

# It's Time to Play Safe: Shield Synthesis for Timed Systems

Roderick Bloem<sup>1</sup>, Peter Gjøøl Jensen<sup>3</sup>, Bettina Knighofer<sup>1,2</sup>, Kim Guldstrand  
Larsen<sup>3</sup>, Florian Lorber<sup>3</sup>, and Alexander Palmisano<sup>1</sup>

<sup>1</sup> Graz University of Technology, Institute IAIK, Austria

<sup>2</sup> Silicon Austria Labs, TU-Graz SAL DES Lab, Austria

<sup>3</sup> Department of Computer Science, Aalborg University, Denmark

**Abstract.** Erroneous behaviour in safety critical real-time systems may inflict serious consequences. In this paper, we show how to synthesize *timed shields* from timed safety properties given as timed automata. A timed shield enforces the safety of a running system while interfering with the system as little as possible. We present *timed post-shields* and *timed pre-shields*. A timed pre-shield is placed *before* the system and provides a set of safe outputs. This set restricts the choices of the system. A timed post-shield is implemented *after* the system. It monitors the system and corrects the system's output only if necessary. We further extend the timed post-shield construction to provide a guarantee on the recovery phase, i.e., the time between a specification violation and the point at which full control can be handed back to the system. In our experimental results, we use timed post-shields to ensure the safety in a reinforcement learning setting for controlling a platoon of cars, during the learning and execution phase, and study the effect.

## 1 Introduction

Today's systems are becoming increasingly sophisticated and powerful. At the same time, systems have to perform highly safety critical tasks, e.g., in the domain of self-driving cars, and make extensive use of machine learning, such as reinforcement learning [15]. As a result, complete offline verification is rarely possible. This holds especially true for safety critical real-time systems, where a deadline violation comes with serious consequences. An alternative is runtime verification [3–5, 7]. Runtime enforcement (RE) [13, 19, 27] extends this by enforcing the expected behavior of a system at runtime.

In this paper, we focus on the enforcement of regular timed properties for reactive systems and automatically synthesize *timed shields* from *timed automata specifications*. A timed shield can be attached to a system in two alternative ways. A *timed post-shield* (see Fig. 1(a)) is implemented after the system. It monitors the system and corrects the system's output if necessary. A *timed pre-shield* (see Fig. 1(b)) is placed before the system and provides a list of safe outputs to the system. This list restricts the choices of the system to safe actions.

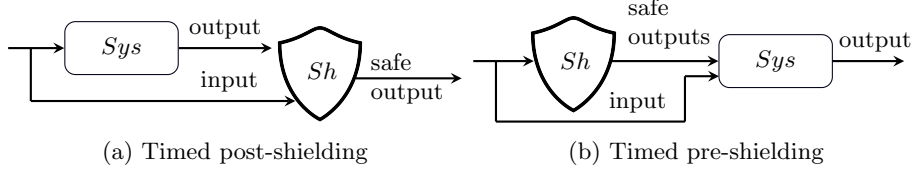


Fig. 1: Types of shielding.

Timed post-shields guarantee the following two properties. (1) *Correctness*: the shielded system satisfies the safety specification, and (2) *No-Unnecessary-Deviation*: the shield intervenes with the system only if safe system behavior would be endangered otherwise. We extend timed post-shields to shields that additionally provide guarantees on the recovery time, after which control is handed back to the system; i.e., from that point on the shield forwards the outputs to the environment and does not deviate anymore. (3a) *Guaranteed-Recovery*: the recovery phase ends within a *finite time*, and (3b) *Guaranteed-Time-Bounded-Recovery*: the recovery phase ends after a given *bounded time*.

Timed pre-shields guarantee the following two properties. (1) *Correctness*: the shield provides only safe outputs to the system, and (2) *No-Unnecessary-Restriction*: the shield provides all safe outputs to the system.

Shields can be employed during reinforcement learning (RL) to ensure safety by enforcement both during a system’s learning and execution phases [1, 9]. We introduce a timed post-shield after the learning agent, as depicted in Fig. 2. The shield monitors the actions selected by the learning agent and corrects them if and only if the chosen action is unsafe.

To sum up, we make the following contributions:

- We propose to synthesize timed shields from timed automata specifications.
- We discuss two basic types of timed shields: pre-shields and post-shields.
- We propose timed post-shields with the ability to recover.
- Our experiments show the potential of timed shields to enforce safety in RL.

**Related Work.** In most work about RE, an enforcer monitors a program that outputs events and can either terminate the program once it detects an error [24], or alter the event in order to guarantee, for example, safety [16] or privacy [18, 28]. Renard et al. [23] and Falcone et al. [14] considered RE for timed properties. The similarities between these enforcers and a shield is in their ability to alter events. Their work only considers static programs whereas we consider enforcing correctness for reactive systems.

The term runtime assurance [25] is often used if there exist a switching mechanism that alternates between running a high-performance system and a provably safe one. These concept is similar to post-shielding, with the difference that we synthesize the entire provable safe system (the shield) including the switching mechanism while proving guarantees on recovery.

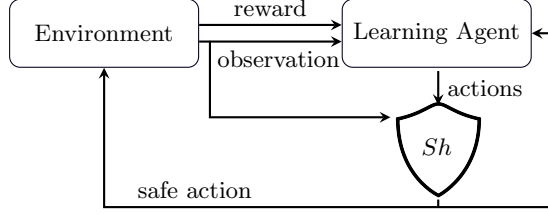


Fig. 2: A timed post-shield in a reinforcement learning setting.

Knighofer et. al. introduced [19] shield synthesis from LTL specifications, which, while related, are not expressible enough to capture real time behavior and thus cannot shield against timing related faults in real timed systems. Wu et al. [26] extend shields for boolean signals to real-valued shields to enforce the safety of cyber-physical systems. The concept of shield synthesis for RL from (probabilistic) LTL specification was discussed in [1, 17] and early work on combining (basic) shielding with RL was demonstrated in [9].

We use the specification theory of Timed In-put/Output Automaton (TIOA) used by UPPAAL ECDAR [11]. To synthesize our timed shields we use UPPAAL TIGA [8], a tool which implements algorithms for solving games based on timed game automata with respect to reachability and safety properties, producing non-deterministic safety strategies. UPPAAL STRATEGO [9, 10] extends UPPAAL TIGA by the capability of optimizing these strategies with respect to desired performance measures. While UPPAAL STRATEGO contains a RL component, we utilize a third-party, off-the-shelf, RL-system to demonstrate the applicability of the proposed method in a generic RL setting.

**Outline.** In Sec. 2 we give the formal notation and constructions. Definition and construction of timed post-shields (with recoverability guarantees under different fault models) are introduced in Sec. 3 (Sec. 4). Timed pre-shields are introduced in Sec. 5. We discuss our experimental findings from a car platooning problem in Sec. 6 followed by a conclusion in Sec. 7.

## 2 Specification Theory for Real-Time Systems

Let us recall definitions of Timed (Input/Output) Automata. Let  $X$  be a finite set of real-valued *clocks*. Let  $\mathcal{V}(X) \mapsto \mathbb{R}_{\geq 0}$  be the valuations over  $X$  and let  $\vec{0}$  be the valuation that assigns 0 to each clock. For the value of a single clock of a given valuation  $\nu \in \mathcal{V}(X)$ , we write  $\nu(x)$  and  $\nu[x]$  denotes the reset of the clock  $x$  in the valuation  $\nu$ . We extend the notion of reset to sets, i.e., for some  $Y \subseteq X$  let  $\nu[Y]$  be the valuation after resetting all values of clocks in  $Y$  to zero and otherwise retaining their value. If  $\delta \in \mathbb{R}_{\geq 0}$  is a positive delay, then we denote by  $\nu + \delta$  the valuation s.t. for all  $x \in X$ ,  $(\nu + \delta)(x) = \nu(x) + \delta$ .

