Taming AI-CPS
(Introduction of JST CREST CyPhAI project)

Kohei Suenaga（Kyoto University）

Joint work with
Masaki Waga, Amit Gurung, Chiao Hsieh, Chao Gao
Ryotaro Banno, Kotaro Matsuoka, Naoki Matsumoto
Junya Shijubo, and Tsubasa Matsumoto

JST ASPIRE “AI-Physical Systems”
Kick-Off Meeting

March 4th, 2026

What this talk is about

- Introducing results of JST CREST project “AI-intensive Cyber-Physical Systems: Formal Analysis and Design Methods”
  - Black-box checking (BBC): Testing a black-box system effectively and efficiently using formal verification
  - Oblivious monitoring: Monitoring sensitive data securely, using fully-homomorphic encryption scheme

Study theories and methods to formally model them to guarantee their safety

CyPhAI: Formal Analysis and Design of AI-intensive Cyber-Physical Syst

... in which AI is an important ingredient

Systems in which software and physical world interacts...

CyPhAI: AI-intensive Cyber-Physical Systems: Formal Analysis and Design Methods

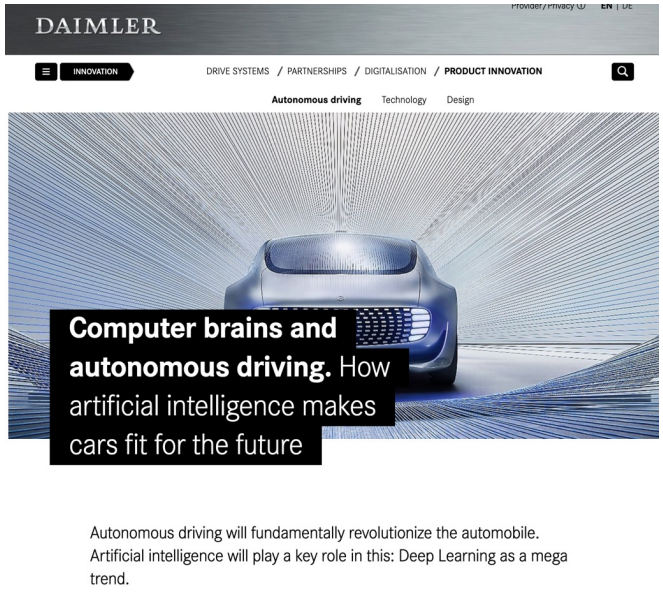
Kohei Suenaga (Kyoto University, Japan)
Thao Dang (CNRS, France)

AI-intensive Cyber-Physical Systems (AI-CPS)

Cyber-Physical Systems (CPS) in which AI plays an important role

Systems in which **software** and **physical world** interacts

Machine-learned software components

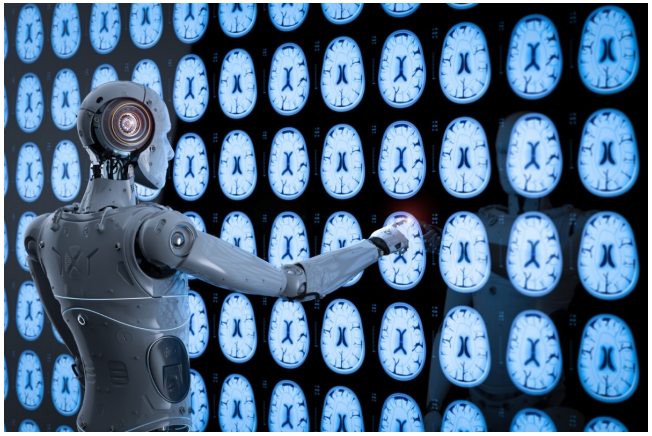


https://www.daimler.com/innovation/case/autonomous/artificial-intelligence.html

Autonomous Vehicle (AV)

https://www.bbc.com/news/business-49019390

Factories



https://newatlas.com/nvidia-uk-nhs-artificial-intelligence-radiological-platform/59637/

Medical devices

Unsafe AI-CPS may cost lives and money

https://www.vox.com/science-and-health/2018/3/19/17139868/self-driving-uber-killed-pedestrian-human-drivers-deadly

Ethiopian Airlines Tragic Crash

Surabhi Ashok, Staff Writer
March 22, 2019

The Ethiopian Airlines flight crashed only a little after its takeoff on Sunday. On it were international experts and humanitarian workers, mostly, who

https://www.falconers-voice.com/2019/03/ethiopian-airlines-tragic-crash/

NEWS

'Turn it Off and On Again Every 149 Hours' Is a Concerning Remedy for a $300 Million Airbus Plane's Software Bug

Andrew Liszewski
Tuesday 11:40am • Filed to: AIRBUS

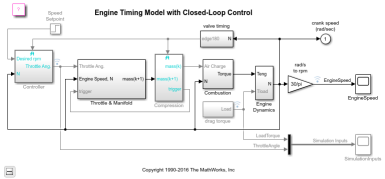
105.8K 162 5

https://gizmodo.com/turn-it-off-and-on-again-every-149-hours-is-a-concernin-1836818094

Current methods to make CPS safer

Design CPS as a model, repeat testing and verification using the model, and then deploy the system

Model M



Correctness specification ϕ: "If the value of gear exceeds 3, then speed should exceed 10 in 3 sec."

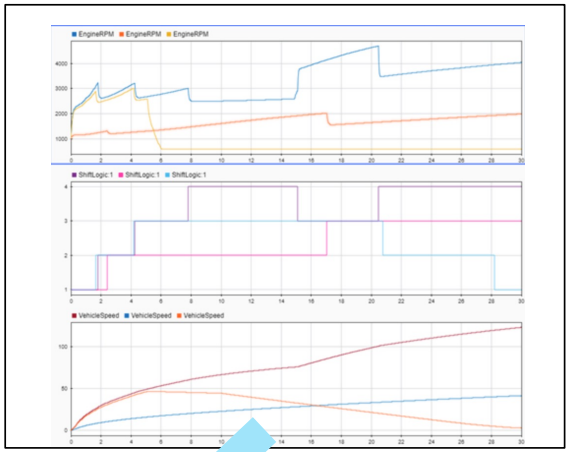
- Model M expresses the design of CPS
- Specification ϕ expresses the properties that M has to satisfy

Testing and verification (mainly by simulations)

MATLAB SIMULINK

MapleSim

Simulation result



Corrections to M if bug is found

If bug is not found

https://jp.mathworks.com/help/simulink/inspect-and-analyze-simulation-results.html

https://www.avl.com/-/system-engineering-for-assisted-autonomous-driving

Goal of the proposal

Mathematically solid theories and machine-learning-assisted methods for formally modeling and reasoning about AI-CPS to ensure their safety

- **Theory to formally model AI-CPS (i.e., learning + CPS)**
  - Measures for behaviors of AI-CPS
  - Theory for formally modeling learning behaviors of AI-CPS
- **Methods to design and monitor AI-CPS**
  - Effective learning algorithms for AI-CPS and its correctness specification
  - Validation for AI-CPS
  - Monitoring and controlling AI-CPS
- **Experiments and case studies for the proposed framework**

Outline

- Taming black-boxes by Black-Box Checking (BBC)
  - Preliminary
  - Our work related to BBC
- Other work done in our project
  - Oblivious runtime verification
  - Interpretability for image-classification AI

