Formal Methods for Trusted AI

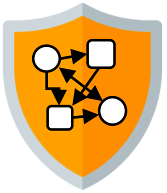
Bettina Könighofer
bettina.koenighofer@iaik.tugraz.at

May 25, 2023

Outline

- What are Formal Methods?

- **Shielding**

Idea  **Methodology**  **Application** 

Formal Methods

- **Formal methods** = Mathematical techniques and tools to model, design and analyze systems
- **Goal**: To **prove/guarantee** correctness
- **3 Categories:**
  1. **Specification**: **WHAT** the system must/must not do?

Specification $\boldsymbol{\varphi}$

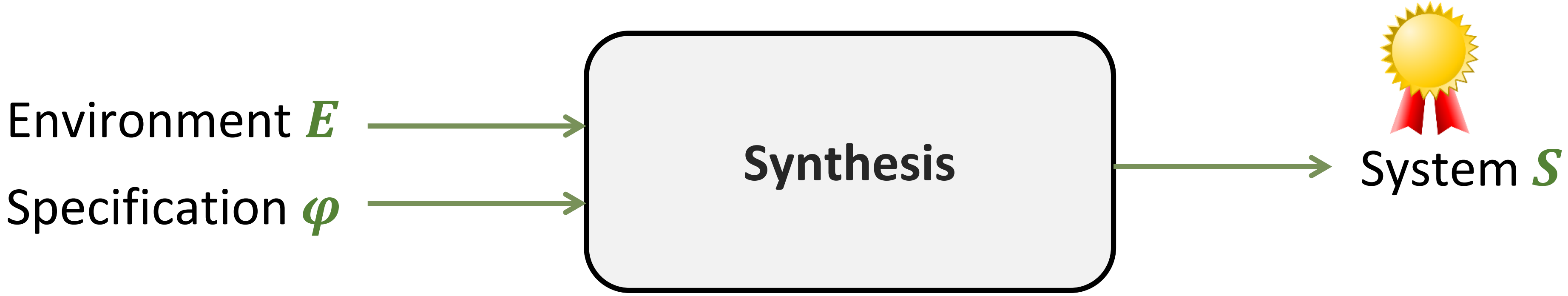
Always $\neg(\,grant_1 \wedge\ grant_2)$
Always $(\,request \rightarrow$ Next $grant)$
….

Formal Methods

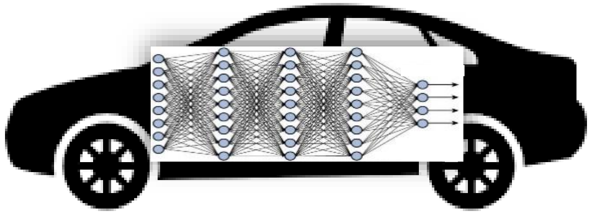
- **Formal methods** = Mathematical techniques and tools to model, design and analyze systems
- **Goal**: To **prove/guarantee** correctness
- **3 Categories:**
  1. **Specification**: **WHAT** the system must/must not do?
  2. **Verification**: DOES the system meet the specification? (and **WHY**?)

Formal Methods

- **Formal methods** = Mathematical techniques and tools to model, design and analyze systems
- **Goal**: To **prove/guarantee** correctness
- **3 Categories:**
  1. **Specification**: **WHAT** the system must/must not do?
  2. **Verification**: DOES the system meet the specification? (and **WHY**?)
  3. **Synthesis**: **HOW** it meets the specification (correct-by-construction design/synthesis)



Challenges of Deep Learning for Formal Methods



Environment $\boldsymbol{E}$ → Verification

System $\boldsymbol{S}$ → Verification

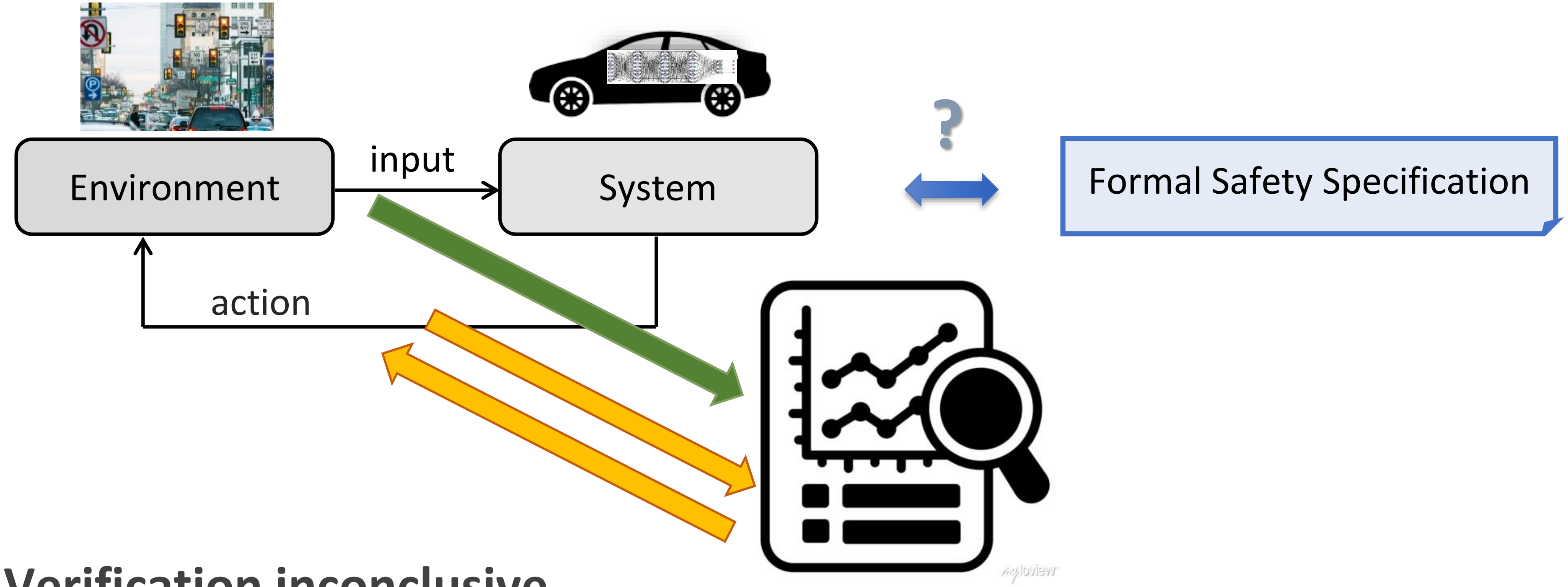
Specification $\boldsymbol{\varphi}$ → Verification

Verification

→ **Yes**

→ **No**

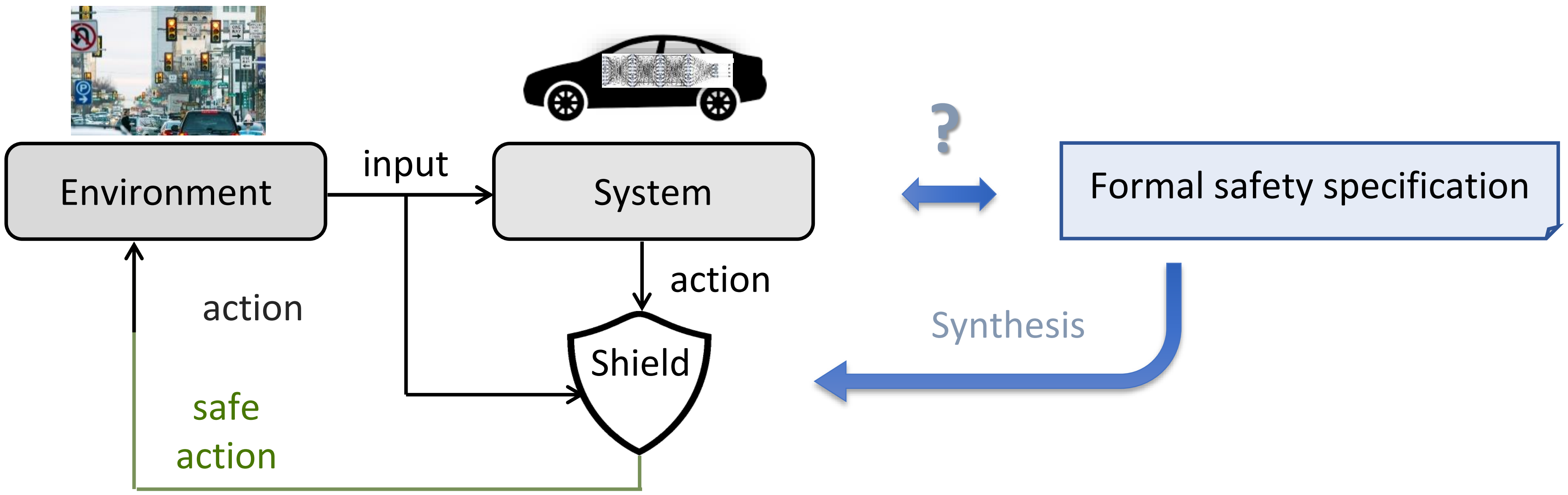
How to guarantee Safety?



Verification inconclusive

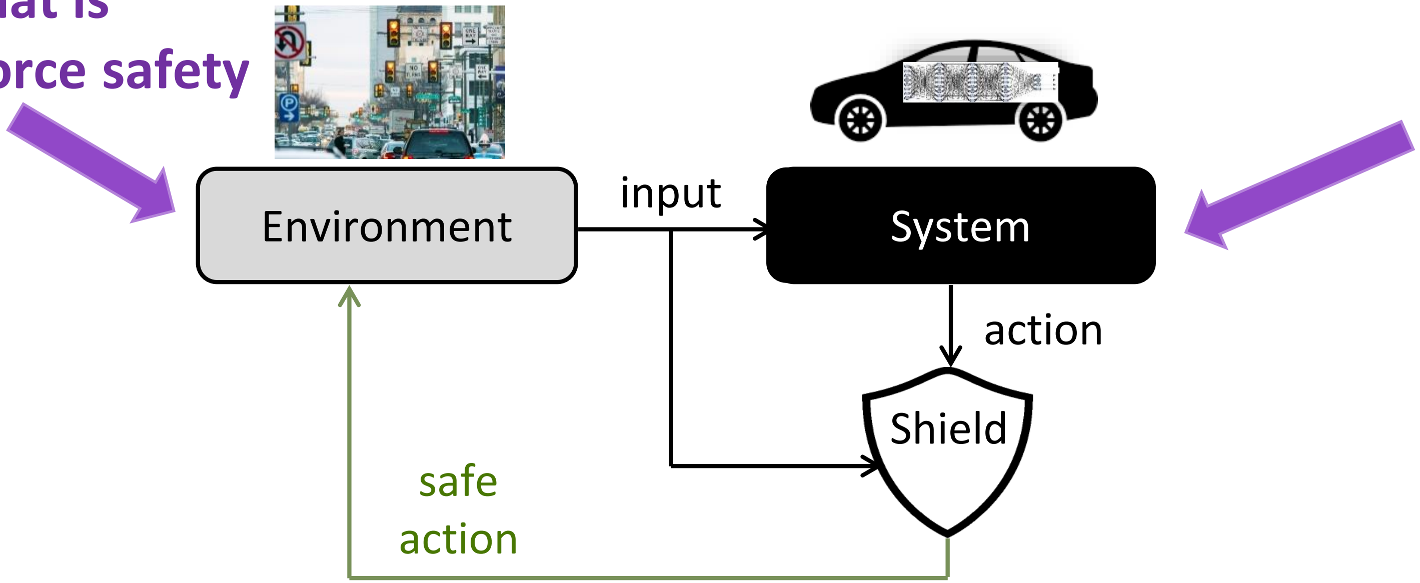
- System too complicated
- … but we need to have absolute certainty

Shielding – Correct-by-Construction Runtime Assurance



Shielding – Scalability

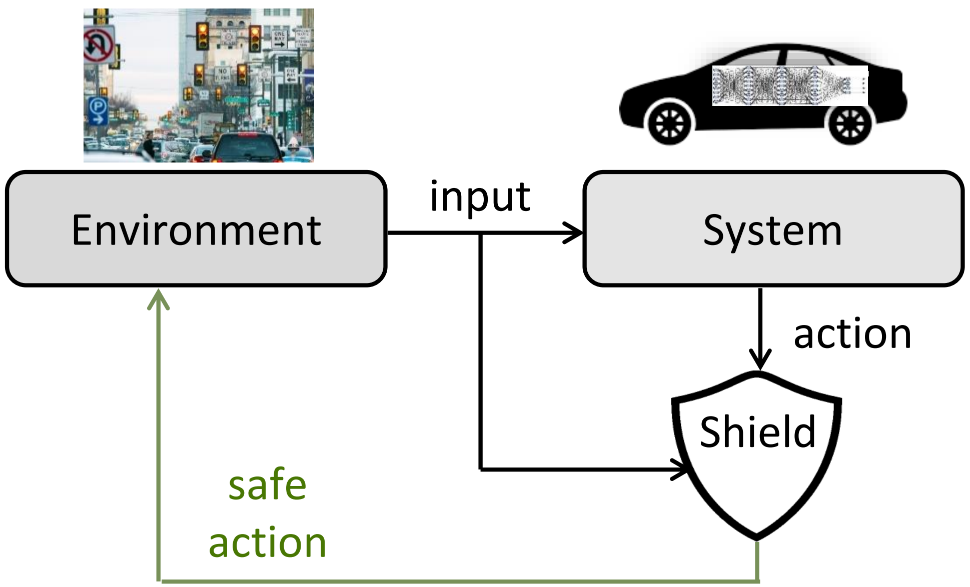
Only model what is needed to enforce safety