Let  $Y \subseteq X$  and  $Z \subseteq X \cup \mathbb{Z}$ . We denote the set of all simple constraints as  $\Phi(Y, Z) = Y \times \{<, \leq, \geq, >\} \times Z$ . The set of all possible clock constraints is

defined by  $\mathcal{C}(X) = 2^{\Phi(X, X \cup \mathbb{Z})}$ . We denote the (conjunctive) subset of restricted clock constraints by  $\mathcal{B}(X) \subseteq \mathcal{C}(X)$  with  $\mathcal{B}(X) = 2^{\Phi(X, \mathbb{Z})}$ .

**Definition 1 (Timed Input/Output Automaton (TIOA) [2,12]).** A Timed Input/Output Automaton (TIOA) is a tuple  $\mathcal{A} = (L, \ell_0, \Sigma^?, \Sigma^!, X, E, I)$  where  $L$  is a finite set of locations,  $\ell_0$  is the initial location,  $\Sigma = \Sigma^? \cup \Sigma^!$  is a finite set of actions partitioned into inputs ( $\Sigma^?$ ) and outputs ( $\Sigma^!$ ),  $X$  is a set of clocks,  $E \subseteq L \times \mathcal{B}(X) \times (\Sigma^! \cup \Sigma^?) \times 2^X \times L$  is a set of edges, and  $I : L \rightarrow \mathcal{B}(X)$  is a function assigning invariants to locations.

States of a TIOA are given as a pair  $(\ell, \nu) \in L \times \mathbb{R}_{\geq 0}^X$  consisting of a discrete location and a real-valued assignment to the clocks. From a given state  $(\ell, \nu) \in L \times \mathbb{R}_{\geq 0}^X$  where  $\nu \models I(\ell)$ , we have two kinds of transitions; (1) **discrete transitions**:  $(\ell, \nu) \xrightarrow{\alpha} (\ell', \nu')$  if there exists  $(\ell, \psi, \alpha, Y, \ell') \in E$  s.t.  $\nu \models \psi$ ,  $\nu' = \nu[Y]$  and  $\nu' \models I(\ell')$ , and (2) **delay transitions** for some  $\delta \in \mathbb{R}_{\geq 0}$  where  $(\ell, \nu) \xrightarrow{\delta} (\ell, \nu')$  if  $\nu' = \nu + \delta$  and  $\nu' \models I(\ell)$ . We define the semantics of a TIOA as a TLTS.

**Definition 2 (Timed Labeled Transitions System (TLTS) [2, 12]).** A TLTS is a tuple  $\llbracket \mathcal{A} \rrbracket = (\mathcal{Q}, q_0, \rightarrow)$  s.t.  $\mathcal{Q} = L \times \mathbb{R}_{\geq 0}^X$ ,  $q_0 = (\ell_0, \vec{0})$ , and  $\rightarrow : \mathcal{Q} \times \Sigma^! \cup \Sigma^? \cup \mathbb{R}_{\geq 0} \times \mathcal{Q}$  is a transition relation defined as above.

A trace  $\sigma$  of an TIOA is a finite sequence of alternating delay and discrete transitions of the form  $(l_0, v_0) \xrightarrow{\delta_1} (l_0, v_0 + \delta_1) \xrightarrow{\tau_1} (l_1, v_1) \xrightarrow{\delta_2} \dots \xrightarrow{\delta_n} (l_{n-1}, v_{n-1} + \delta_n) \xrightarrow{\tau_n} (l_n, v_n)$ , where  $v_0 = \vec{0}$  and  $\tau_i = (l_{i-1}, \alpha_i, g_i, Y, l_i) \in E$ . By agreement we let  $\rightarrow^*$  denote the transitive closure of the transition-function.

The state space of TIOA can be represented via *zones* [2], symbolic sets of states containing the max. set of clock valuations satisfying given constraints.

In this paper, we only consider deterministic and input enabled TIOA.  $\mathcal{A}$  is *deterministic*, if  $\llbracket \mathcal{A} \rrbracket$  satisfies that  $\forall \alpha \in \Sigma, \forall (\ell, \nu) \in \mathcal{Q}. (\ell, \nu) \xrightarrow{\alpha} (\ell', \nu') \wedge (\ell, \nu) \xrightarrow{\alpha} (\ell'', \nu'') \implies \ell' = \ell'' \wedge \nu' = \nu''$ . Thus, in every state, transitions of a given label will lead to a unique state.  $\mathcal{A}$  is *input-enabled*, if  $\llbracket \mathcal{A} \rrbracket$  satisfies that  $\forall \alpha \in \Sigma^?, \forall (\ell, \nu) \in \mathcal{Q}. (\ell, \nu) \xrightarrow{\alpha} (\ell', \nu')$ . Thus, every state can receive any input. For simplicity, we assume implicit input-enabledness in our figures.

TIOA can be used as *specifications*. Specifications are not necessarily exact on timing behaviour (e.g., on the output of a label) but allow for ranges of timing. An *implementation* satisfies a specification as long as any behavior is included in that of the specification.

**Example.** Fig. 3(a) shows a specification  $\text{Spec}_1$  of a light switch: whenever the light is switched **ON**, this setting has to be kept for 1 to 5 time units. Whenever the light is switched **OFF**, the light switch has to blink at least once every 3 time units. The timing is tracked via the clock  $x$ .

We say that a TIOA  $\mathcal{A}_T$  *refines* a TIOA  $\mathcal{A}_S$  if the corresponding TLTS  $\llbracket \mathcal{A}_T \rrbracket$  refines the TLTS  $\llbracket \mathcal{A}_S \rrbracket$ , let us formally defined this relationship.

**Definition 3 (Refinement [12]).** A TLTS  $\mathcal{I} = (\mathcal{Q}_I, q_0^I, \rightarrow_I)$  refines a TLTS  $\mathcal{S} = (\mathcal{Q}_S, q_0^S, \rightarrow_S)$ , written  $\mathcal{I} \leq \mathcal{S}$ , iff there exists a binary relation  $R \subseteq \mathcal{Q}_I \times \mathcal{Q}_S$  containing  $(q_0^I, q_0^S)$  such that for each pair of states  $(q_I, q_S) \in R$  we have:

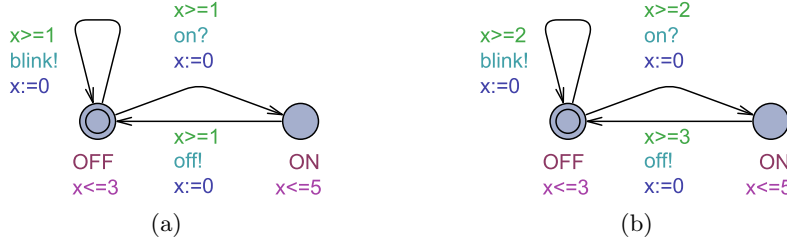


Fig. 3: Specification  $\text{Spec}_1$  (a) of a light switch; specification  $\text{Spec}_2$  refines it (b).

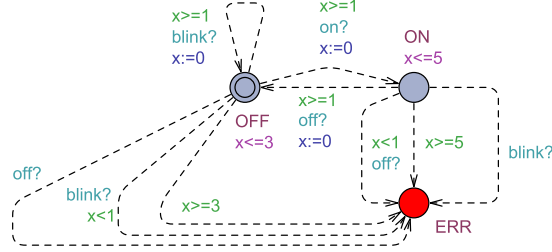


Fig. 4: A monitor  $\text{mSpec}$  of the light switch specification  $\text{Spec}_1$  of Fig. 3(a).

