Quick Bug Detection Through Black-Box Checking: a Systematic Evaluation

Anonymous Author(s)

ABSTRACT

Combinations of active automata learning, model-based testing and model checking have been successfully used in numerous applications, e.g., for spotting bugs in implementations of major network protocols and to support refactoring of embedded controllers. However, in the large majority of these applications, model checking is only used at the very end, when no counterexample can be found anymore for the latest hypothesis model. This contrasts with the original proposal of black-box checking (BBC) by Peled, Vardi & Yannakakis, which applies model checking for *all* hypotheses, also the intermediate ones. In this article, we present the first systematic evaluation of the ability of BBC to find bugs quickly, based on 77 benchmark models from real protocol implementations and controllers for which specifications of safety properties are available. Our main finding are: (a) In cases where the full model can be learned, BBC detects violations of the specifications with just 3% of the queries needed by an approach in which model checking is only used for the full model. (b) Even when the full model cannot be learned, BBC is still able to detect many violations of the specification. In particular, BBC manages to detect 96% of the safety property violations in the challenging RERS 2019 industrial LTL benchmarks. (c) Our results also confirm that BBC is way more effective than existing MBT algorithms in finding deep bugs in implementations.

CCS CONCEPTS

• **Software and its engineering → Formal methods**; **Software verification and validation**.

KEYWORDS

Black-box checking, Active automata learning, Model learning, Model checking, Model-based testing, Runtime monitoring

ACM Reference Format:

Anonymous Author(s). 2026. Quick Bug Detection Through Black-Box Checking: a Systematic Evaluation. In *Proceedings of IEEE/ACM Automated Software Engineering (ASE) Conference.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

IEEE/ACM Automated Software Engineering (ASE) Conference, October 12–16, 2026, Munich, Germany

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Testing whether a software system conforms to a high-level specification is challenging without access to source code or a detailed, low-level model. Without guidance for the selection of tests inputs, testers typically resort to random test generation algorithms, possibly in combination with user-provided input grammars and evolutionary mechanisms [29, 64]. However, such methods can easily fail to expose software bugs whose occurrence depends upon specific sequences of input events.

Recently, some powerful learning-based testing techniques have been proposed that address these limitations by combining testing and learning [4, 39, 40, 44, 45, 49]. These techniques are based on ideas of [10, 60] about the duality between inductive inference and conformance testing: where inductive inference aims to construct a hypothesis model based on the outcomes of a finite number of experiments, the goal of conformance testing is to find a counterexample for this hypothesis by performing experiments. The central idea of learning-based testing is to create a feedback loop in which counterexamples found during testing help to make the learned models more accurate, whereas the learned models drive the test generation to obtain increasingly powerful test suites.

Learning-based testing has been successfully used in numerous applications, e.g., for spotting security vulnerabilities and bugs in implementations of major network protocols [18, 20–22, 24, 50] and to support refactoring of embedded controllers [51, 52, 61]. We refer to [2, 14, 31, 57] for surveys and further references. As a result, learning-based testing has become a standard tool in the toolbox of software engineers who are trying to find deep bugs in protocol implementations or state-machine-based embedded controllers.

In their seminal paper from 1999 on learning-based testing, Peled, Vardi & Yannakakis [44, 45] combine three approaches for the verification of computer-based systems into what they call *black-box checking (BBC)*: *model checking*, which checks if a known state diagram model conforms to a specification, *conformance testing* (a.k.a. *model-based testing (MBT)*), which checks if a black-box system conforms with an abstract design, and *active automata learning* (a.k.a. *model learning*), which constructs a state diagram model of a black-box system by providing inputs and observing outputs. By combining these approaches, the authors obtain a method to check whether a black-box implementation satisfies a high-level specification. Conceptually, BBC is just a form of model-based testing. However, by learning the state-transition behavior of the SUT during testing, BBC is more effective than MBT algorithms in finding deep bugs in implementations for which only a high-level specification is available. In the rest of this paper, we will use the term black-box checking, but one could argue that the phrase *learning-based testing* coined by Meinke [39, 40] is more appropriate.

Figure 1 schematically shows how BBC works. First of all, it assumes an implementation (of a black-box system), referred to as

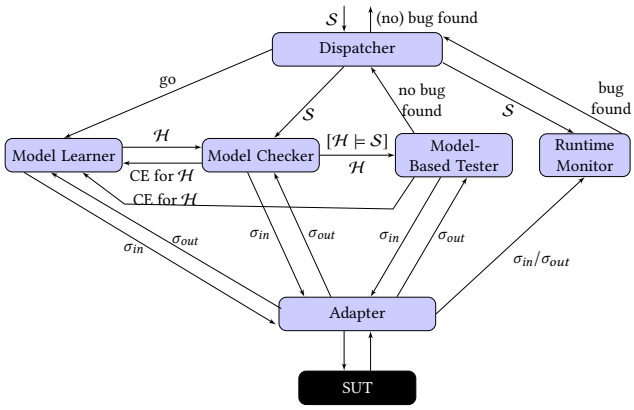

Figure 1: Black-box checking for specification $\mathcal{S}$ and system under test SUT.

the *System Under Test (SUT)*, that needs to be analyzed. It also assumes a (partial) specification $\mathcal{S}$, usually a conjunction of a number of requirements on the behavior of the SUT. An *Adapter* translates the abstract input and output symbols from the specification to concrete inputs and outputs that are accepted/produced by the SUT. The adapter makes BBC scalable to realistic applications, by using powerful abstractions [1, 56]. A *Dispatcher* orchestrates the activities of the different analysis tools used in the BBC approach. The dispatcher starts the analysis by activating the *Model Learner* [4, 31, 57], which builds a state-transition model $\mathcal{H}$ (an *hypothesis*) from results of *learning queries* (a.k.a. *output queries*): the learner sends sequences σ_{in} of inputs to the adapter/SUT, and in return receives sequences σ_{out} of outputs. The hypothesis $\mathcal{H}$ aims to describe how the SUT actually works. Whereas specification model $\mathcal{S}$ will often be small (e.g., a two state model describing that a *response* may only occur after a *request*), the hypothesis $\mathcal{H}$ will typically be much bigger, because the learning algorithm discovers — through systematic exploration — that states of the SUT reached via certain input sequences have different input-output behavior. Subsequently, BBC runs a *Model Checker* [6, 13] to check if hypothesis $\mathcal{H}$ *satisfies* specification $\mathcal{S}$, writen $\mathcal{H} \models \mathcal{S}$, meaning that all behaviors of $\mathcal{H}$ are allowed by $\mathcal{S}$. Two cases are considered:

(1) If $\mathcal{H} \models \mathcal{S}$ then a *Model-Based Tester* [9, 35, 53] will run *testing queries* to check whether $\mathcal{H}$ also is a correct model of the SUT: it sends sequences σ_{in} of inputs to the adapter/SUT, and checks if the resulting sequence σ_{out} of outputs agree with the prediction of $\mathcal{H}$. If a test fails, then the model learner uses this test to improve $\mathcal{H}$. If all tests pass then BBC terminates and reports that no evidence was found that the SUT violates specification $\mathcal{S}$.

(2) If $\mathcal{H} \not\models \mathcal{S}$ then there is a counterexample σ_{in} for which $\mathcal{H}$ produces an output not allowed by $\mathcal{S}$. BBC then performs an additional learning query σ_{in} on the adapter/SUT. If the resulting output σ_{out} agrees with the prediction by $\mathcal{H}$ then the SUT violates $\mathcal{S}$. Otherwise, $\mathcal{H}$ is not a correct model of the SUT, and the model learner uses the counterexample to further improve $\mathcal{H}$.

Finally, we included a *Runtime Monitor* [36] in Figure 1, as a minor extension of the BBC framework presented in [44, 45].[1] The runtime monitor checks, for all (learning and testing) queries performed on the SUT, whether the output σ_{out} produced in response to an input σ_{in} is allowed by specification $\mathcal{S}$. If an output violates $\mathcal{S}$ then a bug is reported. Use of a runtime monitor allows us to detect bugs faster: when a bug is encountered, it is reported immediately, rather than using it to further improve hypothesis model $\mathcal{H}$.

Surprisingly, in the large majority of applications of BBC, model checking is only used at the very end, when no counterexample can be found anymore for the latest hypothesis model (exceptions are, e.g., [38–40]). This contrasts with the original BBC proposal of [44, 45] which applies model checking for *all* hypotheses, also the intermediate ones. This is remarkable since, as observed by [44, 45], "intuitively it is clear that in many cases this method [only checking the final hypothesis] can be wasteful in that it does not take advantage of the property to avoid doing a complete identification". In a black-box setting we can never be sure about the correctness of learned models unless we are willing to make strong (typically unrealistic) assumptions, such as a bound on the number of states of the system. Therefore one might argue that, rather than learning models, the main goal of BBC is to find bugs in implementations as quickly as possible.

Some serious studies have been carried out to benchmark combinations of learning and testing algorithms for BBC [3, 27], but (surprisingly) there is no systematic evaluation of the ability of BBC to find bugs quickly. Only a few isolated results have been reported. Meinke & Sindhu [40], for instance, concluded for a single benchmark of an elevator system that the time required by BBC to discover a bug in the SUT is between 0.003% and 7% of the total time needed to completely learn the full model. The objective of this article is to provide a systematic evaluation of the effectiveness of BBC. In particular, we aim to answer the following questions:

RQ1 How effective is BBC (in terms of number of required queries) in detecting specification violations when compared with the effort required to learn the full model?
RQ2 How effective is BBC in detecting specification violations in situations where the full model cannot be learned?
RQ3 Does it help to add runtime monitoring to the BBC toolbox?
RQ4 How effective is BBC in detecting bugs when compared with traditional model-based testing algorithms.

