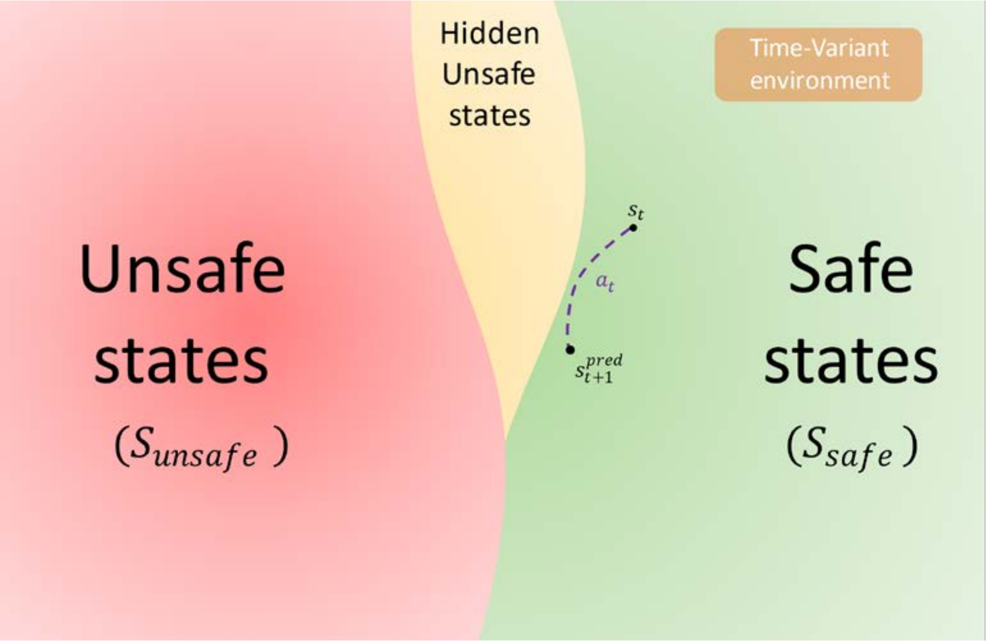


- Commonly, the shield uses a model of the environment dynamic.
- In time-variant environments, Shielded RL shows a shortcoming regarding its robustness.*

*Odriozola-Olalde, H., Zamalloa, M., & Arana-Arexolaleiba, N. (2023, January). Shielded Reinforcement Learning: A review of reactive methods for safe learning. In *2023 IEEE/SICE International Symposium on System Integration (SII)* (pp. 1-8). IEEE.

Shielded RL

Shielded Reinforcement Learning.