1. if  $\exists q'_S \in \mathcal{Q}_S. q_S \xrightarrow{i?}_S q'_S$  then  $\exists q'_I \in \mathcal{Q}_I. q_I \xrightarrow{i?}_I q'_I$  and  $(q'_I, q'_S) \in R$
2. if  $\exists q'_I \in \mathcal{Q}_I. q_I \xrightarrow{o!}_I q'_I$  then  $\exists q'_S \in \mathcal{Q}_S. q_S \xrightarrow{o!}_S q'_S$  and  $(q'_I, q'_S) \in R$  and
3. if  $\exists \delta \in \mathbb{R}_{\geq 0}. q_I \xrightarrow{\delta}_I q'_I$  then  $q_S \xrightarrow{\delta}_S q'_S$  and  $(q'_I, q'_S) \in R$ .

**Example.** Fig. 3(b) shows a specification  $\text{Spec}_2$  that refines the specification  $\text{Spec}_1$  of Fig 3(a) by restricting the timings of some of the signals.

Given a specification  $\text{Spec}$  and a system  $\text{Sys}$ , a **monitor**  $\text{mSpec}$  observes  $\text{Sys}$  w.r.t  $\text{Spec}$  (i.e., all transitions in  $\text{mSpec}$  are inputs) and enters an error-state **ERR** whenever non-conformance between  $\text{Sys}$  and  $\text{Spec}$  is observed, that is, whenever an output of  $\text{Sys}$  is observed that is not allowed by  $\text{Spec}$ .

**Example.** Fig. 4 depicts a monitor for  $\text{Spec}_1$  of Fig. 3(a).

We use *Networks of Timed Automata* to enable parallel composition of TIOA.

**Definition 4 (Networks of Timed Automata (NTA) [2]).** Let  $\Sigma$  be a set of actions and let  $\Sigma_1, \dots, \Sigma_n$  be a partitioning of  $\Sigma$ . Let  $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n$  be TIOAs, where  $\llbracket \mathcal{A}_i \rrbracket = (\mathcal{Q}^i, q_0^i, \rightarrow^i)$  and  $\mathcal{A}_i$  has  $\Sigma_i$  as its output and  $\Sigma$  as its input alphabet. A network of TIOA  $\mathcal{A}_1 \parallel \mathcal{A}_2 \parallel \dots \parallel \mathcal{A}_n$  is defined via the TLTS  $(\mathcal{Q}^1 \times \dots \times \mathcal{Q}^n, (q_0^1, \dots, q_0^n), \rightarrow)$  with:

$$\frac{[s_i \xrightarrow{\delta}_i s'_i]_{i=1..n}}{(s_1, \dots, s_n) \xrightarrow{\delta} (s'_1, \dots, s'_n)} \quad \text{with } \delta \in \mathbb{R}_{\geq 0}, \text{ and}$$

$$\frac{[s_j \xrightarrow{\alpha!}_j s'_j][s_i \xrightarrow{\alpha?}_i s'_i]_{i \neq j}}{(s_1, \dots, s_n) \xrightarrow{\alpha!} (s'_1, \dots, s'_n)} \quad \text{with } \alpha \in \Sigma_j.$$

For deterministic specifications, the following important theorem, given in the syntax of UPPAAL queries [6], holds, stating that **a system refines a specification iff the parallel product of the system and the specification monitor cannot reach the error-state.**

**Theorem 1.**  $\forall \text{ Sys. Sys} \leq \text{Spec} \iff (\text{Sys} \parallel \text{mSpec}) \models A\Box \neg \text{mSpec.ERR}$

**Definition 5 (Timed Game Automaton (TGA) [21]).** A *Timed Game Automaton (TGA)* is a TIOA in which the set of output actions  $\Sigma^!$  is partitioned into controllable actions ( $\Sigma_C$ ) and uncontrollable actions ( $\Sigma_U$ ).

The definition of NTA trivially extends to Networks of TGA (NTGA).

Given a TGA  $\mathcal{G} = (\langle L, \ell_0, \Sigma, X, E, I, \rangle, \Sigma_U, \Sigma_C)$ , a *memoryless strategy*  $\omega : L \times \mathbb{R}_{\geq 0}^X \rightarrow 2^{\{\lambda\} \cup \Sigma_C}$  is a function over the states of  $\mathcal{G}$  to the set of controllable actions or a special nothing-symbol  $\tau$ .

**Definition 6 (Strategy Composition).** Given a TGA  $\mathcal{G} = (\langle L, \ell_0, \Sigma, X, E, I, \rangle, \Sigma_U, \Sigma_C)$  and a memoryless strategy  $\omega$ , the composition  $\mathcal{G} \parallel \omega$  provides a restriction of the transition-system of  $\mathcal{G}$ ;

- if  $(\ell, \nu) \xrightarrow{\alpha} (\ell', \nu')$  then either we have  $\alpha \in \Sigma^?$  or we have  $\alpha \in \Sigma^!$  and  $\alpha \notin \Sigma_C$  or  $\alpha \in \omega(\ell, \nu)$ , and
- if  $(\ell, \nu) \xrightarrow{\delta} (\ell', \nu')$  for  $\delta \in \mathbb{R}_{\geq 0}$  then  $\forall \delta' < \delta$  it holds that  $\lambda \in \omega(\ell, \nu + \delta')$ .

Let  $\varphi \subseteq L$  be a set of losing locations for a TGA  $\mathcal{G}$ . The *safety control problem* consists in finding a strategy  $\omega$ , s.t.  $\mathcal{G} \parallel \omega$  constantly avoids  $\varphi$ . A trace  $\sigma = (l_0, v_0) \xrightarrow{\delta_1} (l_0, v_0 + \delta_1) \cdots \xrightarrow{\delta_n} (l_{n-1}, v_{n-1} + \delta_n) \xrightarrow{\tau_n} (l_n, v_n)$  is *winning* if  $\forall k \leq n : l_k \notin \varphi$ . A strategy  $\omega$  is winning, if all traces of  $\mathcal{G} \parallel \omega$  are winning.

We denote by  $\text{Wr} \subseteq L \times \mathbb{R}_{\geq 0}^X$  the *winning region*, i.e., the set of all states  $s$  such that there exists a winning strategy from  $s$ . We denote by *correct (wrong)* outputs for a state  $s$  the outputs that lead to an  $s' \in \text{Wr}$  ( $s' \notin \text{Wr}$ , respectively). We denote by  $\mathcal{S}_{wr}$  the set of states in  $\text{Wr}$ , such that any delay would leave  $\text{Wr}$ .

### 3 Timed Post-Shields

This section defines and gives the construction of timed post-shields, illustrated in Fig. 1(a). A timed post-shield is attached after the system, monitors its inputs and outputs, corrects the system's output if necessary and forwards the correct output to the environment.

#### 3.1 Definition of Timed Post-Shields

In this section, we define a timed post-shield based on its two desired properties:

**Definition 7 (Correctness for Post-Shields.).** Let  $\text{Spec}$  be a specification, and let  $\text{Sys}$  be a timed system. We say that a shield  $\text{Sh}$  ensures correctness if and only if it holds that

$$(\text{Sys} \mid \text{Sh}) \leq \text{Spec}.$$

That is, for any (faulty) system  $\text{Sys}$ , if it is placed in parallel with the shield, the shielded system is guaranteed to satisfy the specification.