In our experiments we simulate the SUT from benchmark models that were obtained using automata learning in previous works. From the Automata Wiki [42][2] repository, we selected 77 benchmark models with the following characteristics:

- All models were obtained via automata learning from real protocol implementations and real embedded controllers.
- All models are deterministic Mealy machines in which each input from a finite set I triggers a sequence of outputs taken from a finite set O.

[1] Monitors have also been added to BBC by Meijer & Van de Pol [38] but with a different purpose. They study BBC for general LTL formulas and propose to let the model checker consider the safety portion of an LTL property first and derive simpler counterexamples using monitors. So [38] uses monitors as part of the model checker, and not for runtime monitoring of learning and testing queries.
[2] https://automata.cs.ru.nl/

- For all models, a specification was available consisting of a number of safety properties, specified either as a DFA over alphabet $I \cup O$, or as an LTL formula that we converted to a DFA over alphabet $I \times O^*$ using Spot [17].[3] The specifications that we consider are DFAs, but multiple outputs may be enabled in a single state, meaning that they allow for *observable nondeterminism* in the sense of [7, 46].

The rest of this article is structured as follows. Section 2 recalls the basic notations and definitions related to DFAs and Mealy machines that we use. Section 3 briefly introduces the case studies from which we obtained our benchmark models and specifications. Section 4 discusses our experimental setup, both for BBC and model-based testing. The results from our experiments are presented in Section 5. Finally, Section 6 presents our conclusions, and Section 7 discusses limitations of our approach and directions for future research.

2 PRELIMINARIES

In this preliminary section, we first fix notation for partial maps and sequences, and then for DFAs and Mealy machines. Next, we describe how DFAs can be used to specify properties of Mealy machines, and a simple model checking approach for this setting. Finally, we briefly discuss our MBT approach.

2.1 Partial maps and sequences

We write $f\colon X \rightharpoonup Y$ to denote that f is a partial function from X to Y and write $f(x)\downarrow$ to mean that f is defined on x, that is, $\exists y \in Y\colon f(x) = y$, and conversely write $f(x)\uparrow$ if f is undefined for x. Function $f\colon X \rightharpoonup Y$ is *total* if $f(x)\downarrow$, for all $x \in X$. Often, we identify a partial function $f\colon X \rightharpoonup Y$ with the set $\{(x, y) \in X \times Y \mid f(x) = y\}$. The composition of partial maps $f\colon X \rightharpoonup Y$ and $g\colon Y \rightharpoonup Z$ is denoted by $g \circ f\colon X \rightharpoonup Z$, and we have $(g \circ f)(x)\downarrow$ iff $f(x)\downarrow$ and $g(f(x))\downarrow$. We use the Kleene equality on partial functions, which states that on a given argument either both functions are undefined, or both are defined with equal values for that argument.

An **alphabet** Σ is a finite set of symbols. A **word** over alphabet Σ is a finite sequence of symbols from Σ. We write ϵ to denote the empty word, and a to denote the word consisting of a symbol $a \in \Sigma$. If w and w' are words over Σ then we write $w\ w'$ to denote their **concatenation**. We write Σ^* to denote the set of all words over Σ, and Σ^+ to denote set $\Sigma^* \setminus \{\epsilon\}$. A word v is a **prefix** of a word w if there exists a word u such that $w = v\ u$. Similarly, v is a **suffix** of w if there exists a word u such that $w = u\ v$. A **language** over Σ is a subset of Σ^*. If L is a language over Σ, then we define $(L)^c$ as the language $\Sigma^* \setminus L$. A language L is **prefix closed** if $w \in L$ implies $v \in L$, for any prefix v of w.

2.2 DFAs and Mealy machines

Definition 2.1 (DFA). A (partial) **deterministic finite automaton (DFA)** is a five-tuple $\mathcal{A} = (Q, \Sigma, \delta, q^0, F)$, where Q is a finite set of **states**, Σ is a set of **input symbols**, $q^0 \in Q$ is the **initial** state, $F \subseteq Q$ is a set of **final** states, and $\delta : Q \times \Sigma \rightharpoonup Q$ is a (partial) **transition function**. The transition function is inductively extended to words over Σ in the standard way, (for $w \in \Sigma^*$ and $a \in \Sigma$):

$$\begin{aligned} \delta(q, \epsilon) &= q \\ \delta(q, w\ a) &= \delta(\delta(q, w), a) \end{aligned}$$

A word $w \in \Sigma^*$ is **accepted** by $\mathcal{A}$ if $\delta(q^0, w) \in F$, and **rejected** by $\mathcal{A}$ if $\delta(q^0, w) \notin F$. The **language** accepted by $\mathcal{A}$, denoted $\mathcal{L}(\mathcal{A})$, is the set of all words accepted by $\mathcal{A}$. Two DFAs $\mathcal{A}$ and $\mathcal{A}'$ are **equivalent**, denoted $\mathcal{A} \approx \mathcal{A}'$, if they accept the same language. A DFA $\mathcal{A}$ is **prefix closed** if, for each state q and input a, $\delta(q, a) \in F$ implies $q \in F$. DFA $\mathcal{A}$ is **complete** if transition function δ is total.

It is easy to see that if $\mathcal{A}$ is prefix closed then $\mathcal{L}(\mathcal{A})$ is prefix closed. Each DFA can be turned into a complete DFA by adding a fresh (nonfinal) **sink state** and adding, for each missing transition, a transition to that sink state.

Definition 2.2 (Completion of a DFA). Let $\mathcal{A} = (Q, \Sigma, \delta, q^0, F)$ be a DFA. Then Complete($\mathcal{A}$) is the complete DFA $(Q', \Sigma, \delta', q^0, F)$, where $Q' = Q \cup \{q_s\}$ and, for all $q \in Q'$ and $a \in \Sigma$,

$$\delta'(q, a) = \begin{cases} \delta(q, a) & \text{if } \delta(q, a)\downarrow \\ q_s & \text{otherwise} \end{cases}$$

It is straightforward to verify that $\mathcal{A} \approx$ Complete($\mathcal{A}$).

Whereas the output of a DFA is limited to a binary signal (accept/reject), a Mealy machine associates a more general output to each transition. The definition below, adapted from [23], associates a sequence of outputs from some alphabet to each transition. Such machines are quite useful to model the behavior of protocol entities.

Definition 2.3 (Mealy machine). A (partial) **Mealy machine** is a tuple $\mathcal{M} = (I, O, Q, q_0, \delta, \lambda)$, where I and O are alphabets of **input** and **output symbols**, respectively, Q is a set of **states** containing the **initial** state q_0, $\delta : Q \times I \rightharpoonup Q$ is a (partial) **transition function**, and $\lambda : Q \times I \rightharpoonup O^*$ is a (partial) **output function**. We require, for all $q \in Q$ and $i \in I$, that $\delta(q, i)\downarrow$ iff $\lambda(q, i)\downarrow$. A Mealy machine is **complete** if δ (and hence λ) is total.

2.3 Model checking

As part of BBC, the model checker verifies if a hypothesis $\mathcal{H}$ satisfies specification $\mathcal{S}$. Our hypotheses are Mealy machines, while we use Deterministic Finite Automata (DFA) for specifications. Inspired by [23], we translate our Mealy machines into DFAs, such that satisfaction of $\mathcal{H}$ to $\mathcal{S}$ can be checked using standards available for DFAs [30].

To translate from Mealy machines to DFAs, the idea is to insert auxiliary states (w, q) in between the source q' and target q of a transition, in which first the outputs from w are performed before jumping to state q.

Definition 2.4 (Mealy machine to DFA). Let $\mathcal{M} = (I, O, Q, q_0, \delta, \lambda)$ be a Mealy machine, with $I \cap O = \emptyset$. Then MealyToDFA($\mathcal{M}$) is the partial DFA $(Q', I \cup O, \delta', q^0, Q')$ where, for all $q \in Q$, $i \in I$,

[3] In the original paper of Peled, Vardi & Yannakakis [44, 45], Büchi automata are used as specifications and liveness properties are considered. However, in order to establish soundness of their BBC procedure, they need to assume a known bound on the number of states of the SUT. As pointed out by Meijer & Van de Pol [38] this can be either dangerous (if the guessed bound is too low) or inefficient (if the bound is too high). Meijer & Van de Pol [38] propose an alternative BBC procedure for liveness properties that requires the ability to test for equality of SUT states, but this is not truly black-box. We therefore decided to focus on safety properties in our study.

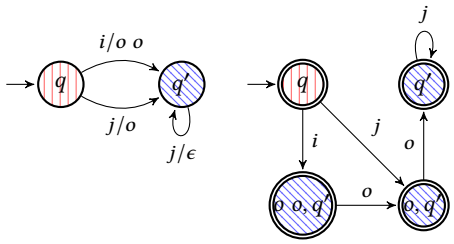

Figure 2: A Mealy machine (left) and the corresponding DFA (right).

$o, o' \in O$ and $w \in O^*$: $\delta'(q, o)\uparrow$, $\delta'((o\ w, q), i)\uparrow$ and

$$\begin{aligned} Q' &= Q \cup Q_{aux} \\ Q_{aux} &= \{(v, q) \in O^+ \times Q \mid \exists q' \in Q, j \in I : \\ &\qquad \delta(q', j) = q \text{ and} \\ &\qquad v \text{ suffix of } \lambda(q', j)\} \end{aligned}$$

$$\delta'(q, i) = \begin{cases} \delta(q, i) & \text{if } \lambda(q, i) = \epsilon \\ (\lambda(q, i), \delta(q, i)) & \lambda(q, i) \in O^+ \\ \text{undefined} & \lambda(q, i)\uparrow \end{cases}$$

$$\delta'((o\ w, q), o') = \begin{cases} \text{undefined} & \text{if } o \neq o' \\ q & \text{if } o = o' \text{ and } w = \epsilon \\ (w, q) & \text{otherwise} \end{cases}$$

Figure 2 gives an example of our translation from Mealy machines to DFAs. Note that MealyToDFA($\mathcal{M}$) is a labelled transition system with inputs and outputs in the sense of Tretmans [53], but of a restricted form. States of the DFA either have a *single* outgoing output transition, or zero or more outgoing input transitions. Moreover, at most finitely many consecutive output transitions are possible from any state. It is easy to see that, given I and O, we may fully retrieve $\mathcal{M}$ from MealyToDFA($\mathcal{M}$). Our translation improves on the one presented in [23], which turns the Mealy machine of Figure 2 into a DFA with 5 states.