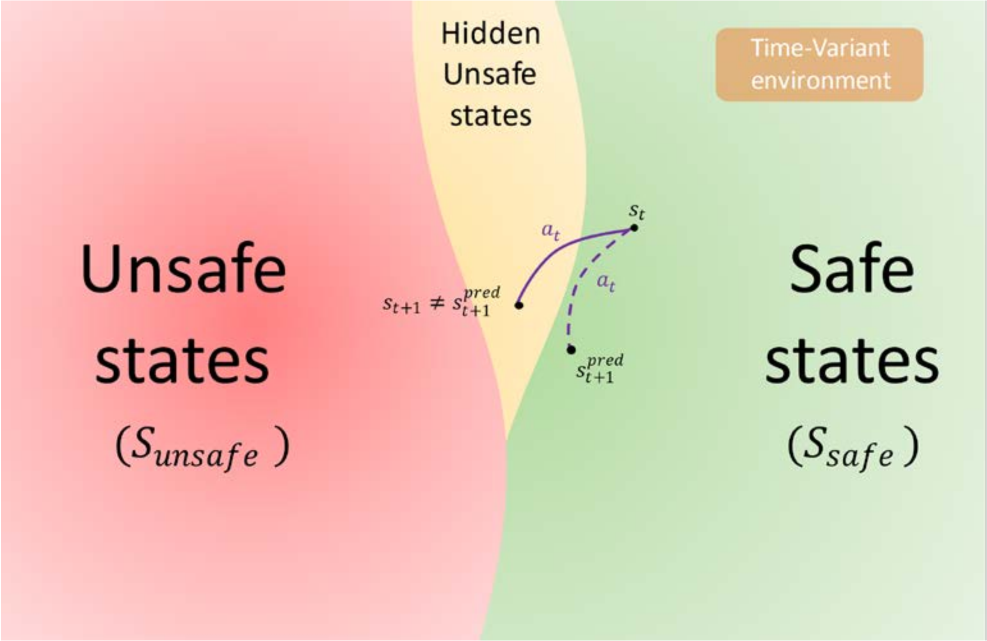


- Outdated environment dynamic model.

Odriozola-Olalde, H., Zamalloa, M., & Arana-Arexolaleiba, N. (2023, January). Shielded Reinforcement Learning: A review of reactive methods for safe learning. In *2023 IEEE/SICE International Symposium on System Integration (SII)* (pp. 1-8). IEEE.

Shielded RL

Shielded Reinforcement Learning.



- Outdated environment dynamic model.
- Until the model is updated, the predictions and the reached states can differ.
- During this period, previous safety guarantees are lost.

Odriozola-Olalde, H., Zamalloa, M., & Arana-Arexolaleiba, N. (2023, January). Shielded Reinforcement Learning: A review of reactive methods for safe learning. In *2023 IEEE/SICE International Symposium on System Integration (SII)* (pp. 1-8). IEEE.