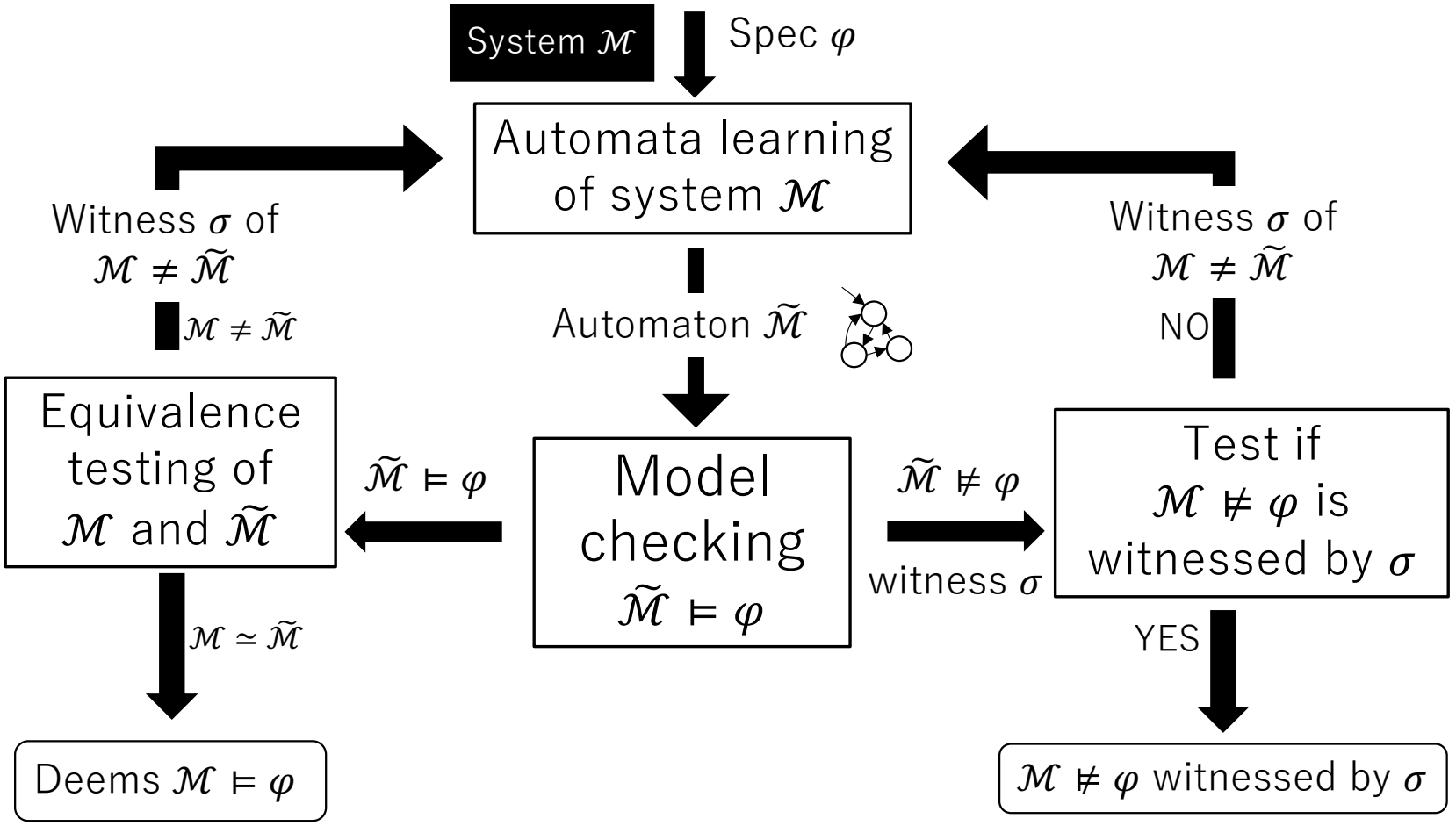
Q: What is the difficulty of AI-CPS from the viewpoint of formal methods?

A: AI-CPS contains black boxes

Then, we can create a formal method than can handle black boxes!!

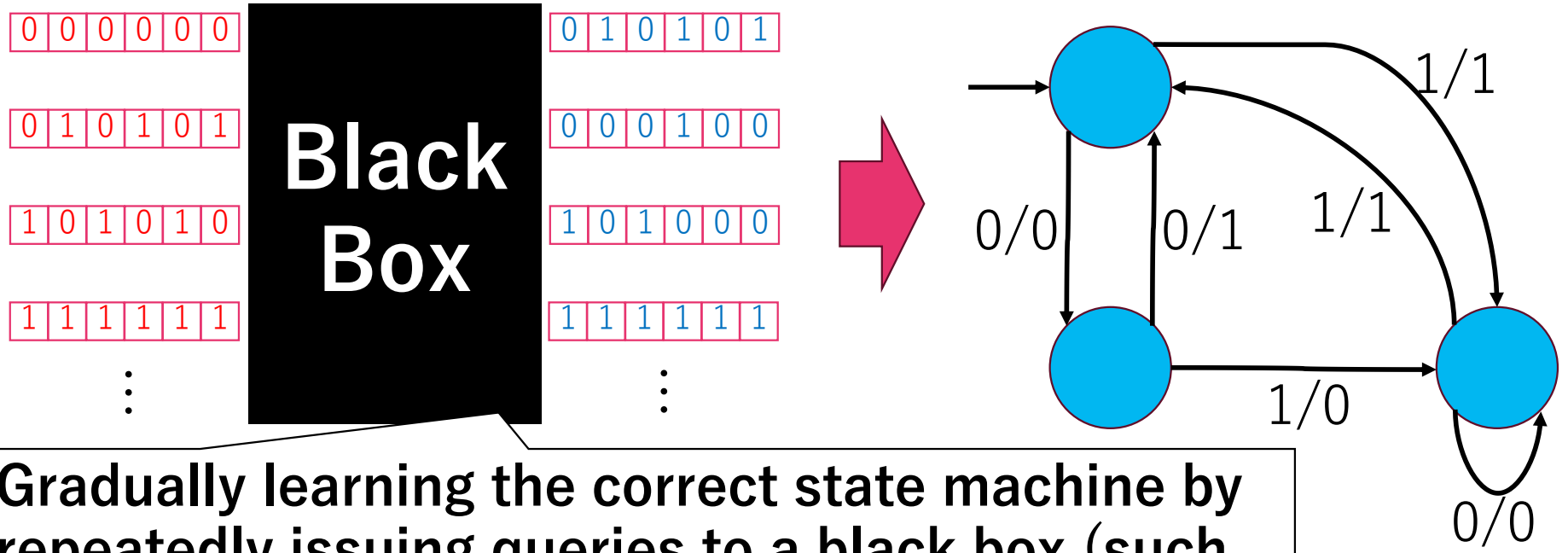
Black-Box Checking (BBC) [Peled et al. PSTV'99]

Effective and efficient testing of black-box systems using automata learning



Automata Learning [Angluin'87, etc.]

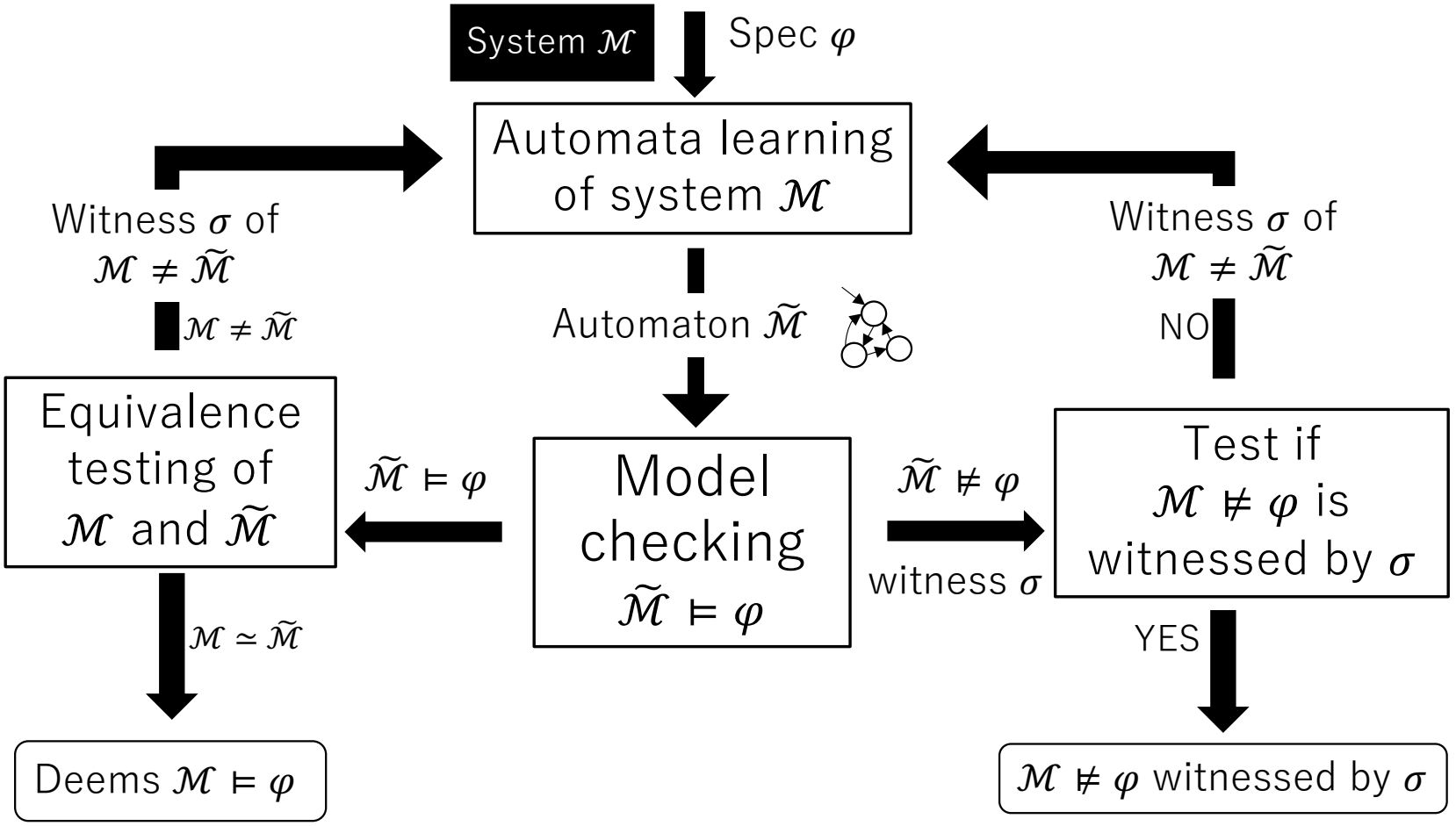
- Automatically learning state machines (automaton/Mealy machine/Moore machine) approximating behavior of a black box from its I/O behavior



Gradually learning the correct state machine by repeatedly issuing queries to a black box (such as outputs for input strings, equivalence with the learned state machine, etc.)