Definition 2.5 (Specification). Consider a Mealy machine $\mathcal{M}$ with inputs I and outputs O. A **specification** for $\mathcal{M}$ is a prefix closed DFA $\mathcal{S}$ over alphabet $I \cup O$. We say that $\mathcal{M}$ *satisfies* specification $\mathcal{S}$ if $\mathcal{L}(\text{MealyToDFA}(\mathcal{M})) \subseteq \mathcal{L}(\mathcal{S})$.

The specifications for our benchmark models need to be massaged a bit to turn them into the above format. For the BLE and RERS case studies the specifications are LTL formulas. Using the Spot tool [17], we translate these to equivalent DFAs over alphabet $I \times O$. By splitting each (i, o)-transition into an i-transition followed by an o-transition, we can translate these DFA into (prefix-closed) DFAs over alphabet $I \cup O$. For the SSH and DTLS case studies, the specifications are given in terms of "'bug automata". These are DFAs over alphabet $I \cup O$, but with the roles of final/nonfinal states interchanged: accepting runs correspond to undesired behavior of the SUT. By complementing these bug automata, we obtain specifications in the sense of our Definition 2.5.

A model checker can efficiently check that $\mathcal{M}$ satisfies $\mathcal{S}$, using the fact that for all DFAs $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{L}(\mathcal{A}) \subseteq \mathcal{L}(\mathcal{B})$ iff $\mathcal{L}(\mathcal{A}) \cap (\mathcal{L}(\mathcal{B}))^c = \emptyset$. Both intersection and complement can be easily defined for DFAs:

Definition 2.6 (Complement). Let $\mathcal{A} = (Q, \Sigma, \delta, q^0, F)$ be a complete DFA. Then the **complement** of $\mathcal{A}$, denoted $(\mathcal{A})^c$, is the DFA $\mathcal{A} = (Q, \Sigma, \delta, q^0, Q \setminus F)$.

Definition 2.7 (Product). Let $\mathcal{A}_1 = (Q_1, \Sigma, \delta_1, q_1^0, F_1)$ and $\mathcal{A}_2 = (Q_2, \Sigma, \delta_2, q_2^0, F_2)$ be two DFAs. Then the **product** of $\mathcal{A}_1$ and $\mathcal{A}_2$, denoted $\mathcal{A}_1 \parallel \mathcal{A}_2$, is the DFA $(Q_1 \times Q_2, \Sigma, \delta, (q_1^0, q_2^0), F_1 \times F_2)$, where for all $q_1 \in Q_1$, $q_2 \in Q_2$ and $a \in \Sigma$,

$$\delta((q_1, q_2), a) = (\delta_1(q_1, a), \delta_2(q_2, a)).$$

It is well-known that for complete $\mathcal{A}$, $\mathcal{L}((\mathcal{A})^c) = (\mathcal{L}(\mathcal{A}))^c$ and, for DFAs $\mathcal{A}_1$ and $\mathcal{A}_2$ with the same alphabet, $\mathcal{L}(\mathcal{A}_1 \parallel \mathcal{A}_2) = \mathcal{L}(\mathcal{A}_1) \cap \mathcal{L}(\mathcal{A}_2)$, see, e.g., [30]. Thus $\mathcal{M}$ satisfies $\mathcal{S}$ iff

$$\mathcal{L}(\text{MealyToDFA}(\mathcal{M}) \parallel (\text{Complete}(\mathcal{S}))^c) = \emptyset.$$

Emptiness of the language accepted by a DFA can be decided in time linear in the size of the DFA [30].

2.4 Model-Based Testing

In this paper, Model-Based Testing (MBT) is used in two different ways. First, MBT is used as part of the BBC procedure, namely to check conformance of the current hypothesis to the SUT. Secondly, we use MBT as a standalone method to find bugs in the SUT. This way, we can compare the bug finding capabilities of BBC with standalone MBT. In MBT, tests are derived from a model. In MBT as part of BBC this model is the hypothesis. Standalone MBT uses the specification as its model: it uses the prefix-closed DFAs over the alphabet $I \cup O$, where a transition is either labeled with an input or an output label. Therefore these DFAs can be trivially translated to Labeled Transition Systems, for which standard test derivation is defined as usual [53], i.e. a test either: (i) supplies an input specified by the model, (ii) waits to observe an output – if the received output is not the output specified by the model, the test stops with verdict fail, or (iii) stops with verdict pass, when the stop condition, e.g., number of test steps, has been satisfied. While testing one keeps track of the current state of the model according to the supplied inputs and observed outputs. For selecting the next input of a test, existing MBT test generation strategies can be used, e.g., to select an input randomly, or to select a different input than already tried.

3 CASE STUDIES

We selected case studies from several published papers for which both models of SUTs were available, as well as properties that could be converted to DFAs. The SUTs were available on the automata wiki; corresponding specifications were obtained from the paper and corresponding artefacts. The case studies vary in size and application area, so that we have a good basis to draw conclusions.

SSH (DFA). In [23], Fiterau-Brostean et al. describe expected bugs and vulnerabilities as DFAs, and check these during automata learning. They apply their method to three more recent versions of the SSH server implementations (i.e., BitVise 8.49, DropBear v2020.81 and OpenSSH 8.8p1).

DTLS (Client & Server). Fiterau-Brostean et al., also apply their method from [23] to 16 DTLS client implementations and 17 DTLS server implementations (there were multiple versions of the same DTLS implementations).

BLE. In [48], Pferscher et al. use an active automata learning approach to detect anomalies and security vulnerabilities in five distinct System on Chip (SoC) Bluetooth Low Energy (BLE) devices, all of which implement the BLE 5 standard. Pferscher et al. found several crashes, anomalies, and a security vulnerability with their fuzzing-enhanced learning process.

SSH (LTL). In [24], Fiterau-Brostean et al. use active automata learning and model checking to verify several security and functional requirements against three SSH server implementations (Bitvise 7.23, Dropbear v2014.65 and OpenSSH 6.9p1-2). They use the NuSMV [12] model checker to automatically check their LTL formalizations of 12 requirements imposed by RFCs that describe the second version of the SSH protocol (i.e., SSHv2) against the final Mealy machine models they learned for their selected implementations. They found that each of the implementations under test violated at least one of their selected functional requirements.

RERS. The Rigorous Examination of Reactive Systems (RERS)-challenges are semi-annual open competitions in which participants attempt to solve a given set of reachability and verification problems. The goal behind these competitions is to encourage the use of new combinations of various tools and approaches and to provide a basis of comparison for the approaches used by the participants. The 2019-edition of the RERS challenge [32] contained an industrial track with benchmark programs that are based on Mealy machine models of controller software provided by the company ASML. We used all thirty SUTs of the RERS 2019 industrial LTL challenge, along with their accompanying LTL specifications.[4]

4 EXPERIMENTAL SETUP

We first explain our experimental setup to evaluate BBC, and its variant in which only the final hypothesis is model-checked. Then we explain our experimental setup for (traditional) MBT.

4.1 Black-box checking and model learning

In this work, we determine the efficacy of black-box checking (BBC) by comparing its efficiency to that of "basic" model learning. The main difference between basic variant and BBC is that the former does not involve the use of a model checker[4, 31, 44, 45, 57], except possibly at the very end when learning has finished. In our experiments, basic model learning finishes once the Model-Based Tester concludes that the current hypothesis $\mathcal{H}$ is equivalent to the SUT. The basic model learning setup also does not involve the use of a Runtime monitor.

Our BBC and model learning approaches both offer support for the use of multiple properties at once. When BBC or model learning uses the Model Checker, it checks every hypothesis that it receives against each property. For each counterexample that it finds, the Model Checker checks the output against that of the SUT. It passes any counterexamples for which these outputs are equal to the Dispatcher, along with their associated properties, and it will stop checking hypotheses against the associated properties.

Our approach to using multiple properties has the consequence that all properties may affect one another: any spurious Model Checker counterexample is used to refine the current hypothesis, the hypothesis then changes, which affects the way in which BBC will look for counterexamples for all remaining properties.

We implemented the BBC and basic model learning approaches in a single codebase written in Python 3.11.14 on top of version 1.5.1 of the AALpy [41] library for active automata learning. We used version 2.13 of the LTL and ω-automata manipulation and model checking library Spot [17] to convert the LTL properties into monitors. For performing experiments, we used Docker[5].

We used the active automata learning algorithm $L^{\#}$ because its performance is competitive with that of several alternatives [54]. We use separating sequences for $L^{\#}$'s extension rule because this is the default in AALpy. For most experiments, we use Adaptive Distinguishing Sequences (ADS) for the separation rule as this is AALpy's default setting. The exception is RERS, where we used separating sequences because we found that using ADS with experiments that were otherwise unchanged made the hypothesis refinements and thereby the BBC process in the RERS models considerably slower.

We fixed the model-based tester for BBC to Hybrid-ADS [52] because it is a recent, state-of-the-art approach that is available as a tool. We based our configuration for Hybrid-ADS on [54]. As such, we set the operation mode to "random", the prefix mode to "buggy", the number of extra states to check for (minus 1) to 10, and the expected number of random infix symbols to 10. We repeat each of our experiments for 50 random seeds to account for the randomness introduced by our use of Hybrid-ADS. We measure the time that each seed takes with the `perf_counter_ns`-function from the `time`-module of Python's standard library.