**Definition 8 (No-Unnecessary-Deviation.).** *Let  $\text{Spec}$  be a specification, let  $\text{mSpec}$  be its corresponding monitor, let  $\text{Sys}$  be a timed system, and let  $\text{Sh}$  be a shield. Let  $\sigma$  be any correct timed trace of  $\text{Sys}$ , i.e., every action in  $\sigma$  is correct, and no state along  $\sigma$  is in  $\mathcal{S}_{wr}$ . We say that  $\text{Sh}$  does not deviate from  $\text{Sys}$  unnecessarily, if  $(\text{Sys}|\text{Sh})$  keeps the output for  $\sigma$  intact.*

In other words if  $\text{Sys}$  does not violate  $\text{Spec}$ ,  $\text{Sh}$  simply forwards the outputs of  $\text{Sys}$  to the environment without altering them. Once we reach a state in  $\mathcal{S}_{wr}$ , there is no way to know whether the system would produce an output on time, and the shield is allowed to deviate.

**Definition 9 (Timed Post-Shields.).** *Given a specification  $\text{Spec}$ ,  $\text{Sh}$  is a timed post-shield if for any timed system  $\text{Sys}$ , it holds that  $\text{Sh}$  enforces correctness of the shielded system w.r.t.  $\text{Spec}$  (Def. 7) and  $\text{Sh}$  does not deviate from  $\text{Sys}$  unnecessarily (Def. 8).*

### 3.2 Construction of Timed Post-Shields

In this section we discuss the synthesis procedure of timed post-shields without guarantees on the recovery time.

**Algorithm 1.** Let  $\text{Spec}$  be a specification,  $\text{mSpec}$  a monitor for  $\text{Spec}$ , and  $\text{Sys}$  a timed system. We construct a timed post-shield  $\text{Sh}$  via the following steps.

**Step 1: Construction of the monitor  $\text{mSpec}'$ .** To differentiate the outputs given by the system and those given by the shield, we prime the outputs of the shield. The monitor  $\text{mSpec}'$  is a copy of  $\text{mSpec}$ , where all outputs are primed.  $\text{mSpec}'$  is used to ensure that the outputs of the shield are correct.

**Step 2: Construction of the automaton  $\text{Ctr}$ .** The automaton  $\text{Ctr}$  is the only component that contains controllable transitions and depicts the control options for the shield, i.e.,  $\text{Ctr}$  produces the primed outputs.  $\text{Ctr}$  is constructed such that the no-unnecessary-deviation property is satisfied by the shield. Therefore,  $\text{mSpec}$  informs  $\text{Ctr}$  whether the system's output is wrong or the current state is in  $\mathcal{S}_{wr}$ , i.e., the winning region of  $\text{mSpec}$  is left.  $\text{Ctr}$  can perform three types of actions: (1) *pre-fault actions*: before an error was detected,  $\text{Ctr}$  can only mirror actions produced by  $\text{Sys}$ ; (2) *post-fault actions*: after an error was detected,  $\text{Ctr}$  has full control and can choose any action at any time; and (3) *last-chance actions*: whenever the current state is in  $\mathcal{S}_{wr}$  and any delay would leave the winning region, the shield can prevent that possibility and can choose any action that is allowed in the current state of  $\text{mSpec}'$ .

**Example.** Fig. 5 depicts the  $\text{Ctr}$  component for the light switch of Fig. 3(a). The two edges on the left (marked in blue) are pre-fault actions, that copy the system behaviour if no fault was detected, i.e.,  $\text{error} == 0$  (The error variable is set when  $\text{Sys}$  produces an output that would leave the winning region.). The two transitions in the top right corner (green) are post-fault actions: if an error was

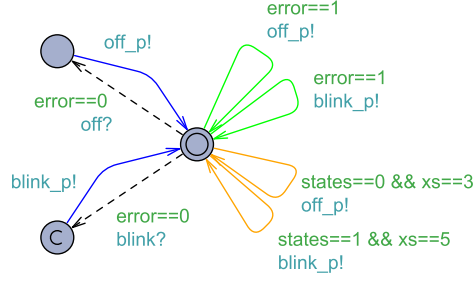


Fig. 5: An automaton  $\text{Ctr}$ , defining control options of a shield for the light switch.

detected, i.e.,  $\text{error} == 1$ , they will be enabled, and  $\text{Ctr}$  can choose any output. The two transitions in the bottom right (orange) show last-chance actions, for the case where the specification reached a time bound, in which case  $\text{Ctr}$  is allowed to take over, as a fault of the system may be imminent.

**Step 3: Construct the timed safety game  $\mathcal{G}$ .** We construct the timed game  $\mathcal{G}$  by the following composition:

$$\mathcal{G} = \text{mSpec} \mid \text{mSpec}' \mid \text{Ctr}$$

In this game the monitor  $\text{mSpec}$  observes whether the system  $\text{Sys}$  satisfies the specification  $\text{Spec}$ , the monitor  $\text{mSpec}'$  observes whether the outputs of the shields are correct, and  $\text{Ctr}$  enforces that the shield does not deviate unnecessarily and is in charge of producing the primed outputs.

**Step 4: Compute a strategy  $\omega_S$  of  $\mathcal{G}$ .** Post-shields ensure correctness (safety), i.e., the control objective is to ensure that  $\text{Spec}'$  is never violated by the shielded system. This can be expressed via the following safety query specifying that the error state should never be reached, given in UPPAAL TIGA syntax.

$$\text{control} : A\Box \neg \text{mSpec}'.\text{ERR}$$

Solving the safety game w.r.t. this query produces a strategy  $\omega_S$ , which we use in the next step to produce a timed post-shield.

**Step 5: Construction of the timed post-shield  $\text{Sh}_S$ .** From the strategy  $\omega_S$ , we construct the shield  $\text{Sh}_S$  in the following way: A shield  $\text{Sh}_S$  is the network of timed automata received by composing  $\mathcal{G}$  with the derived strategy  $\omega_S$ , denoted by  $\mathcal{G}|\omega_S$ , meaning that all unsafe transitions (or transitions that would not lead to recovery) are restricted. This shield may still permit multiple outputs in a given state, any of which ensures safety.

**Theorem 2.** A shield  $\text{Sh}_S$  constructed according to Alg. 1 is a timed post-shield.

*Proof.* We have to proof that  $\text{Sh}_S$  satisfies the correctness property and the no-unnecessary-deviation property.  $\text{Sh}_S$  satisfies the correctness property, since all transitions leaving the winning region were removed from  $\mathcal{G}|\omega_S$ . It thus holds that  $(\text{Sys} \parallel \text{Sh}) \leq \text{Spec}'$ . Thus the primed outputs of the shield satisfy the



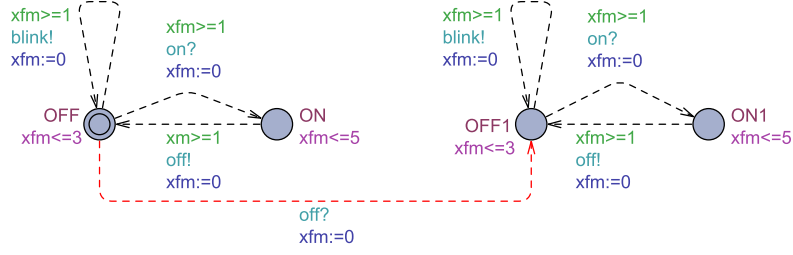


Fig. 6: A fault model  $\text{Spec}^{fi}$  for a transient fault that captures an unexpected output that resets a clock.

specification. Additionally, the construction  $\mathcal{G}$  via the automaton  $\text{Ctr}$  ensures that the shield cannot alter an action before a fault occurs, thereby ensuring that the shield cannot deviate unnecessarily  $\square$

## 4 Timed Post-Shields with Recovery Guarantees

In this section, we first discuss the challenges that we face when synthesizing shields with the ability to recover from system faults and discuss assumptions that we make on the system and its faults that are necessary for our synthesis procedure. Next, we define and construct timed post-shields with *guaranteed recovery* and *guaranteed time-bounded recovery*.