Black box

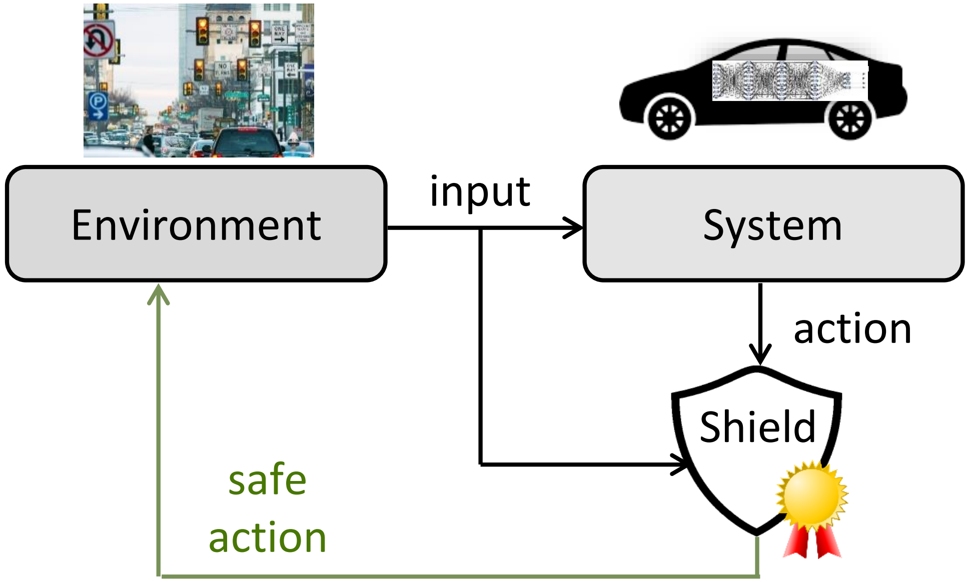
Shielding – Properties

1. **Shields guarantee correctness**
2. **Shields are minimal interfering**



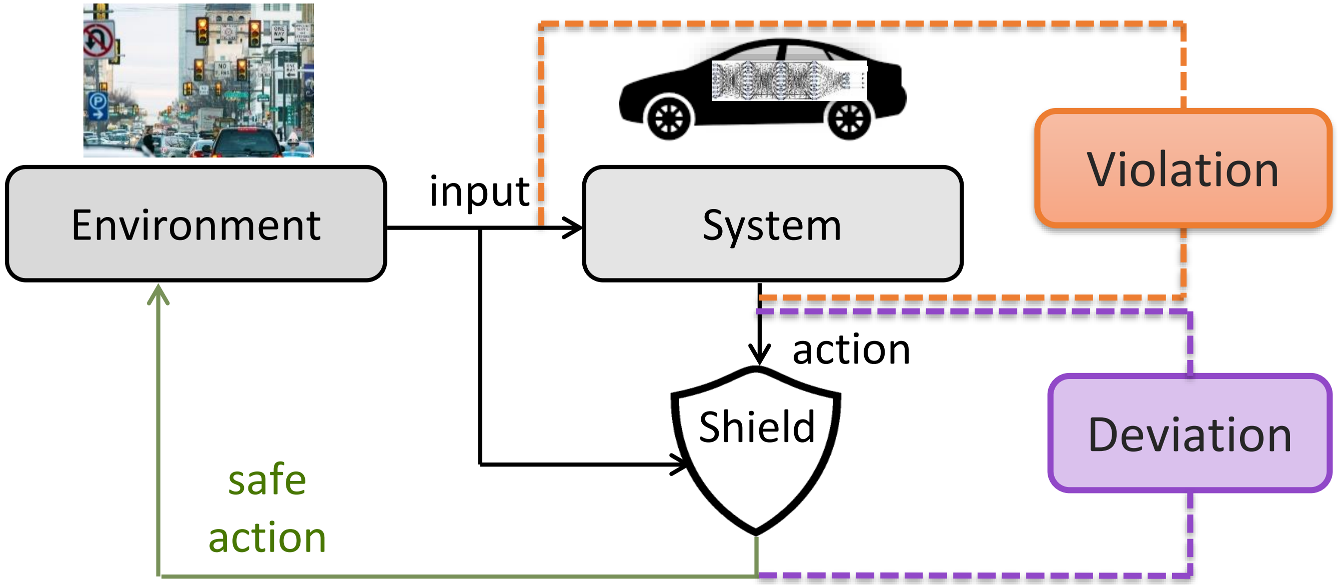
Shielding – Properties

1. **Shields guarantee correctness**
   - Correct-by-construction
   - Predictive

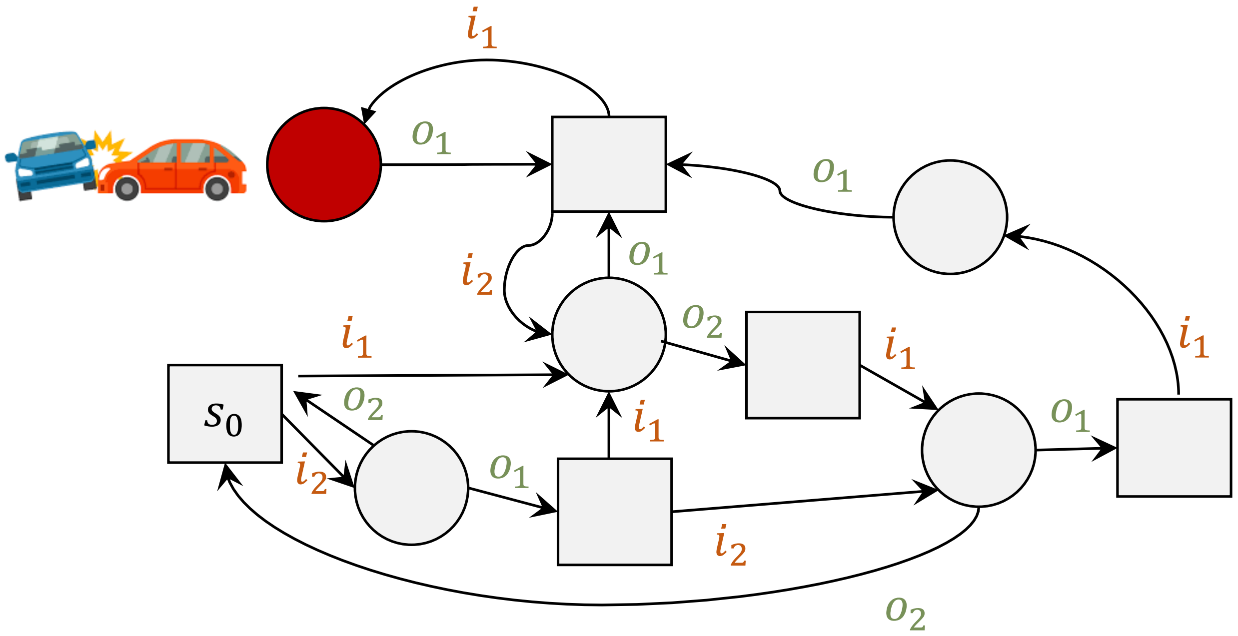


Shielding – Properties

1. **Shields guarantee correctness**
   - Correct-by-Construction
   - Predictive
2. **Shields are minimal interfering**
   - Only interfere when absolute necessary…
   - … and as little as possible



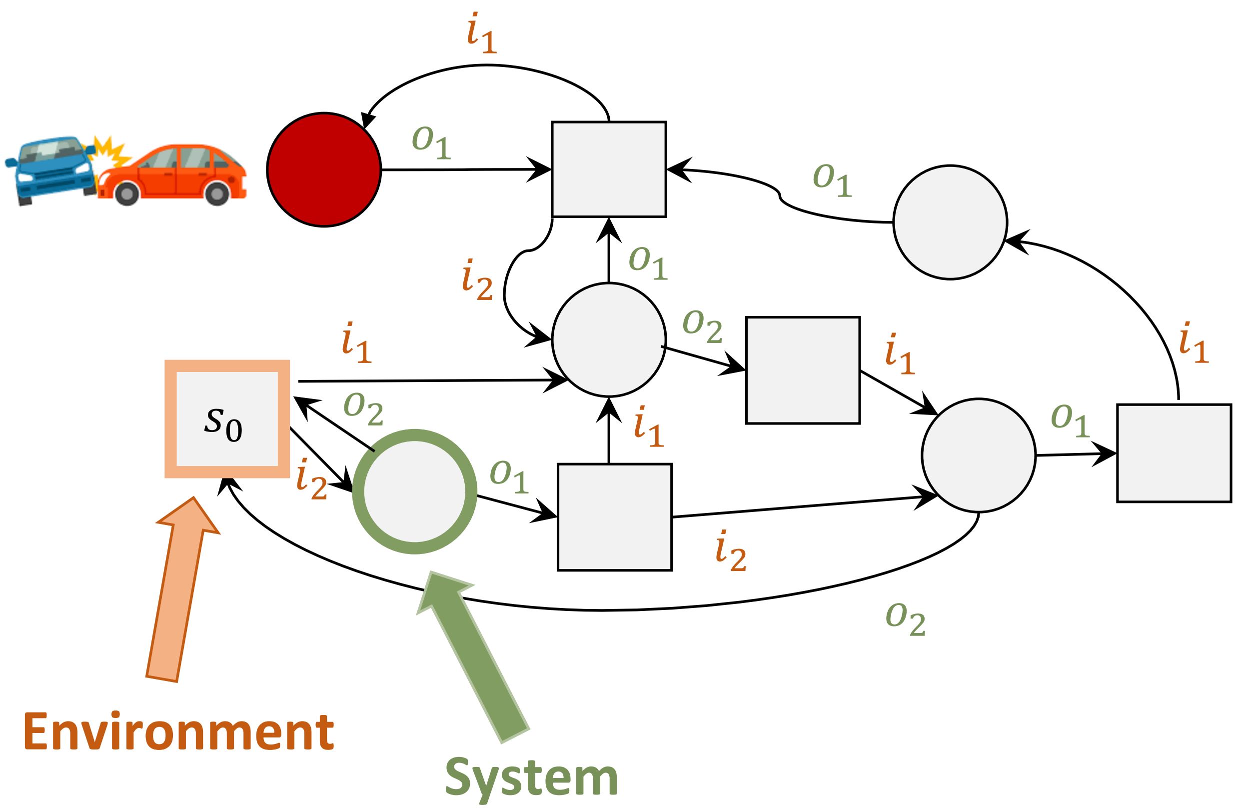
Shield Construction – Synthesis is a Game

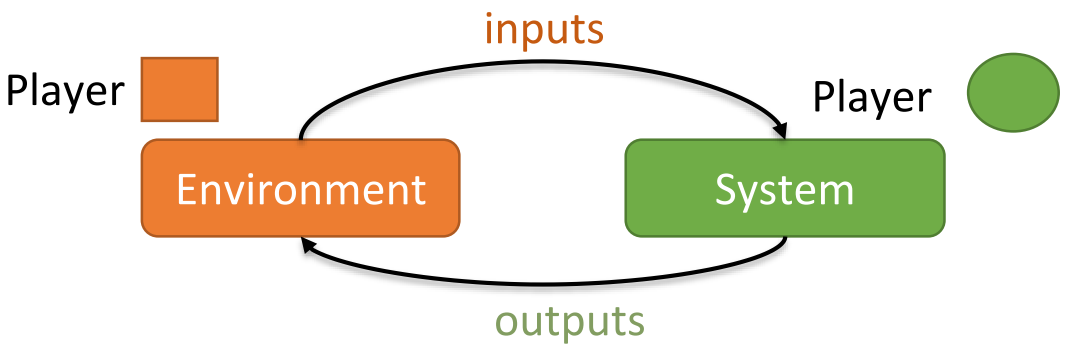


Formal safety specification

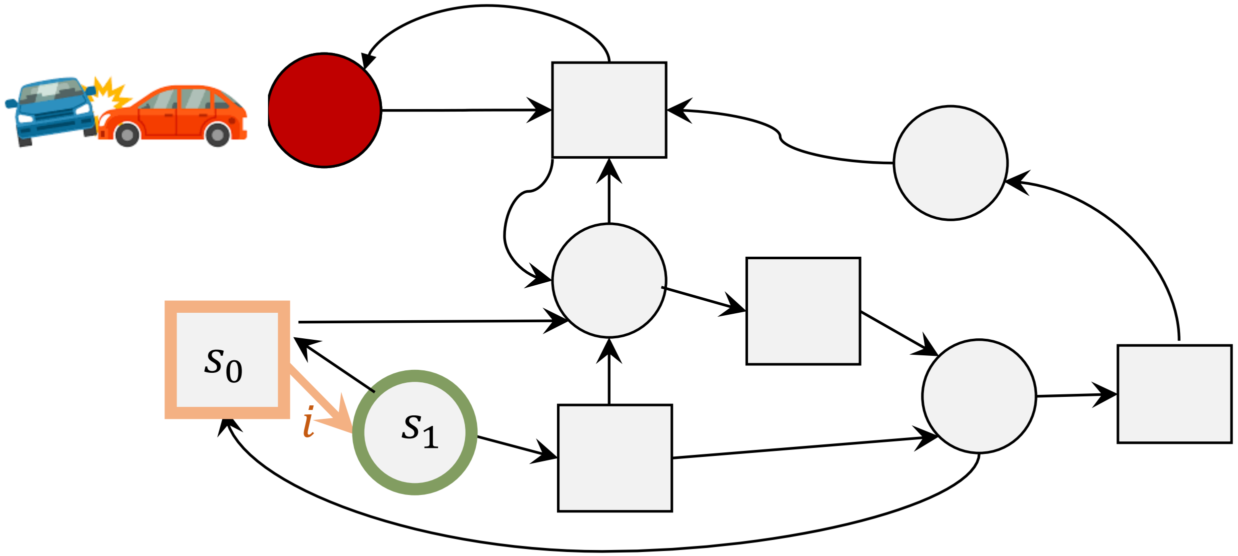
Model of environment

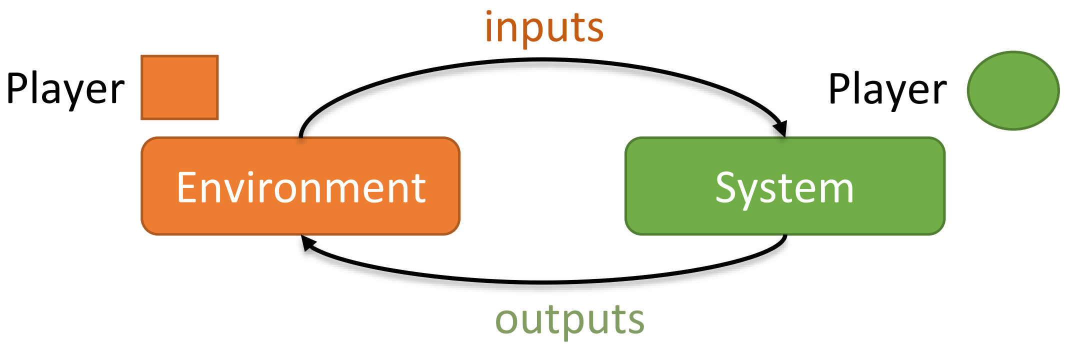
Shield Construction – Synthesis is a Game



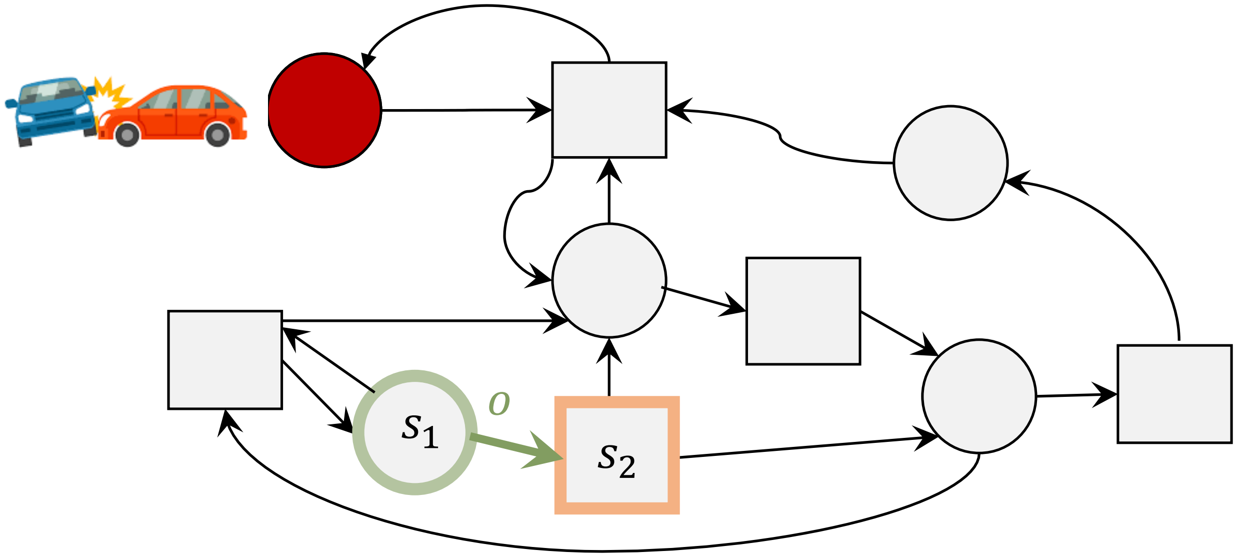


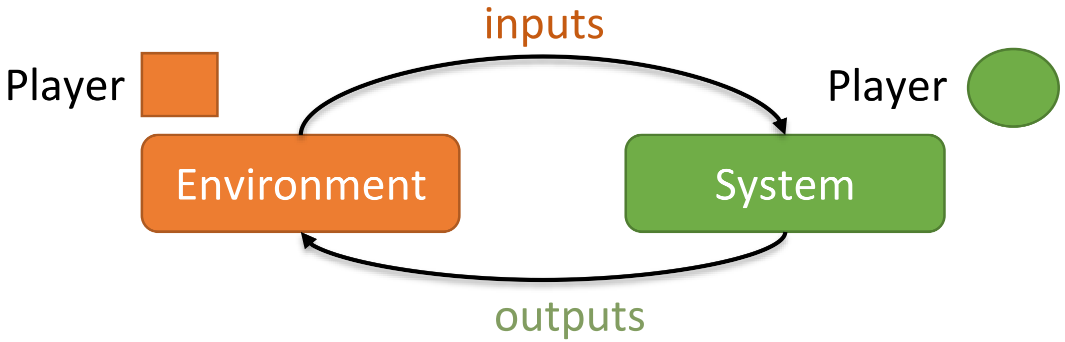
Shield Construction – Synthesis is a Game