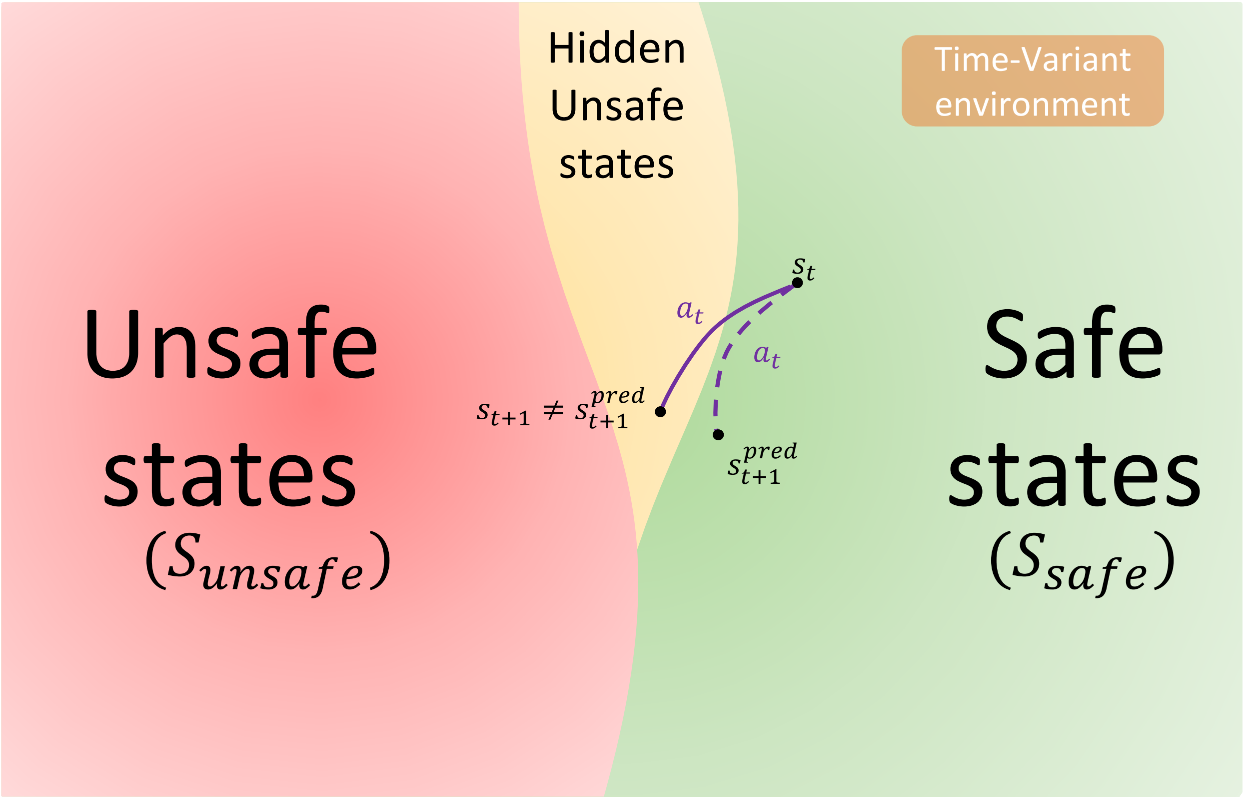
Proposed solution

Fear Field: Adaptive constraints in time-variant environments.

As humans adapt the caution measures according to our confidence and knowledge of the environment, Fear Field proposes adapting the safety constraints depending on the shield's confidence in the environment's model accuracy.

Proposed solution

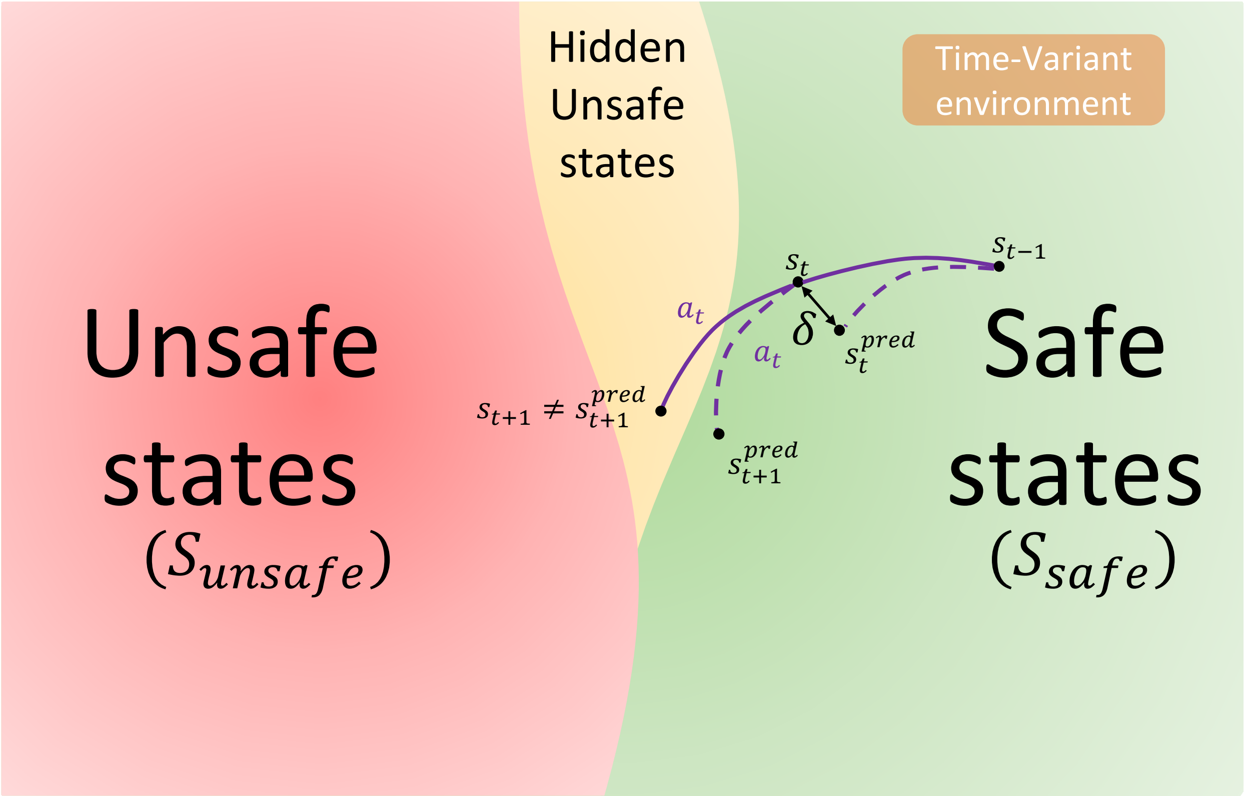
Fear Field: Adaptive constraints in time-variant environments.