### 4.1 Recovery under Fault Models

In shield synthesis, we consider the system to be shielded as a black box, which brings a huge scalability advantage, especially when shielding complex timed implementations. Therefore, a shield has to ensure correctness for any system.

In order to end the recovery phase and to hand back control to the system, the shield needs to *resolve* the error that occurred, i.e., the state of the shield needs to align with the actual state of the system. The tricky part is, as mentioned before, that the system is considered as a black box and the shield can only observe the system's outputs but not its internal state.

In this paper we assume that the only violations that happen are due to *transient errors*.

**Definition 10 (Transient Error).** *A transient error is an error that happens only once, and has correct pre-error and post-error behavior.*

To determine the state of a system in case of a fault, we launch several *fault models* and assume that one of this fault models captures the fault.

**Definition 11 (Fault Model).** *Let  $\text{Spec}$  be a specification. A fault model  $\text{Spec}^{fi}$  for  $\text{Spec}$  consists of two copies of the specification  $\text{Spec}$ , one copy for the pre-fault behavior and one copy for the post-fault behavior. The two copies are connected with a single transient error.*

**Example.** Fig. 6 gives an example of a fault model. The fault captures the situation, in which an *off* signal is produced in the **OFF** location and this faulty signal additionally resets the clock *xfm*, but does not change the location of the automaton. In the fault model, this transient fault leads from the **OFF** location from the pre-fault part to the **OFF** in the post-fault part.

Since we cannot observe the internal state of the system, we do not know which fault model captured the fault that occurred. Thus, the shield can only end the recovery phase if all fault models and the specification *align*; i.e., if all  $\text{Spec}^{f_i}$  and  $\text{mSpec}$  reach the same state. To achieve this, in the recovery phase, we monitor the behavior of the system and update the fault models accordingly. If a fault model can not follow the output of the system (including not allowing a delay that is possible in the system), it was not the correct fault model for the observed fault and is discarded. Only if all *active*, i.e., non-discarded, fault models agree on the same state, the shield and the system synchronized again.

**Types of transient faults.** We consider fault models covering the following fault types; categorized in location-faults, clock-faults, and their combination.

- Location-faults:
  - *Go-to-any-location* faults: the system goes to an arbitrary location. To track the its location, we need a fault model for every location in  $\text{Spec}$ .
  - *Go-to-next-location* faults: the system gives an incorrect output, but continues in a correct successor location. Thus, the fault models only need to cover the valid successor locations.
- Clock-faults:
  - *Wrong-reset* faults: the system illegally resets a clock to zero. Such faults occur, e.g., if the system gives a planned reset at a wrong point in time, or the systems resets the wrong clock.
  - *Swapped-clocks* faults: the system swaps the values of several selected clocks. This might be a *binary* swaps between two clocks, or *permutations* between several clocks.
  - *Missing-reset* faults: the system ignores a planned reset of a clock, resulting in a clock value that is too high.

**Example.** The fault shown in Fig. 6(a) is a *wrong-reset* fault, i.e., the fault does not change the location of the model, but only resets the clock *xfm*.

## 4.2 Definition of Timed Post-Shields with Recovery Guarantees

We now define shields which satisfy an additional property: *guaranteed-recovery* or *guaranteed-time-bounded-recovery*.

**The Guaranteed-Recovery Property.** In this paper, we synthesize shields with guaranteed (time-bounded) recovery under the assumption that the system satisfies the specification except for a single transient fault and that this fault is covered by one of the fault models  $\text{Spec}^{f_i}$  with  $i \in \{1 \dots n\}$ .

**Definition 12 (Guaranteed Recovery).** Let  $\text{Spec}$  be a specification, let  $\text{mSpec}$  be its monitor, let  $\text{Spec}^f = \{\text{Spec}^{f_1} \dots \text{Spec}^{f_n}\}$  be a set of fault models, let  $\text{Sys}$  be

a timed system with  $\text{Sys} \leq \text{Spec}^{f_i}$  for some  $\text{Spec}^{f_i} \in \text{Spec}^f$ . We say that  $\text{Sh}$  guarantees recovery if for every trace  $\sigma$  containing a single transient fault, i.e., there exists a point in time  $t_1$  such that  $\text{mSpec}$  reaches the error location **ERR** at  $t_1$ , there exists a time  $t_2 > t_1$  such that after  $t_2$   $(\text{Sys}|\text{Sh})$  keeps  $\sigma$  intact.

That is, if the system refines any of our considered fault models, we guarantee that an observed error will lead to recovery and the system and the shield give the same output again.

**The Guaranteed-Time-Bounded-Recovery Property.** This property guarantees that the recovery phase lasts for at most  $T$  time units after a fault.

**Definition 13 (Guaranteed Time-Bounded Recovery).** Let  $\text{Spec}$  be a specification, let  $\text{mSpec}$  be its monitor, let  $\text{Spec}^f = \{\text{Spec}^{f_1} \dots \text{Spec}^{f_n}\}$  be a set of fault models, let  $\text{Sys}$  be a timed system with  $\text{Sys} \leq \text{Spec}^{f_i}$  for some  $\text{Spec}^{f_i} \in \text{Spec}^f$ . We say that  $\text{Sh}$  guarantees recovery within a bound  $T$  if for every trace  $\sigma$  containing a single transient fault, i.e., there exists a point in time  $t_1$  such that  $\text{mSpec}$  reaches the error location **ERR** at  $t_1$ , we have that after time  $t_1 + T$ ,  $(\text{Sys}|\text{Sh})$  keeps  $\sigma$  intact.

### 4.3 Construction of Timed Post-Shields with Recovery Guarantees

In this section we discuss the synthesis procedure of timed post-shields with *guaranteed recovery* and with *guaranteed time-bounded recovery*.

**Algorithm 2.** Let  $\text{Spec}$  be a specification,  $\text{mSpec}$  its monitor, and  $\text{Spec}^f = \{\text{Spec}^{f_1} \dots \text{Spec}^{f_n}\}$  a set of fault models. Starting from  $\text{Spec}$ ,  $\text{mSpec}$ ,  $\text{Spec}^f$ , we construct a timed post-shield with guaranteed-recovery ( $\text{Sh}_G$ ), or with guaranteed-time-bounded-recovery ( $\text{Sh}_{GT}$ ) via the following steps.

**Steps 1 and 2.** Perform as in Section 3.2.

**Step 3: Construct the monitors  $\text{mSpec}^{f_i}$ .** Transform the fault models  $\text{Spec}^{f_i}$  into monitors  $\text{mSpec}^{f_i}$  for  $i \in \{1 \dots n\}$ .

**Step 4: Construct the timed Game  $\mathcal{G}$ .** Now we can consider the timed game given by the following composition:

$$\mathcal{G} = \text{mSpec}^{f_1} \mid \dots \mid \text{mSpec}^{f_i} \mid \text{mSpec} \mid \text{mSpec}' \mid \text{Ctr}$$

In this game, we observe conformance of the system with respect to the fault models  $\text{mSpec}^{f_i}$  and the original specification via  $\text{mSpec}$ , enforce the correctness of the shield via  $\text{mSpec}'$ , and ensure no unnecessary deviations via **Ctr**. Next, we compute the winning strategies  $\omega_g$  and  $\omega_{gt}$  of  $\mathcal{G}$  such that the corresponding shields guarantee recovery and guarantee time-bounded recovery, respectively.

