Journal Title
XX(X):2–43

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Temporal Neuro-Symbolic Reasoning: from architectures to verifiable and auditable systems

Aurélien Vannieuwenhuyze[1], Nada Mimouni[1] and Cedric Du Mouza[1]

Abstract

Temporal reasoning is a central challenge in artificial intelligence when decisions depend on causal relations, ordering constraints, or evolving dynamics between events. Temporal neuro-symbolic approaches have been proposed to combine learning capabilities with formal guarantees in this context. This survey analyzes these works by proposing a unified structuring of integration paradigms and by highlighting the fundamental trade-offs between logical expressiveness, differentiability, and scalability. Beyond a state-of-the-art review, we argue that the current limitations of temporal neuro-symbolic reasoning are primarily structural. The non-locality of time, the irreversibility of constraints, and the dependence on complete history raise difficulties that cannot be resolved through computational optimizations. We also show that dominant evaluation practices favor point-wise performance and do not assess the global consistency of trajectories. Finally, this survey offers a forward-looking perspective structured around a roadmap organized into maturity levels, linking architectures, explainability, and evaluation protocols, and aiming to support the development of models that are verifiable and auditable by design, as well as evaluation and governance methods suited to reliable reasoning.

Keywords

Neuro-Symbolic AI, Temporal Reasoning, Temporal Logic, Neuro-Symbolic Integration, Explainable AI, Trustworthy AI

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

Introduction

Reasoning about temporal phenomena cannot be reduced to producing instantaneous predictions, but requires guaranteeing the validity of a decision over an entire evolving trajectory. In many application contexts, producing a correct prediction at a given time is insufficient to ensure the global validity of reasoning. Decisions must be consistent with past event trajectories, respect explicit constraints, and remain stable in the face of evolving contexts. The central challenge is therefore not only to learn from temporal data, but to guarantee the coherence, verifiability, and controllability of reasoning.

Deep learning approaches have demonstrated remarkable effectiveness in exploiting large-scale sequential data. Recurrent architectures and attention mechanisms capture complex dependencies, but rely primarily on learned statistical correlations. This inductive and often opaque nature limits their ability to explicitly guarantee compliance with formal constraints, such as event ordering, causality, or critical delays.

Conversely, symbolic approaches based on temporal logics provide precise tools to specify and verify time-related properties. However, their expressiveness comes with significant computational limitations and difficulties in handling real-world, continuous, or noisy data. When the temporal horizon extends, these approaches impose structural trade-offs between reasoning richness, operational feasibility, and generalization capacity.

Temporal neuro-symbolic reasoning emerges in this context as an attempt to overcome these limitations by articulating neural learning with explicit symbolic knowledge. The challenge is no longer only to improve predictive performance, but to understand how temporal and causal relations can be represented, manipulated, and possibly extracted as rules or symbolic structures from learned models. The goal is to design systems capable of reasoning in a controllable, verifiable, and inspectable manner, by making explicit the inference trajectories and dependencies that underlie decisions.

This survey analyzes works dedicated to learning, representing, and reasoning about temporal and causal relations in neuro-symbolic approaches. It highlights structural and methodological obstacles that go beyond local algorithmic choices, particularly regarding the definition, extraction, and evaluation of explicit reasoning structures. Beyond a descriptive analysis, this work aims to clarify the criteria by which reasoning can be interpreted, controlled, and compared, and to provide guidance elements for the design and evaluation of temporal neuro-symbolic systems. On this basis, we propose a roadmap

[1]Conservatoire national des arts et métiers, Paris (France)

Corresponding author:
Aurélien Vannieuwenhuyze Conservatoire national des arts et métiers, Laboratoire CEDRIC, Equipe ISID, 292 Rue Saint-Martin, 75003 Paris
Email: aurelien.vannieuwenhuyze@lecnam.net, nada.mimouni@lecnam.net, cedric.dumouza@lecnam.net

Prepared using sagej.cls

structured into successive maturity levels, intended to guide the evolution of the field toward systems that are auditable, explainable, and governable by design.

Methodology

The methodological protocol implemented for this survey follows the PRISMA 2020 guidelines to ensure transparency and traceability of the publication selection process, while supporting a critical conceptual analysis (Page et al. 2021).

The study is structured around a central research question concerning how neuro-symbolic approaches learn, represent, and make explicit forms of temporal and causal reasoning. This question is divided into five thematic axes that structure the survey: (i) theoretical and computational formalisms of temporal reasoning, (ii) integration modalities between neural learning and temporal logic, (iii) interpretability mechanisms and extraction of temporal structures or rules, (iv) evaluation practices for temporal reasoning, and (v) open challenges and research perspectives. The literature search covers the period 1980–2025, to integrate both foundational works and contemporary approaches, while acknowledging the associated conceptual heterogeneity.

Publications were identified from complementary sources, including arXiv, Semantic Scholar, OpenAlex, and Crossref, with the first three providing the majority of relevant contributions. Queries were constructed from thematic axes and key concepts, with filtering aimed at improving the signal-to-noise ratio, at the acknowledged risk of excluding some works that are atypical in terminology.

The corpus thus constituted comprises 2,576 publications, supplemented by 47 works integrated in a reasoned manner for their structural role. After applying the PRISMA process, as well as deduplication and selection steps, 73 articles were retained for the final analysis.