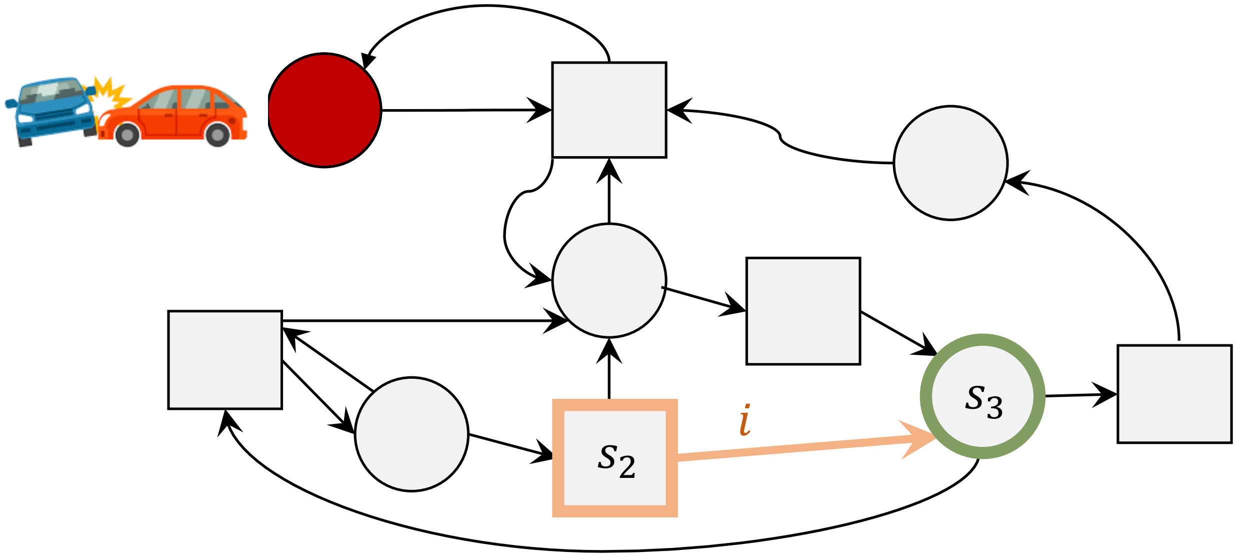


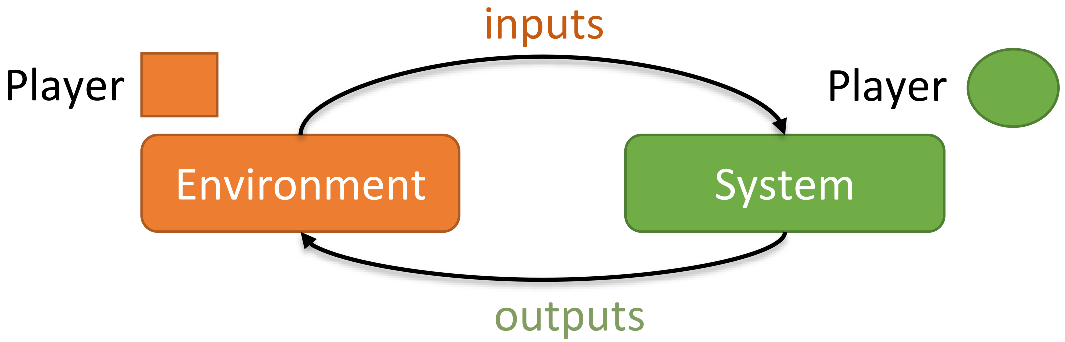
Shield Construction – Synthesis is a Game



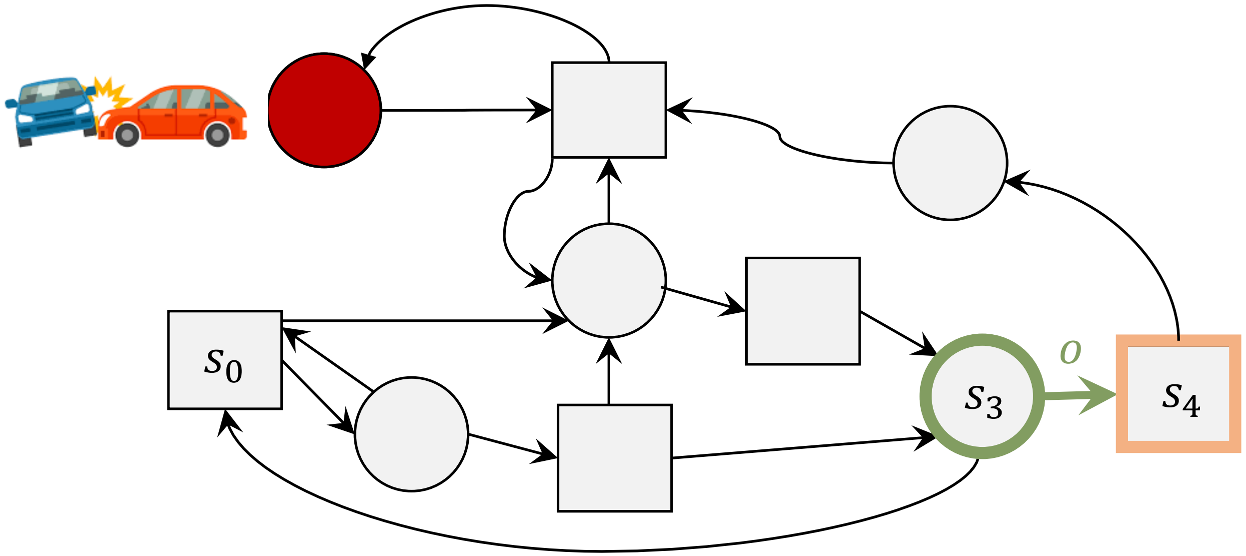


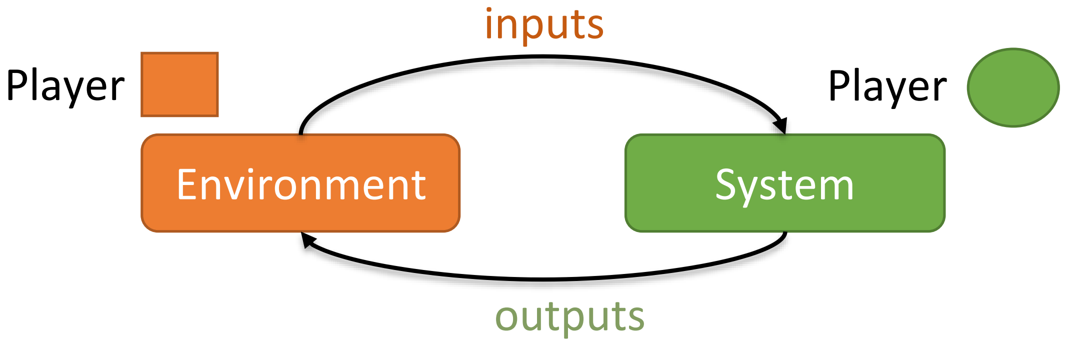
Shield Construction – Synthesis is a Game



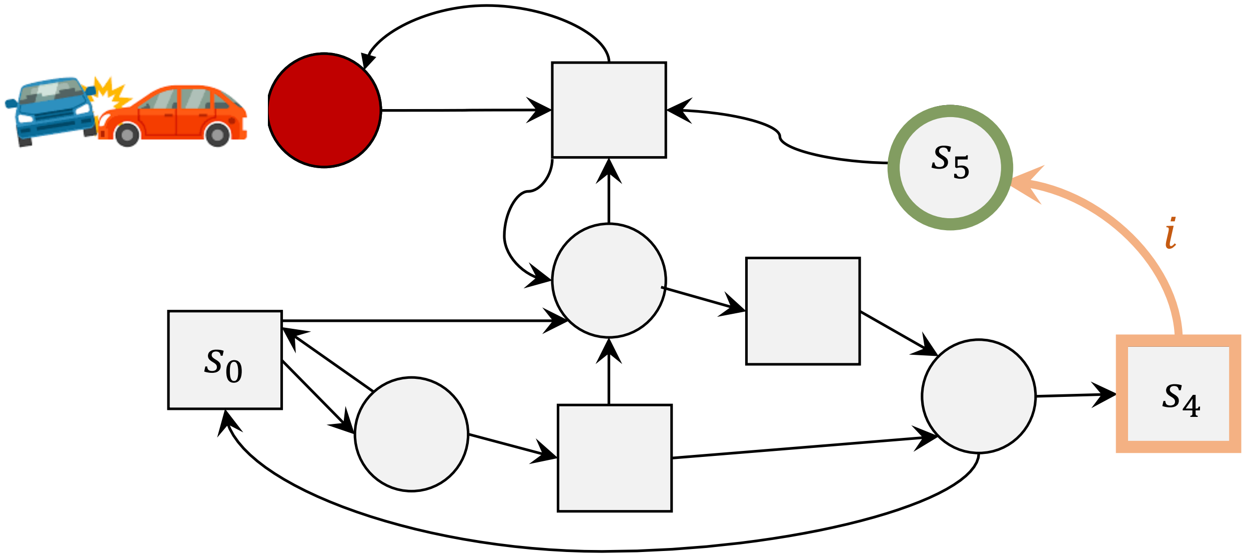


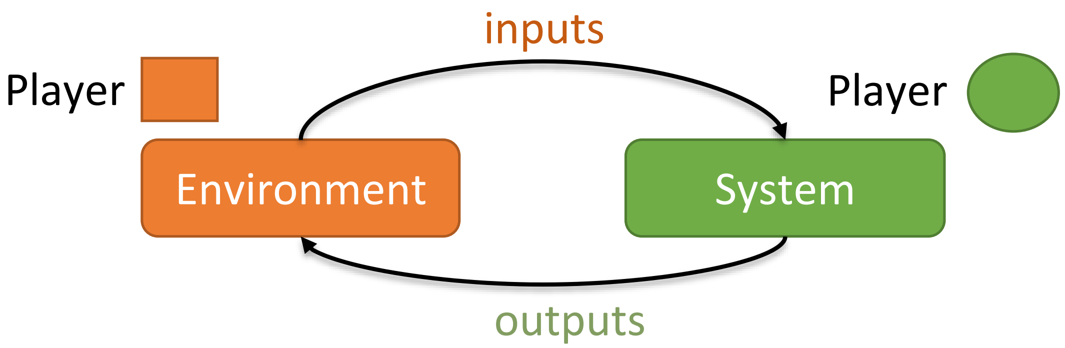
Shield Construction – Synthesis is a Game



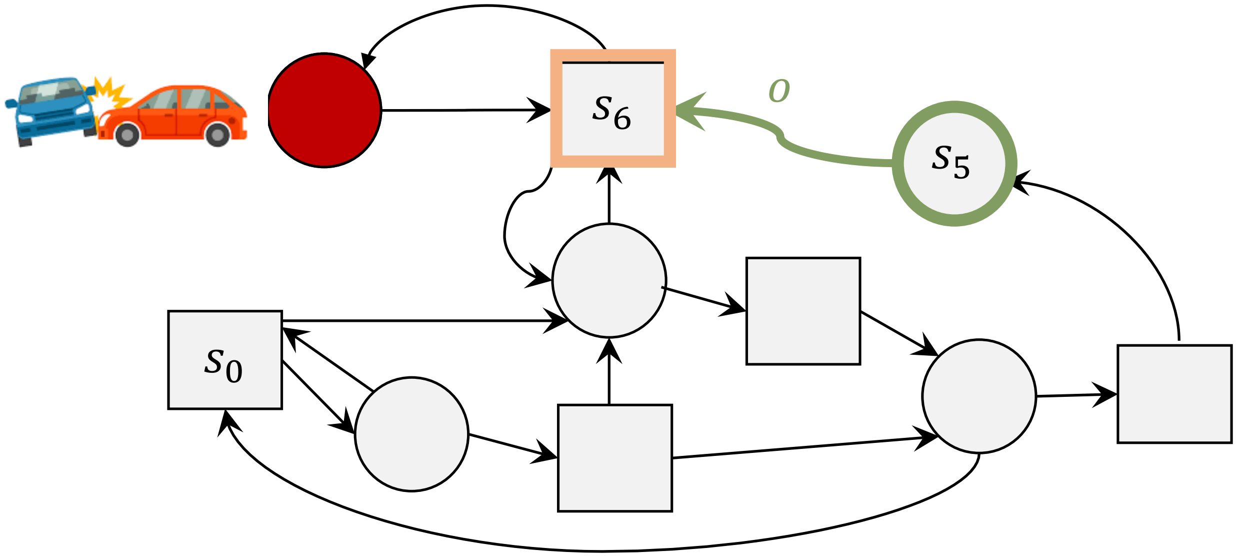


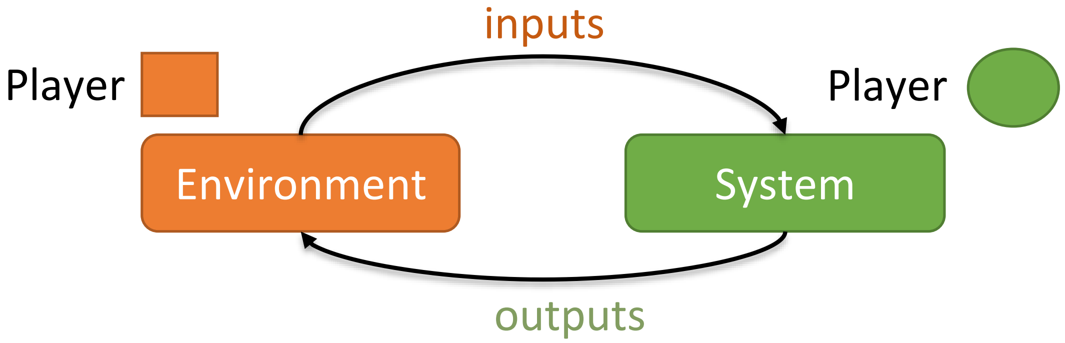
Shield Construction – Synthesis is a Game