**Step 5a: Compute a strategy  $\omega_g$  of  $\mathcal{G}$  for guaranteed-recovery.** For *guaranteed recovery*, we need to establish a state where all active fault models agree with each other and the  $\text{Spec}'$ , that is, they are all in the same location with same clock values. This can be achieved with the following leads-to property [6], specifying that if we observe a fault, this will eventually lead to recovery.

$$\begin{aligned} &\text{control} : A \Box \neg \text{mSpec}'.\text{ERR} \wedge \text{mSpec}.\text{ERR} \text{ leadsto} \\ &(\forall i. [\text{mSpec}^{f_i}.\text{ERR} \vee (\text{mSpec}^{f_i}.l == \text{mSpec}'.l \wedge \text{mSpec}^{f_i}.x = \text{mSpec}'.x)]) \end{aligned}$$

Solving  $\mathcal{G}$  w.r.t. this query will produce a strategy  $\omega_g$ . (Note, that if a fault model is inactive, i.e., it reached its error-state, this means that  $\text{Sys}$  performed an output that was not valid in the fault model.)

**Step 5b: Compute a strategy  $\omega_{gt}$  of  $\mathcal{G}$  for guaranteed-time-bounded-recovery.** We slightly change the timed leads-to property to compute  $\omega_{gt}$  for guaranteed recovery within a time bound  $T$ :

$$\begin{aligned} \text{control} : \mathbf{A} \Box \neg \text{mSpec}'.\text{ERR} \wedge \text{mSpec}.\text{ERR} \text{ leadsto}_{\leq T} \\ (\forall i. [\text{mSpec}^{f_i}.\text{ERR} \vee (\text{mSpec}^{f_i}.l == \text{mSpec}'.l \wedge \text{mSpec}^{f_i}.x = \text{mSpec}'.x)]) \end{aligned}$$

**Step 6: Construction of the timed post-shields  $\text{Sh}_G$  and  $\text{Sh}_{GT}$ .** (We construct a shield with guaranteed-recovery  $\text{Sh}_G$  by  $\mathcal{G} \parallel \omega_g$ , and a shield with guaranteed-time-bounded-recovery  $\text{Sh}_{GT}$  by  $\mathcal{G} \parallel \omega_{gt}$ .)

**Theorem 3.** *The shields  $\text{Sh}_G$  and  $\text{Sh}_{GT}$ , constructed by Alg. 2, are timed post-shields with guaranteed-recovery and guaranteed-time-bounded-recovery, resp.*

*Proof.* Correctness and no-unnecessary-deviation are given as for regular timed post-shields.  $\text{Sh}_G$  ( $\text{Sh}_{GT}$ ) is a shield with guaranteed-(time-bounded)-recovery, simply by the query it was produced from, which can only be satisfied if all traces in  $\text{Sh}_G$  ( $\text{Sh}_{GT}$ ) either do not encounter a fault, or the fault will lead to recovery at some point (within  $T$  time units) along the trace.  $\square$

**Discussion: timed post-shields.** (Both types of shields, with and without recovery, have pros and cons. Obviously, the ability of guaranteed recovery after a system fault is highly desirable.) This is especially true, if the system to be shielded is highly optimized and performs complex tasks that are not captured in the specification of the shield. Nevertheless, it may not be feasible to synthesize post-shields with the ability to recover. First, guaranteed recovery is not always possible and therefore, no shield that guarantees recovery may exist. An obvious example that demonstrates this fact is a shield for a system that never resets one of its clocks. If a fault occurred that changes the value of this clock, synchronization is never possible. Second, to capture any transient fault that is possible under a given fault category, an exponential number of fault models is needed. This results in an exponential blowup in the state space and synthesis time. Therefore, it may not be feasible to synthesize shields with guaranteed recovery for large specifications considering a large number of fault models.

## 5 Timed Pre-Shields

In this section we define and construct timed pre-shields. A timed pre-shield is attached before the system as illustrated in Figure 1(a). At any point in time, a timed pre-shield provides a *set of actions*  $\text{Act}$  for the system to choose from. This set of action can contain a delay action. If this is the case, the system is permitted to wait without performing any discrete action. If the set does not include a delay action, the system has to produce an output immediately. If the system picks the outputs according to this list, then it is guaranteed that the shield and the timed system together satisfy the specification.

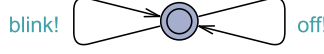


Fig. 7: The control options for a timed-pre shield of the light switch.

### 5.1 Definition of Timed Pre-Shields

In this section we define timed pre-shields based on two properties; correctness and no unnecessary restriction.

**Definition 14 (Correctness for Pre-Shields).** *Let  $\text{Spec}$  be a specification, and let  $\text{Sh}$  be a pre-shield. For any state  $s$  from  $\text{Sh}$ , let  $\text{Act} = \alpha_1, \dots, \alpha_n$  be the set of enabled actions sent by  $\text{Sh}$ . A pre-shield is correct, if it holds that if  $\alpha$  is a wrong action for the given situation, then  $\alpha \notin \text{Act}$ .*

**Definition 15 (No-Unnecessary-Restriction).** *Let  $\text{Spec}$  be a specification, and let  $\text{Sh}$  be a pre-shield. For any state  $s$  from  $\text{Sh}$ , let  $\text{Act} = \alpha_1, \dots, \alpha_n$  be the set of enabled actions sent by  $\text{Sh}$ . A pre-shield is not unnecessarily restrictive, if it holds that if  $\alpha$  is a correct action for the given situation, then  $\alpha \in \text{Act}$ .*

**Definition 16 (Timed Pre-Shields.).** *Given a specification  $\text{Spec}$ .  $\text{Sh}$  is a timed pre-shield if it holds that  $\text{Sh}$  enforces correctness for any timed system  $\text{Sys}$  w.r.t.  $\text{Spec}$  (Def. 14) and  $\text{Sh}$  is not unnecessarily restrictive (Def. 15).*

### 5.2 Construction of Timed Pre-Shields

We construct timed pre-shields according to the following algorithm.

**Algorithm 3.** Starting from  $\text{Spec}$ ,  $\text{mSpec}$ , we construct a timed pre-shield.

**Step 1. Construction of the automaton  $\text{Ctr}$ .** Since there is no concept of minimal deviation in pre-shields, the control options of the  $\text{Ctr}$  component are not restricted. Instead, in this setting, we build  $\text{Ctr}$  such that it can fire any output at any time, i.e., it has a single location and unguarded self-loops for every output. Note that timed pre-shields do not need primed outputs.

**Example.** The component  $\text{Ctr}$  for the light switch is depicted in Fig. 7.

**Steps 2. Construct the timed safety game  $\mathcal{G}$ .** We construct the timed game  $\mathcal{G}$  by the following composition:

$$\mathcal{G} = \text{mSpec} \mid \text{Ctr}$$

**Step 3: Compute a strategy  $\omega_{pre}$  of  $\mathcal{G}$ .** Pre-shields ensure correctness (safety), i.e., the control objective is to ensure that  $\text{Spec}$  is never violated. Solving the safety game w.r.t. the following query produces a strategy  $\omega_{pre}$ .

$$\text{control} : A \Box \neg \text{mSpec.ERR}$$

**Step 4: Computing the set of enabled actions via zones.** For a given state, the set of enabled actions  $\text{Act}$  is computed via zones. From any given state, its zones can be calculated straightforwardly, see [2].

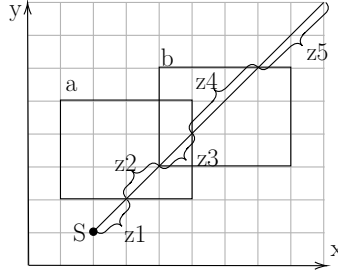


Fig. 8: The zones reachable by delay from a state  $s$  in a TIOA with two clocks,  $x$  and  $y$ . The squares represent constraints for two transitions with label  $a$  and  $b$ .