The selection relies on a weighted multidimensional evaluation of the relevance of contributions, formalized by a global score on a scale from 0 to 2. This score aggregates several complementary dimensions, including the centrality of temporal reasoning as an explicit object of inference (beyond simple sequence processing), the methodological contribution to temporal neuro-symbolic integration, representation or evaluation, as well as the structural role or influence of the work in the literature. Each dimension is evaluated individually then combined to obtain a composite score. Publications with a score strictly below a threshold of 1.75 were excluded during the screening phase, as shown in the PRISMA diagram (Figure 1).

These dimensions were not treated as independent exclusion criteria, but jointly informed a relevance judgment based on expertise. As is frequently the case in integrative and conceptual reviews, this process does not aim for strict reproducibility of selection decisions, but rather transparency of the criteria guiding inclusion.

Language assistance tools based on language models were used exclusively for exploratory and editorial purposes, particularly for translation from French to English and linguistic revision, without intervention in the selection, scientific evaluation, or interpretation of results. All scientific content, analyses, and arguments remain under the exclusive responsibility of the authors. The comparative analysis aims to identify

cross-cutting regularities according to representation modes, reasoning forms, and explainability properties, rather than establishing an exhaustive ranking.

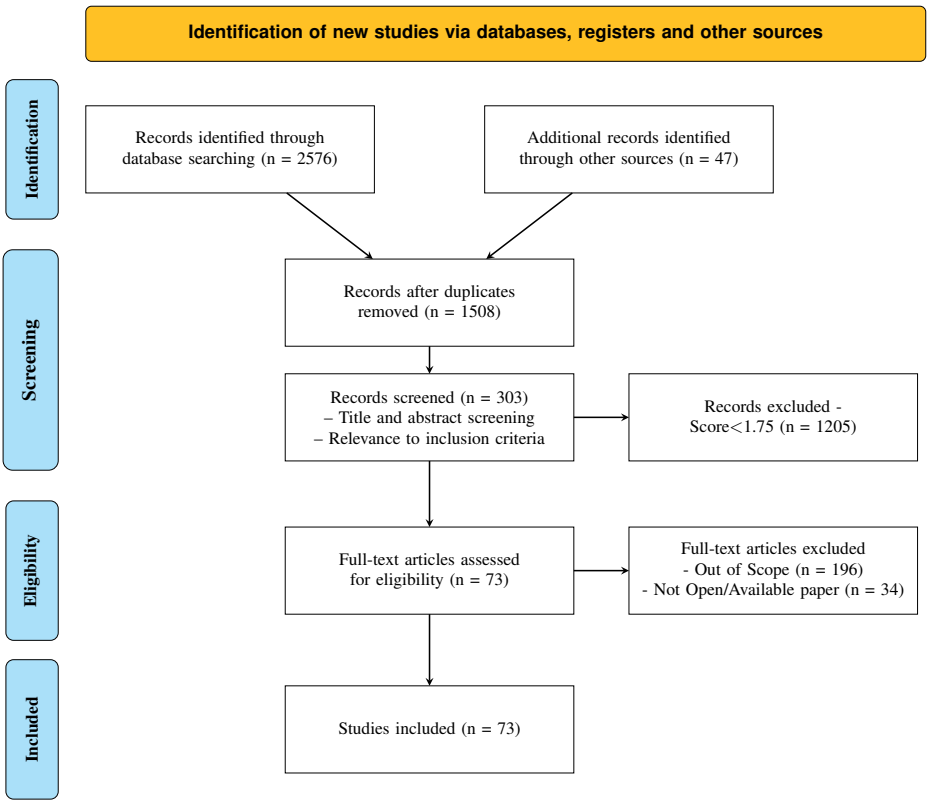


Figure 1. PRISMA 2020 flow diagram of the literature selection process

The article is structured as follows. Section 1 presents the formal and computational foundations of temporal reasoning, introducing the main logical formalisms, constraint models, and symbolic representations used in the literature. Section 2 then analyzes the fusion of neural models and temporal logic, distinguishing the modalities by which neural and symbolic components interact within the inference process. Section 3 is devoted to interpretability and the extraction of temporal rules, examining the extent to which these approaches make the reasoning produced explicit and inspectable. Section 4 discusses the evaluation of temporal reasoning, including benchmark datasets and the criteria used to assess the coherence, validity, and robustness of decisions. Finally, Section 5 offers a synthesis and roadmap, identifying current limitations, key open challenges, and research perspectives for the development of controllable, explainable, and governable temporal neuro-symbolic systems.

Foundations of temporal reasoning in A.I.

Temporal reasoning constitutes a central element in the modeling of evolving systems in artificial intelligence. Temporal formalisms provide the formal tools to represent the order, duration, and synchronization of events, whether these are conceptualized as instantaneous occurrences or as states maintained over a determined temporal interval. These formalisms play a particularly important role at the interface between continuous neural perception and discrete symbolic reasoning. They establish a correspondence between observations from continuous flows and temporally ordered symbolic representations. This mediation capacity proves decisive for developing systems that integrate learning and formal reasoning.