- A significant change in the environment's dynamic makes the actual model outdated.

Proposed solution

Fear Field: Adaptive constraints in time-variant environments.

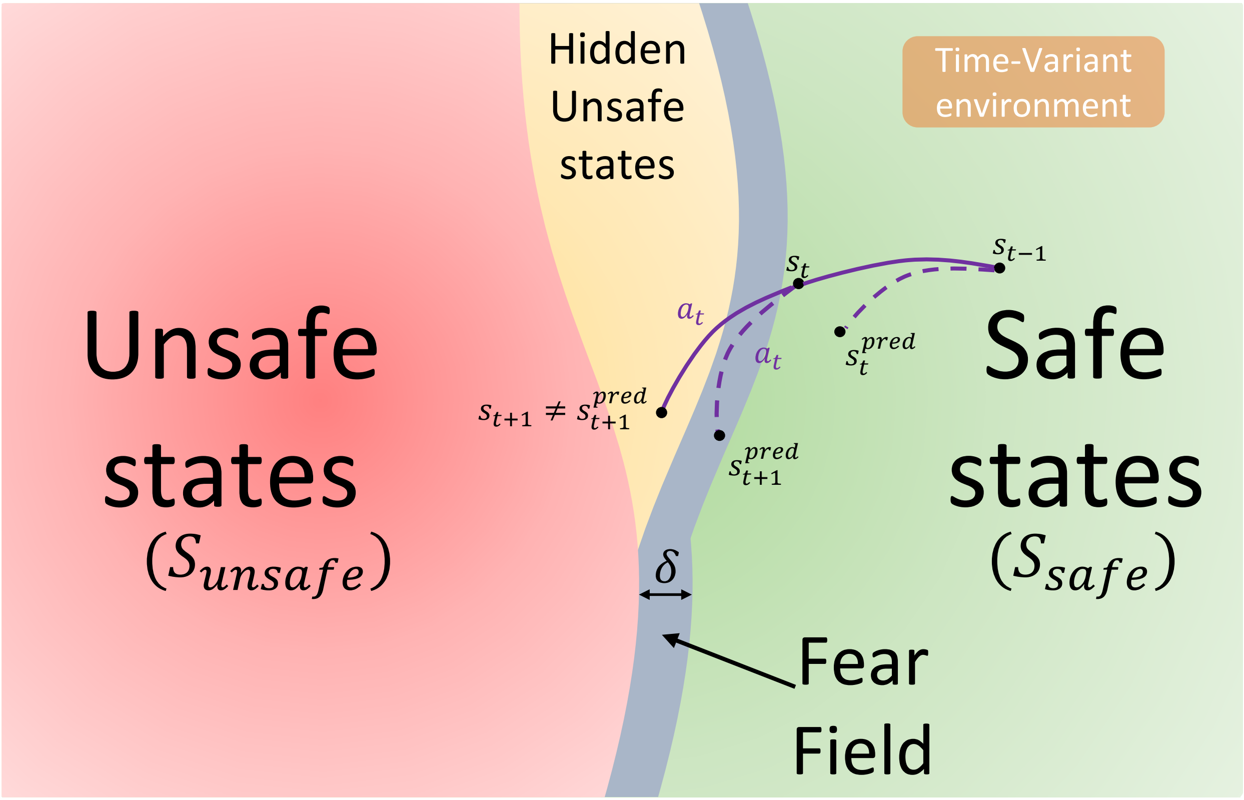


- A significant change in the environment's dynamic makes the actual model become outdated.
- The Euclidean distance between the previous step's predicted and reached states is computed.

$$\delta_{t+1}(s) \propto \left| s_t - s_t^{pred} \right|$$

Proposed solution

Fear Field: Adaptive constraints in time-variant environments.