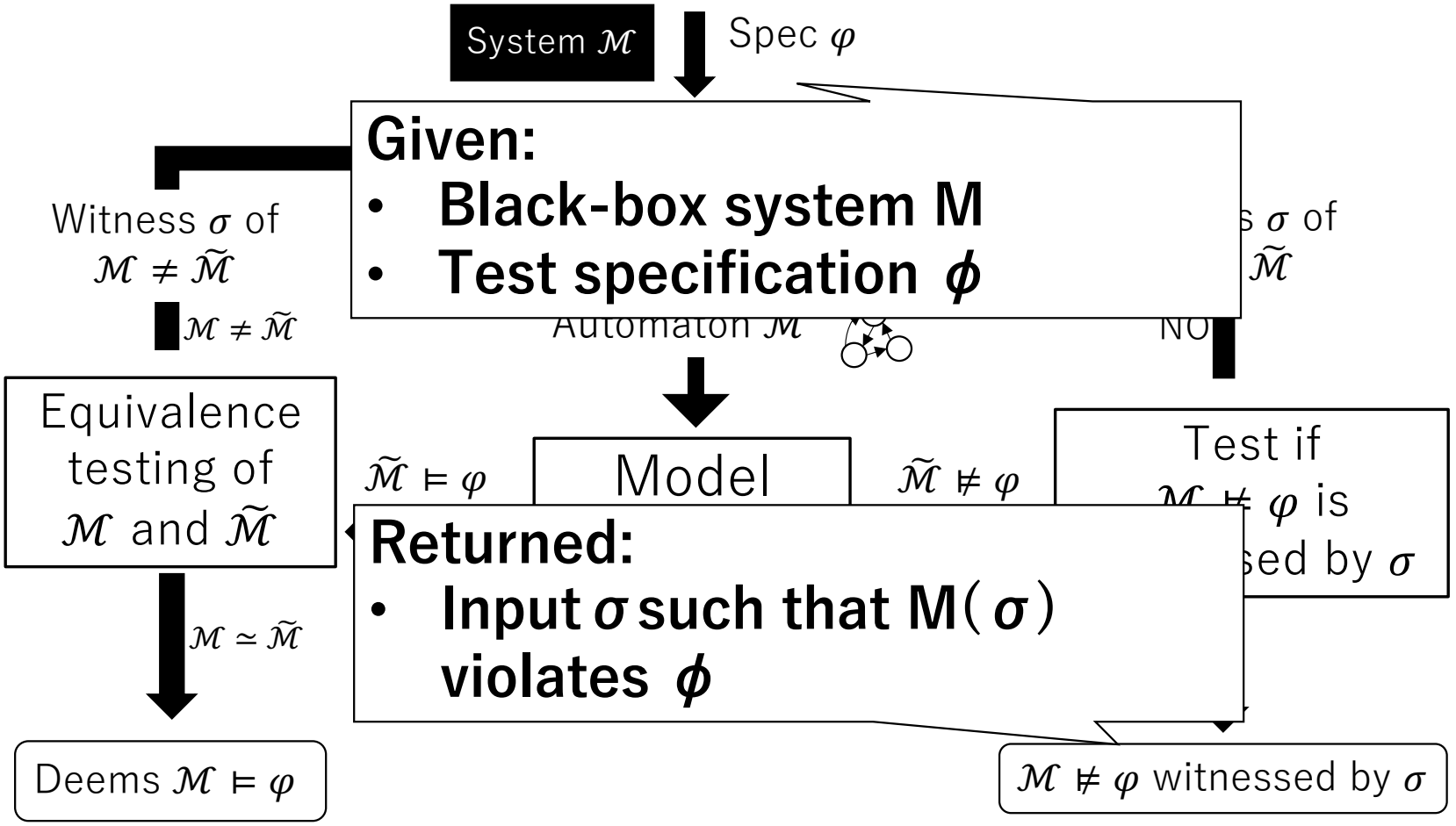
Black-Box Checking (BBC) [Peled et al. PSTV'99]

Effective and efficient testing of black-box systems using automata learning



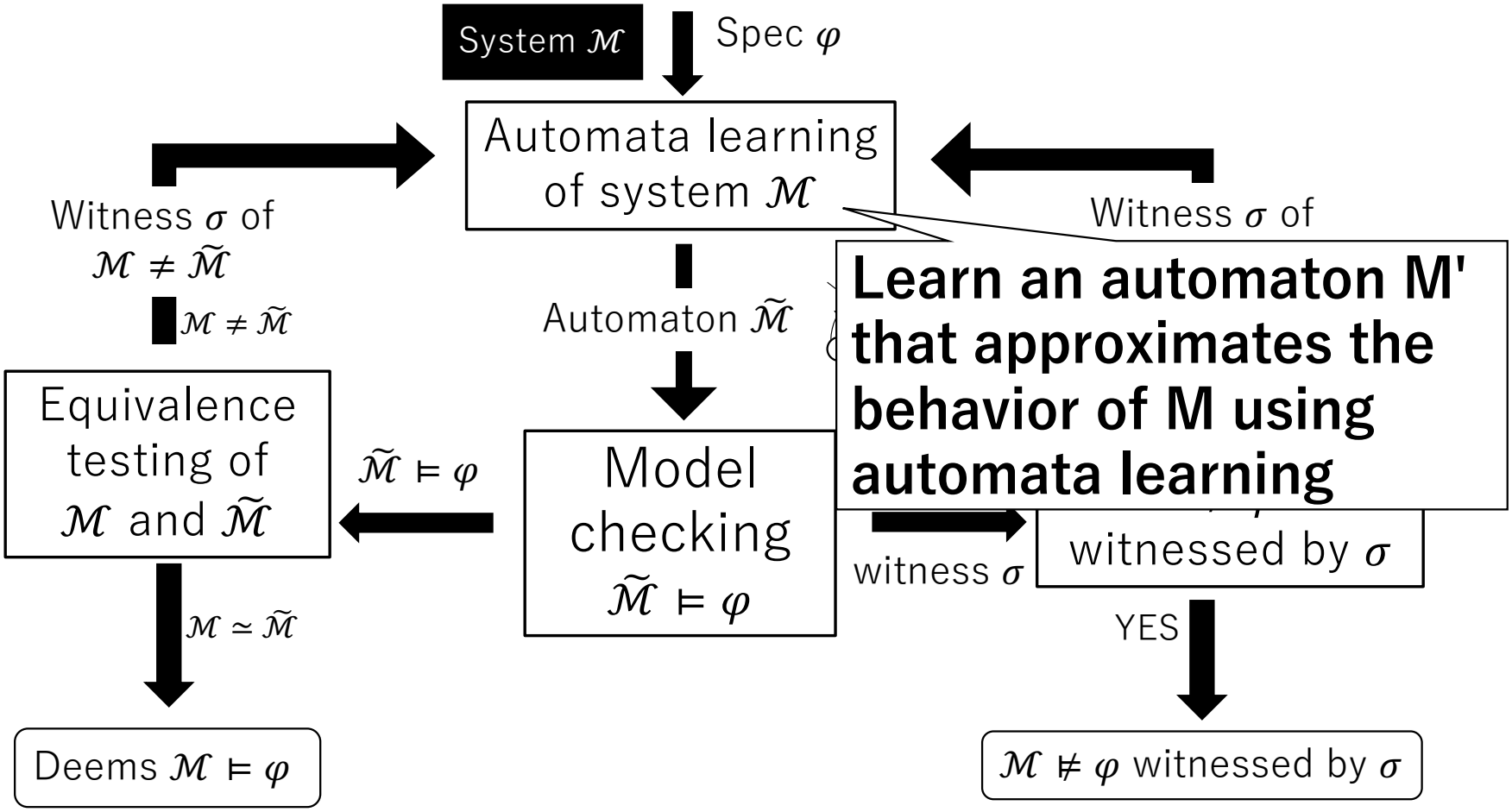
Black-Box Checking (BBC) [Peled et al. PSTV'99]