This section examines the main logical formalisms, representation models, and inference mechanisms that underlie temporal reasoning. We highlight the current limitations of these approaches, limitations that constitute an essential motivation for the development of neuro-symbolic architectures.

Temporal logic formalisms

Linear temporal logics constitute a major formal framework for reasoning about trajectories of ordered events, typically interpreted as system executions observed over a given temporal horizon (Clarke et al. 2001). Linear Temporal Logic (LTL) represents system behavior through discrete traces, considered as ordered sequences of states describing its temporal evolution. Logical formulas express global properties over an entire trajectory, such as exclusion of particular configurations or guaranteed occurrence of expected events.

These properties classically correspond to the notions of safety and liveness. As an illustration, consider an industrial control system in which an operator transmits a request to an automaton to trigger a critical action, for example stopping a machine or opening a valve. A liveness requirement then stipulates that any request actually received must be followed, at a later time, by a confirmation signal attesting that the action has been taken into account. This constraint is expressed in LTL by the formula $G(req \rightarrow F\ ack)$, which requires that any occurrence of *req* be followed, over all admissible trajectories, by a future state in which *ack* holds (Clarke et al. 2001). Such specifications enable automatic verification of compliance with ordering and temporal dependency constraints, independently of the mechanism that produced these trajectories.

The introduction of past operators extends the expressiveness of this formalism by allowing the formulation of properties dependent on history. It enables the characterization of situations where the validity of a current state explicitly depends on prior events or conditions, without altering the linear semantics of the model (Lichtenstein et al. 1985).

When discrete modeling proves inadequate, particularly for representing continuous physical dynamics, Signal Temporal Logic (STL) generalizes this approach to dense-time, real-valued signals whose evolution cannot be correctly captured by a simple succession of discrete instants. Its main contribution lies in introducing robustness semantics, which associates with each property a quantitative measure of its degree of

satisfaction. In a control system where a physical variable such as distance, velocity, or temperature must remain above a safety threshold, robustness not only identifies a potential constraint violation but also quantifies the observed deviation or, conversely, the available safety margin. This quantitative interpretation enables direct integration of temporal constraints into differentiable optimization processes, by providing a continuous signal that can be exploited during learning or adjustment of neural models (Leung et al. 2021).

Interval logics adopt a complementary perspective by taking temporal segments rather than isolated instants as primitives, in order to directly represent durative phenomena whose duration and temporal relations constitute essential aspects. Allen's algebra defines a set of qualitative relations to characterize temporal interactions between intervals, such as precedence, coincidence, or exact succession. However, direct integration of these relations into fully expressive formalisms, such as Halpern–Shoham logic, leads to undecidability results in many configurations (Halpern and Shoham 1991). To preserve decidability of reasoning, trade-offs have been proposed, notably by encoding Allen's relations in linear frameworks whose complexity remains comparable to that of LTL, such as Allen's Linear Logic (Roşu and Bensalem 2006).

Non-classical and hybrid logics

Several extensions of temporal logics have been developed to handle situations that escape the assumptions of classical linear approaches, particularly when global coherence, uniqueness of the temporal sequence, or strict time discretization cannot be guaranteed. These extensions have emerged in distinct application contexts and address specific needs, such as explicit structuring of temporal information, inconsistency management, or modeling of continuous and hybrid dynamics (Hou 2012; Maria et al. 2014; Kamide 2015, 2017).

Multi-sequence temporal logics, such as SLTL, introduce operators that explicitly organize multiple information sequences associated with the same temporal trace, by allowing different levels or reading orders of a given temporal evolution (Kamide 2017). This structuring enables reasoning about ordered or hierarchical information as represented in the model, without imposing uniqueness of the observation sequence.

Paraconsistent extensions, such as PSLTL, aim to control the impact of inconsistency by preventing logical explosion, particularly in situations where distinct sources provide contradictory information regarding the occurrence or temporality of the same event (Kamide 2015). This type of situation arises, for example, in multi-source critical monitoring systems, where heterogeneous sensors may simultaneously produce incompatible observations about the state of the same process, such as indication of physiological arrest by a biomedical signal while other sensors suggest persistent activity. Unlike classical logics, which invalidate the entire reasoning in the presence of such contradiction, paraconsistent approaches isolate these local inconsistencies while preserving the reasoning capacity on coherent parts of the temporal trace. They thus allow intermediate decisions based on identification of the conflict itself, without invalidating all temporal inferences.

Other formalisms rely on introducing strong structural constraints intended to control the complexity of temporal reasoning. TLX logic thus imposes exclusivity relations between propositions, modeled by XOR sets, notably to represent mutually exclusive operating modes, such as *active* and *inactive* states of a component, while guaranteeing decidability and manageable algorithmic complexity for this fragment (Dixon 2007).

Certain extensions have been proposed to account for continuous or hybrid dynamics, inadequately represented in purely discrete approaches. HyLL logic relies on indexing judgments by worlds representing time or continuous resources, which allows conditioning the validity of a transition on the effective availability of these resources (Maria et al. 2014). This type of formalization proves particularly suited to situations where action execution depends simultaneously on continuous evolution and discrete constraints, for example when a mobile robot is authorized to continue moving only as long as its battery level remains above a minimal threshold, and must interrupt or modify its behavior as soon as this resource becomes insufficient.