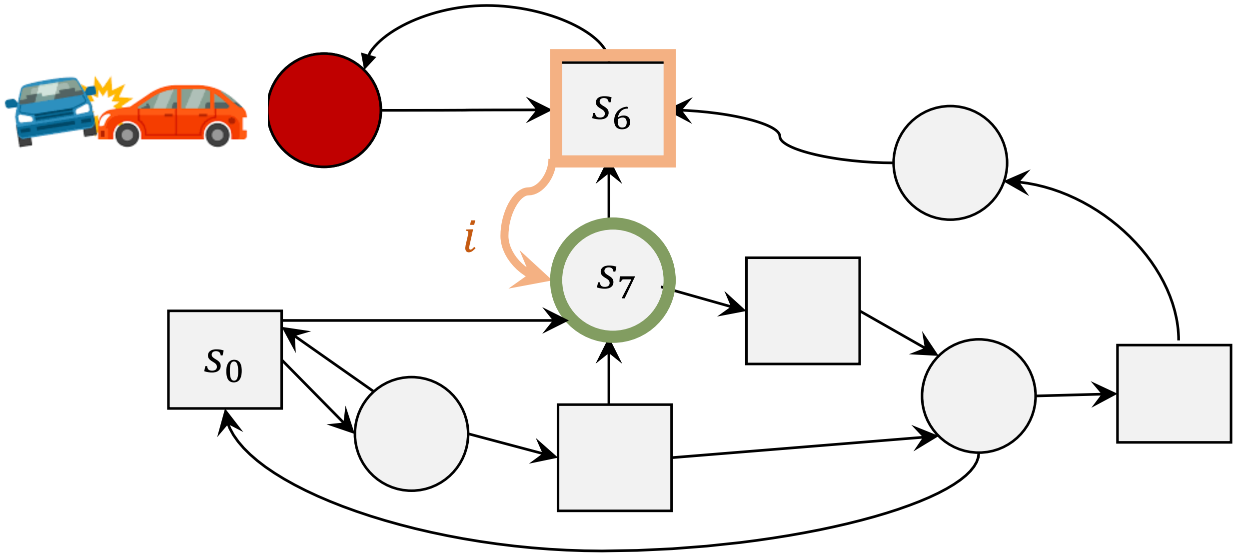


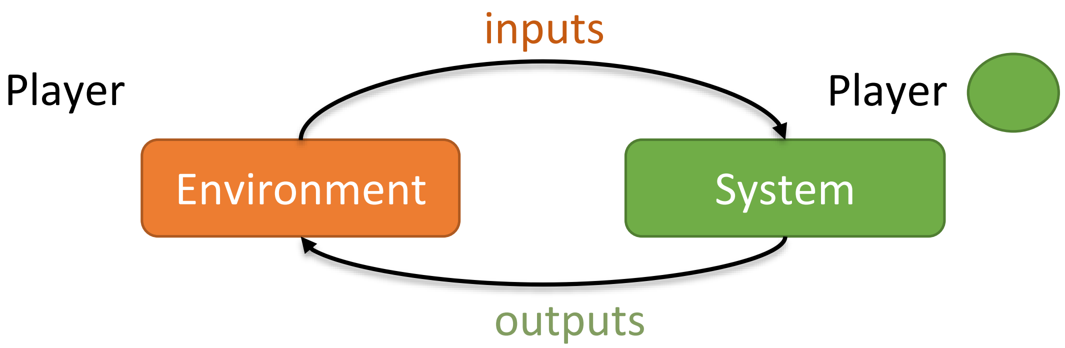
Shield Construction – Synthesis is a Game





Shield Construction – Synthesis is a Game

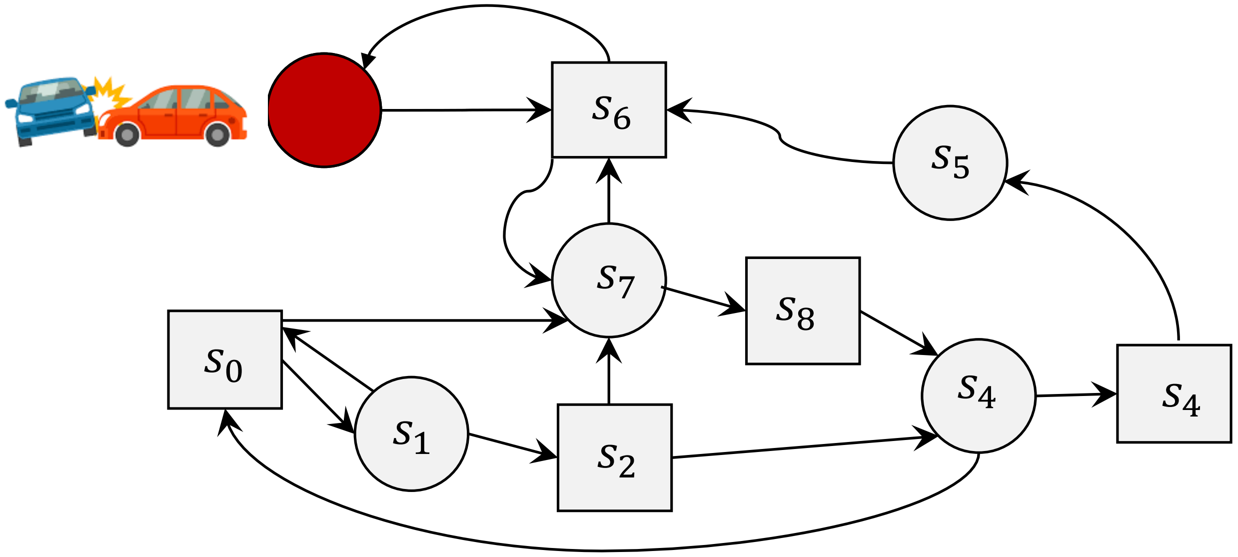




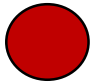
System Player wins, if ● is **never** visited

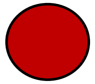
Winning Region: States from which the system can enforce that ● is **never** visited

Shield Construction – Synthesis is a Game

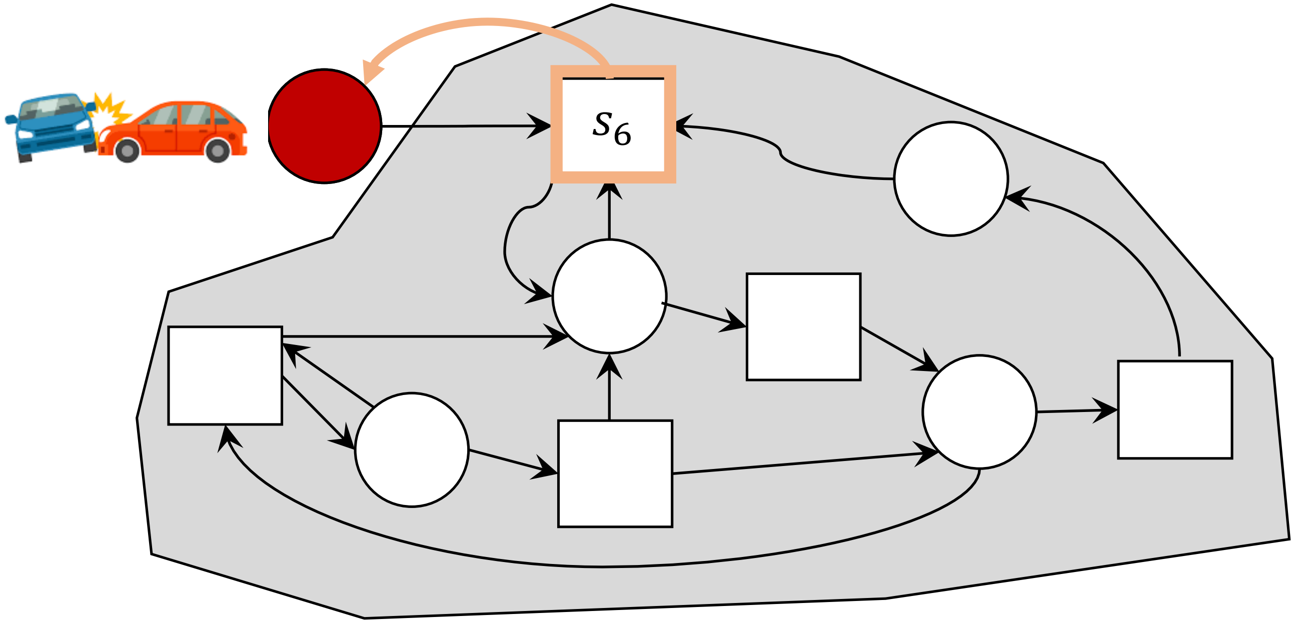


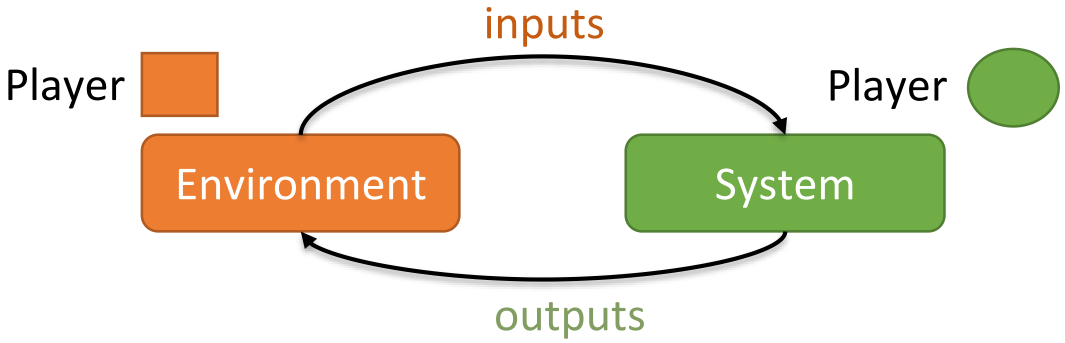
- **Question: Winning Region for this example?**

System Player wins, if  is **never** visited

Winning Region: States from which the system can enforce that  is **never** visited

Shield Construction – Synthesis is a Game

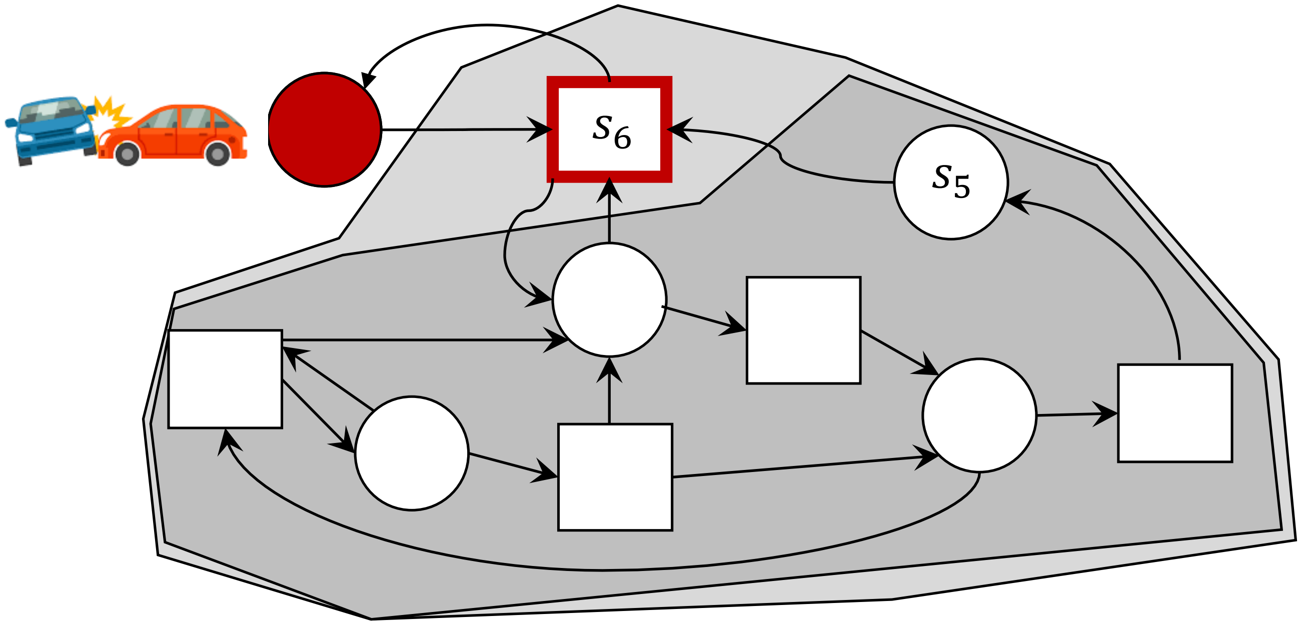


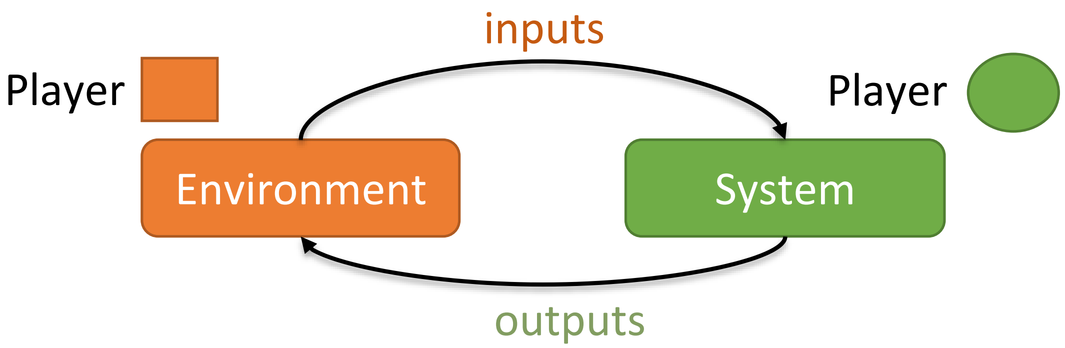


System Player wins, if ● is **never** visited

Winning Region: States from which the system can enforce that ● is **never** visited