Effective and efficient testing of black-box systems using automata learning



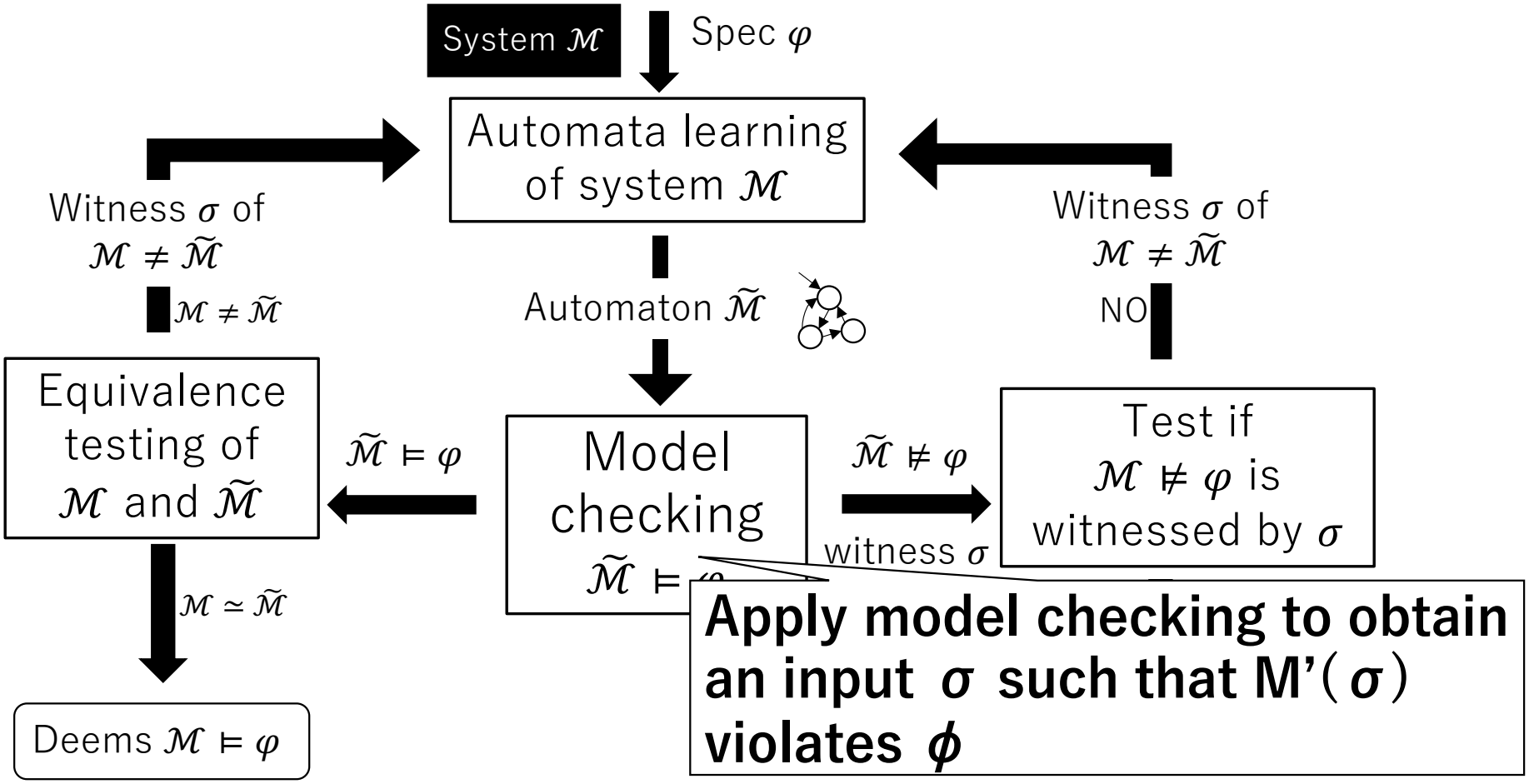
Black-Box Checking (BBC) [Peled et al. PSTV'99]

Effective and efficient testing of black-box systems using automata learning



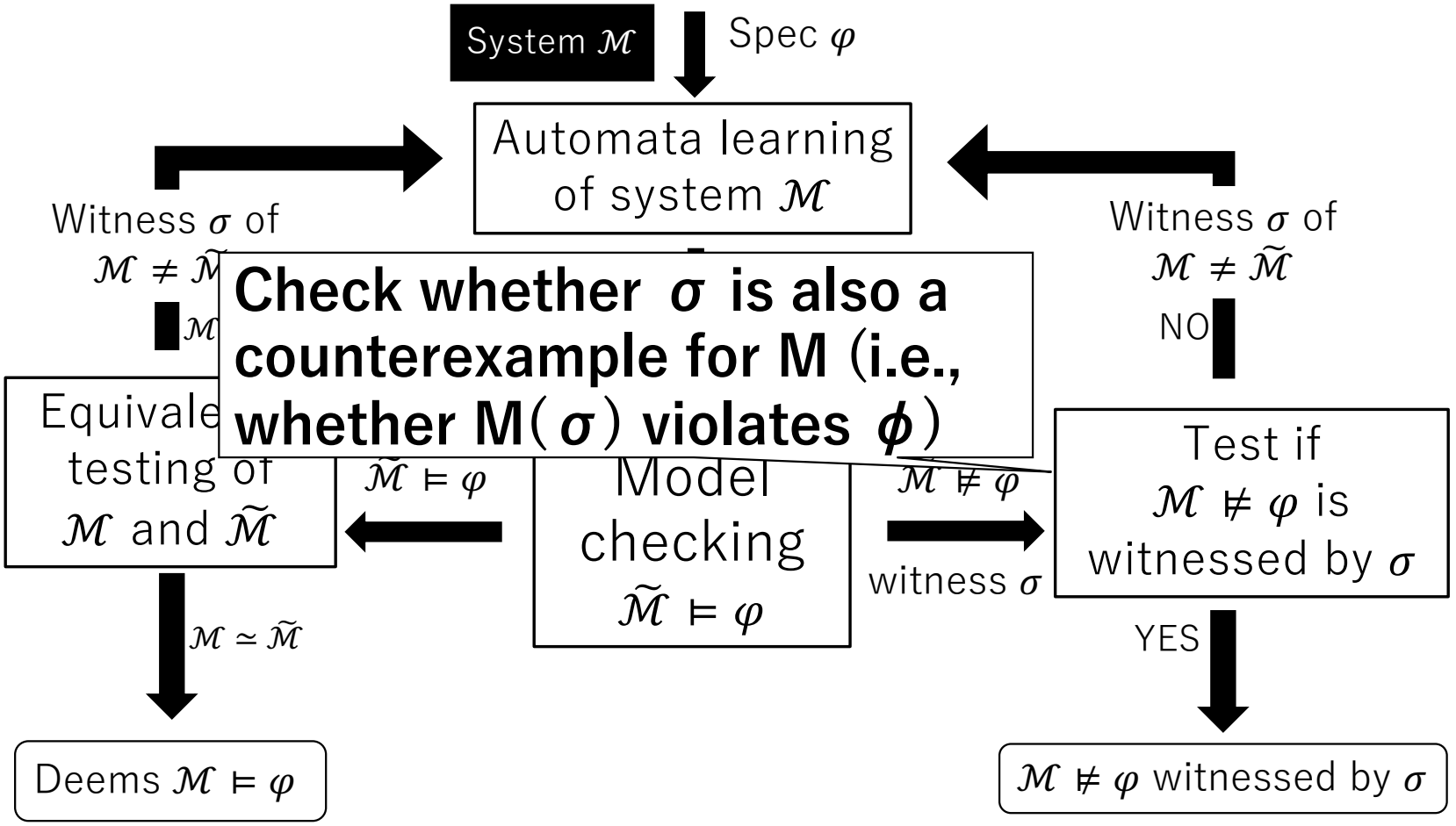
Black-Box Checking (BBC) [Peled et al. PSTV'99]

Effective and efficient testing of black-box systems using automata learning



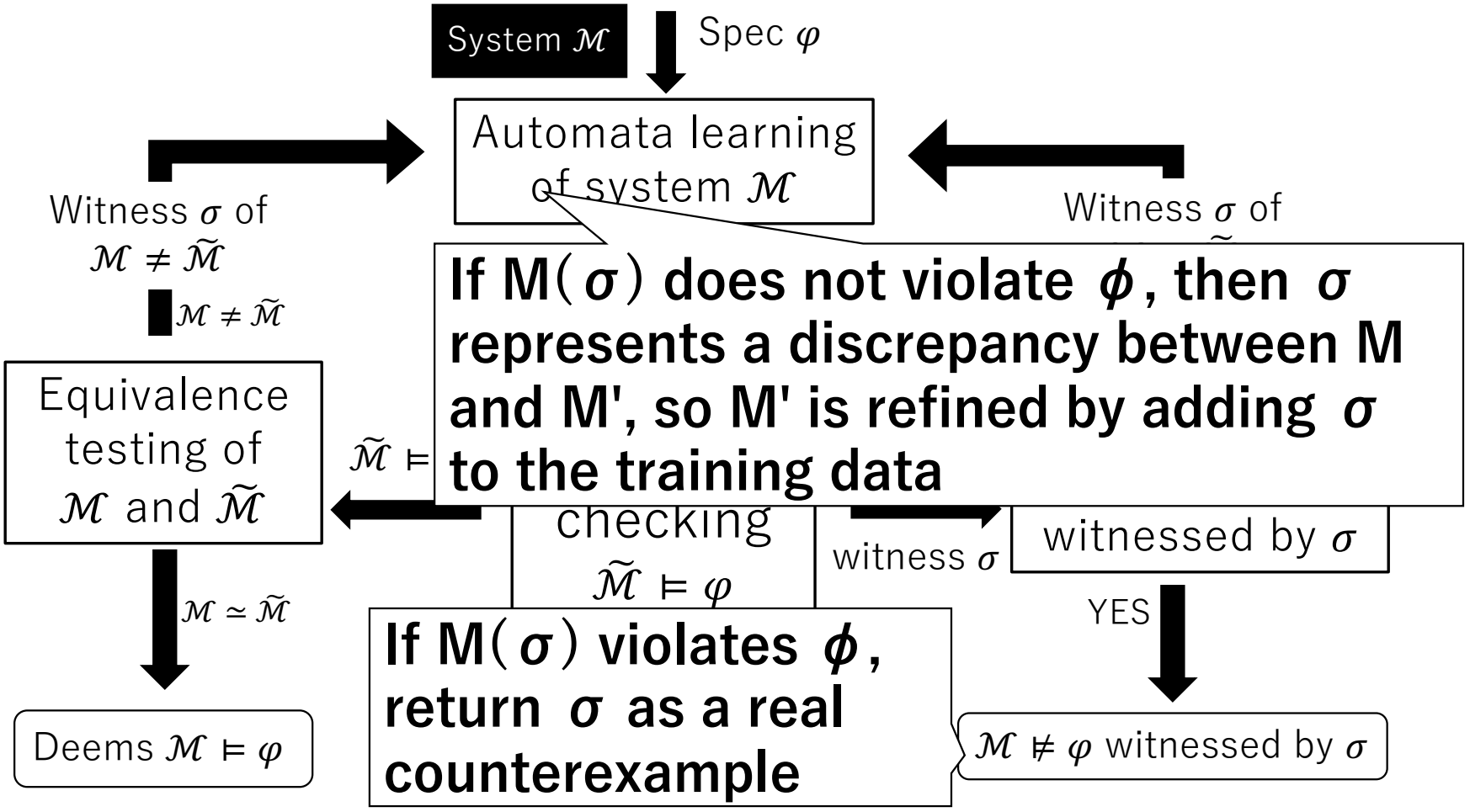
Black-Box Checking (BBC) [Peled et al. PSTV'99]