**Example.** The concept of zones is illustrated in Fig. 8, where the  $X$  and  $Y$  axis depict different clocks, and the squares represent the constraints in which different controllable actions  $a$  and  $b$  are enabled. We have for  $a$  :  $\{1 < x < 5, 2 < y < 5\}$  and for  $b$  :  $\{4 < x < 8, 3 < y < 6\}$ .

The set of enabled actions is kept up to date by monitoring the current state of  $\text{Sh}_{pre}$ . Whenever a new input or output from the system is received, the state is updated. From the new state, we calculate all zones that can be reached via delay and the actions enabled in each zone. The set of actions  $\text{Act}$  for the current zone is sent to the system. In case the end of the zone is not met yet, this includes a delay action. If enough time passes so that the end of the current zone is met, the shield needs to check whether future zones permit actions. If so, the set of actions of the next zone is passed to the system. Otherwise, the current set of actions is transmitted again, this time without a delay action. The system may choose any of the permitted actions, including delay if possible.

**Example.** In Fig. 8, the actions enabled per zone are  $z1 = \{\}$ ,  $z2 = \{a\}$ ,  $z3 = \{a, b\}$ ,  $z4 = \{b\}$ ,  $z5 = \{\}$ . Thus, in the state  $S$  which is in  $z1$ ,  $\text{Act}$  is empty, after one time unit  $\text{Act} = \{a\}$ , and so on.

**Theorem 4.** A shield  $\text{Sh}_{pre}$  constructed according to Alg. 3 is a timed pre-shield.

*Proof.* In order for  $\text{Sh}_{pre}$  to be a timed pre-shield,  $\text{Sh}_{pre}$  needs to satisfy correctness for pre-shields and the no-unnecessary-restriction property.  $\text{Sh}_{pre}$  is correct, as the safety game removes every transition leaving the winning region. It has no-unnecessary-restriction, as  $\text{Ctr}$  is not restricted when producing outputs, and all correct actions are kept when producing the strategy.  $\square$

**Discussion: Pre-shielding vs. Post-shielding.** Post-Shielding has the advantage, that we treat the system as a total back box. In order for pre-shielding to work, we need a system that chooses the actions w.r.t. the suggestions of the shield. Usually, we have this setting in RL where the shield can easily influence the set of available actions from the agent. Instead of overruling the system, a pre-shield leaves the choice of which action should be executed always to the system, i.e., the system can do anything as long as it is safe. Thus, the overall efficiency of the system is kept as high as possible. It has already been shown that

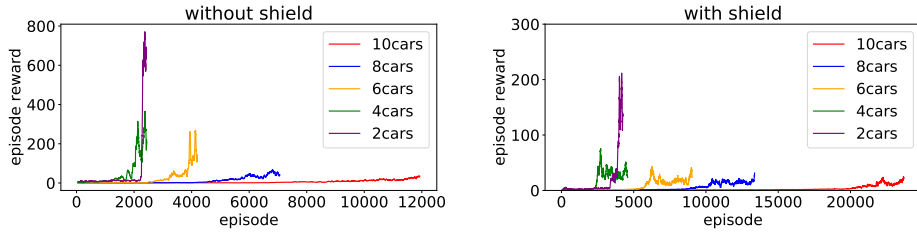


Fig. 9: Results training phase.

removing unsafe options during learning can significantly speed up the learning process [1].

## 6 Experiments

To validate our approach, we extend the case-study of Larsen et. al. [20] to a platoon of multiple cars<sup>4</sup>. In the case study, an RL agent controls  $n$  *follower* vehicles in the platoon following an (environment controlled) *leader* vehicle. All vehicles can drive a maximum of 20m/s and have three different possible accelerations modes:  $-2\text{m/s}^2$ ,  $0\text{m/s}^2$  and  $2\text{m/s}^2$  which can be changed at every time unit. The goal of the RL agent is to control the followers in the platoon such that the total distance between all vehicles is minimized. Furthermore, the RL agent receives a negative reward if the distance between two cars is outside a safe region ( $\leq 5\text{m}$ ) or is too large (above 200m). The hyper-parameters of the RL setting can be found in Appendix A. We used the models from [20] and synthesized timed post-shields with UPPAAL TIGA, as discussed in Sec. 3. We study the behaviour of RL agents in the context of 1. no shielding, 2. post-shielding during execution, and 3. post-shielding during both training and execution. We report the learning curves during the training phase and the performances in the execution phase for  $n \in \{2, 4, 6, 8, 10\}$  where  $n$  denotes the number of cars.

*Training.* Each training episode starts with random but safe initial distances and velocities of all cars. During the simulation, the environment picks the accelerations of the leading car via a uniform distribution. A training episode lasts for 2000 time units, or until the distance between two cars gets smaller than 5m or larger than 200m. Note, that with a shield, a training episode always lasts 2000 time units, since safety is always guaranteed.

Fig. 9 compares the learning curves as a mean of 20 training phases, for the unshielded case (left) and the shielded case (right). The reward in the unshielded case is considerably higher than in the shielded setting. We observe that the agent exploits the relatively low risk of a crash and makes potentially unsafe choices. Since the accelerations of the leading car are picked via an uniform distribution, it

<sup>4</sup> The source code, including some demonstrative videos and the running example used in the paper, is available online [22].

	No Shield			Shield E	Shield T+E
#Cars	#Crashes	Time	Reward	Reward	Reward
2	703	1133	747	915	603
4	13	1989	1070	685	393
6	0	2000	638	617	375
8	85	1908	477	495	386
10	983	544	170	608	342

Table 1: Results exploitation phase using 10000 simulations. Number of crashes is given in absolute values over all simulations whereas Reward and Time measures are given as averages. Time and Crash values are omitted when shielding is applied as these are  $> 2000$  and  $0$  respectively. Time denotes the time-units of simulation prior to a first crash.

is unlikely, that e.g., the leading car accelerates to the maximum speed and then immediately hits the break until it reaches zero. Such risk tolerance is not allowed when deploying the shield as even a potential but unlikely future crash should be shielded against. *Execution Phase.* We tested all controller combinations for 1000 simulations, and each simulation lasts until a crash or for a maximum of 2000 time units. Table 1 depicts the results. Note, that we learned a global controller for each number of cars (but use local shields) and that the controllers optimize a local minimum, therefore the controllers performances differ from each other. Interestingly, we observe that the combination of unshielded training (Shield E) provides better results in our setting, compared to a RL agent utilizing the shield also during training (Shield T+E). But more experiments are needed to discern this effect in more detail.

## 7 Conclusion

We presented timed post-shields and timed pre-shields. These are shields for real-time systems which can be attached either before or after a system, correcting the outputs received by the environment. In addition, we discussed how timed shields can be used in reinforcement learning settings. We presented a case study of a platoon of cars, and demonstrated the potential of a timed post-shield in RL. In the future, we would like to extend this work into several different directions. First, we see great potential for shields in RL. In future work, we want to apply timed shields on several challenging RL settings and investigate techniques for speeding up the learning performance in addition to providing safety. Furthermore, we want to exploit techniques from model repair and model refinement to deal with dynamic environments, and adapt the shields during runtime if needed. In this work, we treat the system as adversary. In future work, we plan to study ways to model the spectrum between cooperative and adversarial systems to be shielded.