Hybrid-ADS produces an unbounded number of testing queries when used in random mode. Performing another correctness check when the current hypothesis is equivalent to the SUT would take an infinite amount of time. In [54], the authors skip this final correctness check if they determine that the current hypothesis is already equivalent to the SUT. We follow this approach because the number of testing queries that one would perform to look for a difference between a correct hypothesis and the SUT is always both finite, and ultimately arbitrary.

We followed the approach of Kruger et al. [33], by using a step budget to further reduce the time taken by certain BBC experiments. If an experiment is about to exceed this budget, BBC first terminates its current operation, and then uses the model checker to look for counterexamples for the properties for which it hadn't found one. This use of the model checker does imply that the step budget can be exceeded, since the model checker will still verify any counterexamples that it finds against the SUT. Whenever we used step budgets for an SUT, we did so because we found that it should take us considerably more than a week to obtain all of our results for it. We then used a range of step budgets that spanned from 10^3 to 10^8 and selected the results for the largest budget for which we obtained results in a timely fashion.

[4] https://rers-challenge.org/2019/index.php?page=industrialProblemsLTL#

[5] https://www.docker.com/

To evaluate our approach, we run our experiments on SUTs simulated from Mealy machines (stored as DOT files) obtained with automata learning in previous work (see Section 3). This way our experiments are not hindered by, e.g., response times of a real SUT. For each case study, multiple SUT models were available, as they were obtained from different real implementations. All properties available for the case studies were converted from their format, which ranged from natural language to LTL formulae, to DFAs. We performed experiments for all combinations of SUTs and their associated properties.

We took the following decisions in obtaining the case studies:

SSH (DFA) & DTLS (Client & Server). We obtained the SUTs and associated properties used in Fiterau et al.'s [23] from their archived software artifact on [19]. We only used the SUTs for which the artifact specified which of their properties hold and which do not. These were precisely the SUTs covered in [23].

BLE. We used all available SUTs and properties which we could convert into LTL properties.

SSH (LTL). We use nine of the eleven LTL properties for our evaluation. Property 3, property 4 and property 9 are excluded, because they each use either NuSMV's *O* (once), or NuSMV's *S* (since) LTL operator, neither of which are supported by Spot[6]. Property 8 relies on the ability to check whether the SUT is in its initial state, a requirement that cannot be satisfied in our black-box setting. We therefore replaced this property with an LTL property that we based on a DFA from the SSH (DFA) case study. Fiterau et al. designed that DFA to capture the same requirement of the SSHv2 specification as Property 8. It does so in a way that doesn't rely on explicit knowledge of the SUT's current state, which allows it to work in our black-box setting.

RERS. We used the thirty provided models, and the safety properties among their accompanying sets of twenty LTL properties. The number of safety properties, and thereby of the properties that we used per model ranges from 9 to 18, with a mean of 13.7 and a standard deviation of 1.9.

Table 1 shows some general information about the case studies. For each case study, it shows the number of SUTs, the means and standard deviations of the number of SUT states and the total number of properties that are violated by the SUTs, followed by the means and standard deviations of the number of violated properties for each individual SUT. BBC is used to determine whether a black-box SUT satisfies a given set of properties. The user wouldn't know beforehand which of the properties are satisfied by the SUT, and which are not. We therefore included the properties that are satisfied by the SUT in our experimentation. The inclusion of these properties can affect the results, since any model checking counterexamples that are found for these properties will trigger a hypothesis refinement.

To make for a fair comparison, we use the same settings for the model learning experiments as for their corresponding BBC experiments. The only difference between them is in the way that the Model Learner, Model-Based Tester, Model Checker and Runtime Monitor are (or are not) used to follow either the BBC schematic of Figure 1 or the model learning variant.

[6]https://spot.lre.epita.fr/concepts.html#ltl

Case study	# Models	# SUT states		# SUT Violations		
		Mean	Stdev	Total	Mean	Stdev
BLE	8	7.6	4.5	1	0.1	0.4
DTLS Client	16	84.6	122.8	66	4.1	3.4
DTLS Server	17	90.7	238.0	44	2.6	3.1
RERS	30	432.7	835.6	182	6.1	2.4
SSH (DFA)	3	33.7	11.4	11	3.7	1.5
SSH (LTL)	3	42	20.8	6	2	1.0

Table 1: The total number of SUT models, mean and standard deviation of the number of states of the SUT models, and the total number of properties that violate the SUTs, followed by the mean and standard deviation of the number of properties that are violated by each individual SUT, per case study

We performed the black-box checking and model learning experiments on a computer with an AMD Epyc 7642 processor. Every experiment was constrained to 2 processing cores and 4Gb of RAM.

4.2 Model-based Testing

In order to evaluate BBC's performance against a baseline, we applied (standalone) Model-Based Testing on a subset of the evaluated benchmarks. We selected a subset of experiments focusing on bugs that require only low to moderate effort for BBC to produce counterexamples. This restriction is motivated by preliminary results indicating significantly worse performance of MBT.

For the MBT implementation, we use Lattest, a Haskell library for MBT. As a test selection strategy, we implemented a tester that combines a random walk with the aim of increasing transition coverage. Specifically, we add memory to store the input sequences that were tested already by some test of the test suite. At each step of a test, a new input is chosen randomly from the set of inputs that has not been tried yet after this sequence. The test stops either when it fails or when it reaches the configured maximum number of steps, which we defined as twice the number of states of the specifications.

We considered an alternative bound for the stopping criteria of a test, by looking at the length of traces for BBC to find bugs, but since we then would use white-box knowledge, we decided that a generic bound based on the specification would be fairer since that is truly black-box. On hind-sight we checked that this practical bound exceeded the number of steps that BBC needed to find bugs.

We chose our random walk strategy with memory for transition coverage as an optimum in between a simple random walk and more sophisticated techniques, such as n-complete test suites [8, 16, 35]. Given that a test from an n-complete test suite always starts with an access sequences to a specification state, we noted, also from initial experiments where we tried tests of the n-complete test suite, that the access sequences of the specification where not matching with access sequences to a respective state in an implementation. The reason for this is that the specification consists of properties that typically do not consider a trace from the initial state, but only the few relevant inputs and outputs considered by

that property. For example, it was expecting to see a response after a certain input, but that response was only observed if before the input some other actions were done. Consequently, a direct access sequence which just provides the respective input would not work to reach the state where the response transition is enabled. Summarizing, the high-level and partial nature of the specification omits details that sophisticated algorithms need to efficiently explore the implementation and thus efficiently find bugs.

Initial experiments also confirmed that random walk could find few bugs compared to BBC. With our memory test strategy, by remembering sequences, we have a somewhat course representation of traces of the implementation (still much courser than hypothesis models), so that MBT is able to somewhat compete with BBC, while still representing a straightforward MBT test strategy.

As done in BBC, multiple properties can be tested on the same SUT simultaneously. Given a set of properties, the tester constructs a conjunction of the corresponding specifications. When a bug in the SUT is detected, the violated property is identified, and the specification is regenerated excluding that property, to continue testing. Shadowing of deeper bugs by shallower ones is thus avoided.

The subset of experiments executed for MBT consists of 48 combinations of SUTs and properties, therefore, 48 potential bugs to be found. For each experiment, the number of test cases for the test suites is ten times the number of queries it takes for the BBC implementation to find the corresponding bug. The number of test steps, within those tests, is defined as twice the number of states of the specification, to aid the identification of deeper bugs. Each test suite is repeated fifty times, with different seeds.

The specifications for these experiments were constructed from LTL formulae as described in Theorem 2.5, or from the DFAs, depending on the benchmark. As in BBC, the implementations described in the DOT files are simulated by a Python script. Both the tester and the Python-based SUT are executed inside Docker containers. MBT experiments were executed on a computer with an Intel i7 processor. Each instance of the tester was constrained to 1 processing core and 2Gb of RAM.

5 RESULTS

In this section, we present the results of our experiments, addressing each of the questions posed in Section 1.

5.1 BBC effectiveness evaluation

We consider 77 SUT models, for a total of 310 pairs of models with accompanying violated properties. Our use of step budgets kept us from finding all 310 bugs. In total, our BBC approach found 304 (98%) of the bugs across all 50 random seeds. The RERS case study was the only case study for which not all bugs were found in all seeds: BBC found 96% of the bugs on average across all seeds, with a standard deviation of 10%. On average, our BBC found 278.2 (90%) of the 310 bugs per seed, with a standard deviation of 5.28 bugs.

Of the 65 models across which BBC found 304 bugs, 32 (49%) were fully learned by model learning. This accounts for 118/310 (38%) of the bugs. Figure 3 shows for each of these 118 properties the mean number of queries that BBC and model learning required to find their bugs across all 50 seeds. The figure shows that for all bugs, the mean number of queries required by BBC was lower than that

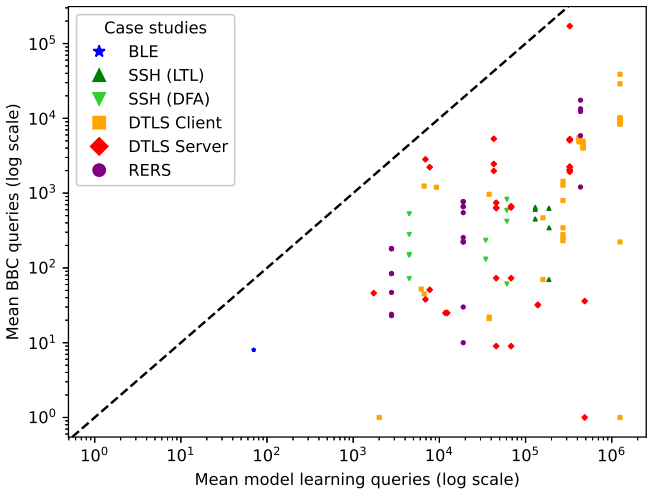

Figure 3: Scatter plot showing the number of queries needed for BBC to find a bug, relative to the number of queries needed to find the bug with model learning, in log scale.