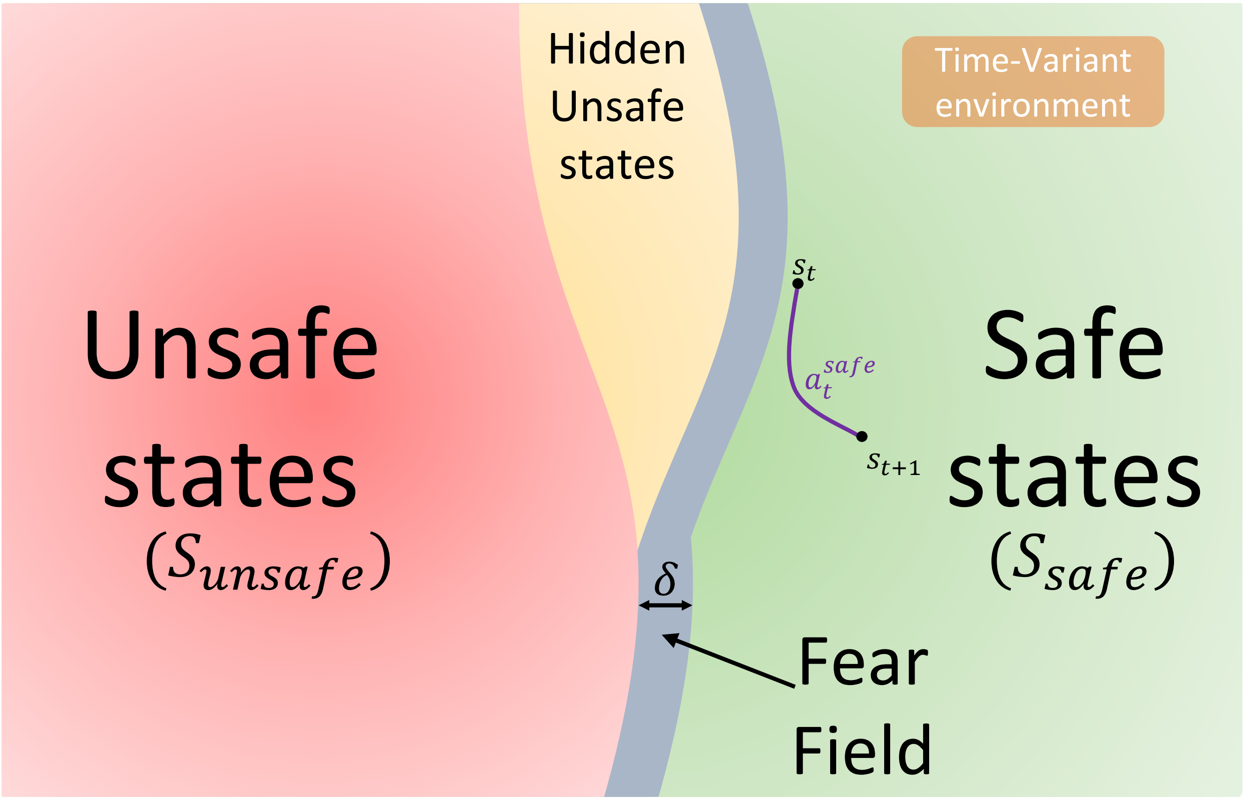
- A significant change in the environment's dynamic makes the actual model become outdated.
- The Euclidean distance (δ) between the previous step's predicted and reached states is computed.
- Safe state space is shrunked the computed δ distance, generating the Fear Field subspace.
- With the outdated dynamic model, the shield predicted state is now within the Fear Field.
- Even if the predicted state is safe, it is undesirable due to its proximity to an unsafe state.

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

MONDRAGON UNIBERTSITATEA GOI ESKOLA POLITEKNIKOA FACULTY OF ENGINEERING

Proposed solution

Fear Field: Adaptive constraints in time-variant environments.



- Therefore, the shield proposes an action a_t^{safe} that makes the environment transit to a safe state.

Proposed solution

Fear Field: Adaptive constraints in time-variant environments.

