Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Doctoral Consortium Track

Recovering Formal Guarantees for Neural Network-Controlled Systems Under Changing Dynamics

Sterre Lutz
Delft University of Technology, the Netherlands
S.Lutz@tudelft.nl

Abstract

Neural networks enable high-performance decision-making, but their deployment in safety-critical settings requires formal guarantees that may become invalid after changes in the system dynamics. This research studies how such guarantees can be recovered using information from the initial verification, enabling post-deployment analysis to assess preservation, quantify degradation, and restore compliant behavior without recomputing verification from scratch.

1 Introduction

Neural network-controlled systems are increasingly deployed in safety-critical domains such as autonomous driving, healthcare, and robotics, where they enable decision-making in complex, high-dimensional environments. By leveraging techniques such as reinforcement learning, neural-network controllers can achieve strong performance in settings where classical control synthesis methods fail to scale.

However, deployment of neural-network controllers requires formal guarantees on safety and performance. Such guarantees are typically established with respect to a fixed model of the system and its environment. In practice, these assumptions are fragile: environmental variations, mechanical degradation, or system updates can introduce discrepancies between the verified model and the deployed system, potentially invalidating previously established guarantees.

Such discrepancies in dynamics raise the question: *how should we reason about safety and performance once a verified system changes after deployment?* Our goal is to determine whether guarantees still hold, quantify their degradation, and restore compliant behavior without restarting verification from scratch. In our projects, the initial verification result plays a central role.

2 Problem Setting

We consider stochastic dynamical systems of the form

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t), \quad \mathbf{u}_t := \pi(\mathbf{x}_t), \quad \mathbf{w}_t \sim \mu, \qquad (1)$$

where at time t, $\mathbf{x}_t \in \mathbb{X}$ is the state, $\mathbf{u}_t \in \mathbb{U}$ the action, and $\mathbf{w}_t \in \mathbb{W}$ a disturbance drawn from a known distribution μ. The dynamics function f in (1) is assumed to be Lipschitz continuous, but may be non-linear. The action at each timestep is chosen by a neural-network policy $\pi : \mathbb{X} \to \mathbb{U}$ based on the current state of the system.

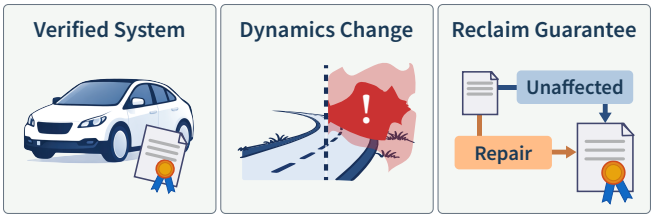

Figure 1: Guarantee recovery under changed dynamics: the original guarantee informs the assessment of behavior after changes.

Desirable behavior of such systems can be (partially) captured by the reach-avoid probability, i.e., the probability that a trajectory starting in $\mathbb{X}_0$ reaches a target set $\mathbb{X}_\star$ while avoiding an unsafe set $\mathbb{X}_\oslash$. A typical quantitative guarantee is

$$\mathbb{P}_{\mathbf{x}_0}\big(\text{reach } \mathbb{X}_\star \text{ before } \mathbb{X}_\oslash\big) \geq \rho, \quad \text{for all } \mathbf{x}_0 \in \mathbb{X}_0, \qquad (2)$$

where $\rho \in [0, 1]$. Such guarantees are established with respect to f, the system dynamics. After deployment, the system may change, resulting in modified dynamics f'. We develop theory and methods to answer the following questions:

- **Preservation:** Does the original guarantee still hold?
- **Degradation:** If not, does the guarantee hold for a relaxed threshold $\rho' < \rho$?
- **Repair:** Can we change policy π to restore compliance?

We are specifically interested in approaches that do *not* reduce these questions to new, isolated verification or synthesis problems, but instead exploit existing verification results.

3 Certificate-Based Perspective

To exploit verification results obtained for the original dynamics f, we consider certificate-based verification methods. These methods provide a guarantee by constructing a function $C : \mathbb{X} \to \mathbb{R}$ whose values satisfy constraints on relevant parts of the state space and along the system dynamics. The constraints are devised exactly so that such a function cannot exist unless the property of interest holds. Thus, any function C satisfying the constraints is a *certificate* of the property.

The certificate function itself may also be represented by a neural network, provided that sound verification of certificate constraints is performed [Barrett *et al.*, 2026].

To obtain quantitative reach-avoid guarantees as in (2), the certificate can be constrained so that its values at initial, target, and unsafe states, together with a one-step condition in expectation under f, imply a bound ρ on the probability of reaching $X_\star$ before $X_\oslash$; see, e.g., Žikelić *et al.* [2023] on neural reach-avoid certificates for stochastic systems.