Case study	Mean	Stdev
BLE	12%	—
DTLS Client	2%	3%
DTLS Server	5%	12%
RERS	3%	2%
SSH (DFA)	3%	4%
SSH (LTL)	0%	0%

Table 2: The mean and standard deviation of the percentages of the number of queries BBC took when compared to model learning for the bugs from Figure 3, per case study

required by model learning. Model learning required an average of 251732.67 queries for these 118 properties, compared to an average of 4047.32 (2%) queries for BBC. For the 118 properties, BBC took an average of 3% of the queries required by model learning, with a standard deviation of 7%. Table 2 shows the means and standard deviations of the percentages for the properties per case study.

Note how in Figure 3, some data points for bugs from the same case studies appear in vertical lines. These lines are formed by properties that belong to the same model; with model learning, all bugs for the same model take an almost equal number of queries to find, since they are only found once the models are fully learned[7].

We performed an ablation study to determine the effect of our runtime monitoring on BBC's performance. To this end, we repeated each BBC experiment with runtime monitoring disabled. Unmonitored BBC found the same 304 bugs found by (monitored) BBC. Figure 4 shows the mean number of queries that BBC and unmonitored BBC needed to find these bugs across the 50 seeds. This took unmonitored BBC 279199.84 queries on average, while BBC required an average of 274859.52 (98%) queries. For the 304

[7]We say that they take an *almost* equal number of queries to find, because the Model Checker checks every counterexample found after model learning against the SUT with an additional learning query, as we explain in Section 1.

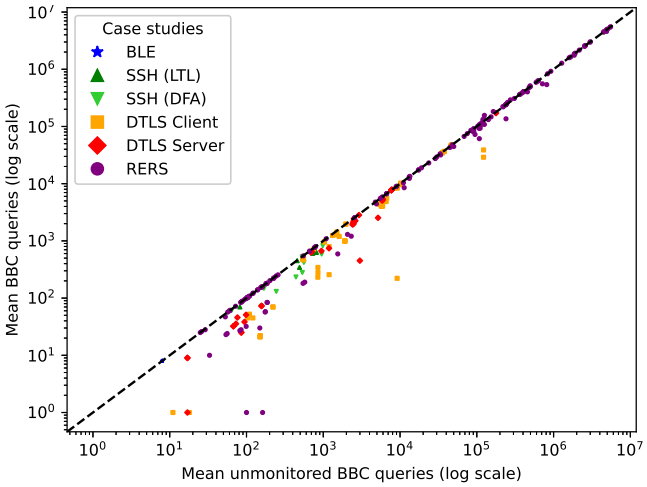

Figure 4: Scatter plot showing the number of queries needed for BBC to find a bug with runtime monitoring enabled, relative to the number of queries BBC required with runtime monitoring disabled, in log scale.

Case study	Mean	Stdev
BLE	100%	—
DTLS Client	69%	31%
DTLS Server	71%	28%
RERS	85%	31%
SSH (DFA)	73%	20%
SSH (LTL)	82%	10%

Table 3: The mean and standard deviation of the percentages of the number of queries BBC took when compared to unmonitored BBC for the properties of Figure 4, per case study.

properties, BBC took an average of 79% of the queries required by unmonitored BBC, with a standard deviation of 30%. Table 3 shows the means and the standard deviations of the percentages of the properties per case study.

Figure 5 shows for each case study the number of states the hypotheses had when BBC found the bugs. The colored markers indicate the mean number of states across all 50 seeds; the colored lines that cross the markers show the standard deviations. The gray bars indicate the mean number of states of the SUT models. On average, the SUT models have 209.94 states. BBC needed to learn an average of 16.62 (8%) states to find the bugs, with a standard deviation of 19.41 states.

5.2 BBC vs. MBT

Finally, we compare BBC's performance for bug finding with standalone MBT, for a subset of the benchmarks. Across all 50 seeds, MBT was able to find 32 (66.6%) of the 48 bugs, while BBC found all of them.

Figure 6 shows the mean numbers of tests that MBT required, and the mean numbers of queries that BBC required to find these 32 bugs. MBT used an average of 1357.71 tests to find them, compared

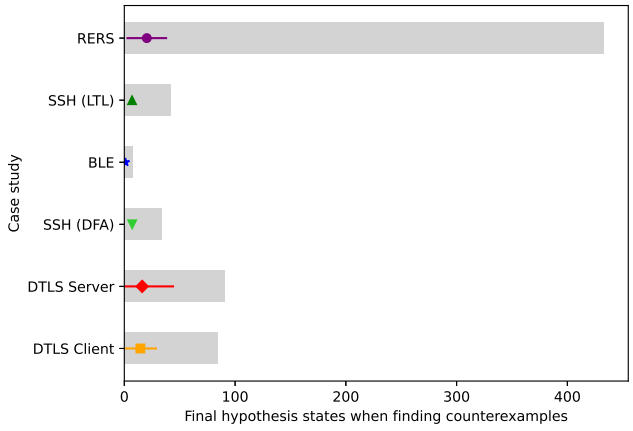

Figure 5: Bar plot comparing the mean and standard deviation of the number of states BBC needed to learn to find the bugs for the six case studies with the mean number of states in the SUTs.

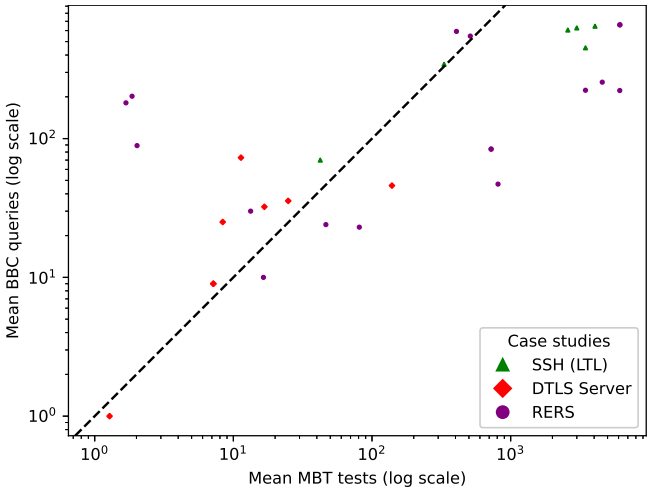

Figure 6: Scatter plot showing the mean number of queries needed for BBC to find a bug, compared to the mean number of tests MBT needs to find the same bug.

to an average of 215.87 (16%) queries for BBC. For the 32 bugs, BBC took an average of 897.08% of the queries, compared to the number of tests required by MBT, with a standard deviation of 2714.41%.

We also compared the efficiency of two testing methods on the number of steps (i.e., input events) used to find the bugs. Figure 7 shows the mean number of steps that MBT and BBC required to find the 32 bugs found by MBT, across all 50 seeds. The figure shows that for all of the 32 bugs, the mean number of steps required by BBC was lower than that required by MBT. MBT required an average of 127136.48 steps to find these bugs, compared to an average of 1211.75 (1%) for BBC. For the 32 bugs, BBC took an average of 6.08% of the steps required by MBT, with a standard deviation of 11.70%.

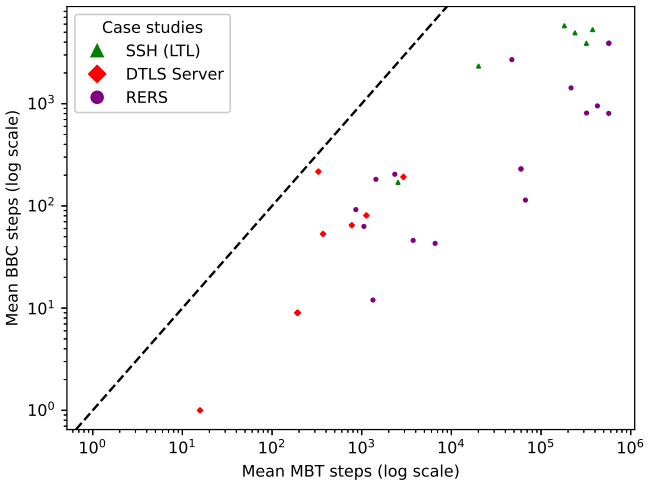

Figure 7: Scatter plot showing the mean number of steps needed for BBC to find a bug, compared to the mean number of steps MBT needs to find the same bug.

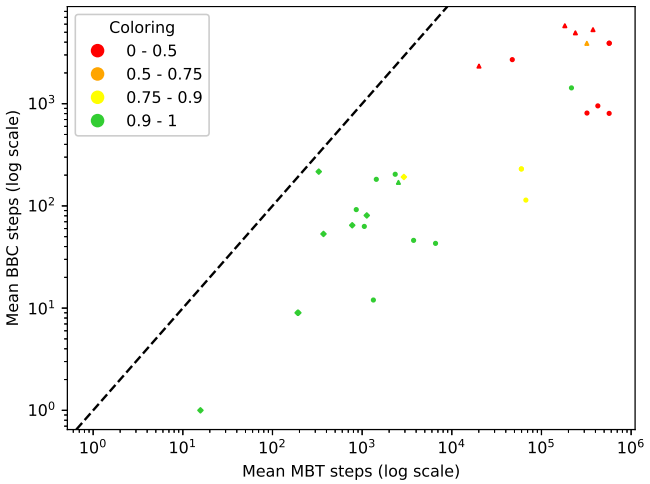

Figure 8: Scatter plot showing the mean number of steps needed for BBC to find a bug, compared to the mean number of steps MBT needs to find the same bug. Colors of each dot denote the fraction of experiments where MBT found the bug, ranging from green (all experiments) to red (less than half of the experiments). BBC reported these bugs in all experiments.

BBC not only found more bugs, quicker, but also did so consistently. MBT, only reported 17 (35.4%) of 48 bugs consistently across all 50 repetitions of the experiment, whereas BBC managed to find the bugs on every iteration. Figure 8 presents again the comparison of number of steps required for both BBC and MBT to find each bug, but this time, the coloring of each point represents how consistent was MBT in finding such bug; colors range from green, meaning the bug was found in all experiments, to red, meaning that the bug was found in less than half of the experiments. The results shows clearly that MBT consistently finds shallow bugs (i.e. those that take few test steps), while deep bugs are only occasionally reported.