Effective and efficient testing of black-box systems using automata learning



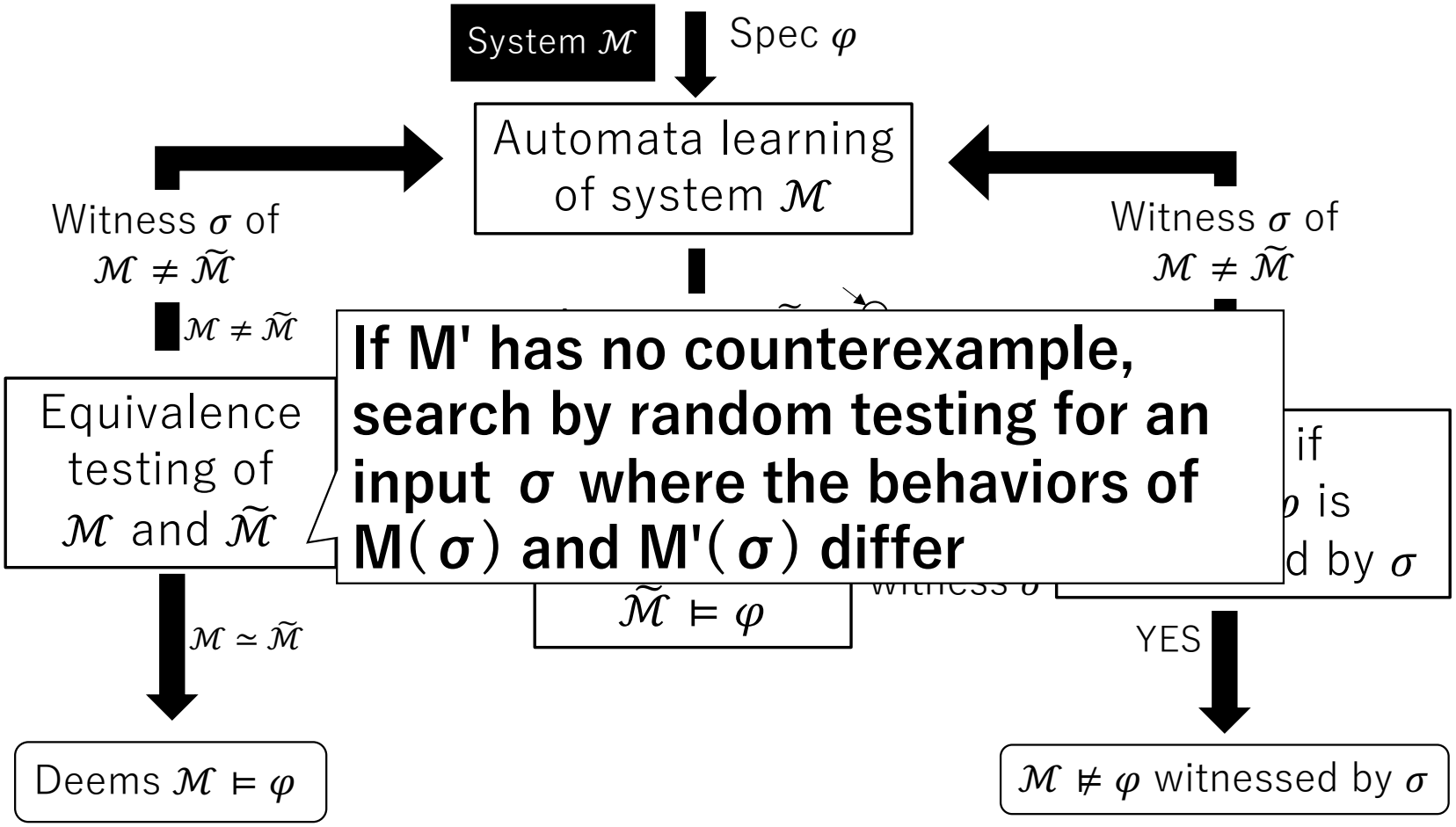
Black-Box Checking (BBC) [Peled et al. PSTV'99]

Effective and efficient testing of black-box systems using automata learning

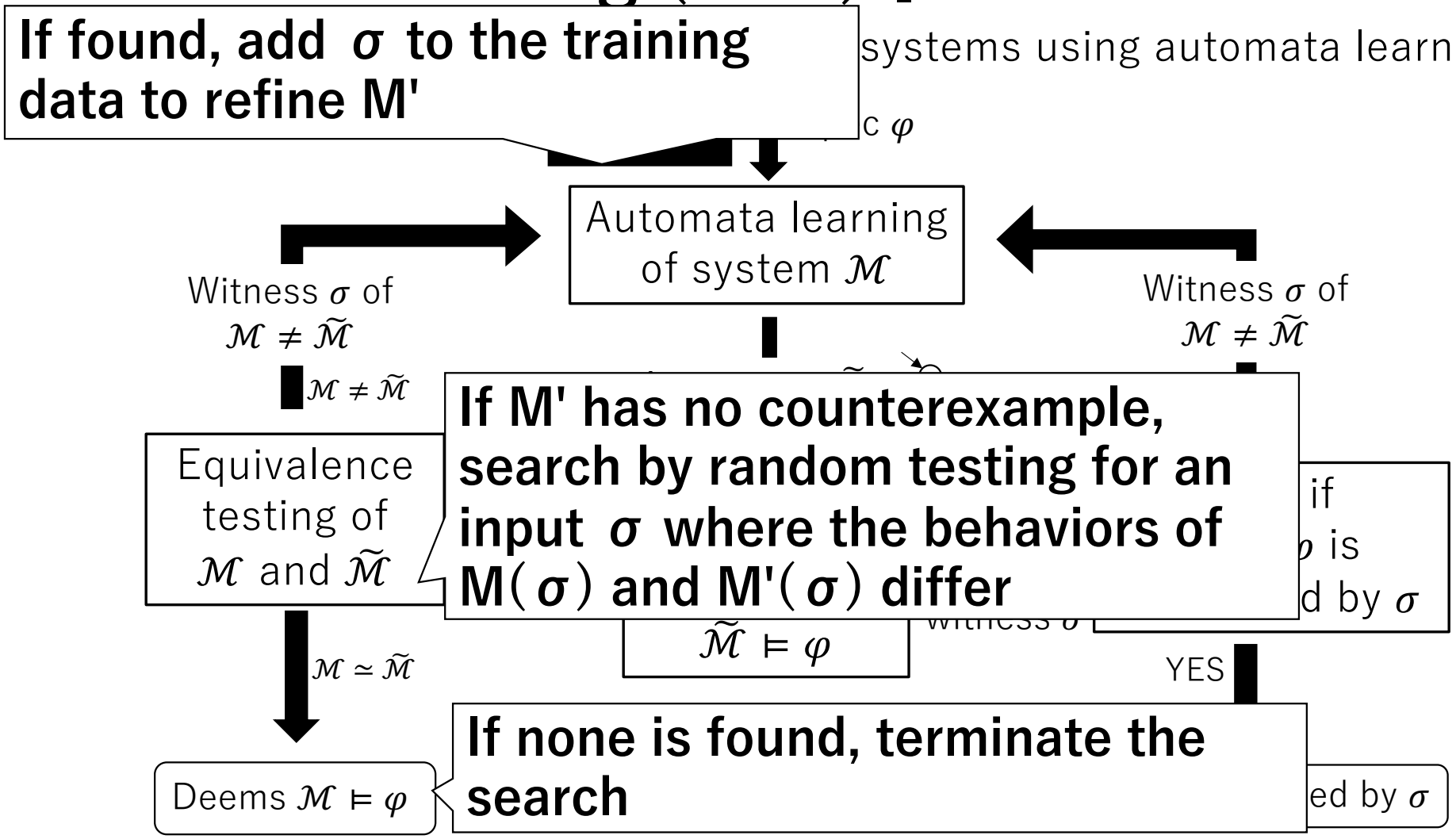


Black-Box Checking (BBC) [Peled et al. PSTV'99]