Shield Construction – Synthesis is a Game

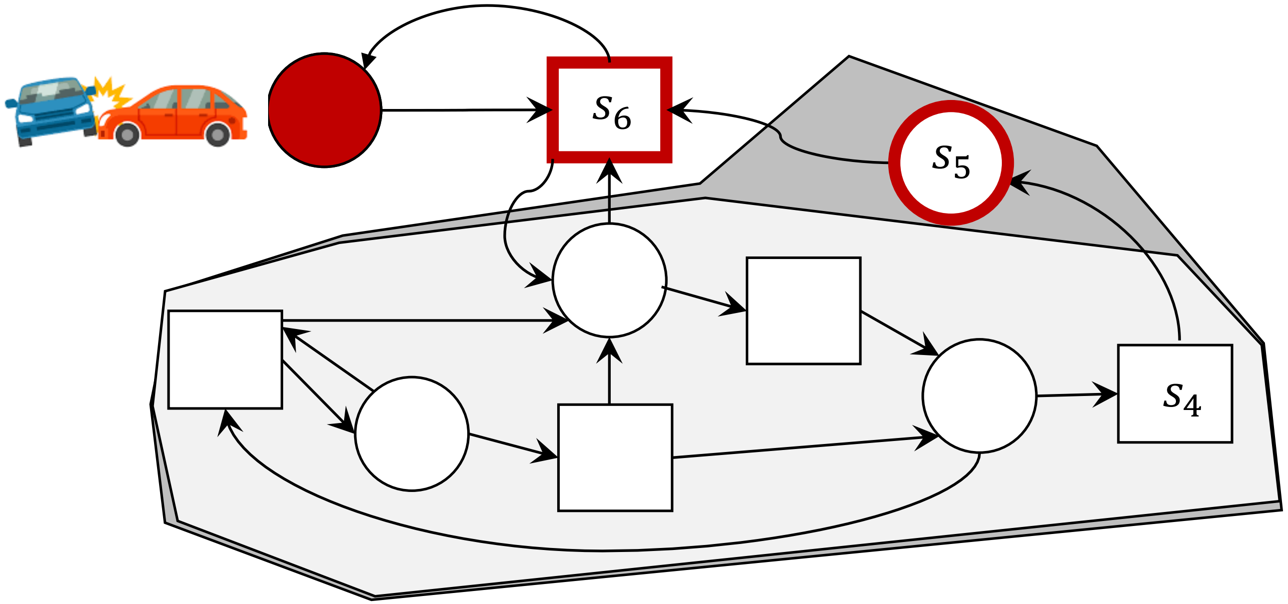


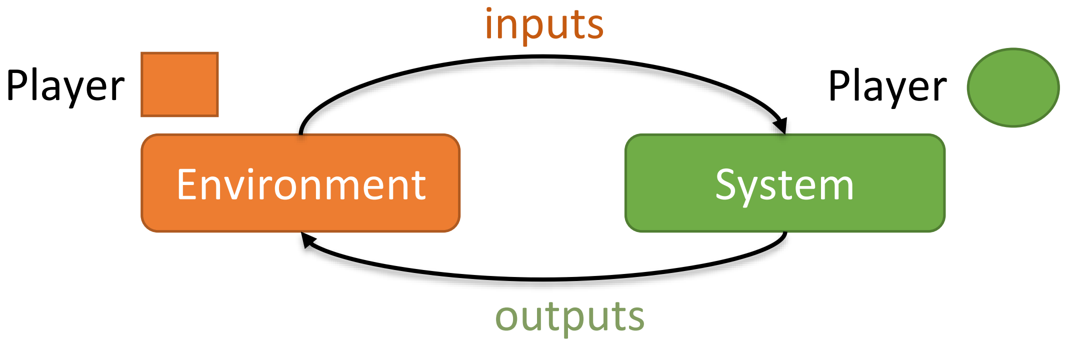


System Player wins,
if is **never** visited

Winning Region: States from which the system can enforce that is **never** visited

Shield Construction – Synthesis is a Game

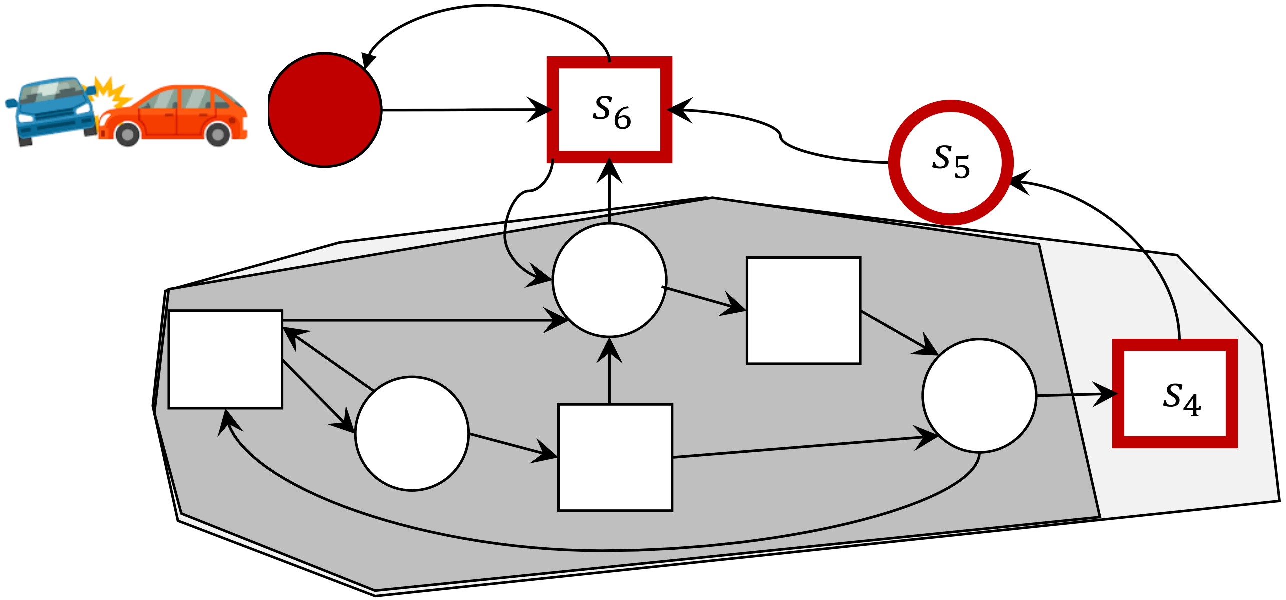


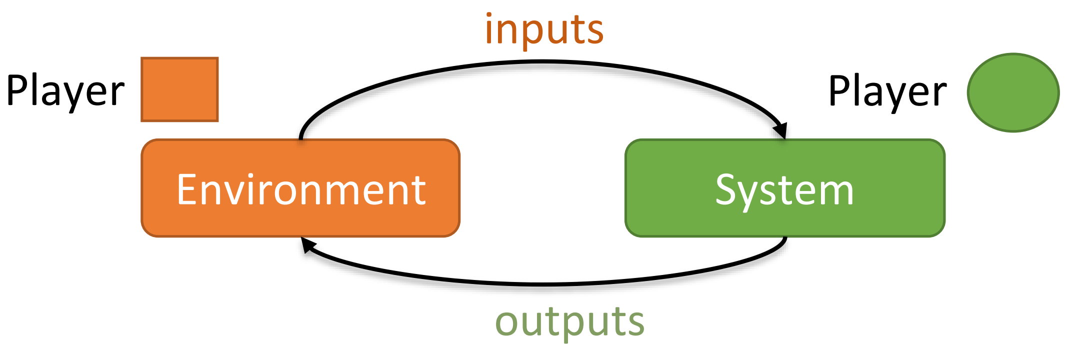


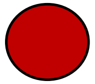
System Player wins, if ● is **never** visited

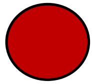
Winning Region: States from which the system can enforce that ● is **never** visited

Shield Construction – Synthesis is a Game

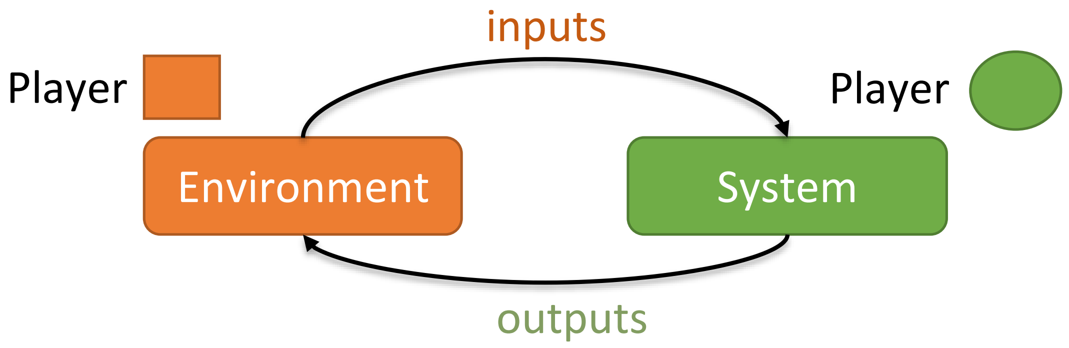


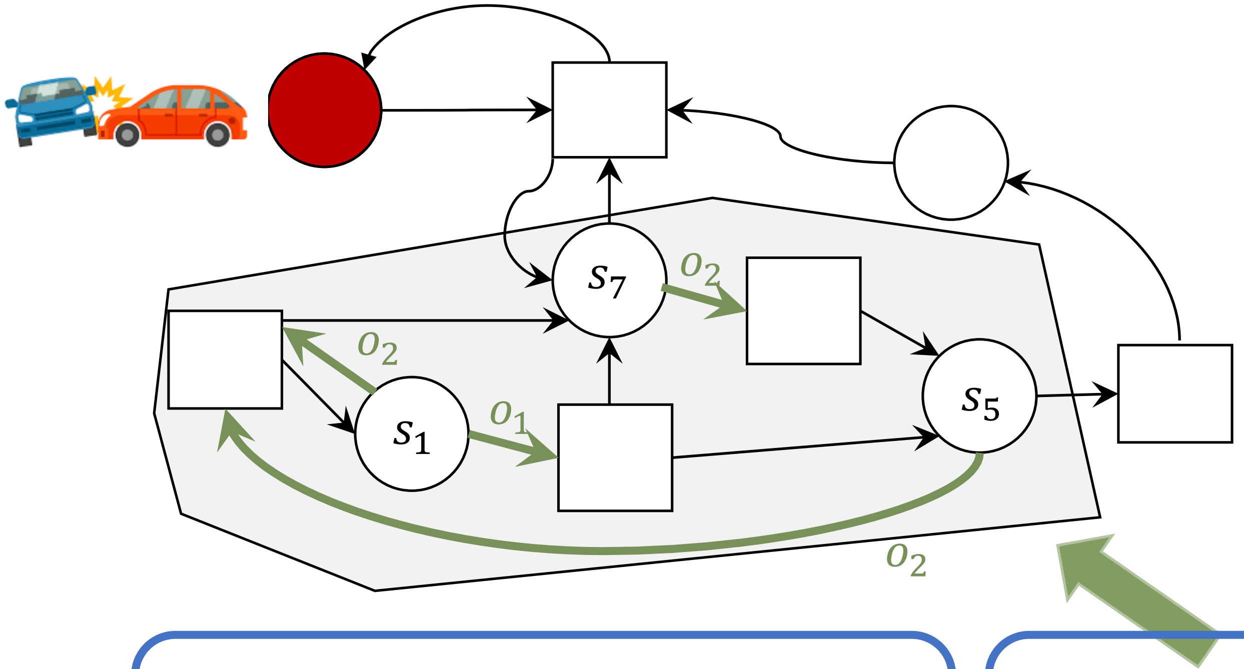


System Player wins,
if  is **never** visited

Winning Region: States from which the system can enforce that  is **never** visited

Shield Construction – Synthesis is a Game

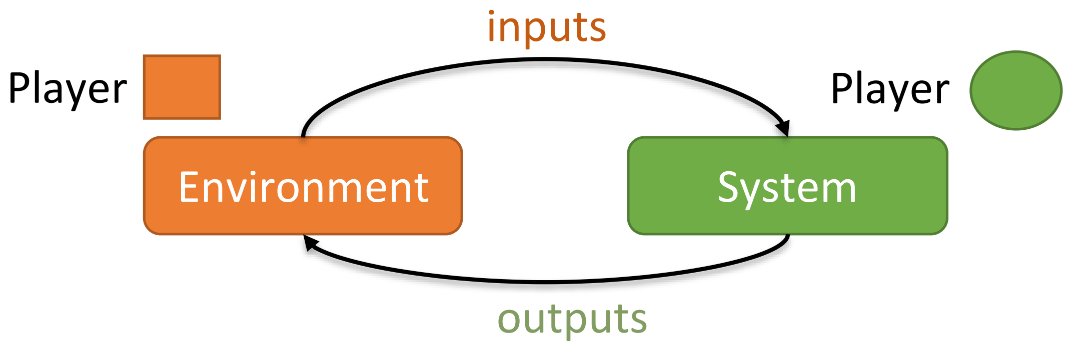


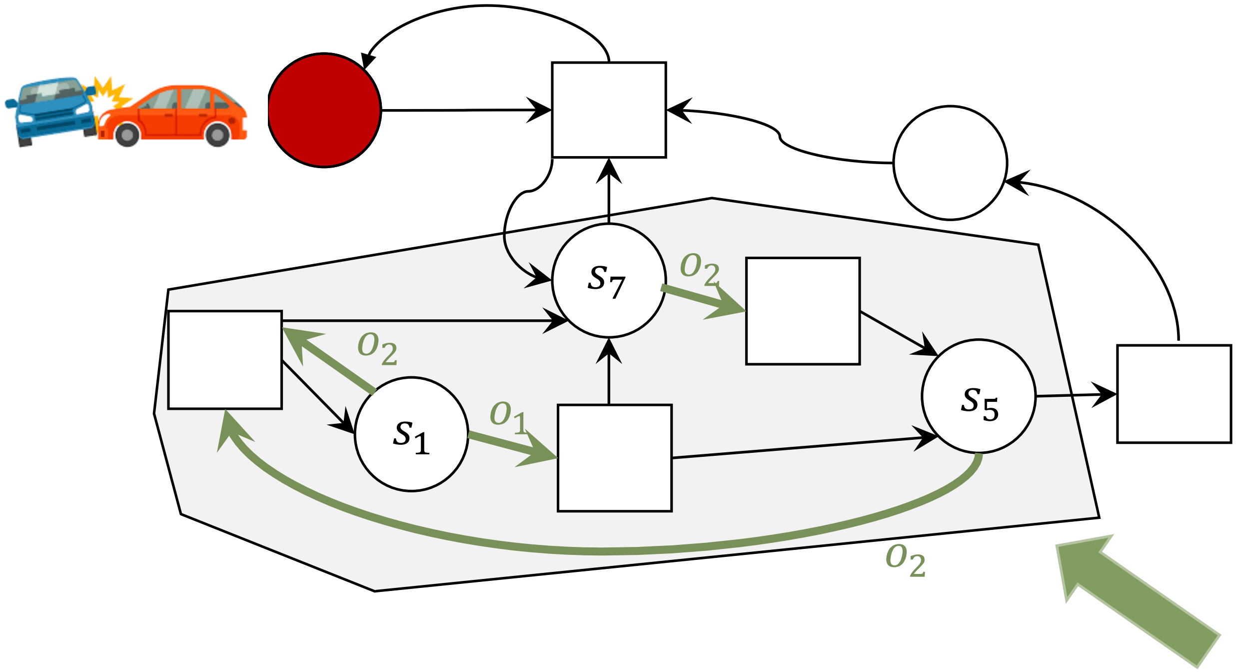


System Player wins, if ● is **never** visited

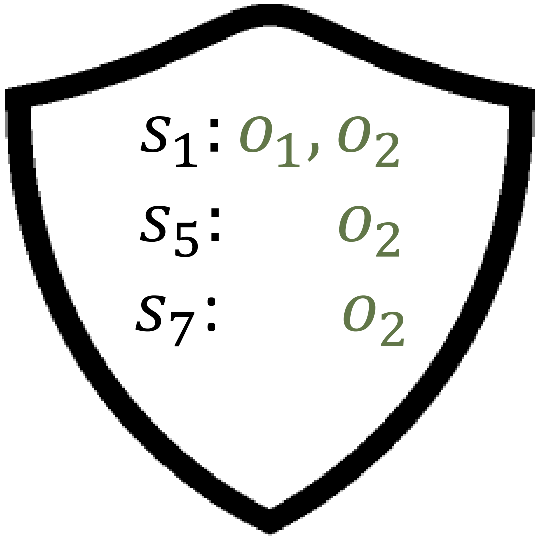
Winning Region: States from which the system can enforce that ● is **never** visited

