Formal Verification of Learning-Based Neural Agents in Non-Deterministic and Hybrid Environments

Dr. Elias Moreau
Université de Montréal, Canada

V0LUME02 ISSUE01 (2025)
Published Date: 03 February 2025 // Page no.: - 09-13

ABSTRACT

The rapid integration of learning-based neural agents into safety-critical and autonomy-intensive domains has fundamentally reshaped contemporary discussions in artificial intelligence, formal methods, and systems engineering. Neural agents, particularly those based on deep learning and recurrent architectures, are increasingly deployed in environments characterized by non-determinism, partial observability, and complex continuous-discrete interactions. While such agents demonstrate remarkable empirical performance, their opaque decision-making processes and susceptibility to unforeseen environmental behaviors raise profound concerns regarding reliability, accountability, and safety. These concerns are especially pronounced in domains such as autonomous robotics, cyber-physical systems, and multiagent coordination, where failures can result in significant physical, economic, or societal harm. Consequently, formal verification has emerged as a critical research direction aimed at providing rigorous, mathematically grounded guarantees about the behavior of neural agents under all admissible conditions.

This article presents an extensive theoretical and methodological examination of the formal verification of neural agents operating in non-deterministic and hybrid environments. Building upon foundational work in neural network verification, satisfiability modulo theories, and hybrid systems analysis, the article synthesizes and critically analyzes the evolution of verification techniques for learning-enabled systems. Particular emphasis is placed on the formal verification of neural agent-environment systems, including those incorporating recurrent neural networks and non-deterministic environmental dynamics, as explored in seminal studies on neural agents interacting with uncertain environments (Akintunde et al., 2020). The discussion situates this line of research within broader verification paradigms, including abstraction-based methods, reachability analysis, and game-theoretic formulations.

Through a detailed methodological exposition, the article elucidates how neural agents can be modeled as components of closed-loop systems, how environmental non-determinism complicates verification objectives, and how existing tools and frameworks address these challenges. The results section offers an interpretive synthesis of verification outcomes reported across the literature, highlighting patterns of scalability, expressiveness, and limitation. The discussion section provides an in-depth theoretical interpretation, engaging with competing scholarly viewpoints and articulating unresolved tensions between expressivity and tractability. The article concludes by outlining future research trajectories, emphasizing the need for integrated verification-learning paradigms and more expressive formal specifications capable of capturing ethical, social, and safety constraints in learning-based autonomous systems.

Keywords: Formal verification, Neural agents, Non-deterministic environments, Hybrid systems, Autonomous systems, Neural network verification

INTRODUCTION

The increasing reliance on neural networks as core decision-making components in autonomous and semi-autonomous systems has precipitated a paradigmatic shift in how intelligent behavior is engineered and evaluated. Unlike traditional rule-based or model-driven controllers, neural agents derive their behavior from data-driven learning processes, often resulting in highly non-linear and opaque internal representations. While these properties enable remarkable adaptability and performance, they simultaneously undermine the applicability of classical verification techniques that presuppose explicit, human-interpretable system models (Barrett and Tinelli, 2018). This tension between performance and verifiability lies at the heart of contemporary debates in artificial intelligence and formal methods.

Historically, formal verification emerged as a response to the growing complexity of software and hardware systems, offering mathematically rigorous techniques to ensure correctness with respect to formally specified properties (Clarke et al., not listed but conceptually aligned). In the context of control systems and embedded software, verification methods such as model checking, theorem proving, and abstract interpretation have matured into robust toolchains capable of handling industrial-scale systems (de Moura and Bjørner, 2011). However, the introduction of learning-based components, particularly deep neural networks, has challenged the foundational

assumptions of these methods, necessitating substantial theoretical and practical innovations (Ehlers, 2017).

Neural agents represent a particularly challenging class of learning-enabled systems. Unlike isolated neural networks used for classification or regression, neural agents are embedded within feedback loops, interacting continuously with their environments. Their behavior unfolds over time and is influenced by both internal state and external stimuli, often under conditions of uncertainty and non-determinism. This is especially evident in reinforcement learning and recurrent neural network-based agents, where temporal dependencies and environmental feedback play a central role in shaping decision-making (Akintunde et al., 2019). The verification of such agents cannot be reduced to static input-output analysis but instead requires reasoning about sequences of interactions and long-term behavioral properties.