In a verification-oriented perspective, QdTL is defined on hybrid trace semantics (Hou 2012), allowing reasoning about cyber-physical systems combining continuous dynamics and discrete transitions. It notably enables expressing and verifying that safety constraints on continuous variables, such as bounds on physical quantities, are satisfied for all possible states and trajectories of a hybrid system, despite alternation between continuous evolution and discrete updates.

Temporal representation models

Alongside logical formalisms, specific representation structures have been developed to organize temporal information in a manner exploitable by reasoning and analysis mechanisms. These structures constitute intermediate supports between raw data and inference processes. Notable distinctions include annotation standards such as TimeML (Pustejovsky et al. 2005) and temporal constraint networks such as STNs (Dechter et al. 1990). These approaches rely on distinct modeling choices, adapted to specific types of data, dynamics, or constraints, without aiming to establish a unified theoretical framework.

Certain formalisms introduce explicit organization of information according to distinct sequences or contexts, allowing reasoning about ordered temporal descriptions and distinguishing multiple temporal viewpoints associated with the same observed evolution (Kamide 2017). This structuring facilitates exploitation of temporal information during reasoning, according to their order or context of introduction.

Other approaches rely on an event-based modeling of time, where occurrences and durations of phenomena constitute the primitives of reasoning, such as appearance, repetition, or cessation of events during temporal observation. This perspective naturally leads to considering events whose temporal presence may be discontinuous or fragmented. To overcome the limitations of classical convex intervals, extensions such as n-intervals have been proposed to represent non-convex events, particularly discontinuous or recurring ones (Shoaff 1993).

This type of representation proves particularly relevant in situations where an activity manifests through brief and repeated occurrences rather than continuous presence, for example when medical treatment is administered at distinct times during the same hospitalization period. Unlike a single convex interval, n-intervals allow precise reasoning about these fragmented occurrences without introducing false temporal continuities. Figure 2 illustrates the difference between a convex interval in Allen's algebra sense and a non-convex event represented as a union of disjoint segments. These models thus offer an expressive framework for describing fragmented phenomena while preserving consistency mechanisms inherited from interval networks.

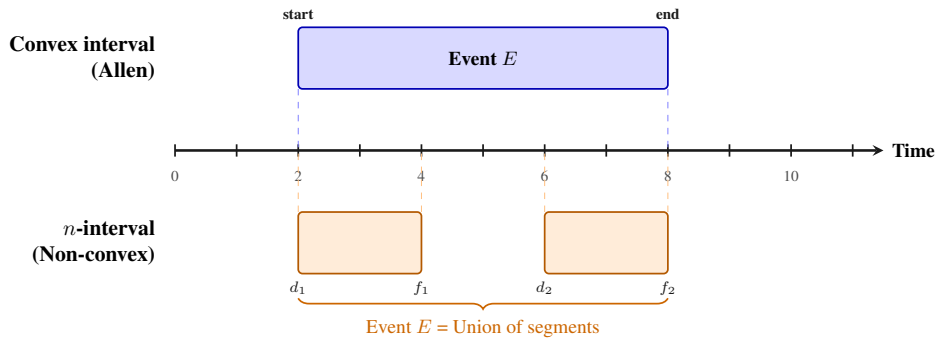


Figure 2. Comparison between a convex interval in the sense of Allen (top) and a non-convex n-interval (bottom). In the latter case, the event E is temporally discontinuous, and is defined as the union of disjoint segments ($[d_1, f_1] \cup [d_2, f_2]$).

The representation of durative actions in distributed environments has led to the introduction of finer temporal structures in certain multi-agent modeling approaches. These approaches describe activities as sequences of explicitly synchronizable temporal phases, allowing reasoning about interactions between distinct entities. In a cooperative multi-agent surveillance or maintenance system, an intervention may require one agent to collect data while another executes corrective action, with these different phases needing to respect precise ordering or overlap constraints. Such representations enable analysis of temporal dependencies between agents and verification of global coherence of their coordinated behaviors (Kummari et al. 2025).

Hybrid system modeling aims to articulate, within a single framework, continuous dynamics and discrete transitions, such as continuous evolution of physical variables and state changes triggered by events. Associated logical formalisms explicitly define this type of system and allow expressing properties over all their evolutions (Hou 2012; Maria et al. 2014). Figure 3 illustrates a typical hybrid trace, in which continuous evolution phases alternate with discrete jumps. In this framework, logics such as QdTL enable formulating properties valid over all possible hybrid trajectories of a system (Hou 2012).