Model update

While the Fear Field is enabled, a dataset is sampled to update the environment's dynamic model. ($n_{Dataset}$)

Fear Field deactivation

If all the model predictions match the reached states in a defined continuous number of steps, the Fear Field is disabled. (n_{Steps})

Fear Field

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Experiments

Experiment setup.

- Modified versión (10x10 grid) of Frozen Lake, an OpenAI Gym-based GridWorld.
- Bi-dimensional discrete reach-avoid problem.
- Periodically, the world is slippery: the robot will move an additional square for the same action.

Experiments

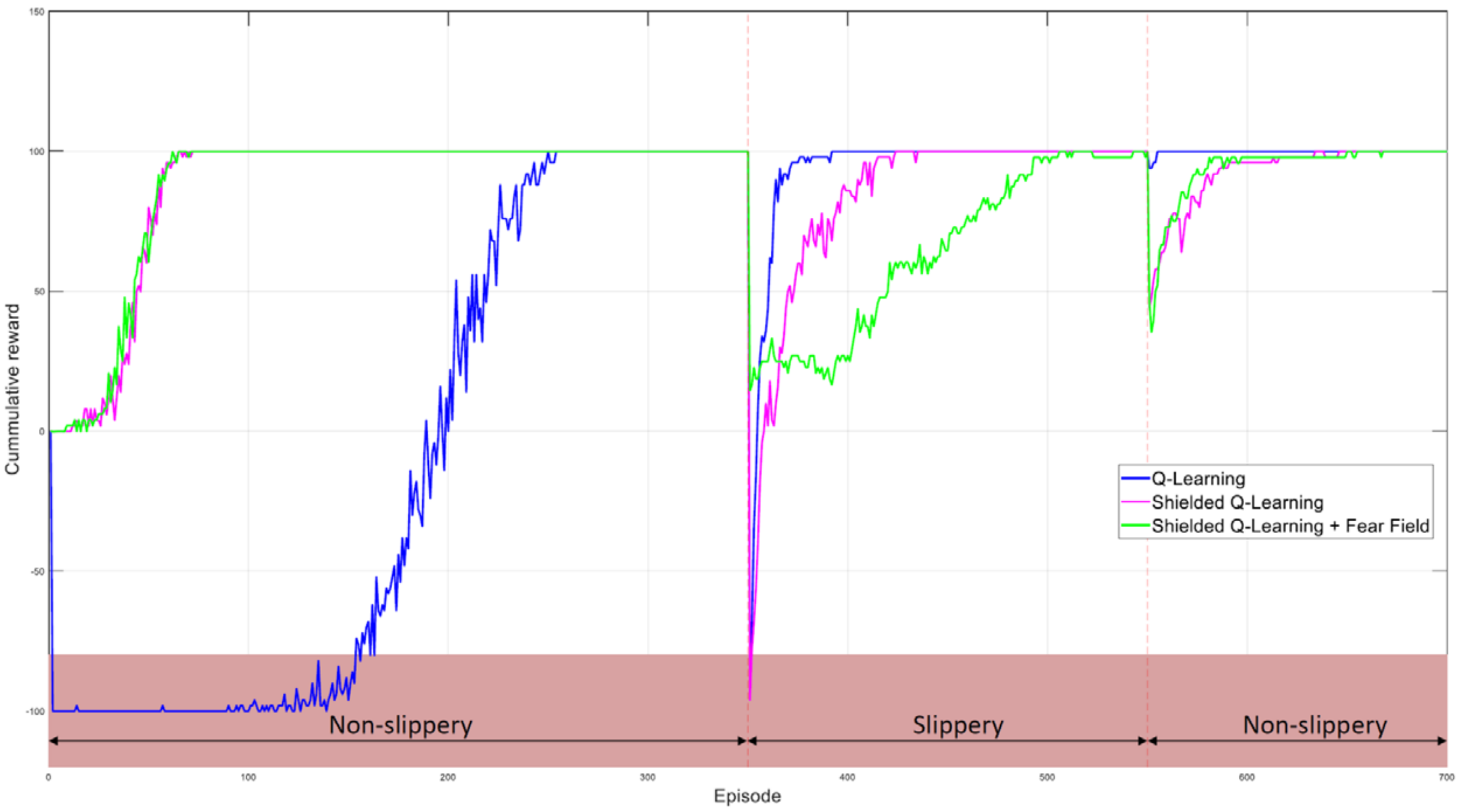
Experiment setup.