## References

1. M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, and U. Topcu. Safe reinforcement learning via shielding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2669–2678, 2018.
2. R. Alur and D. L. Dill. A theory of timed automata. *Theoretical computer science*, 126(2):183–235, 1994.
3. E. Bartocci, Y. Falcone, B. Bonakdarpour, C. Colombo, N. Decker, K. Havelund, Y. Joshi, F. Klaedtke, R. Milewicz, G. Reger, G. Rosu, J. Signoles, D. Thoma, E. Zalinescu, and Y. Zhang. First international competition on runtime verification: rules, benchmarks, tools, and final results of CRV 2014. *Int. J. Softw. Tools Technol. Transf.*, 21(1):31–70, 2019.
4. E. Bartocci, Y. Falcone, A. Francalanza, and G. Reger. Introduction to runtime verification. In *Lectures on Runtime Verification - Introductory and Advanced Topics*, pages 1–33, 2018.
5. A. Bauer, M. Leucker, and C. Schallhart. Runtime verification for LTL and TLTL. *ACM Trans. Softw. Eng. Methodol.*, 20(4):14:1–14:64, 2011.
6. G. Behrmann, A. David, and K. G. Larsen. A tutorial on uppaal. In *Formal methods for the design of real-time systems*, pages 200–236. Springer, 2004.
7. J. O. Blech, Y. Falcone, and K. Becker. Towards certified runtime verification. In *Formal Methods and Software Engineering - 14th International Conference on Formal Engineering Methods, ICFEM 2012, Kyoto, Japan, November 12-16, 2012. Proceedings*, pages 494–509, 2012.
8. F. Cassez, A. David, E. Fleury, K. G. Larsen, and D. Lime. Efficient on-the-fly algorithms for the analysis of timed games. In *CONCUR 2005 - Concurrency Theory, 16th International Conference, CONCUR 2005, San Francisco, CA, USA, August 23-26, 2005, Proceedings*, pages 66–80, 2005.
9. A. David, P. G. Jensen, K. G. Larsen, A. Legay, D. Lime, M. G. Sørensen, and J. H. Taankvist. On time with minimal expected cost! In F. Cassez and J.-F. Raskin, editors, *Automated Technology for Verification and Analysis*, pages 129–145, Cham, 2014. Springer International Publishing.
10. A. David, P. G. Jensen, K. G. Larsen, M. Mikucionis, and J. H. Taankvist. Uppaal stratego. In *Tools and Algorithms for the Construction and Analysis of Systems - 21st International Conference, TACAS 2015, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2015, London, UK, April 11-18, 2015. Proceedings*, pages 206–211, 2015.
11. A. David, K. G. Larsen, A. Legay, U. Nyman, and A. Wasowski. Ecdar: An environment for compositional design and analysis of real time systems. In *International Symposium on Automated Technology for Verification and Analysis*, pages 365–370. Springer, 2010.
12. A. David, K. G. Larsen, A. Legay, U. Nyman, and A. Wasowski. Timed i/o automata: a complete specification theory for real-time systems. In *Proceedings of the 13th ACM international conference on Hybrid systems: computation and control*, pages 91–100, 2010.
13. Y. Falcone, L. Mounier, J. Fernandez, and J. Richier. Runtime enforcement monitors: composition, synthesis, and enforcement abilities. *Formal Methods Syst. Des.*, 38(3):223–262, 2011.

14. Y. Falcone and S. Pinisetty. On the runtime enforcement of timed properties. In *Runtime Verification - 19th International Conference, RV 2019, Porto, Portugal, October 8-11, 2019, Proceedings*, pages 48–69, 2019.
15. N. Fulton and A. Platzer. Verifiably safe off-model reinforcement learning. In *Tools and Algorithms for the Construction and Analysis of Systems - 25th International Conference, TACAS 2019, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2019, Prague, Czech Republic, April 6-11, 2019, Proceedings, Part I*, pages 413–430, 2019.
16. K. W. Hamlen, J. G. Morrisett, and F. B. Schneider. Computability classes for enforcement mechanisms. *ACM Trans. Program. Lang. Syst.*, 28(1):175–205, 2006.
17. N. Jansen, B. Könighofer, S. Junges, and R. Bloem. Shielded decision-making in mdps. *CoRR*, abs/1807.06096, 2018.
18. Y. Ji and S. Lafortune. Enforcing opacity by publicly known edit functions. In *56th IEEE Annual Conference on Decision and Control, CDC 2017, Melbourne, Australia, December 12-15, 2017*, pages 4866–4871, 2017.
19. B. Könighofer, M. Alshiekh, R. Bloem, L. Humphrey, R. Könighofer, U. Topcu, and C. Wang. Shield synthesis. *Formal Methods in System Design*, 51(2):332–361, 2017.
20. K. G. Larsen, M. Mikucionis, and J. H. Taankvist. Safe and optimal adaptive cruise control. In *Correct System Design - Symposium in Honor of Ernst-Rüdiger Olderog on the Occasion of His 60th Birthday, Oldenburg, Germany, September 8-9, 2015. Proceedings*, pages 260–277, 2015.
21. O. Maler, A. Pnueli, and J. Sifakis. On the synthesis of discrete controllers for timed systems. In *Annual Symposium on Theoretical Aspects of Computer Science*, pages 229–242. Springer, 1995.
22. A. Palmisano, F. Lorber, B. Knighofer, P. G. Jensen, R. Bloem, and K. Guldstrand. Experiments for "It's Time to Play Safe: Shield Synthesis for Timed Systems", June 2020. <https://doi.org/10.5281/zenodo.3903227>.
23. M. Renard, Y. Falcone, A. Rollet, T. Jéron, and H. Marchand. Optimal enforcement of (timed) properties with uncontrollable events. *Math. Struct. Comput. Sci.*, 29(1):169–214, 2019.
24. F. B. Schneider. Enforceable security policies. *ACM Trans. Inf. Syst. Secur.*, 3(1):30–50, 2000.
25. L. Sha. Using simplicity to control complexity. *IEEE Softw.*, 18(4):20–28, 2001.
26. M. Wu, J. Wang, J. Deshmukh, and C. Wang. Shield synthesis for real: Enforcing safety in cyber-physical systems. In *2019 Formal Methods in Computer Aided Design, FMCAD 2019, San Jose, CA, USA, October 22-25, 2019*, pages 129–137, 2019.
27. M. Wu, H. Zeng, and C. Wang. Synthesizing runtime enforcer of safety properties under burst error. In *NASA Formal Methods - 8th International Symposium, NFM 2016, Minneapolis, MN, USA, June 7-9, 2016, Proceedings*, pages 65–81, 2016.
28. Y. Wu, V. Raman, B. C. Rawlings, S. Lafortune, and S. A. Seshia. Synthesis of obfuscation policies to ensure privacy and utility. *J. Autom. Reasoning*, 60(1):107–131, 2018.

## Appendix A Reinforcement Learning Configuration

Our hyper-parameters for the DQN were chosen in the following way. The input features consists of the distances between the cars and the velocities of the cars.

Therefore, for  $n$  follower cars in the platoon, the input layer has the size  $2*n+1$ . We have DNNs for actor and critic, containing 3 hidden layers with Rectified Linear Units and a linear layer for the output. Networks were optimized with an Adamax optimizer. We used 16 units in the hidden layers. We used the learning rate  $\alpha = 0.002$  and the exponential decay rates  $\beta_1 = 0.9$  and  $\beta_2 = 0.9999$ . The output layer is  $3^n$ , since the RL agent can pick one of three different possible accelerations for each follower car. The reward function is designed such that the total spacing between the vehicle is minimized. If the distance between any two cars is either  $\leq 5\text{m}$  or  $\geq 200\text{m}$ , then the reward is set to  $-1$ . In all other cases, the distances between the cars are used within a logarithmic scale to determine the reward  $0 \leq r \leq 1$  per step.