Temporal graphs and databases provide a relational support for organizing and exploiting these representations at larger scale. Annotation schemes such as TimeML enable transforming narrative or event data into partially ordered event graphs, when

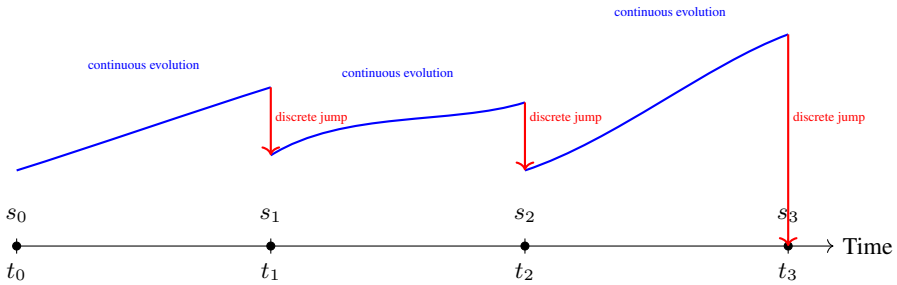


Figure 3. Example of a hybrid trace for an autonomous vehicle: phases of continuous evolution correspond to the physical dynamics of the vehicle (motion governed by differential equations), while discrete jumps model instantaneous commands such as *turn left*,*brake*, or *change lane*. This type of alternation between continuous dynamics and discrete updates is characteristic of hybrid systems modeled and analyzed by logics such as QdTL.

precedence or simultaneity relations must be made explicit (Figure 4) (Pustejovsky et al. 2005).

Temporal constraint networks, notably STNs and TCSPs, offer a structured means to ensure metric coherence of temporal graphs, by explicitly representing duration or delay constraints between events, at the cost of increased algorithmic complexity when disjunctive constraints are introduced (Dechter et al. 1990). If we consider the example of an emergency coordination procedure following a fire alert, events such as smoke detection, team departure, and their arrival on site are not associated with fixed times, but connected by constraints, for example requiring that departure occur within a maximum delay after the alert, then that on-site intervention respect a bounded travel time. A constraint network then automatically propagates these requirements and detects any inconsistency, for example when estimated delays render the procedure infeasible before its execution.

These structures can also be combined with other modeling dimensions to enrich reasoning. Integration of qualitative spatial relations, such as those formalized in RCC-8 algebra, notably enables joint reasoning about temporal and spatial constraints, for example to verify coherence of events distributed in space and time (Wolter and Zakharyaschev 2000). Furthermore, constraint networks can be used with history management mechanisms from temporal databases, notably via bitemporal models distinguishing valid time and transaction time, as well as regular path queries allowing interrogation of event or state sequences over time (Jensen et al. 1998; Barceló 2013). These combinations thus offer an expressive approach for analyzing complex evolutions while preserving structural coherence guarantees.

Inference and complexity

Complexity results associated with temporal formalisms show that expressiveness gains can come with high algorithmic costs, particularly when properties to verify

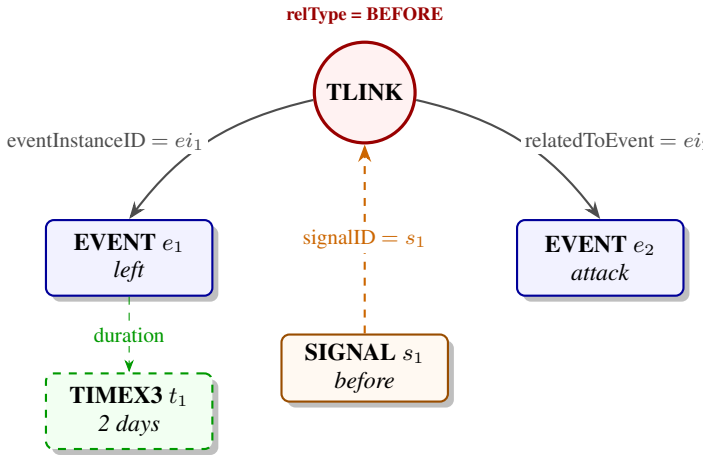


Figure 4. Example of a TimeML structure inspired by the sentence *John left 2 days before the attack*. The nodes represent events (`EVENT`), a temporal expression (`TIMEX3`), and a signal (`SIGNAL`). The temporal relation is encoded by a `TLINK` of type `BEFORE`.

concern extended behaviors or combine multiple temporal dimensions. LTL satisfiability is PSPACE-complete, and several linear approaches discussed in this section have encodings or translations to LTL that place them within a comparable complexity bound (Sistla and Clarke 1985; Kamide 2015, 2017). In contrast, significant enrichment of syntax through combination of additional qualitative dimensions or complete internalization of interval relations leads to formalisms whose complexity can reach EXPSPACE or become undecidable (Halpern and Shoham 1991; Wolter and Zakharyaschev 2000). These contrasts highlight a recurring trade-off between expressiveness and algorithmic cost that directly conditions the choice of formalisms exploitable in practice, without uniform progression from one formalism to another Sistla and Clarke (1985); Halpern and Shoham (1991); Wolter and Zakharyaschev (2000).

On the operational level, temporal inference relies on several families of proof and verification methods, whose objective is to automatically establish satisfaction of temporal properties by a given model.

Approaches based on model checking compile temporal specifications into automaton structures, typically Büchi automata, and constitute the foundation of numerous verification tools (Clarke et al. 2001). These mechanisms are also used in application evaluation protocols aimed at testing the temporal reasoning capabilities of neural models. Applied to the domain of large language models, a model checker can be used as an oracle to determine whether a trajectory described or generated in natural language satisfies a given LTL formula, which allows systematic comparison of model outputs to a formal temporal specification. This approach is illustrated by the LTLBench benchmark Tang et al. (2026).