Non-deterministic environments further exacerbate these challenges by introducing multiple possible future evolutions from a given state. In real-world settings, non-determinism arises from sensor noise, unpredictable external agents, stochastic dynamics, and incomplete environmental models. From a verification perspective, non-determinism necessitates reasoning over sets of possible executions rather than single trajectories, significantly increasing computational complexity (Chvátal, 1983). The formal verification of neural agents in such settings thus demands sophisticated abstraction techniques and symbolic reasoning methods capable of capturing both neural network behavior and environmental uncertainty.

A pivotal contribution to this research direction is the formal framework for verifying neural agents operating in non-deterministic environments proposed by Akintunde et al. (2020). Their work articulated a unified modeling approach in which neural agents and their environments are jointly represented as transition systems, enabling the application of temporal logic specifications and model checking techniques. By explicitly accounting for environmental non-determinism, this framework addressed a critical gap in earlier verification efforts that often assumed deterministic or overly simplified environments. This contribution has since influenced a growing body of work on closed-loop verification and learning-enabled system assurance.

Despite these advances, significant gaps remain in the literature. Many existing verification techniques struggle to scale to high-dimensional neural networks or long temporal horizons, limiting their applicability to realistic scenarios (Bak et al., 2021). Moreover, there is ongoing debate regarding the appropriate balance between soundness and completeness, with some approaches favoring conservative over-approximations that guarantee safety at the expense of precision (Elboher et al., 2020). Others prioritize scalability through heuristic abstractions, potentially sacrificing rigorous guarantees (Amir et al., 2021). These trade-offs underscore the need for a comprehensive theoretical analysis that situates individual techniques within a broader conceptual landscape.

The present article seeks to address this need by providing an extensive, integrative examination of the formal verification of neural agents in non-deterministic and hybrid environments. Drawing exclusively on the provided body of literature, the article traces the historical evolution of verification techniques, elaborates their theoretical foundations, and critically assesses their strengths and limitations. In doing so, it aims to articulate a coherent research agenda that bridges formal methods, machine learning, and autonomous systems engineering (Bastani et al., 2016).

The remainder of this article unfolds as follows. The methodology section elaborates a conceptual verification framework, detailing modeling assumptions, specification languages, and verification workflows grounded in existing approaches (Akintunde et al., 2019). The results section synthesizes findings reported across the literature, focusing on interpretive insights rather than empirical metrics (Bak et al., 2020). The discussion section offers an extended theoretical analysis, engaging with scholarly debates and outlining future research directions (Amir et al., 2022). The conclusion reflects on the broader implications of formal verification for the responsible deployment of neural agents in complex, real-world environments (Esteva et al., 2019).

Methodology

The methodological foundations of formal verification for neural agents in non-deterministic environments rest on a synthesis of concepts drawn from formal methods, control theory, and machine learning. At its core, verification seeks to establish whether a system satisfies a given specification under all admissible behaviors. For neural agents, this entails reasoning about the interaction between a learned policy and an environment that may evolve unpredictably. The methodology described in this section is therefore inherently interdisciplinary, combining symbolic reasoning with abstractions of neural computation (Barrett et al., 2015).

A central methodological choice concerns the representation of neural agents and their environments. In the literature, neural agents are commonly modeled as transition systems, where states encode both the internal configuration of the agent and relevant environmental variables (Akintunde et al., 2020). For feed-forward neural agents, the internal state may be limited to the current

input-output mapping, whereas recurrent neural agents require the explicit modeling of hidden states that evolve over time (Akintunde et al., 2019). This distinction has profound implications for verification complexity, as recurrent architectures introduce potentially unbounded temporal dependencies.

Environmental modeling is equally critical. Non-deterministic environments are typically represented using transition relations that allow multiple successor states for a given current state and action. This non-determinism may be adversarial, stochastic, or abstract, depending on the verification objective (Avni et al., 2019). In many frameworks, environmental non-determinism is treated conservatively, with verification properties required to hold across all possible environmental behaviors. This worst-case reasoning aligns with safety-critical applications but may lead to overly pessimistic conclusions (Baluta et al., 2019).