- Tabular Q-Learning algorithm (Skrl library) and shield with time horizon $h = 1$.
- A Neural Network based environment dynamic model.
- The Fear Field width is directly proportional to the difference between the predicted and reached states.

$$\delta_t(s) = \left| s_{t-1} - s_{t-1}^{pred} \right|$$

Experiments

Results.

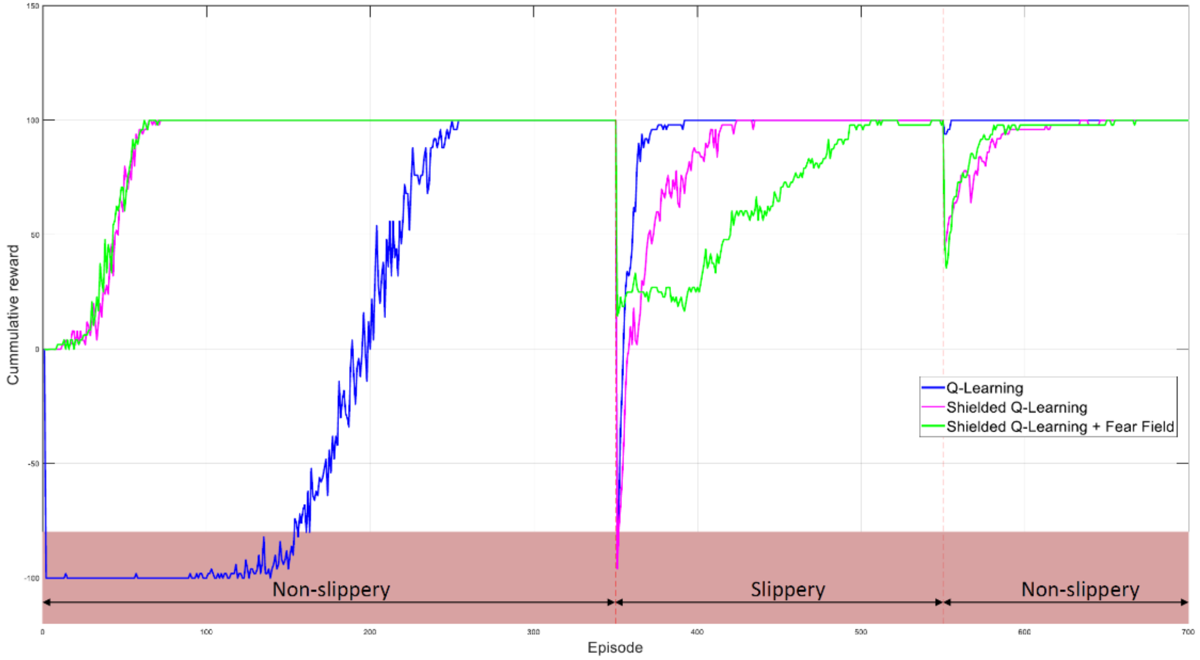


50 trials

Max. 100 steps per episode

Experiments

Results.