Temporal resolution methods exploit normalized forms and specialized rules to mechanize the search for models or contradictions, while controlling the search space. For example, the TLX fragment introduces exclusivity constraints between propositions to model mutually exclusive operating modes, which enables tractable deductive reasoning, notably in multi-agent system contexts or protocols where an agent can occupy only one state at each instant(Dixon 2007). Finally, dedicated engines explicitly integrating time as a first-order parameter or combining logic and uncertainty provide specialized alternatives according to application needs (Panayiotopoulos and Gergatsoulis 2002; Lu et al. 2010; Hou 2012).

Toward neuro-symbolic integration

This body of work highlights the maturity and rigor of symbolic temporal reasoning, but also its structural limitations, linked to the necessity of having explicit and fully specified representations. The high expressiveness of temporal logics and constraint models requires explicit and precise modeling of events, states, and their temporal relations, which is difficult to reconcile with noisy, continuous, or unstructured data. Conversely, neural models learn efficiently from such data, without however providing formal guarantees on temporal coherence or global validity of inferred trajectories.

Neuro-symbolic integration thus appears as a natural response to this duality, by combining statistical learning mechanisms with symbolic components capable of expressing explicit temporal constraints. Logical formalisms and constraint structures provide normative frameworks to guide, constrain, and verify neural predictions, while neural models construct intermediate representations from raw data. Quantitative semantics, notably those from Signal Temporal Logic, play a central role by establishing a direct link between logical specification and differentiable optimization, which enables integrating temporal requirements at the core of learning processes.

This normative conceptual framework justifies the study of temporal neuro-symbolic architectures. The following section examines how these principles are concretely implemented in hybrid systems, where learning, representation, and temporal reasoning interact closely within unified processes.

Integration of neural models and temporal logic

Motivations and principles of temporal neuro-symbolic integration

Deep learning models applied to temporal data are now widely used for modeling sequences and evolving signals. Recurrent, temporal convolutional, and attention-based architectures constitute the dominant families for sequential data processing, and have been successfully used in tasks such as complex event detection, temporal relation extraction, or sequential prediction (Ma et al. 2020; Knez and Žitnik 2023; Han and Srivastava 2024). These architectures exploit temporal regularities when relevant dependencies are encoded in the data, but they do not have, by construction, explicit mechanisms guaranteeing compliance with formal temporal constraints or global

coherence of produced trajectories, particularly when these constraints concern distant relations in time or an entire sequence (Marconato et al. 2024; Marín 2025).

Several empirical works show that locally correct predictions can be obtained through contingent correlations or indirect reasoning, for example when satisfaction of a global constraint is approximated by local regularities, while underlying temporal or causal violations remain difficult to detect using standard predictive metrics (Marconato et al. 2024; Marín 2025). These limitations are particularly evident when reasoning depends on global constraints over complete temporal trajectories, such as respecting event ordering over an extended horizon.

Conversely, temporal logics provide formalisms to explicitly specify and verify properties concerning event ordering, delays, and certain causal relations. They rely on semantics defined over complete traces, which allows expressing global properties valid for an entire trajectory, and enable exact reasoning relative to formulated constraints. However, these guarantees come with well-identified structural limitations. The grounding of temporal dependencies and the cost of logical reasoning grow rapidly with temporal horizon and formalism expressiveness, which makes exhaustive inference difficult to apply at large scale (Dean 1989; Mukherji et al. 2025).

Certain works examine temporal neuro-symbolic approaches combining neural models and logical formalisms to more finely analyze produced reasoning trajectories, by confronting learned representations with explicit temporal constraints. These approaches notably focus on identifying temporal inconsistencies or forms of opportunistic reasoning, for example when predictions locally satisfy data without respecting a specified global constraint (Marconato et al. 2024; Marín 2025). The pursued objective is not to assume systematic elimination of these phenomena, but to make visible the temporal dependencies actually used by the model.

In several approaches, temporal logic is thus not only used for a posteriori verification, but as a support for analyzing reasoning during inference, by allowing intermediate decisions to be associated with identifiable rules or constraints. Works based on rule traces and explicit inference chains show that this association facilitates inspection of reasoning trajectories and analysis of produced temporal behaviors. However, these mechanisms do not imply that all specified temporal properties are systematically satisfied, and trajectory inspection can also reveal persistent failures of temporal reasoning, despite the presence of explicit symbolic constraints (Aditya et al. 2023).

Temporal neuro-symbolic integration is thus studied in several contributions as an experimental perspective to explore different modalities of articulation between statistical learning and symbolic reasoning, particularly for complex temporal phenomena (Garcez and Lamb 2020; Lee et al. 2023; Lorello et al. 2025a; Liang et al. 2025).

While these approaches introduce explicit symbolic structures that facilitate understanding and control of temporal reasoning, they do not however provide universal guarantees of coherence or optimality, due to approximations, uncertainty, and reasoning shortcuts observed empirically (Ma et al. 2020; Aditya et al. 2023; Marconato et al. 2024) (Mukherji et al. 2025).

Neuro-Symbolic Temporal Reasoning Architectures

Temporal neuro-symbolic approaches can be analyzed with regard to how neural and symbolic components are related within reasoning. This articulation constitutes a recurring challenge of neuro-symbolic integration, particularly for temporal tasks where dependencies extend over complete trajectories and require coordination between continuous representations and discrete reasoning, and when local regularities learned from signals must be reconciled with global constraints on event ordering or duration (Garcez and Lamb 2020; Lee et al. 2023; Lorello et al. 2025b).