Shield Construction – Synthesis is a Game

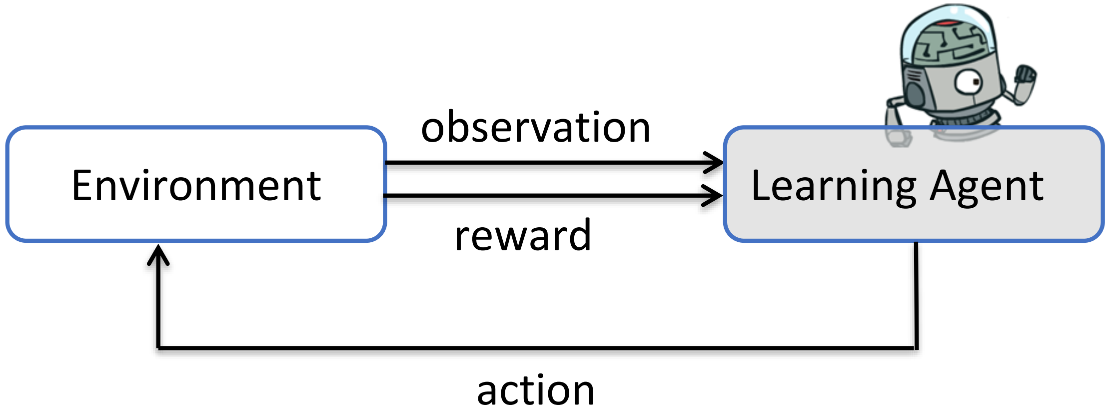




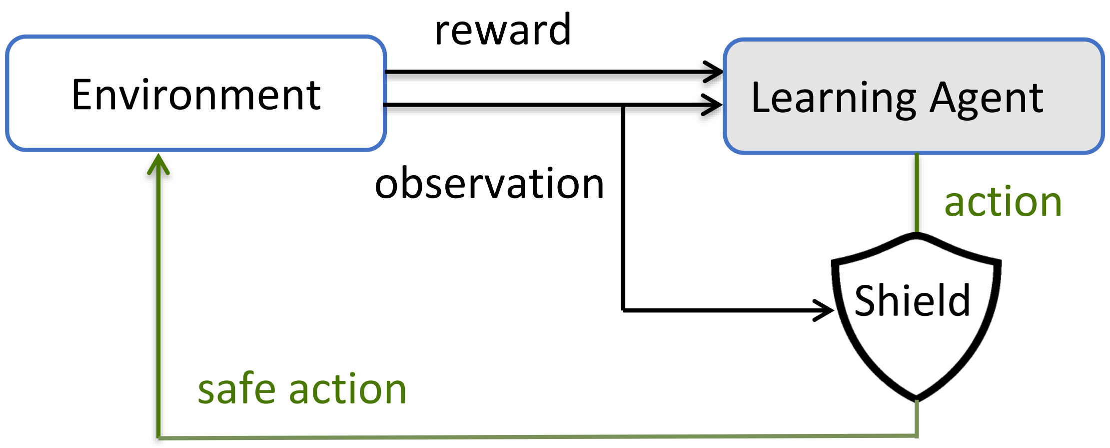
Winning region



Shielded Reinforcement Learning

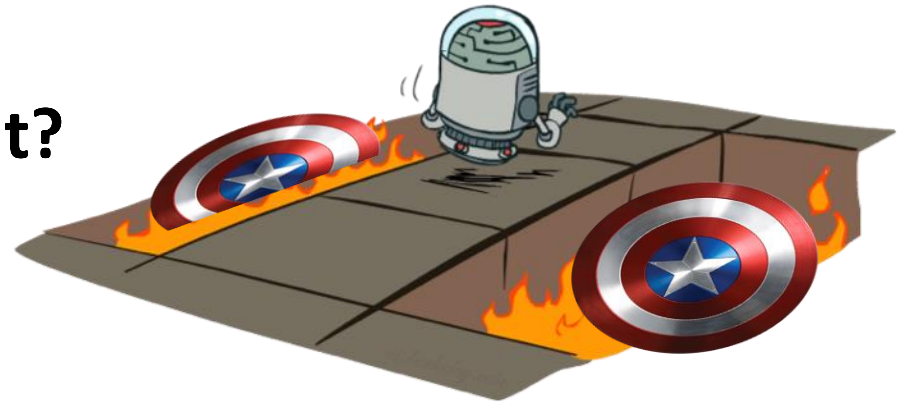


(Post) Shielding of Reinforcement Learning

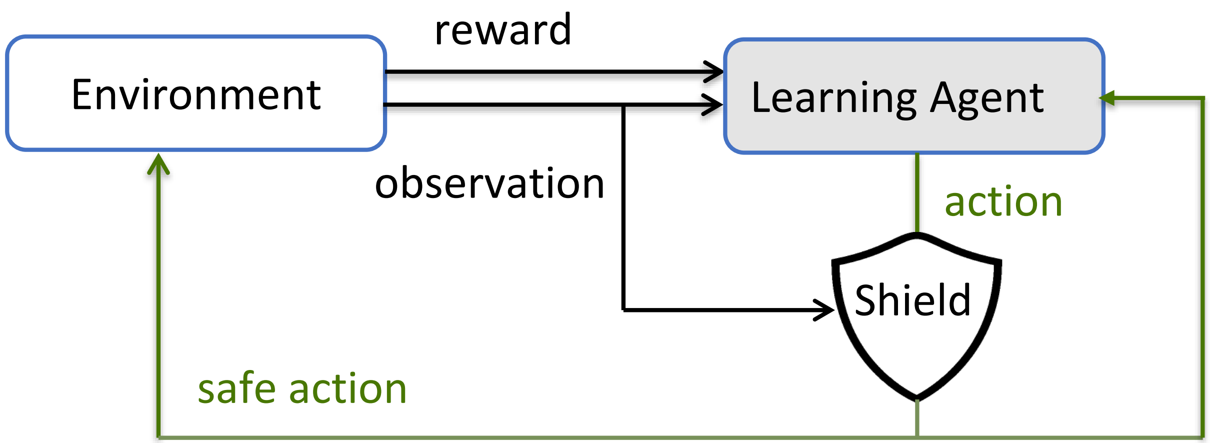


- **Shielding during and after learning**

- **Question: How to update the policy of the agent?**

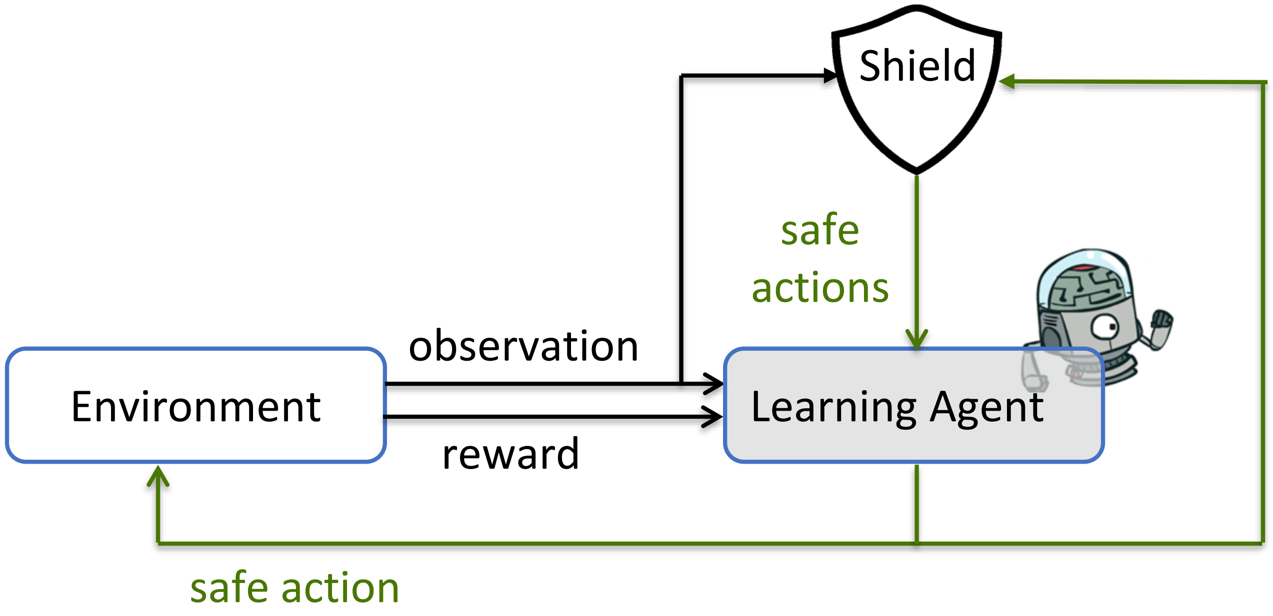


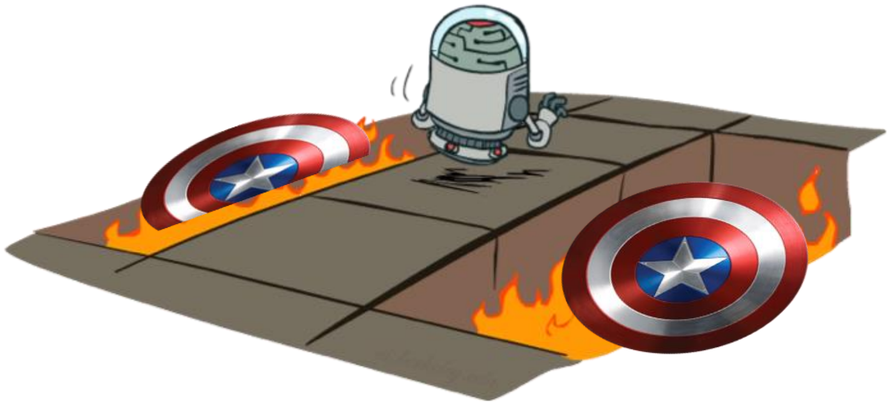
(Post) Shielding of Reinforcement Learning



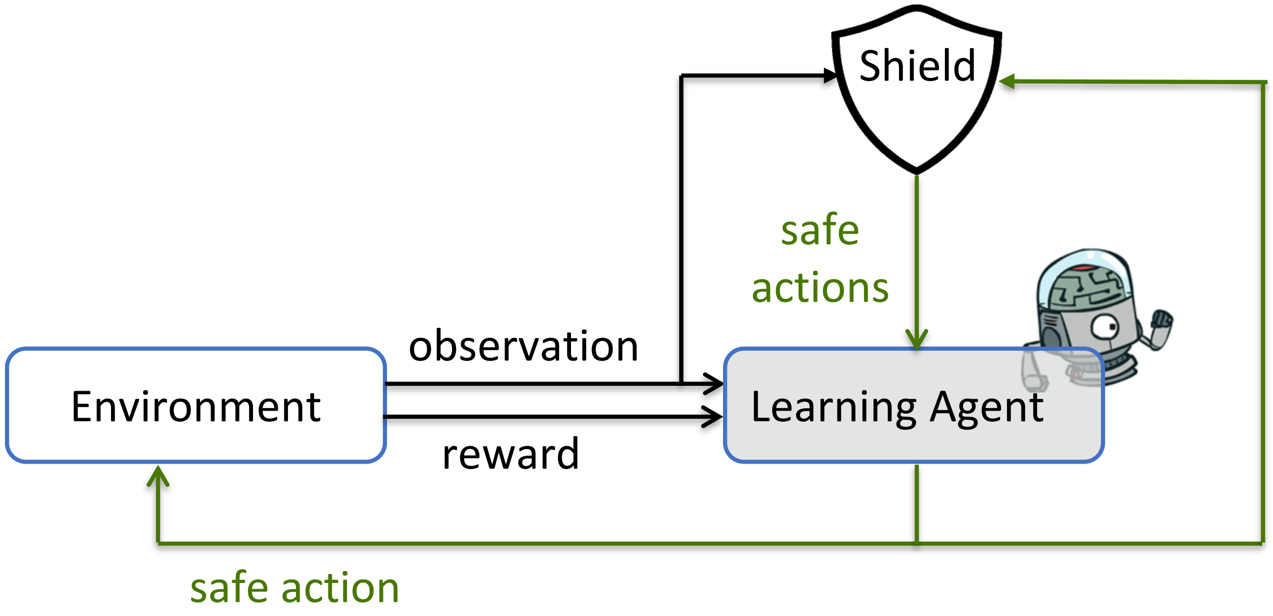
- **Shielding during and after learning**
- **Policy update for action chosen by shield**
- **Policy update for unsafe actions:**
  1. Update with negative reward **or**
  2. Update with gained reward

(Pre) Shielding of Reinforcement Learning

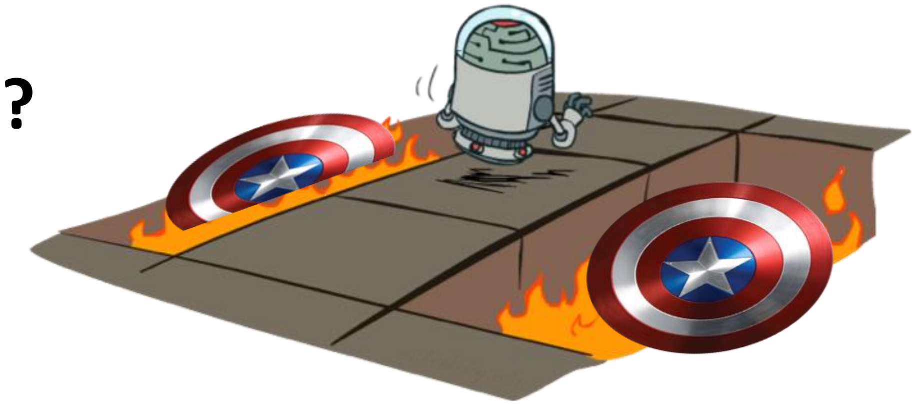




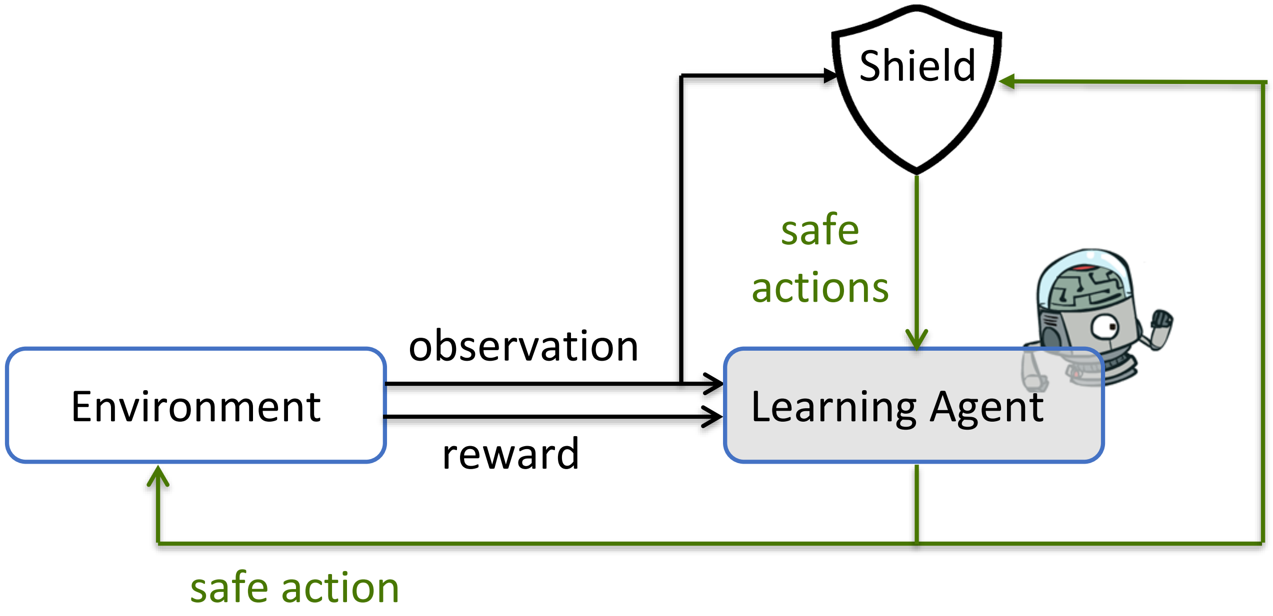
(Pre) Shielding of Reinforcement Learning