The key advantage of certificates for our research objectives is that they expose *how* a global guarantee was obtained from local conditions. When the dynamics change, the certificate can therefore be reused as more than a yes/no proof artifact: it provides a state-dependent reference for checking which parts of the original argument remain valid and where recovery may be needed.

4 VeRecycle

Our algorithm `VeRecycle` [Lutz *et al.*, 2025] applies this perspective to a setting with partially changed dynamics, where deviations from the original dynamics f are confined to a subset of the state space denoted $\mathbb{X}_\Delta$. Formally, we consider dynamics of the form

$$f'(\mathbf{x}, \mathbf{u}, \mathbf{w}) = \begin{cases} \text{unknown} & \text{if } \mathbf{x} \in \mathbb{X}_\Delta \\ f(\mathbf{x}, \mathbf{u}, \mathbf{w}) & \text{otherwise.} \end{cases}$$

Dynamics may change in a specific region due to new obstacles or terrain variations, or model inaccuracies may appear only under certain operating conditions. In such cases, much of the state space remains unchanged, so it is not necessary to recompute guarantees globally.

We have shown that for changed dynamics f', the strongest post-change threshold that can be certified using the original certificate depends only on the original threshold and the lowest value of the original certificate on the changed region. In particular, if $\inf_{x \in \mathbb{X}_\Delta} C(x) \geq 1$, then the maximum threshold that can still be certified with C for any realization of the changed dynamics is

$$\rho' = \min\left(\rho,\, 1 - \frac{1}{\inf_{x \in \mathbb{X}_\Delta} C(x)}\right).$$

Otherwise, no reclaimable threshold exists. Thus, VeRecycle reduces the question of guarantee preservation under unknown localized change to estimating a single quantity: the infimum of the original certificate on the affected region.

In practice, VeRecycle computes sound lower and upper bounds on this infimum, for example using interval bound propagation, and therefore obtains a sound guaranteed threshold without retraining the certificate. This makes the method lightweight: in the original evaluation, VeRecycle achieved competitive guarantees while reducing computational cost by up to three orders of magnitude compared to re-certification, and in compositional settings these updated local thresholds can be used to recompose a globally compliant policy.

5 Related Work

In **robust control synthesis**, policies are designed to perform reliably under a predefined set of dynamics variations. This includes extensions of Markov Decision Processes (MDPs) to uncertain or multi-environment settings [Ou and Bi, 2025]. While these approaches can handle a range of modeled changes, they typically require prior knowledge of possible variations or repeated solution of modified models when changes occur. In contrast, this work focuses on reusing information from an existing verification result to assess and recover guarantees after unforeseen changes, without requiring a priori modeling of all possible dynamics variations.

In **transfer learning for control**, policies trained on a source system are adapted to a target system with modified dynamics or specifications. Typically, safety properties of the original system are not guaranteed to be preserved. One approach to address this is to transfer or re-learn certificate functions alongside the policy, for example by adapting barrier or Lyapunov functions to new dynamics [Nadali *et al.*, 2025]. We view this line of work as complementary, and instead investigate how existing certificates can be reused directly, *without* additional learning.

6 Evaluation and Outlook

We evaluate this research on control tasks with discrepancies between the verification model and the true system dynamics. Criteria are the quality of the obtained guarantees, the computational cost compared to treating it as an isolated verification question, and the success in restoring compliant behavior when guarantees are violated.

Current and future work aims to go beyond localized changes, including parametric variations that affect the entire state space. In addition, we aim to develop methods for adapting NN controllers based on certificate information and to establish more benchmark problems for evaluating post-change verification and repair. Finally, we investigate how the initial verification or certification procedure itself can be designed to better support post-deployment analysis, by guiding it toward certificates that are more informative for assessing preservation, degradation, and repair.

References

[Barrett *et al.*, 2026] Clark Barrett, Thomas A Henzinger, and Sanjit A Seshia. Certificates in ai: Learn but verify. *Communications of the ACM*, 69(1):66–75, 2026.

[Lutz *et al.*, 2025] Sterre Lutz, Matthijs T. J. Spaan, and Anna Lukina. VeRecycle: Reclaiming guarantees from probabilistic certificates for stochastic dynamical systems after change. In *IJCAI*, pages 457–465, 2025.

[Nadali *et al.*, 2025] Alireza Nadali, Ashutosh Trivedi, and Majid Zamani. Stochastic neural simulation relations for control transfer. In *NeuS*, pages 597–620, 2025.

[Ou and Bi, 2025] Wenfan Ou and Sheng Bi. Sequential decision-making under uncertainty: A robust mdps review. *Ann. Oper. Res.*, 353(3):1239–1285, 2025.

[Žikelić *et al.*, 2023] Đorđe Žikelić, Mathias Lechner, Thomas A. Henzinger, and Krishnendu Chatterjee. Learning control policies for stochastic systems with reach-avoid guarantees. In *AAAI*, pages 11926–11935, 2023.