These results show that BBC consistently outperforms MBT in both speed and reliability. Even when MBT approaches BBC's performance in certain benchmarks, it lacks consistency. This variability is originated in our implementation's strong dependence on randomness, i.e. the initial seed, even when augmented with memory. As a result, MBT may occasionally discover faults quickly, but such outcomes are not consistent or reproducible.

6 CONCLUSIONS

BBC is quicker than learning the full model before model checking (RQ1). Our systematic evaluation confirmed the intuition of Peled, Vardi & Yannakakis [44, 45] that bugs are detected (much) quicker when, as part of the BBC loop, model checking is applied for all intermediate hypotheses, rather than just for the full model. If a hypothesis model does not satisfy specification $\mathcal{S}$, then a *single* additional learning query suffices to either reveal a bug in the SUT or a bug in the hypothesis model. Hence the information gain of this query is significant and one should always do it.

BBC finds bugs when the full model cannot be learned (RQ2). Even when the full model is too large and cannot be learned, BBC is still able to detect many violations of the specification. In particular, using BBC, we managed to detect 96% of the safety properties violations in the challenging RERS 2019 industrial LTL benchmarks. This improves (what we believe to be) the best previous result for this benchmark collection by Kruger, Junges & Rot [33], who succeeded to learn partial models for 23 benchmarks (they did not consider the 7 largest benchmarks), and full models for 15 benchmarks. Here we should emphasize that the goal of [33] was to learn the full models rather than finding specification violations.

Adding runtime monitoring helps (RQ3). We also found that adding runtime monitoring to the BBC toolbox leads to a decrease (21%) of the number of queries required to detect bugs.

BBC is more effective than MBT for high-level specifications (RQ4). Our results show that BBC is more effective than existing MBT algorithms in finding (deep) bugs in implementations when only a high-level model is available. MBT [9, 16, 35, 53] has been extensively researched and a reasonable number of commercial and open-source tools exist. MBT has proven to be quite effective when test are generated from "low level" formal models whose behavior is close to the behavior of the SUT, in the sense that all states of the SUT can be reached by first doing an access sequence for a state in the formal model, followed by a few randomly selected inputs. In fact, Vaandrager & Melse [55] observe that prominent MBT approaches, such as the W-method of Vasilevskii [58] and Chow [11], the Wp-method of Fujiwara et al. [26], and the HSI-method of Luo et al. [37] and Petrenko et al. [47, 62], are k-A-complete for fault domains that contain all FSMs in which any state can be reached by first performing a sequence from a state cover A for the specification model, followed by up to k arbitrary inputs, for some small k. This means that when all states of the SUT can be reached by first doing an access sequence for some state of the specification, followed by a few arbitrary inputs, then these MBT approaches are guaranteed to find bugs in this SUT (when present). Outside

these fault domains they are much less effective. Indeed, for the benchmarks considered in this paper, k-A-complete approaches are ineffective, and we did not use them in our final experiments. Our benchmarks have up to 200 inputs, which means that we can only run the MBT algorithms for values of k up to 2. We conjecture that for many of our benchmarks it is not possible to reach the faulty SUT states by an access sequence of a specification DFA state followed by at most 2 arbitrary inputs. Exploring this conjecture (which might explain the limitations of existing MBT algorithms) is a question for future research.

Specifications help to find bugs faster. Our results show that, when BBC is used to detect bugs in implementations, it really helps to have a high-level specification available. Even in cases where model learning is unable to learn the full SUT model, BBC is able to learn a small hypothesis model that enables the detection of specification violations. In one of our benchmarks (RERS M65), for example, the full SUT model has 3966 states, but hypotheses models with (on average) 14 states are enough to detect that the SUT violates 5 out of the 17 properties in the specification. This provides a nice illustration of the aphorism attributed to George Box that "all models are wrong but some are useful."

7 LIMITATIONS AND FUTURE WORK

All benchmark models in our study have been obtained using model learning from real world protocol implementations and real embedded controllers. All specifications for the protocol benchmarks have been derived from the corresponding protocol standards. The LTL specifications for the RERS benchmarks, however, have been artificially generated using a property mining approach [32]. It would be interesting to further explore the effectiveness of BBC to find specification violations in embedded control software.

In our experiments, we restricted ourselves to a single model learning method, a single equivalence oracle, and a single model-based testing approach. We are confident that the main conclusions of Section 6 will also apply for different algorithms for model learning and model-based testing: model checking hypotheses against the specification provides guidance and helps BBC to find bugs faster. Nevertheless, of course, this needs to be confirmed by further studies.

In our study, we used the total number of learning and testing queries, as well as the total number of inputs and resets, to measure the efficiency of bug finding approaches. In applications there is often a strong correlation between the total number of queries and the time spent on testing/learning. However, performing a query with many input symbols will take longer than a query with just a few symbols. Thus, sometimes the total number of input symbols and resets provides a better measure of the efficiency. For certain applications, resets are impossible or extremely time consuming. Groz et al [28] designed the hW-inference algorithm for learning models using only a *single* query. We expect that it will be rather easy to combine hW-inference and BBC.

Learning-based testing (and thus BBC) belongs to the general area of fuzzing of stateful systems. As observed by [14], "many fuzzers use some form of learning to infer information about the message format, the protocol state machine, or both. Evolution can be regarded as a form of learning because it produces and uses new knowledge about the input format, even though this knowledge is (usually) not expressed in the form of a regular expression, state machine, or context-fee grammar." BBC is a black-box technique but it has been frequently and successfully applied to find bugs in software for which the source code was available. For those applications it would be interesting to compare the effectiveness of BBC with other (grey-box or white-box) fuzzing approaches. Fuzzing tools typically target memory corruption bugs that crash the SUT, which is more restrictive than the main objective of BBC: finding specification violations. However, a recent white-box fuzzing approach that is closely related to our work has been proposed by Asadian *et.al.* [5]. They encode protocol requirements by monitors, and then employ symbolic execution to detect violations of these requirements in protocol implementations.

Learning-based testing, a.k.a. active automata learning, is a vibrant research area that is progressing rapidly, see e.g. [15, 25, 34]. Primary challenges are to extend the learning algorithms to richer model classes involving data, timing information and nondeterminism, and scaling the model-based testing algorithms to larger models, in particular in the presence of many possible inputs to the SUT. Two main tools that support active automata learning and provide implementations of the most efficient algorithms are LearnLib [34] and AALpy [41]. LearnLib already supports BBC, based on the work of [38], using LTL formulas as specifications. The BBC implementation described in this paper has been developed on top of AALpy, and accepts both LTL formulas and DFAs as specifications. We think both BBC implementations can be improved by supporting richer languages for describing specifications that are closer to standard engineering practices, e.g. inspired by the bug automata syntax of [23] or automatically generated from Behavior-Driven Development (BDD) specifications [63].

The design of active automata learning algorithms is a subfield of the general area of machine learning. Thus far, however, we have not used LLMs in our study of BBC. We see three promissing ways in which use of LLMs may potentially leverage the effective application of BBC: (1) Given the ability of LLMs to analyze RFCs (see e.g. [43]) it would be interesting to use LLMs to automatically extract formal specifications (e.g., LTL formulas or DFAs) from RFCs. (2) LLMs may help with the construction of adaptors, which translate the abstract inputs and outputs from the specification to concrete messages that are accepted by the SUT (and vice versa). Building adaptors requires both an understanding of the software that will be analyzed as well as an understanding of the intended level of abstraction for the models that one would like to learn. (3) LLMs may support the conformance testing of the SUT with respect to hypothesis models [59].

8 DATA AVAILABILITY STATEMENT

The code and data used to obtain the data covered in this paper is publicly and permanently available on Zenodo and accessible via the DOI 10.5281/zenodo.19235642 on http://doi.org/10.5281/zenodo.19235642. The artifact is currently published anonymously to comply with ASE's double-anonymous review policy. This paper's author(s) will be added to the artifact if the paper is accepted.

REFERENCES

[1] Aarts, F., Jonsson, B., Uijen, J., Vaandrager, F.: Generating models of infinite-state communication protocols using regular inference with abstraction. Formal Methods in System Design **46**(1), 1–41 (2015). doi:10.1007/s10703-014-0216-x

[2] Aichernig, B.K., Mostowski, W., Mousavi, M.R., Tappler, M., Taromirad, M.: Model learning and model-based testing. In: Bennaceur, A., Hähnle, R., Meinke, K. (eds.) Machine Learning for Dynamic Software Analysis: Potentials and Limits - International Dagstuhl Seminar 16172, Dagstuhl Castle, Germany, April 24-27, 2016, Revised Papers. Lecture Notes in Computer Science, vol. 11026, pp. 74–100. Springer (2018). doi:10.1007/978-3-319-96562-8_3, https://doi.org/10.1007/978-3-319-96562-8_3

[3] Aichernig, B.K., Tappler, M., Wallner, F.: Benchmarking combinations of learning and testing algorithms for automata learning. Formal Aspects Comput. **36**(1), 3:1–3:37 (2024). doi:10.1145/3605360, https://doi.org/10.1145/3605360

[4] Angluin, D.: Learning Regular Sets from Queries and Counterexamples. Information and Computation **75**(2), 87–106 (1987). doi:10.1016/0890-5401(87)90052-6

[5] Asadian, H., Fiterau-Brostean, P., Jonsson, B., Sagonas, K.: Monitor-based testing of network protocol implementations using symbolic execution. In: Proceedings of the 19th International Conference on Availability, Reliability and Security, ARES 2024, Vienna, Austria, 30 July 2024 - 2 August 2024. pp. 17:1–17:12. ACM (2024), https://doi.org/10.1145/3664476.3664521