Effective and efficient testing of black-box systems using automata learning



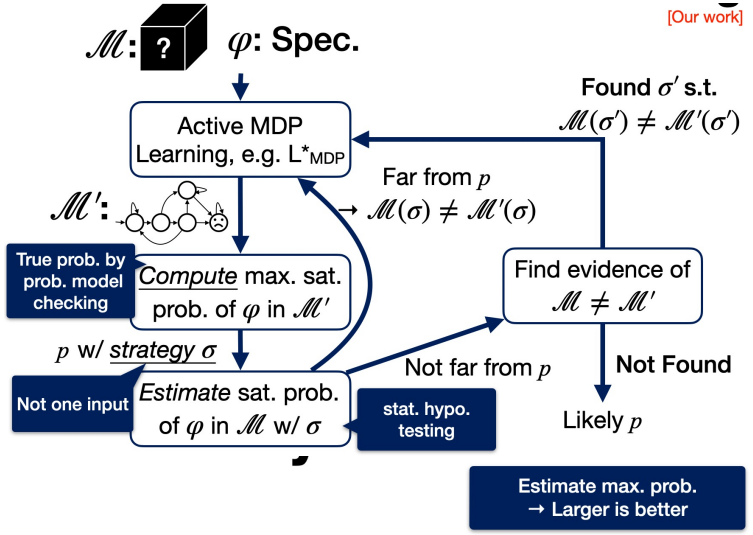
Black-Box Checking (BBC) [Peled et al. PSTV'99]



Outline

- Taming black-boxes by Black-Box Checking (BBC)
  - Preliminary
  - Our work related to BBC
- Other work done in our project
  - Oblivious runtime verification
  - Interpretability for image-classification AI

Extension to Systems with Probabilistic Behaviors [EMSOFT'23]



	Ground Truth	Our Method	ProbBlackReach
Slot machine	0.510	0.507	0.480
Slot machine with limited observation	0.510	0.509	0.448
MQTT	0.815	0.808	0.815
TCP	0.771	0.768	0.771
GridWorld Small	0.618	0.617	0.569
GridWorld Large	0.671	0.672	0.0683
SharedCoin	0.250	0.251	0.218

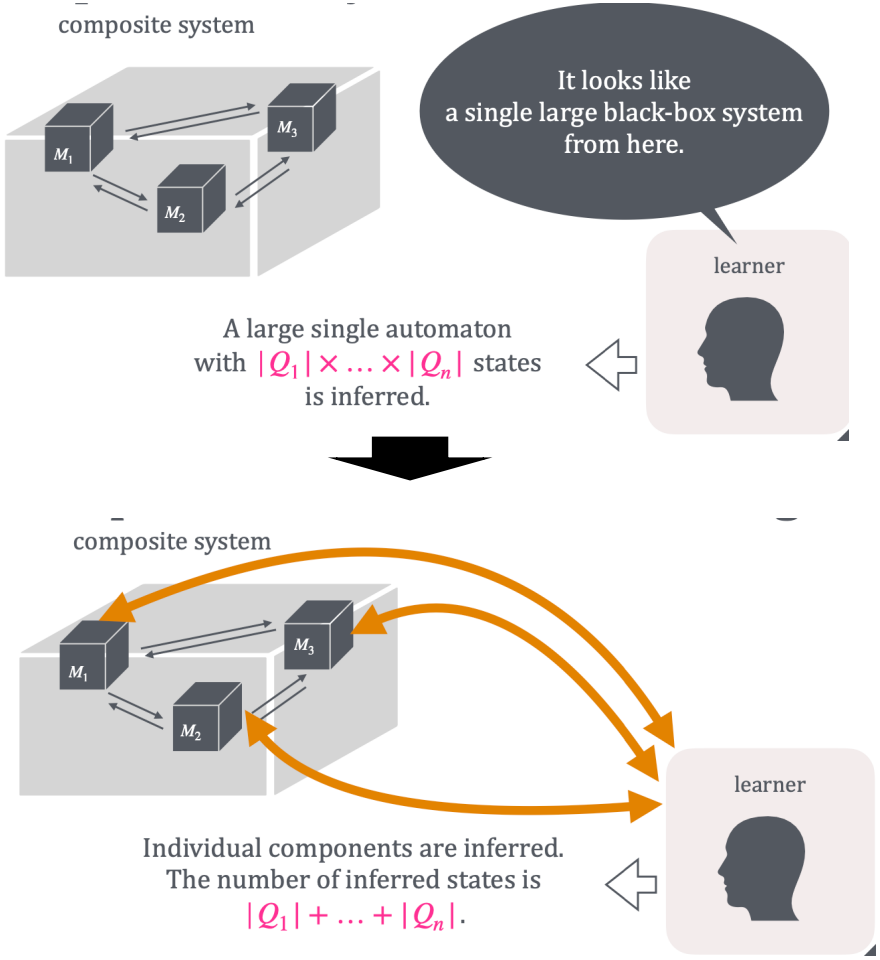
Always close to ground truth

Not much close to ground truth

Far from ground truth

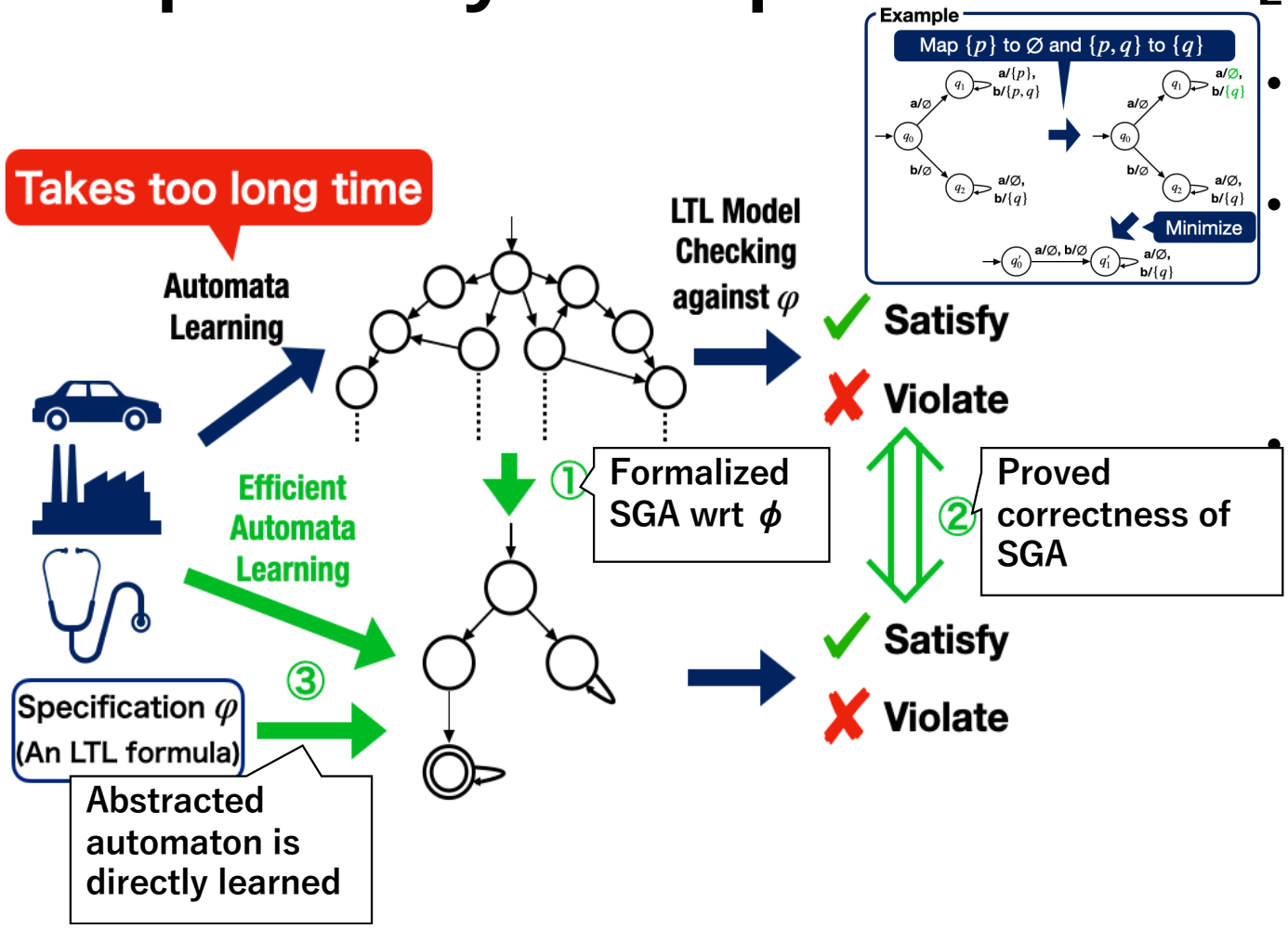
- **Application of BBC to probabilistic systems**
  - Idea: Combination with active MDP learning, probabilistic model checking, and statistical model checking
  - Result: more accurate verification results than conventional methods based on passive MDP learning