- **Question: How to update the policy of the agent?**

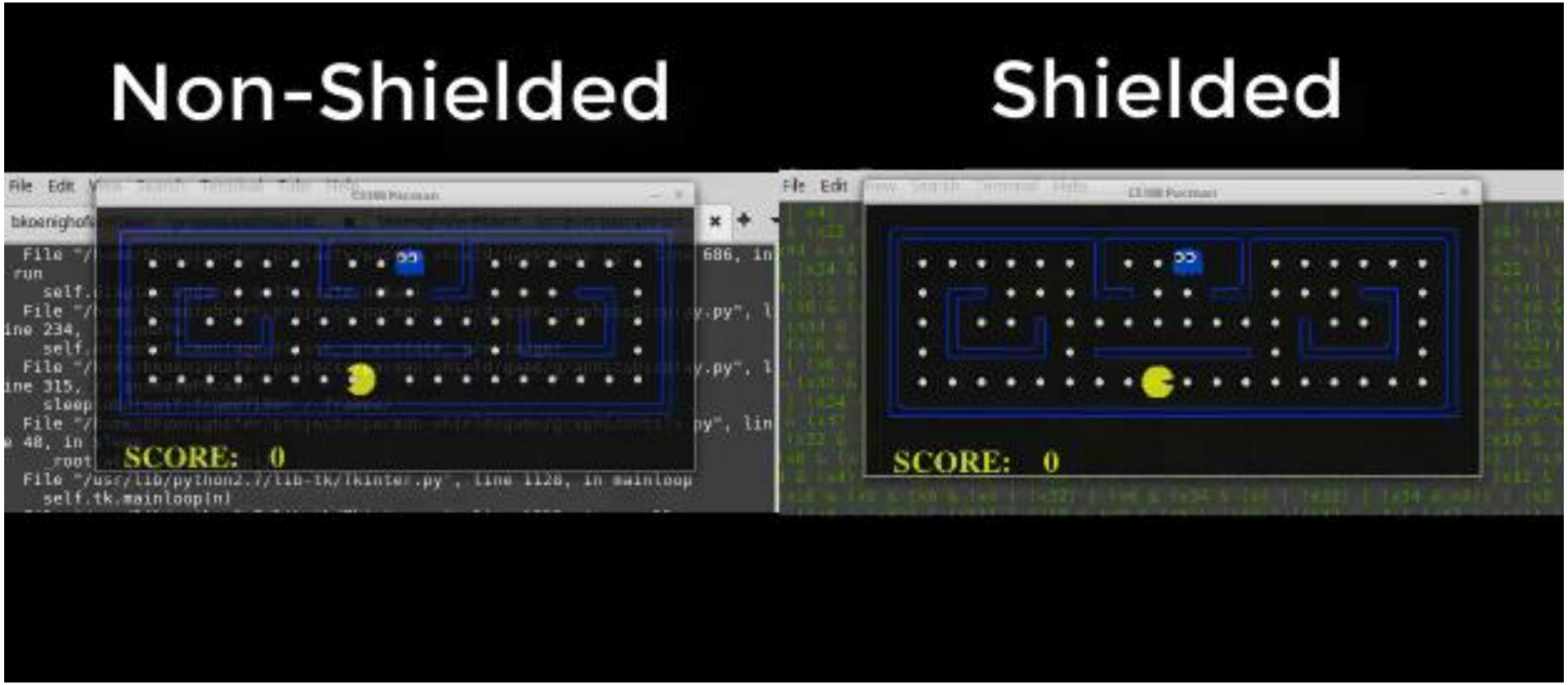


(Pre) Shielding of Reinforcement Learning



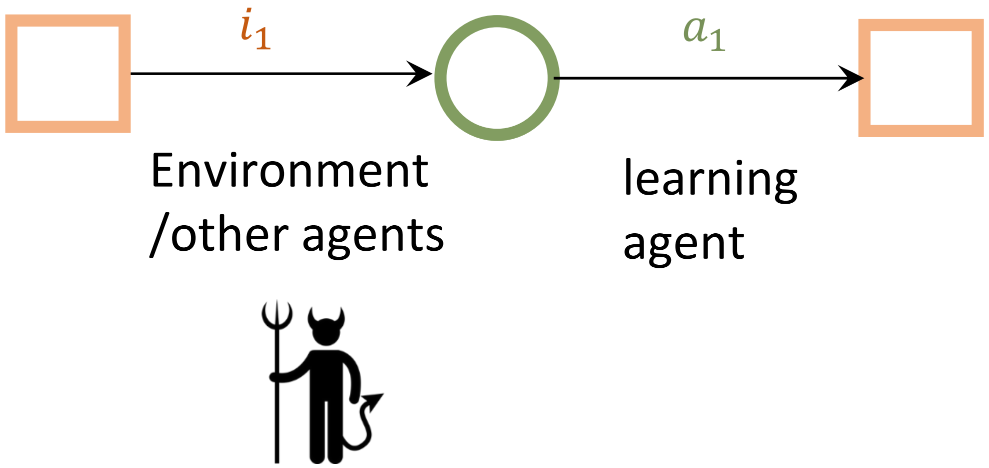
- **Action Masking**
- **Policy update for masked actions**
  - with negative reward

Safe Reinforcement Learning via Shielding



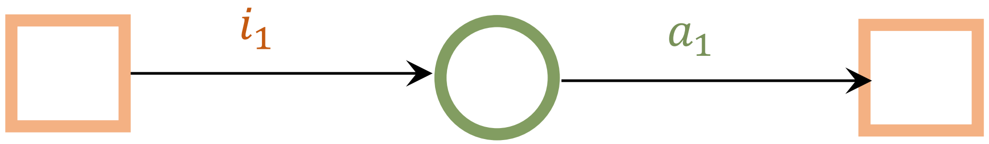
Probabilistic Shielding

2 Player Game – adversarial environment

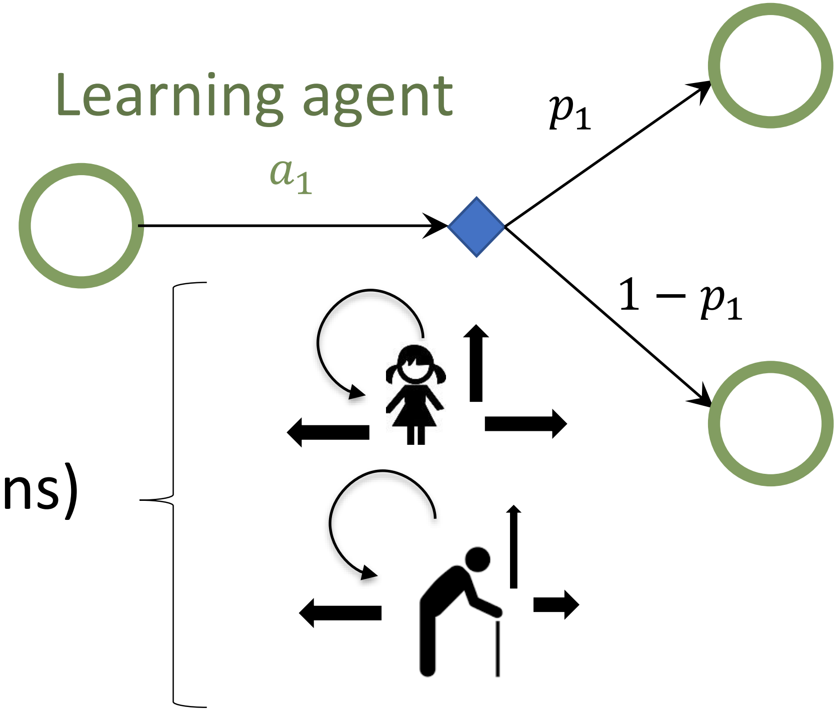


Probabilistic Shielding

2 Player Game – adversarial environment



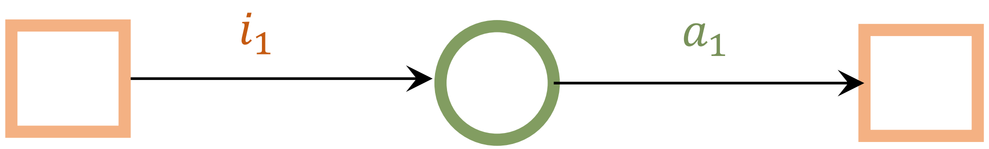
MDP – probabilistic environment



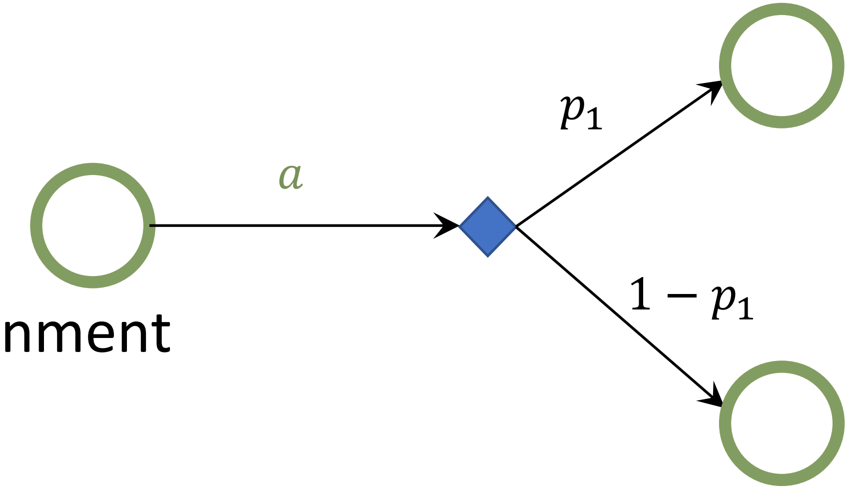
Probabilistic models (Markov chains)
for other agents

Probabilistic Shielding

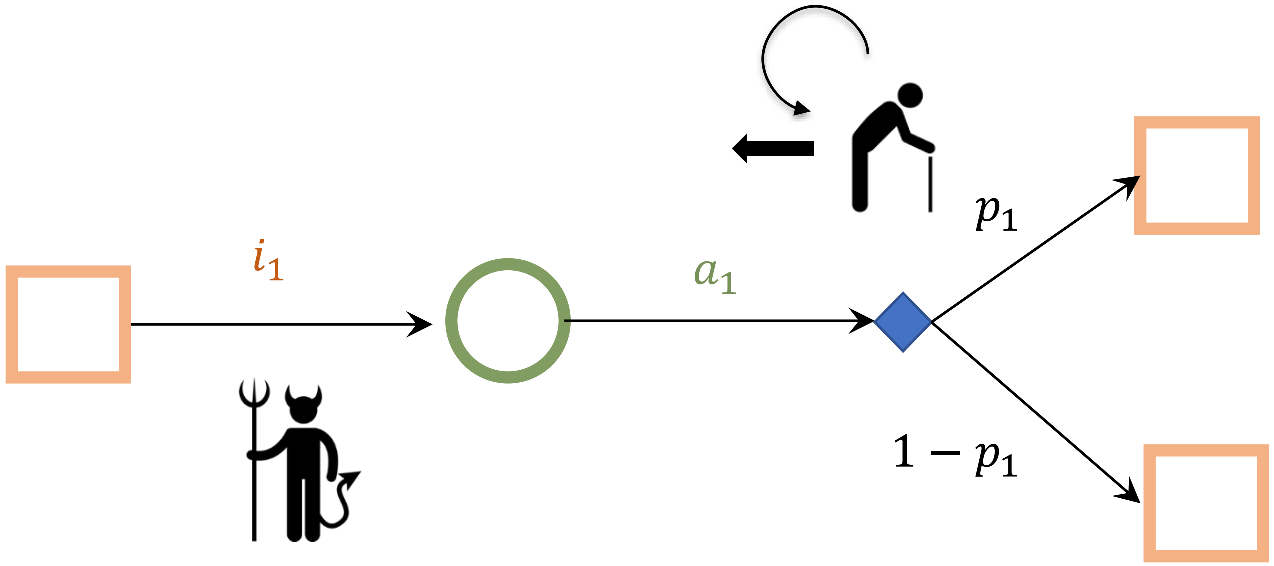
2 Player Game – adversarial environment



MDP – probabilistic environment



2 ½ Player Game – probabilistic & adversarial environment



Property Specification

- **Safety Probabilistic Temporal Logic Specification**
  - Maximal probability to stay safe in the next k steps
  - For all state-actions pairs: Compute **Safety-Value:**
    - $\boldsymbol{P_{max}(s,a) = P_{max}(T(s,a), G^{\leq k} safe)}$