[6] Baier, C., Katoen, J.P.: Principles of Model Checking. MIT Press, Cambridge, Massachusetts (2008)

[7] van den Bos, P., Janssen, R., Moerman, J.: n-complete test suites for IOCO. Softw. Qual. J. **27**(2), 563–588 (2019). doi:10.1007/S11219-018-9422-X, https://doi.org/10.1007/s11219-018-9422-x

[8] van den Bos, P., Vaandrager, F.W.: State identification for labeled transition systems with inputs and outputs. Sci. Comput. Program. **209**, 102678 (2021). doi:10.1016/J.SCICO.2021.102678, https://doi.org/10.1016/j.scico.2021.102678

[9] Broy, M., Jonsson, B., Katoen, J., Leucker, M., Pretschner, A. (eds.): Model-Based Testing of Reactive Systems, Advanced Lectures [The volume is the outcome of a research seminar that was held in Schloss Dagstuhl in January 2004], Lecture Notes in Computer Science, vol. 3472. Springer (2005), https://doi.org/10.1007/b137241

[10] Budd, T.A., Angluin, D.: Two notions of correctness and their relation to testing. Acta Informatica **18**, 31–45 (1982), https://doi.org/10.1007/BF00625279

[11] Chow, T.: Testing software design modeled by finite-state machines. IEEE Trans. Software Eng. **4**(3), 178–187 (1978)

[12] Cimatti, A., Clarke, E.M., Giunchiglia, E., Giunchiglia, F., Pistore, M., Roveri, M., Sebastiani, R., Tacchella, A.: Nusmv 2: An opensource tool for symbolic model checking. In: Brinksma, E., Larsen, K.G. (eds.) Computer Aided Verification, 14th International Conference, CAV 2002,Copenhagen, Denmark, July 27-31, 2002, Proceedings. Lecture Notes in Computer Science, vol. 2404, pp. 359–364. Springer (2002). doi:10.1007/3-540-45657-0_29

[13] Clarke, E.M., Henzinger, T.A., Veith, H., Bloem, R. (eds.): Handbook of Model Checking. Springer (2018). doi:10.1007/978-3-319-10575-8

[14] Daniele, C., Andarzian, S.B., Poll, E.: Fuzzers for stateful systems: Survey and research directions. ACM Comput. Surv. **56**(9), 222:1–222:23 (2024), https://doi.org/10.1145/3648468

[15] Dierl, S., Fiterau-Brostean, P., Howar, F., Jonsson, B., Sagonas, K., Tåquist, F.: Scalable tree-based register automata learning. In: Finkbeiner, B., Kovács, L. (eds.) Tools and Algorithms for the Construction and Analysis of Systems. pp. 87–108. Springer Nature Switzerland, Cham (2024)

[16] Dorofeeva, R., El-Fakih, K., Maag, S., Cavalli, A.R., Yevtushenko, N.: FSM-based conformance testing methods: A survey annotated with experimental evaluation. Information & Software Technology **52**(12), 1286–1297 (2010). doi:10.1016/j.infsof.2010.07.001

[17] Duret-Lutz, A., Renault, E., Colange, M., Renkin, F., Aisse, A.G., Schlehuber-Caissier, P., Medioni, T., Martin, A., Dubois, J., Gillard, C., Lauko, H.: From Spot 2.0 to Spot 2.10: What's new? In: Proceedings of the 34th International Conference on Computer Aided Verification (CAV'22). Lecture Notes in Computer Science, vol. 13372, pp. 174–187. Springer (Aug 2022). doi:10.1007/978-3-031-13188-2_9

[18] Ferreira, T., Brewton, H., D'Antoni, L., Silva, A.: Prognosis: closed-box analysis of network protocol implementations. In: Kuipers, F.A., Caesar, M.C. (eds.) Proceedings of the ACM SIGCOMM 2021 Conference. pp. 762–774. ACM (2021). doi:10.1145/3452296.3472938

[19] Fiterau-Brostean, P., Poll, E., Vaandrager, F., Lenaerts, T., de Ruiter, J., Verleg, P.: Source code and data relevant for the paper 'Model Learning and Model Checking of SSH Implementations' (2018). doi:10.17026/DANS-Z6N-DXQ6

[20] Fiterau-Brostean, P., Howar, F.: Learning-Based Testing the Sliding Window Behavior of TCP Implementations. In: Petrucci, L., Seceleanu, C., Cavalcanti, A. (eds.) Proceedings of the Critical Systems: Formal Methods and Automated Verification - Joint 22nd International Workshop on Formal Methods for Industrial Critical Systems FMICS and 17th International Workshop on Automated Verification of Critical Systems AVoCS 2017. Lecture Notes in Computer Science, vol. 10471, pp. 185–200. Springer (2017). doi:10.1007/978-3-319-67113-0_12

[21] Fiterau-Brostean, P., Janssen, R., Vaandrager, F.W.: Combining Model Learning and Model Checking to Analyze TCP Implementations. In: Chaudhuri, S., Farzan, A. (eds.) Proceedings of the 28th International Conference Computer Aided Verification, CAV 2016. Lecture Notes in Computer Science, vol. 9780, pp. 454–471. Springer (2016). doi:10.1007/978-3-319-41540-6_25

[22] Fiterau-Brostean, P., Jonsson, B., Merget, R., de Ruiter, J., Sagonas, K., Somorovsky, J.: Analysis of DTLS Implementations Using Protocol State Fuzzing. In: Capkun, S., Roesner, F. (eds.) Proceedings of the 29th USENIX Security Symposium, USENIX Security 2020. pp. 2523–2540. USENIX Association (2020), https://www.usenix.org/conference/usenixsecurity20/presentation/fiterau-brostean

[23] Fiterau-Brostean, P., Jonsson, B., Sagonas, K., Tåquist, F.: Automata-based automated detection of state machine bugs in protocol implementations. In: 30th Annual Network and Distributed System Security Symposium, NDSS 2023, San Diego, California, USA, February 27 - March 3, 2023. The Internet Society (2023), https://www.ndss-symposium.org/wp-content/uploads/2023/02/ndss2023_s68_paper.pdf

[24] Fiterau-Brostean, P., Lenaerts, T., Poll, E., de Ruiter, J., Vaandrager, F.W., Verleg, P.: Model learning and model checking of SSH implementations. In: Erdogmus, H., Havelund, K. (eds.) Proceedings of the 24th ACM SIGSOFT International SPIN Symposium on Model Checking of Software, Santa Barbara, CA, USA, July 10-14, 2017. pp. 142–151. ACM (2017). doi:10.1145/3092282.3092289

[25] Frohme, M., Howar, F., Steffen, B.: Learnlib: 10 years later. In: Piskac, R., Rakamaric, Z. (eds.) Computer Aided Verification - 37th International Conference, CAV 2025, Zagreb, Croatia, July 23-25, 2025, Proceedings, Part IV. pp. 141–160. Lecture Notes in Computer Science, Springer (2025), https://doi.org/10.1007/978-3-031-98685-7_7

[26] Fujiwara, S., von Bochmann, G., Khendek, F., Amalou, M., Ghedamsi, A.: Test selection based on finite state models. IEEE Transactions on Software Engineering **17**(6), 591–603 (1991). doi:10.1109/32.87284

[27] Garhewal, B., Damasceno, C.D.N.: An experimental evaluation of conformance testing techniques in active automata learning. In: 26th ACM/IEEE International Conference on Model Driven Engineering Languages and Systems, MODELS 2023, Västerås, Sweden, October 1-6, 2023. pp. 217–227. IEEE (2023), https://doi.org/10.1109/MODELS58315.2023.00012

[28] Groz, R., Brémond, N., da Silva Simão, A., Oriat, C.: hW-inference: A heuristic approach to retrieve models through black box testing. J. Syst. Softw. **159** (2020). doi:10.1016/J.JSS.2019.110426, https://doi.org/10.1016/j.jss.2019.110426

[29] Hamlet, R.: Random testing. Encyclopedia of software Engineering **2**, 971–978 (1994)

[30] Hopcroft, J., Ullman, J.: Introduction to Automata Theory, Languages and Computation. Addison-Wesley (1979)

[31] Howar, F., Steffen, B.: Active Automata Learning in Practice - An Annotated Bibliography of the years 2011 to 2016. In: Bennaceur, A., Hähnle, R., Meinke, K. (eds.) Machine Learning for Dynamic Software Analysis: Potentials and Limits - International Dagstuhl Seminar 16172, Dagstuhl Castle, Germany, April 24-27, 2016, Revised Papers. Lecture Notes in Computer Science, vol. 11026, pp. 123–148. Springer (2018). doi:10.1007/978-3-319-96562-8_5

[32] Jasper, M., Mues, M., Murtovi, A., Schlüter, M., Howar, F., Steffen, B., Schordan, M., Hendriks, D., Schiffelers, R.R.H., Kuppens, H., Vaandrager, F.W.: RERS 2019: Combining synthesis with real-world models. In: Beyer, D., Huisman, M., Kordon, F., Steffen, B. (eds.) Tools and Algorithms for the Construction and Analysis of Systems - 25 Years of TACAS: TOOLympics, Held as Part of ETAPS 2019, Prague, Czech Republic, April 6-11, 2019, Proceedings, Part III. Lecture Notes in Computer Science, vol. 11429, pp. 101–115. Springer (2019). doi:10.1007/978-3-030-17502-3_7, https://doi.org/10.1007/978-3-030-17502-3_7

[33] Kruger, L., Junges, S., Rot, J.: Small test suites for active automata learning. In: Finkbeiner, B., Kovács, L. (eds.) Tools and Algorithms for the Construction and Analysis of Systems - 30th International Conference, TACAS 2024, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2024, Luxembourg City, Luxembourg, April 6-11, 2024, Proceedings, Part II. Lecture Notes in Computer Science, vol. 14571, pp. 109–129. Springer (2024). doi:10.1007/978-3-031-57249-4_6, https://doi.org/10.1007/978-3-031-57249-4_6