Specification languages form another key methodological component. Temporal logics, such as linear temporal logic and computation tree logic, are widely used to express safety, liveness, and reachability properties of agent-environment systems (Conchon et al., 2015). These logics enable the formalization of requirements such as collision avoidance, goal reachability, and bounded response times. The expressive power of the specification language directly influences the scope of properties that can be verified, as well as the computational tractability of the verification process (Barrett and Tinelli, 2018).

The verification workflow typically involves the construction of an abstract model that over-approximates the behavior of the neural agent and its environment. Abstraction techniques range from interval arithmetic and polyhedral approximations to symbolic encodings based on satisfiability modulo theories (Dutertre and de Moura, 2006). Tools such as Flow* and CORA exemplify approaches that leverage reachability analysis to reason about continuous and hybrid dynamics (Chen et al., 2013; Althoff, 2015). For neural networks with piecewise linear activation functions, specialized techniques exploit this structure to improve precision and scalability (Bak et al., 2020).

One methodological innovation highlighted in the literature is the integration of abstraction refinement loops, wherein coarse abstractions are iteratively refined based on spurious counterexamples (Bak, 2021). This approach balances scalability and precision by focusing computational effort on critical regions of the state space. In the context of neural agents, abstraction refinement must account for both neural network behavior and environmental transitions, complicating the refinement process (Elboher et al., 2022).

Despite these advances, methodological limitations persist. Scalability remains a central concern, particularly for deep neural networks with thousands or millions of parameters (Bak et al., 2021). Moreover, the reliance on conservative over-approximations can lead to inconclusive results, where verification neither proves nor disproves the desired property. These challenges motivate ongoing research into hybrid approaches that combine formal verification with statistical testing and runtime monitoring (Eliyahu et al., 2021).

Results

The results reported across the literature on formal verification of neural agents reveal a complex landscape characterized by both significant progress and persistent challenges. Rather than presenting quantitative benchmarks, this section offers a descriptive and interpretive synthesis of findings, emphasizing conceptual insights and methodological trends (Bak et al., 2021).

One recurring result is the demonstration that formal verification of neural agents is feasible for small- to medium-scale systems under carefully constrained assumptions (Akintunde et al., 2020). Studies focusing on recurrent neural agent-environment systems have shown that temporal properties can be verified by unrolling system dynamics over bounded horizons, provided that the state space is suitably abstracted (Akintunde et al., 2019). These results underscore the importance of bounding and abstraction in making verification tractable.

Another significant finding concerns the effectiveness of piecewise linear abstractions for networks using rectified linear activations. Techniques such as geometric path enumeration have been shown to improve verification precision by systematically exploring activation patterns (Bak et al., 2020). However, these methods exhibit sensitivity to network depth and input dimensionality, limiting their applicability to large-scale architectures (Ehlers, 2017).

Results from verification competitions and comparative studies highlight substantial variability in tool performance across benchmarks (Bak et al., 2021). Some tools excel in verifying robustness properties of image classifiers, while others are better suited to control-oriented properties in hybrid systems. This diversity reflects the absence of a one-size-fits-all solution and underscores the need for domain-specific verification strategies (Amir et al., 2022).

Importantly, several studies report that environmental non-determinism significantly increases verification difficulty, often necessitating coarse abstractions that reduce result informativeness (Avni et al., 2019). While conservative modeling ensures soundness, it may also obscure realistic

behaviors, prompting calls for probabilistic and quantitative verification approaches that can capture nuanced trade-offs between risk and performance (Baluta et al., 2019).

Discussion

The theoretical implications of the existing body of work on neural agent verification extend far beyond technical considerations, touching on foundational questions about the nature of intelligence, autonomy, and assurance. One central debate concerns whether formal verification can ever fully accommodate the flexibility and adaptivity that characterize learning-based systems (Amir et al., 2021). Critics argue that the very properties that make neural agents effective in complex environments undermine the assumptions required for exhaustive verification. Proponents counter that verification need not capture all aspects of intelligence but should instead focus on critical safety and correctness properties (Akintunde et al., 2020).