The way this articulation is organized directly affects the possibilities for analyzing reasoning. The literature shows that explicit structuring of temporal constraints, component modularity, and availability of inference traces facilitate evaluation of temporal coherence and inspection of produced behaviors, by making it possible to identify the precise moment in a trajectory where a temporal constraint is violated or circumvented, without however guaranteeing the absence of contradictions (Dean 1989; Aditya et al. 2023; Bazaga et al. 2025). These organizational choices thus condition the practical possibilities for evaluation and interpretation of temporal reasoning, beyond simple implementation decisions (Garcez and Lamb 2020).

Certain approaches propose to distinguish different modes of integration between learning and reasoning, corresponding to increasing degrees of interaction between neural and symbolic components, with the same temporal task thus being approached according to whether logic intervenes before, after, or during neural inference, without claiming to provide an exhaustive typology (Garcez and Lamb 2020; Lee et al. 2023).

A first integration mode (*Logic* $\rightarrow$ *Network*) consists of introducing temporal logic upstream of neural learning, in the form of constraints or regularization mechanisms. Rules then define a space of admissible solutions by penalizing violations of formal properties or excluding predictions incompatible with explicit temporal commitments, as illustrated in schemes (A) and (B) of Figure 5. In these configurations, logic acts either as a differentiable signal integrated into the loss function, or as a structuring constraint on the latent space, guiding learning without directly intervening in the inference dynamics. A trajectory can thus temporarily violate a temporal constraint without being excluded from learning, with the severity of the violation simply reflected by a continuous penalty. Approaches such as Logical Neural Networks or differentiable rule-based probabilistic systems illustrate this strategy, in which symbolic reasoning plays the role of an inductive bias guiding statistical learning (Riegel et al. 2020; Huang et al. 2021).

At a stronger degree of integration, logic intervenes at the execution level and not only as an optimization signal. Scheme (C) illustrates an action or transition filtering mechanism, in which a temporal monitor blocks non-compliant behaviors and guarantees that actually produced trajectories respect safety constraints, independently of any occasional network errors. Scheme (D) corresponds to architectural integration, where an explicit symbolic structure, typically an automaton compiled from a temporal formula, is incorporated into the model dynamics. In this case, logical state becomes an internal variable that constrains computation evolution and allows explicitly tracking progress of specification satisfaction along the trajectory. The transition from schemes (A,B) to

(C,D) thus highlights a continuum ranging from indirect influence of logic on learning to more constraining mechanisms offering reinforced guarantees, at the cost of more direct intervention in neural computation.

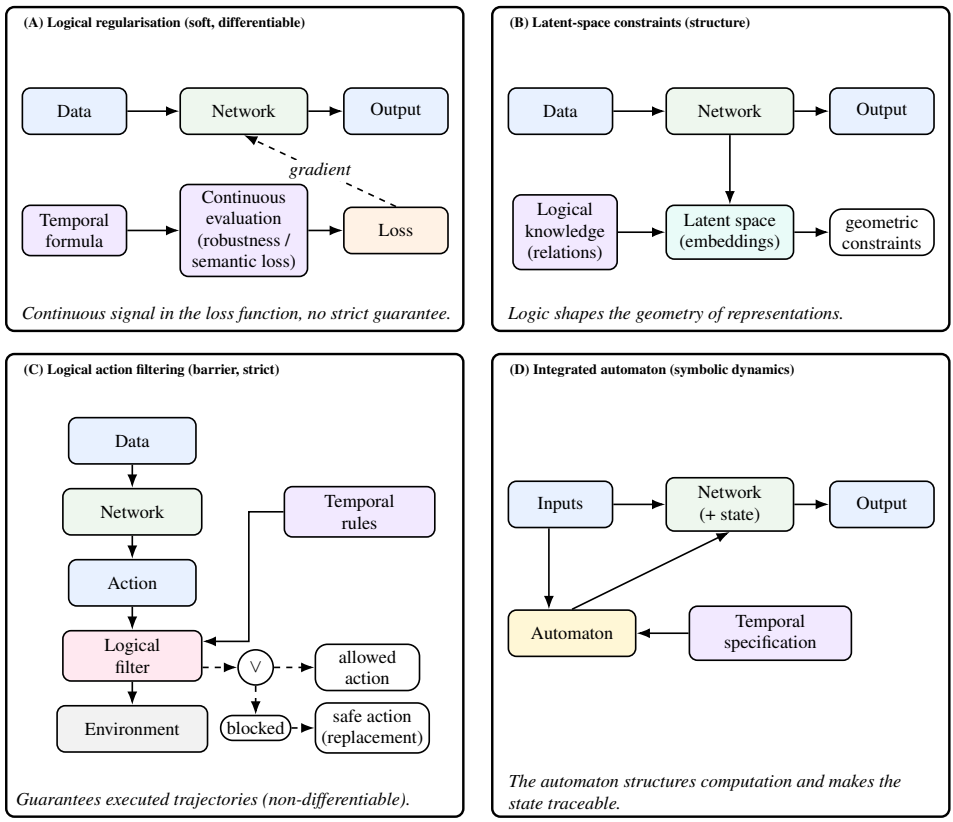


Figure 5. Decomposition of Logic → Network approaches: (A) differentiable regularisation through the loss, (B) geometric constraints over the latent space, (C) strict action filtering, (D) automaton integration within the architecture.