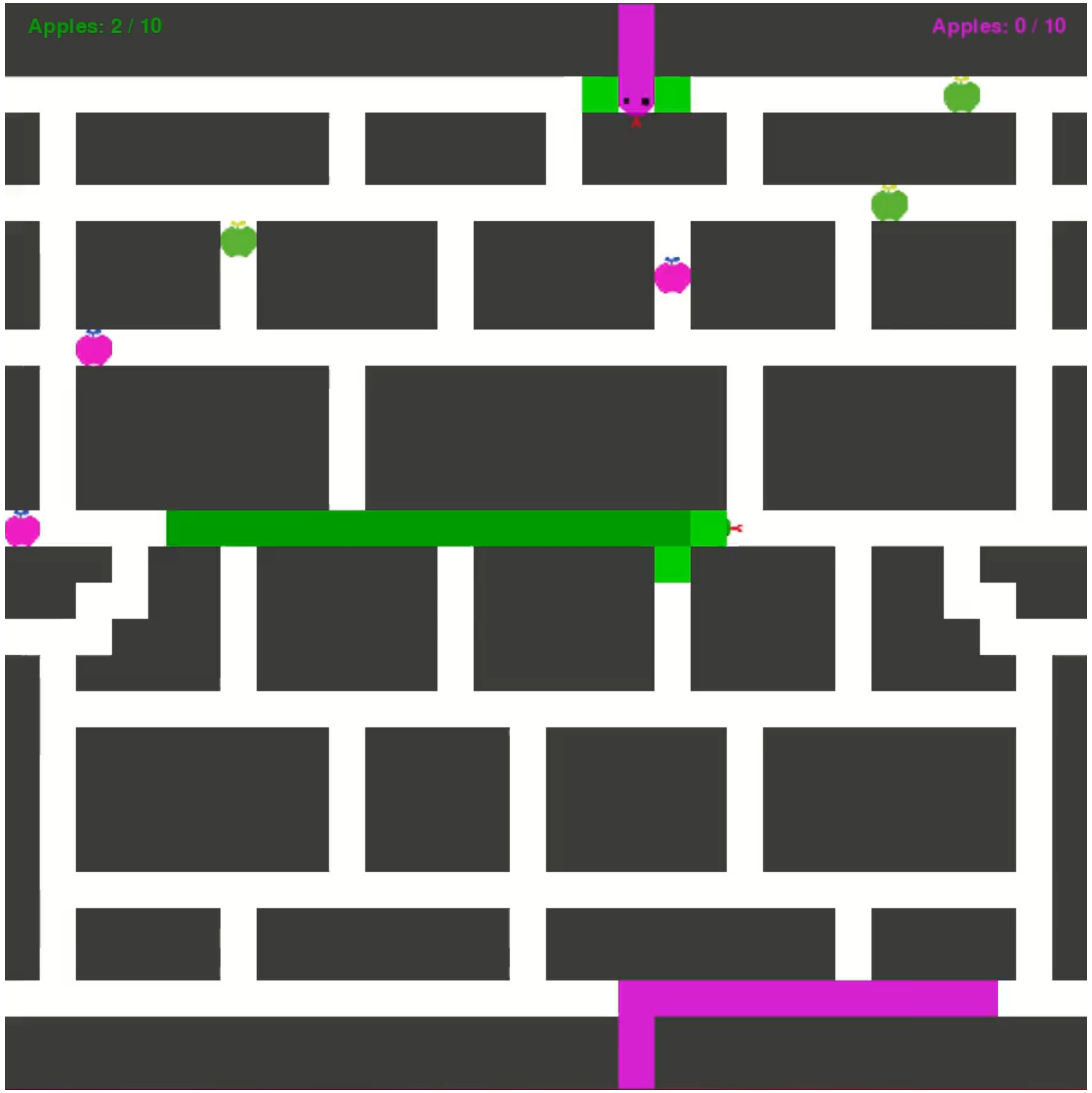
- **Absolute threshold** $\gamma \in [0,1]$
  - If $P_{max}(s,a) < \gamma$ → a is shielded in s
  - **Not deadlock free!**

- **Relative threshold** $\delta \in [0,1]$
  - If $P_{max}(s,a) < \delta \cdot P_{max}(s, a_{opt})$ → a is shielded in s

Property Specification

- **Absolute threshold $\gamma \in [0,1]$**
  - If $P_{max}(s,a) < \gamma \rightarrow a$ is shielded in s
  - **Not deadlock free!**
- **Relative threshold $\delta \in [0,1]$**
  - If $P_{max}(s,a) < \delta \cdot P_{max}(s, a_{opt}) \rightarrow a$ is shielded in s
- **Large γ or $\delta \longrightarrow$ strict shield;**
- **Small γ or $\delta \longrightarrow$ permissive shield**
- **γ and δ can be changed on the fly**

Probabilistic Shielding



Challenges in Shielding

- **Safety specification is typically simple**

  - Invariant properties
    - Do not collide
    - Never jump a red traffic signal
  - Temporal properties:
    - A signal is only allowed to exceed some threshold for t seconds
    - If there is a request, there has to be a grant within the next t seconds
    - …
  - …

Challenges in Shielding

- **Huge model of the environment**
  - Compute safety values for all possible state-action pairs

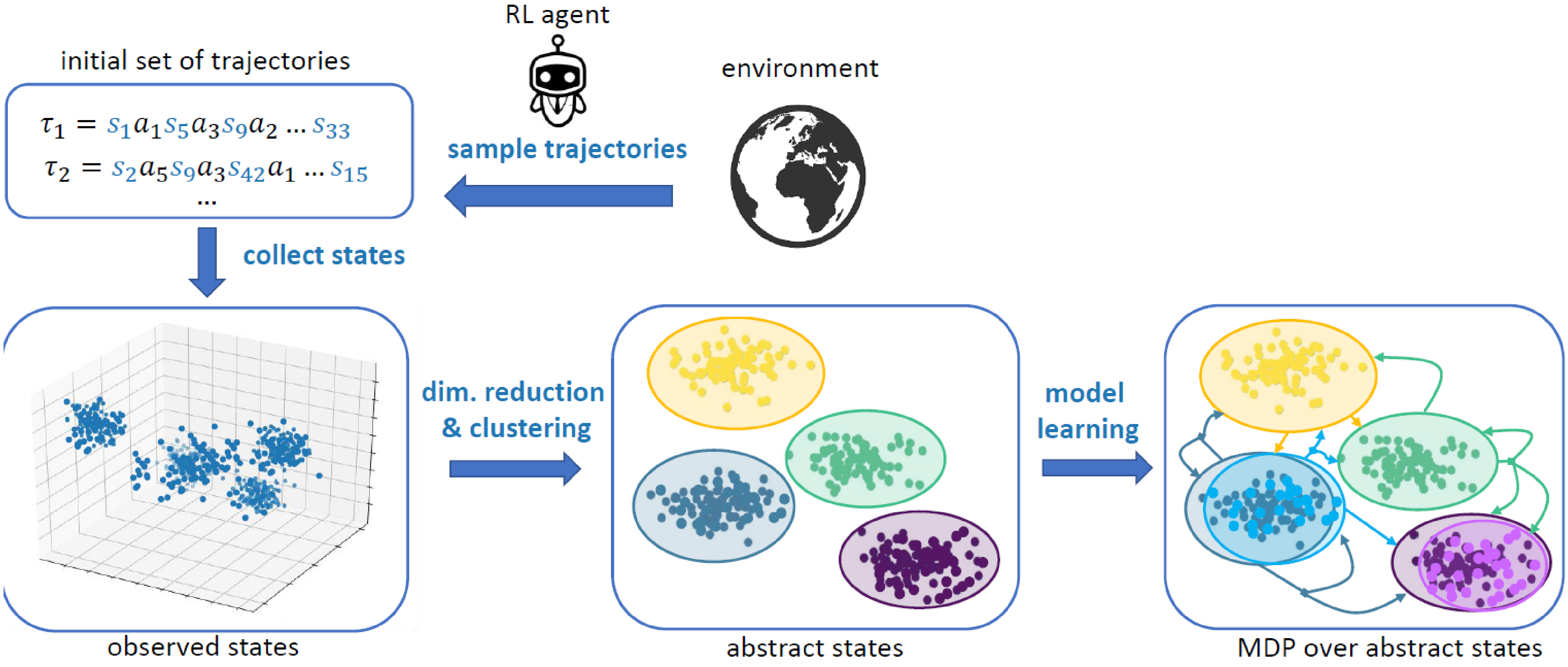
  - Expensive offline pre-computation and huge shielding data bases
  - Limits application of shielding to **small environments**

Challenges in Shielding

- Huge model of the environment
- Environmental model is unknown

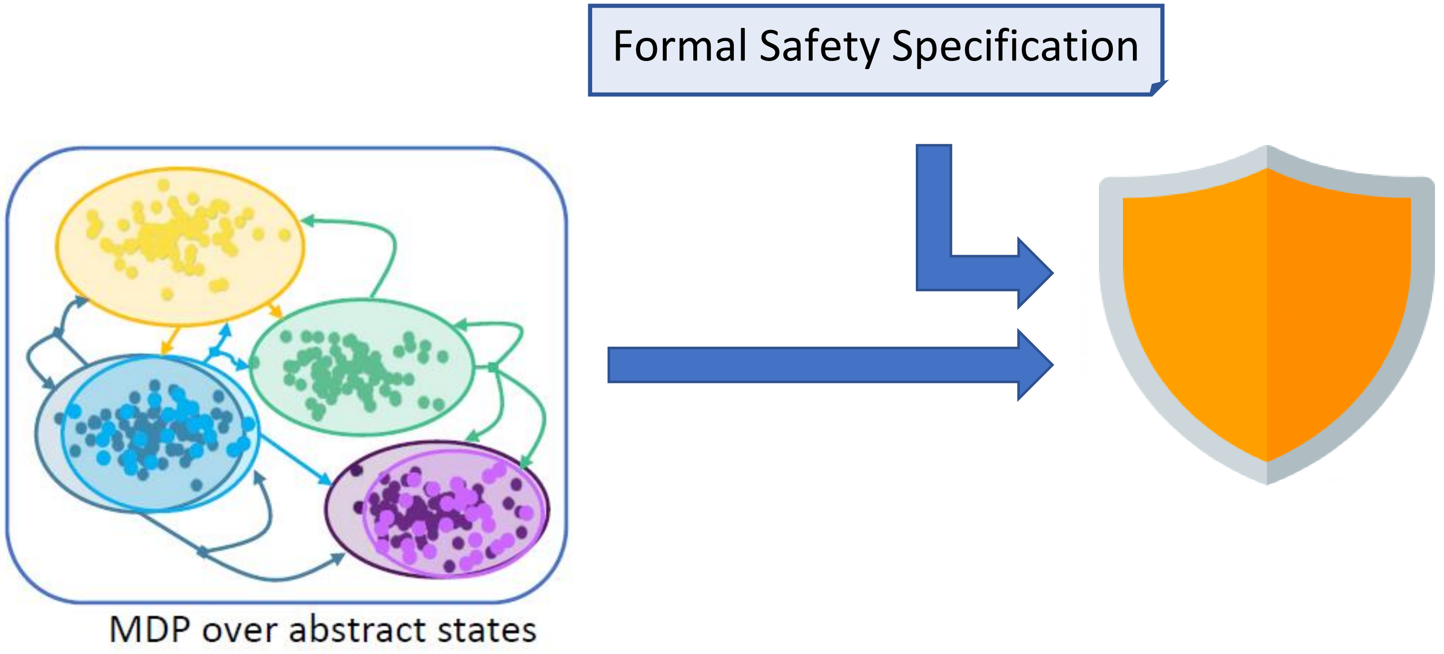
Challenges in Shielding

- Environmental model is unknown
- Idea: Combine shielding with automata learning

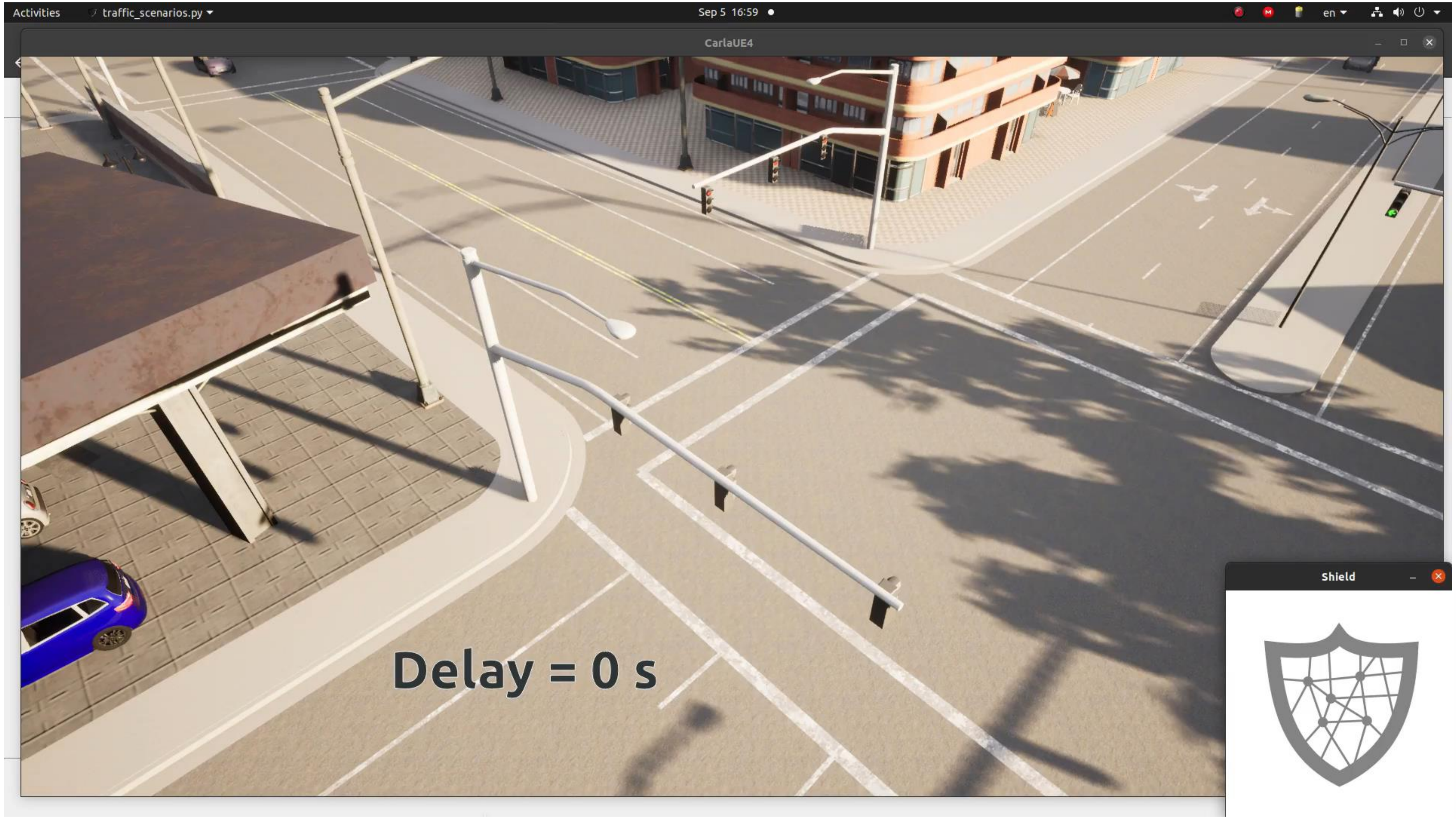


Challenges in Shielding

- Environmental model is unknown
- Idea: Combine shielding with automata learning



Shielding under Delayed Observation

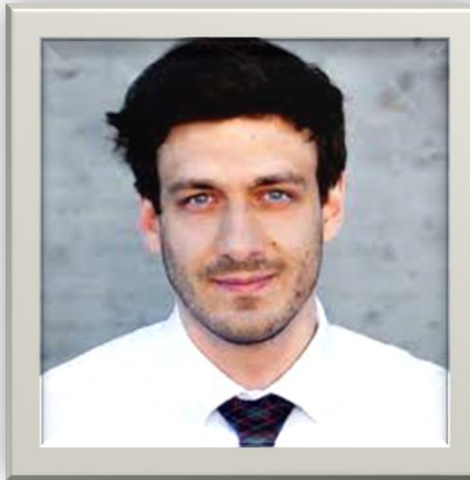


TEMPEST – Synthesis Tool for Reactive Systems and Shields in Probabilistic Environments

- Safety Shields → Guaranteed Safety
- Optimal Shields → Guaranteed Performance

https://tempest-synthesis.org/

Stefan



Shields are great for learned systems

- If you have a correct model

Many possibilities for FM 4 AI

- Verification
- Testing
- Monitoring / Enforcement
- Explainability
- Reward Shaping / Specification Mining…