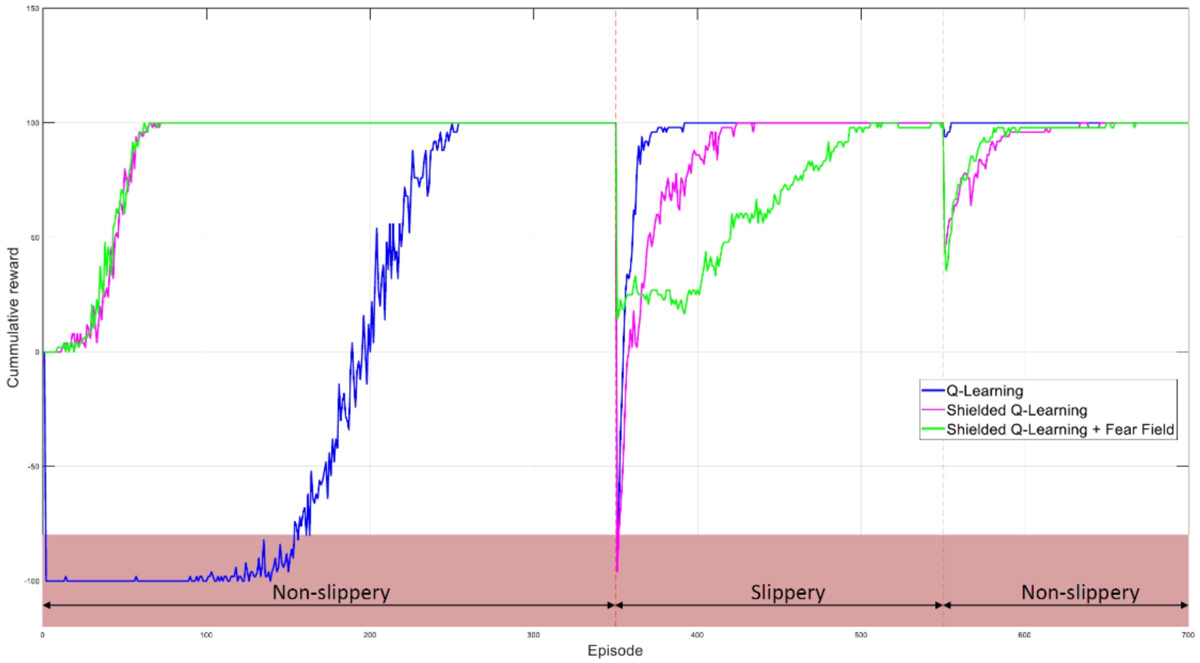


	Q-Learning	Shield	FF
Mean unsafe states	0,77192%	0.0156%	0.00179%

- With Shielded RL no unsafe states were reached during the first training process.
- In the slippery period (episodes 350-550), it can be observed that both Q-Learning and Shielded Q-Learning suffer from reaching unsafe states in the episodes immediately after the change made in the environment's dynamic (episode 350).

Experiments

Results.



- Fear Field significantly reduces, by one order of magnitude approximately, the unsafe state reached after the environment has changed. In the 60% of the trials performed, Fear Field obtained a null number of unsafe states.
- In the remaining 40% of trials, the reason for reaching unsafe states was that the NN was not trained correctly.
- The shortcoming of the Fear Field is that after transitioning to a previously unknown environment, the convergence time is higher than only Shielded RL.

	Q-Learning	Shield	FF
Mean unsafe states	0,77192%	0.0156%	0.00179%

Conclusions

Shielded RL

An interesting method for decision-making controllers due to its effectiveness in avoiding unsafe states. However, in time-variant environments its effectiveness is decreased.

Fear Field

Fear Field showed a significant improvement in reducing the unsafe states reached. Nonetheless, an increase in convergence time is observed.

Complex environments

The Fear Field algorithm must be tested and validated in stochastic, high-dimensional, continuous environments closer to real problems.

Hyperparameters

Further research on how hyperparameters affect the safety constraint violation rate must be studied.

谢谢！

THANK YOU!

Haritz Odriozola Olalde

hodriozola@ikerlan.es

www.ikerlan.es

P.ª José María Arizmendiarrieta, 2 – 20500 Arrasate-Mondragón.

ikerlan

MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

MONDRAGON UNIBERTSITATEA GOI ESKOLA POLITEKNIKOA FACULTY OF ENGINEERING

Questions?

GOI ESKOLA POLITEKNIKOA
FACULTY OF ENGINEERING
MONDRAGON UNIBERTSITATEA

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

www.ikerlan.es
P.ª José María Arizmendiarrieta, 2 – 20500 Arrasate-Mondragón.