Application of BBC to Systems Composed of Multiple Components [ATVA'25]



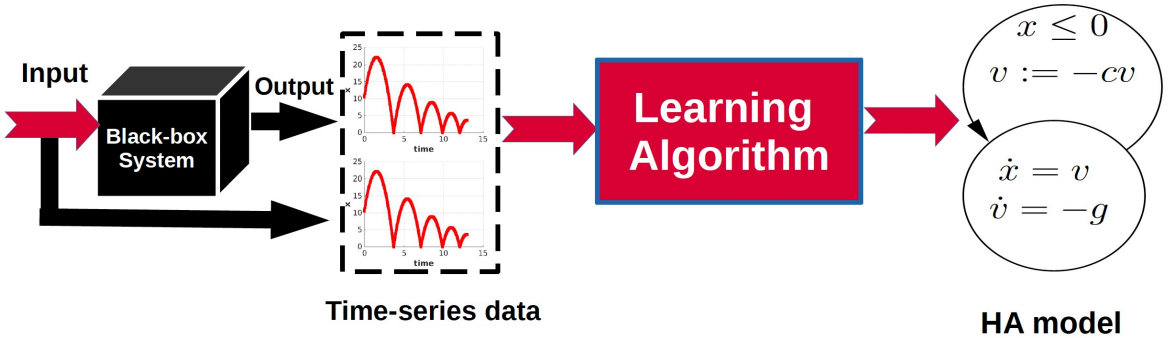
- **Efficient automata learning and BBC for integrated system**
  - Challenge: Integrated system often includes redundant functionalities in each component, resulting in state explosion when automata learning is applied
  - Our work: Scalability by not learning those component functions that are not used by the overall system
  - Result: Effectiveness is demonstrated on examples such as the MQTT protocol, which is widely used in IoT systems

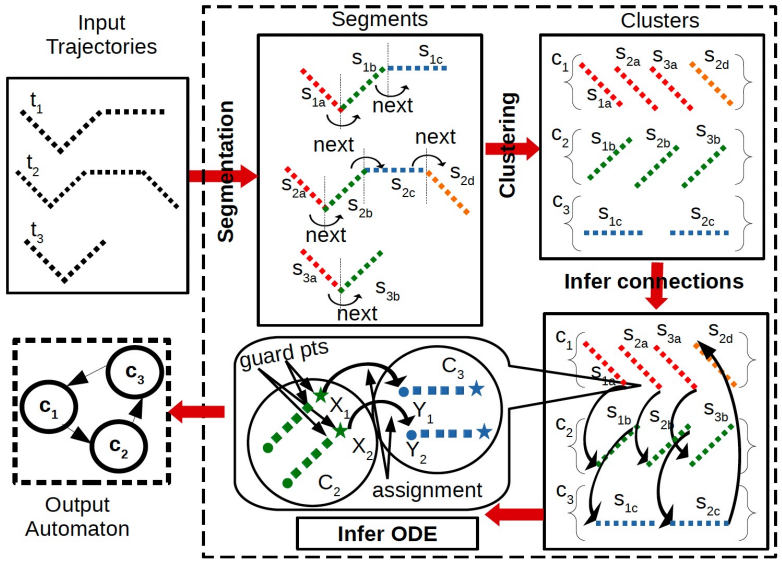
Efficiency through Learning Only the Parts Required by the Specification [EMSOFT'25]



- Achieved efficiency by learning only relevant part
- Observation: Given a specification ϕ, several signals can be identified with each other
- Idea: Specification-guided abstract (SGA)

Automatic Learning for Hybrid Automata [ATVA'22]



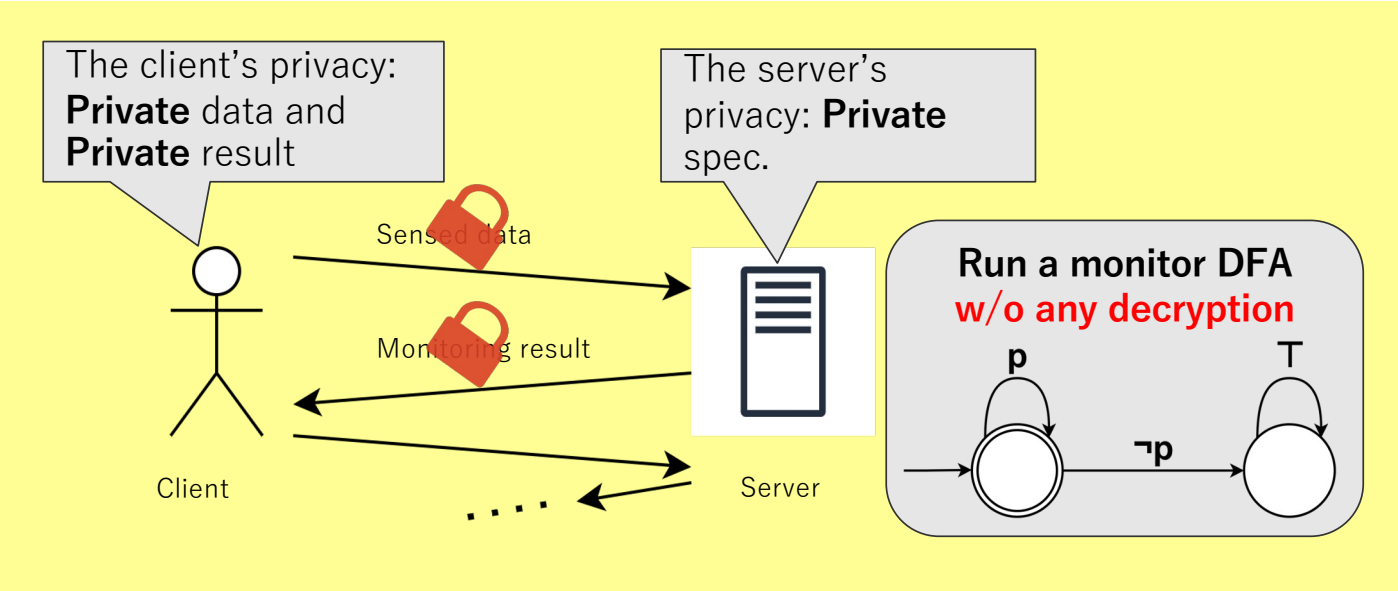


- Automatically generate hybrid automata from only input-output time-series data
- Demonstrate through benchmarks including neural excitation models that the method learns automata more efficiently and accurately than conventional methods
- Ongoing: Experiments on industrial models in joint research and currently co-authoring a paper with the French side
- Ongoing: Application to BBC

Outline

- Taming black-boxes by Black-Box Checking (BBC)
  - Preliminary
  - Our work related to BBC
- Other work done in our project
  - Oblivious runtime verification
  - Interpretability for image-classification AI

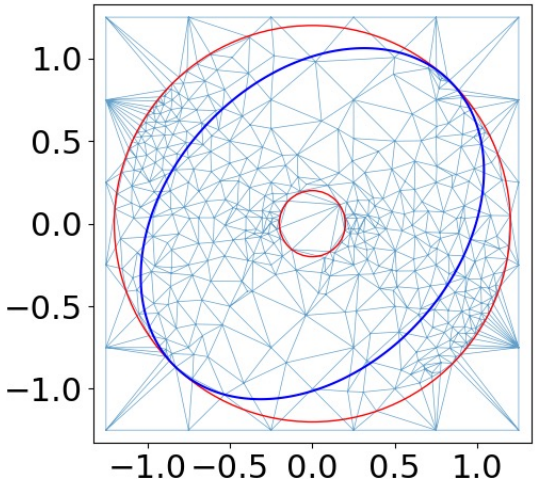
Online Oblivious Monitoring Using Fully Homomorphic Encryption [CAV'22]



- Communication protocol using fully homomorphic encryption to realize monitoring specs in LTL without decrypting sensor data
- Case study detecting hypoglycemia and blood glucose fluctuations
- Extended via joint research with industry to handle more complex numerical computations

Synthesizing Lyapunov Function for Black-Box CDS [HSCC'25]

- Theoretical Outcomes:
  - Sufficient regional verification conditions for arbitrary samples
  - Stability-guided approximation for finding a good candidate V instead of accurately approximating f
- Counterexample Guided Synthesis (CEGIS) algorithm:
  - Lazy sampling with nonuniform regions covering ROI to reduce the number of samples
  - Termination for a class of hypothesis space for V
- Numerical Evaluation:
  - At the best case, we certified a 2D system with few hundreds samples versus 3Kx3K samples in [Zhou, 2022]
  - Verifying Lyapunov criteria on V takes much more samples than learning a good candidate V.