[34] Kruger, L., Junges, S., Rot, J.: Error-awareness accelerates active automata learning (2026), https://arxiv.org/abs/2602.21674, to appear in Proceedings FM'26

[35] Lee, D., Yannakakis, M.: Principles and methods of testing finite state machines — a survey. Proceedings of the IEEE **84**(8), 1090–1123 (1996)

[36] Leucker, M., Schallhart, C.: A brief account of runtime verification. The Journal of Logic and Algebraic Programming **78**(5), 293–303 (2009). doi:https://doi.org/10.1016/j.jlap.2008.08.004, https://www.sciencedirect.com/science/article/pii/S1567832608000775, the 1st Workshop on Formal Languages and Analysis of Contract-Oriented Software (FLACOS'07)

[37] Luo, G., Petrenko, A., von Bochmann, G.: Selecting test sequences for partially-specified nondeterministic finite state machines. In: Mizuno, T., Higashino, T., Shiratori, N. (eds.) Protocol Test Systems: 7th workshop 7th IFIP WG 6.1 international workshop on protocol text systems. pp. 95–110. Springer US, Boston, MA (1995). doi:10.1007/978-0-387-34883-4_6

[38] Meijer, J., van de Pol, J.: Sound black-box checking in the learnlib. Innov. Syst. Softw. Eng. **15**(3-4), 267–287 (2019). doi:10.1007/S11334-019-00342-6, https://doi.

org/10.1007/s11334-019-00342-6
[39] Meinke, K., Niu, F., Sindhu, M.A.: Learning-based software testing: A tutorial. In: Hähnle, R., Knoop, J., Margaria, T., Schreiner, D., Steffen, B. (eds.) Leveraging Applications of Formal Methods, Verification, and Validation - International Workshops, SARS 2011 and MLSC 2011. Revised Selected Papers. Communications in Computer and Information Science, vol. 336, pp. 200–219. Springer (2011)
[40] Meinke, K., Sindhu, M.A.: Incremental learning-based testing for reactive systems. In: Gogolla, M., Wolff, B. (eds.) Tests and Proofs - 5th International Conference, TAP@TOOLS 2011, Zurich, Switzerland, June 30 - July 1, 2011. Proceedings. Lecture Notes in Computer Science, vol. 6706, pp. 134–151. Springer (2011). doi:10.1007/978-3-642-21768-5_11, https://doi.org/10.1007/978-3-642-21768-5_11
[41] Muskardin, E., Aichernig, B., Pferscher, A., Tappler, M.: Aalpy: an active automata learning library. Innovations in Systems and Software Engineering **18**, 1–10 (03 2022). doi:10.1007/s11334-022-00449-3
[42] Neider, D., Smetsers, R., Vaandrager, F.W., Kuppens, H.: Benchmarks for automata learning and conformance testing. In: Margaria, T., Graf, S., Larsen, K.G. (eds.) Models, Mindsets, Meta: The What, the How, and the Why Not? Lecture Notes in Computer Science, vol. 11200, pp. 390–416. Springer (2018)
[43] Pawagi, M., Shao, L., Lee, H., Sun, Y., Wang, W.: Rfcscope: Detecting logical ambiguities in internet protocol specifications. In: 2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE). pp. 1234–1246 (2025). doi:10.1109/ASE63991.2025.00106
[44] Peled, D., Vardi, M.Y., Yannakakis, M.: Black box checking. In: Wu, J., Chanson, S.T., Gao, Q. (eds.) Proceedings FORTE. IFIP Conference Proceedings, vol. 156, pp. 225–240. Kluwer (1999)
[45] Peled, D.A., Vardi, M.Y., Yannakakis, M.: Black box checking. J. Autom. Lang. Comb. **7**(2), 225–246 (2002), https://doi.org/10.25596/jalc-2002-225
[46] Petrenko, A., Yevtushenko, N.: Conformance tests as checking experiments for partial nondeterministic FSM. In: Grieskamp, W., Weise, C. (eds.) Formal Approaches to Software Testing. pp. 118–133. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
[47] Petrenko, A., Yevtushenko, N., Lebedev, A., Das, A.: Nondeterministic state machines in protocol conformance testing. In: Rafiq, O. (ed.) Protocol Test Systems, VI, Proceedings of the IFIP TC6/WG6.1 Sixth International Workshop on Protocol Test systems, Pau, France, 28-30 September, 1993. IFIP Transactions, vol. C-19, pp. 363–378. North-Holland (1993)
[48] Pferscher, A., Aichernig, B.K.: Stateful black-box fuzzing of bluetooth devices using automata learning. In: Deshmukh, J.V., Havelund, K., Perez, I. (eds.) NASA Formal Methods - 14th International Symposium, NFM 2022, Pasadena, CA, USA, May 24-27, 2022, Proceedings. Lecture Notes in Computer Science, vol. 13260, pp. 373–392. Springer (2022). doi:10.1007/978-3-031-06773-0_20
[49] Raffelt, H., Steffen, B.: Learnlib: A library for automata learning and experimentation. In: Baresi, L., Heckel, R. (eds.) Fundamental Approaches to Software Engineering, 9th International Conference, FASE 2006, Held as Part of the Joint European Conferences on Theory and Practice of Software, ETAPS 2006, Vienna, Austria, March 27-28, 2006, Proceedings. pp. 377–380. Lecture Notes in Computer Science, Springer (2006), https://doi.org/10.1007/11693017_28
[50] de Ruiter, J., Poll, E.: Protocol State Fuzzing of TLS Implementations. In: Jung, J., Holz, T. (eds.) Proceedings of the 24th USENIX Security Symposium, USENIX Security 15. pp. 193–206. USENIX Association (2015), https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/de-ruiter
[51] Schuts, M., Hooman, J., Vaandrager, F.: Refactoring of legacy software using model learning and equivalence checking: an industrial experience report. In: Ábrahám, E., Huisman, M. (eds.) Proceedings 12th International Conference on integrated Formal Methods (iFM). LNCS, vol. 9681, pp. 311–325 (2016)
[52] Smeenk, W., Moerman, J., Vaandrager, F.W., Jansen, D.N.: Applying automata learning to embedded control software. In: Butler, M.J., Conchon, S., Zaïdi, F. (eds.) Formal Methods and Software Engineering - 17th International Conference on Formal Engineering Methods, ICFEM 2015, France, 2015, Proceedings. Lecture Notes in Computer Science, vol. 9407, pp. 67–83. Springer (2015). doi:10.1007/978-3-319-25423-4_5
[53] Tretmans, J.: Model based testing with labelled transition systems. In: Hierons, R.M., Bowen, J.P., Harman, M. (eds.) Formal Methods and Testing: An Outcome of the FORTEST Network, Revised Selected Papers. pp. 1–38. Springer, Berlin, Heidelberg (2008), https://doi.org/10.1007/978-3-540-78917-8_1
[54] Vaandrager, F., Garhewal, B., Rot, J., Wißmann, T.: A new approach for active automata learning based on apartness. In: Fisman, D., Rosu, G. (eds.) Tools and Algorithms for the Construction and Analysis of Systems. pp. 223–243. Springer International Publishing, Cham (2022)
[55] Vaandrager, F., Melse, I.: New Fault Domains for Conformance Testing of Finite State Machines. In: Bouyer, P., van de Pol, J. (eds.) 36th International Conference on Concurrency Theory (CONCUR 2025). Leibniz International Proceedings in Informatics (LIPIcs), vol. 348, pp. 34:1–34:22. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2025), https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.CONCUR.2025.34
[56] Vaandrager, F., Wißmann, T.: Action Codes. In: Etessami, K., Feige, U., Puppis, G. (eds.) 50th International Colloquium on Automata, Languages, and Programming (ICALP 2023). Leibniz International Proceedings in Informatics (LIPIcs), vol. 261, pp. 137:1–137:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2023), https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2023.137
[57] Vaandrager, F.W.: Model learning. Commun. ACM **60**(2), 86–95 (2017). doi:10.1145/2967606
[58] Vasilevskii, M.: Failure diagnosis of automata. Cybernetics and System Analysis **9**(4), 653–665 (1973). doi:10.1007/BF01068590, (Translated from Kibernetika, No. 4, pp. 98-108, July-August, 1973.)
[59] Wei, Y., Wei, K., Du, S., Wang, J., Liu, Z., Wang, Y., Li, Z., Miao, C., Xie, X., Cui, Y.: Automated network protocol testing with llm agents (2025), https://arxiv.org/abs/2510.13248
[60] Weyuker, E.J.: Assessing test data adequacy through program inference. ACM Trans. Program. Lang. Syst. **5**(4), 641–655 (1983), https://doi.org/10.1145/69575.357231
[61] Yang, N., Aslam, K., Schiffelers, R.R.H., Lensink, L., Hendriks, D., Cleophas, L., Serebrenik, A.: Improving model inference in industry by combining active and passive learning. In: Wang, X., Lo, D., Shihab, E. (eds.) 26th IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2019, Hangzhou, China, February 24-27, 2019. pp. 253–263. IEEE (2019), https://doi.org/10.1109/SANER.2019.8668007
[62] Yevtushenko, N.V., Petrenko, A.F.: Synthesis of test experiments in some classes of automata. Autom. Control Comput. Sci. **24**(4), 50–55 (apr 1991)
[63] Zameni, T., van den Bos, P., Rensink, A., Tretmans, J.: An intermediate language to integrate behavior-driven development scenarios and model-based testing. In: IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2024 - Companion, Rovaniemi, Finland, March 12, 2024. pp. 199–206. IEEE (2024), https://doi.org/10.1109/SANER-C62648.2024.00033
[64] Zeller, A., Gopinath, R., Böhme, M., Fraser, G., Holler, C.: The Fuzzing Book. CISPA Helmholtz Center for Information Security (2024), https://www.fuzzingbook.org/, retrieved 2024-07-01 16:50:18+02:00