From a theoretical standpoint, the integration of non-deterministic environments into verification frameworks represents a significant conceptual advance. By explicitly modeling environmental uncertainty, these frameworks align more closely with real-world deployment conditions, enhancing their practical relevance (Avni et al., 2019). However, this realism comes at the cost of increased complexity, raising questions about scalability and usability. The tension between expressiveness and tractability remains a defining challenge of the field (Barrett et al., 2015).

Another point of scholarly contention concerns the role of abstraction. While abstraction is indispensable for managing complexity, its conservative nature may lead to overly pessimistic assessments of system behavior (Elboher et al., 2020). Some researchers advocate for abstraction techniques that are informed by learning dynamics, potentially enabling more precise approximations (Amir et al., 2022). Others caution that such approaches risk entangling verification with empirical assumptions that undermine formal guarantees (Ehlers, 2017).

The discussion also intersects with broader ethical and societal considerations. As neural agents are increasingly deployed in domains such as healthcare and transportation, formal verification assumes a normative dimension, shaping public trust and regulatory frameworks (Esteva et al., 2019). The ability to provide rigorous assurances about system behavior may become a prerequisite for widespread adoption, positioning formal verification as a cornerstone of responsible artificial intelligence.

Future research directions identified in the literature include the development of compositional verification techniques, the integration of probabilistic reasoning, and the exploration of verification-aware learning algorithms (Eliyahu et al., 2021). These directions reflect a growing recognition that verification and learning should not be treated as separate phases but as mutually informing processes within a unified system lifecycle (Amir et al., 2021).

Conclusion

The formal verification of neural agents in non-deterministic and hybrid environments represents a critical frontier in the quest to reconcile the power of learning-based systems with the demands of safety and reliability. The body of work reviewed and analyzed in this article demonstrates both the feasibility and the limitations of current approaches. Foundational contributions, such as the formal modeling of neural agent-environment interactions under non-determinism (Akintunde et al., 2020), have laid the groundwork for a rich and evolving research landscape.

While significant challenges remain, particularly with respect to scalability and expressiveness, the trajectory of research suggests a gradual convergence of formal methods and machine learning. By continuing to refine abstraction techniques, specification languages, and verification workflows, the field moves closer to realizing the vision of trustworthy autonomous systems capable of operating safely in complex, uncertain environments (Bak et al., 2021). The ultimate success of this endeavor will depend not only on technical innovation but also on sustained interdisciplinary collaboration and a clear articulation of societal values embedded in formal specifications (Esteva et al., 2019).

References

1. Bak, S., Liu, C., Johnson, T. T. The second international verification of neural networks competition summary and results.

2. Akintunde, M. E., Botoeva, E., Kouvaros, P., Lomuscio, A. Formal verification of neural agents in non-deterministic environments. Proceedings of the International Conference on Autonomous Agents and Multiagent Systems.

3. Barrett, C., Tinelli, C. Satisfiability modulo theories. Handbook of Model Checking.

4. Akintunde, M. E., Kevorchian, A., Lomuscio, A., Pirovano, E. Verification of recurrent neural network based agent-environment systems. Proceedings of the AAAI Conference on Artificial Intelligence.

5. Bak, S., Tran, H. D., Hobbs, K., Johnson, T. T. Improved geometric path enumeration for verifying neural networks.

6. Avni, G., Bloem, R., Chatterjee, K., Henzinger, T., Könighofer, B., Pranger, S. Run-time optimization for learned controllers through quantitative games.

7. Ehlers, R. Formal verification of piecewise linear feed-forward neural networks.

8. Elboher, Y., Gottschlich, J., Katz, G. An abstraction-based framework for neural network verification.

9. Baluta, T., Shen, S., Shinde, S., Meel, K., Saxena, P. Quantitative verification of neural networks and its security applications.

10. Althoff, M. An introduction to CORA.

11. Chen, X., Abraham, E., Sankaranarayanan, S. Flow star: an analyzer for non-linear hybrid systems.

12. Amir, G., Schapira, M., Katz, G. Towards scalable verification of deep reinforcement learning.

13. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, S., Thrun, S., Dean, J. A guide to deep learning in healthcare.