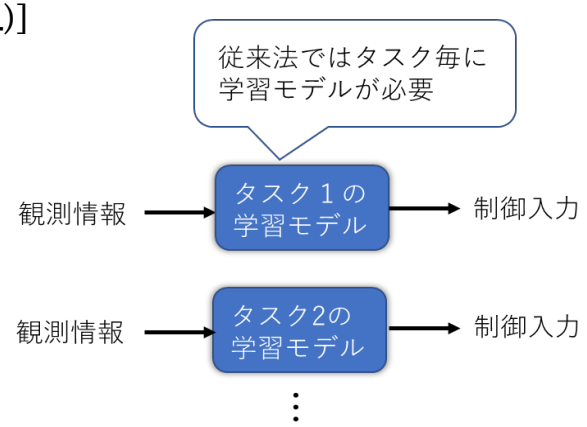


Non-uniform samples and regions for certifying the learned Lyapunov function

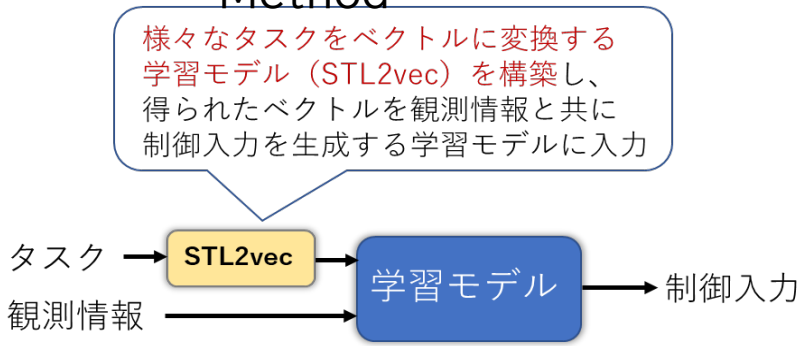
Control Using Vector Embedding of STL

[IEEE Robotics and Automation Letters (RA-L)]

- **Proposed Method: STL2vec**
  - Embeds STL specifications into vectors using a method inspired by **Word2vec (skip-gram)**, trained so that **similar specifications yield close embeddings**.
  - Handling multiple **STL specifications** with a **single RNN** (no retraining needed for specification switching).
    - Reduces **memory consumption**
  - Applied to autonomous mobile robot control, simulating 300+ task types
    - Comparable control performance to conventional methods, successfully reducing memory consumption to 1/24.



Conventional Method



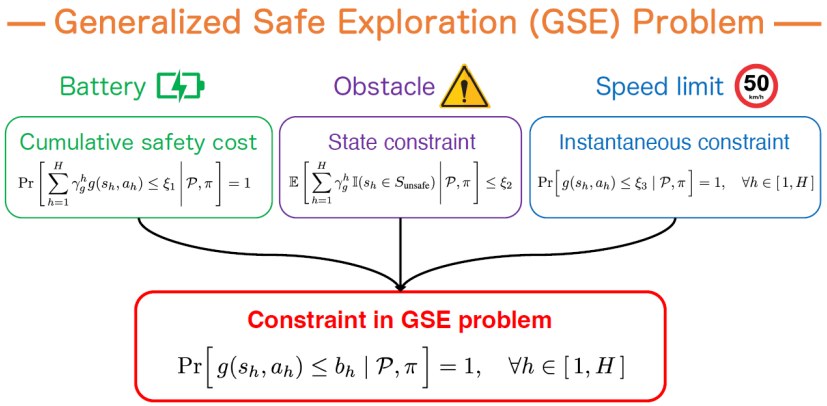
Overview of the Proposed Method

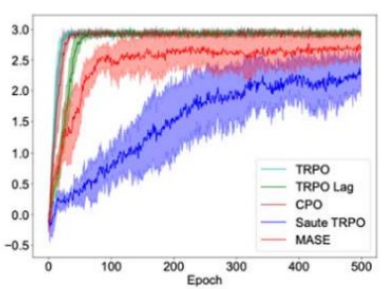
Generalized Safe Reinforcement Learning (RL)

[NeurIPS'23]

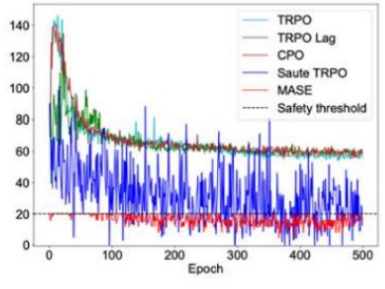
In safe RL research, safety formulations exist with insufficient discussion of their relationships, and few algorithms achieve both theory and application

- Formulated the "Generalized Safe Exploration Problem" that represents various safety constraints
- Proposed the MASE algorithm for solving the generalized safe exploration problem, with theoretical guarantees of obtaining optimal policies while ensuring safety
- **Unification of previously fragmented safe exploration theories**
- **Optimizes policies with high probability of no constraint violations even during training**





(a) Average episode return.



(c) Maximum episode safety.

Experimental evaluation: Comparison of optimality and safety performance between the proposed method (MASE) and existing methods. The proposed method improves safety performance without significantly degrading reward performance.

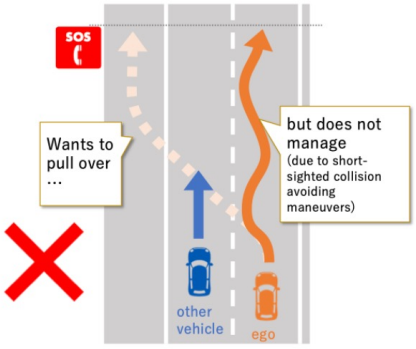
Safe Goal-Achieving Path Planning for Autonomous Vehicles [IV 2024, IEEE Trans. Intell. Veh.]

Responsibility-Sensitive Safety (RSS) widely used as safety spec for autonomous vehicles but only collision avoidance is addressed; goal achievement by path is not guaranteed

↓ Software science research

RSS
Responsibility-Sensitive Safety,
Shalev-Shwartz et al., 2017

- Basic methodology of logical safety rules
- Standardization (IEEE 2846)
- Lack of formal implemantion
  → appl. to complex scenarios is hard
- Guarantees only collision-freedom so far

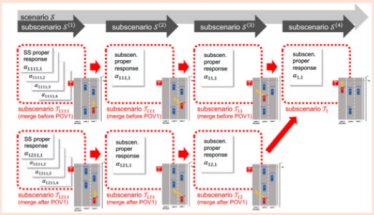


differential program logic dFHL (our contribution)

$$\frac{\begin{array}{ll}\text{inv}: & A \Rightarrow \epsilon_{inv} \sim 0 \quad \epsilon_{var} \geq 0 \wedge \epsilon_{inv} \sim 0 \Rightarrow \mathcal{L}_{\dot{x}=f}\, \epsilon_{inv} \simeq 0 \\ \text{var}: & A \Rightarrow \epsilon_{var} \geq 0 \quad \epsilon_{var} \geq 0 \wedge \epsilon_{inv} \sim 0 \Rightarrow \mathcal{L}_{\dot{x}=f}\, \epsilon_{var} \leq \epsilon_{ter} \\ \text{ter}: & A \Rightarrow \epsilon_{ter} < 0 \quad \epsilon_{var} \geq 0 \wedge \epsilon_{inv} \sim 0 \Rightarrow \mathcal{L}_{\dot{x}=f}\, \epsilon_{ter} \leq 0 \end{array}}{\{A\}\ \text{dwhile}(\epsilon_{var} > 0)\ \dot{x} = f\ \{\epsilon_{var} = 0 \wedge \epsilon_{inv} \sim 0\} : \epsilon_{inv} \sim 0 \wedge \epsilon_{var} \geq 0} \text{(DW}_\text{H}\text{)}$$

- A logical system for deriving and proving safety rules

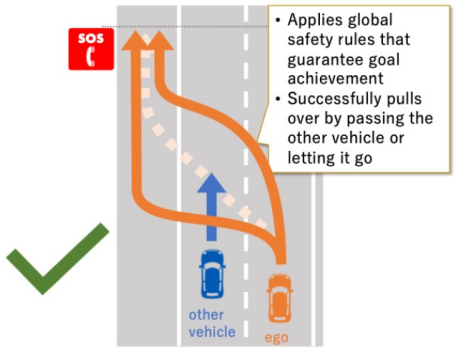
Compositional rule derivation workflow by dFHL (our contribution)



- "Divide and Conquer" complex driving scenarios
- Tool support by autom. reasoning

GA-RSS (our contribution)
Goal-Aware
Responsibility-Sensitive Safety

- Guarantees goal achievement (e.g. successful pull over) and collision-freedom
- Global safety rules that combine mult. maneuvers
- Necessary for real-world complex driving scenarios



- Proposed **GA-RSS** extending RSS to also guarantee goal achievement, with program-verification-based validation
- Validated in highway scenarios where autonomous vehicles safely stop at emergency phones

Conclusion

CyPhAI Project: Taming AI-CPS through Formal Methods

- Rigorous theories and ML-assisted methods for ensuring AI-CPS safety

Key Results

- **Black-Box Checking**: Extended to probabilistic, multi-component, and hybrid systems with spec-guided abstraction
- **Oblivious Monitoring**: Privacy-preserving runtime verification via fully homomorphic encryption [CAV'22]
- **STL2vec**: Scalable multi-task robot control via STL vector embedding; 1/24 memory [IEEE RA-L]
- **Generalized Safe RL**: Unified safe exploration framework (MASE) with theoretical guarantees [NeurIPS'23]
- **Safe Path Planning**: GA-RSS: safety + goal achievement for autonomous vehicles [IV'24, IEEE T-IV]

Takeaway

- Formal methods can tame AI-CPS, even with black-box AI components, bridging theory and real-world safety-critical applications