A second integration mode (*Network* → *Logic*) relies on a posteriori extraction of symbolic structures from trained neural models. The network is freely learned from temporal data, without explicit logical constraints, then rules, graphs, or automata are derived to analyze and explain its behavior, by making explicit the temporal dependencies actually used during inference. This approach improves interpretability and facilitates inspection of produced reasoning trajectories, for example by allowing a final decision to be linked to a succession of learned temporal relations, without offering formal guarantees on the temporal validity of decisions (Aditya et al. 2023; Marconato et al. 2024). Extracted structures thus describe the observed behavior of the model, without intervening on its dynamics or actively correcting its inferences.

Certain architectures nevertheless show that qualitative temporal relations can be operationally exploited in this perspective. For example, NeuSTIP explicitly integrates Allen's algebra to represent relations such as *before*, *overlaps*, or *meets* in the prediction and analysis of temporal links, which allows analyzing temporal coherence of evidence used along a reasoning path, illustrating the practical feasibility of integrating symbolic temporal relations into structured neural models (Figure 6) (Singh et al. 2023).

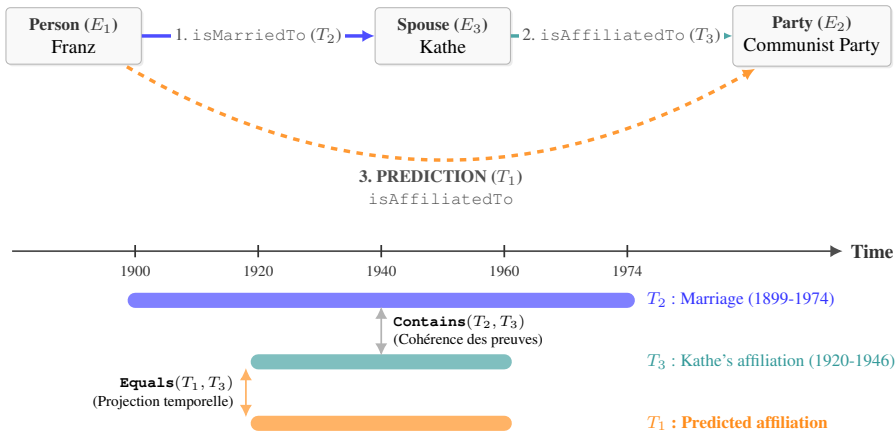


Figure 6. Illustration of neuro-symbolic reasoning (NeuSTIP). The model infers the political affiliation of an individual (T_1) by combining the structural path in the knowledge graph (top) with the coherence of temporal intervals (bottom), validated through Allen's interval algebra.

A third integration mode (*Network* $\leftrightarrow$ *Logic*) aims for closer interaction between neural learning and logical reasoning. Temporal rules are integrated directly at the core of the inference process, in a hybrid or differentiable form, enabling partial co-evolution of learned representations and symbolic constraints, when satisfaction or violation of a temporal rule directly influences internal activations or model gradients.

Works describe architectures in which symbolic representations guide neural inference, detect inconsistencies, and trigger iterative or abductive correction mechanisms, by progressively adjusting temporal hypotheses compatible with observations, according to the architecture considered (Garcez and Lamb 2020; Lorello et al. 2025a; Liang et al. 2025). These approaches do not systematically guarantee satisfaction of all temporal properties, but make visible the dependencies actually used by the model, by exposing the rules or constraints that influence inference at each step, and provide explicit handles for reasoning analysis and evaluation.

Finally, in an even more integrated form, certain methods perform temporal reasoning directly in the internal dynamics of networks via differentiable operations (i.e., formulated continuously to be optimized by gradient descent). Temporal inductive learning approaches, such as TILP or TEILP, jointly learn the structure of logical rules and distributions associated with temporal intervals from noisy data. In these methods,

the temporal rule becomes a learned, revisable component directly involved in inference, empirically demonstrating the feasibility of joint learning of temporal logic and neural representations (Xiong et al. 2024, 2023).

These different forms of integration can be synthesized as a comparative typology, contrasting principles of articulation between learning and temporal reasoning, offered guarantees, and their structural limitations (Table 1).

Table 1. Neuro-symbolic architectures for temporal reasoning and associated guarantees.

Architecture	Integration principle	Role of temporal logic	Main guarantees	Structural limitations
Logic → Network	Symbolic constraints imposed on learning	Regularization or a priori filtering	Partial coherence, local compliance	Rigidity, low adaptability, absence of dynamic correction
Network → Logic	A posteriori extraction of rules or graphs	Explanation and analysis of learned behavior	Interpretability, traceability	Absence of formal guarantees, limited control
Network ↔ Logic	Continuous learning–reasoning interaction	Active constraints and integrated verification	Coherence by construction, auditability	Computational cost, design complexity

A detailed mapping of surveyed approaches, structured according to their integration mode (Logic → Network, Network → Logic, or bidirectional approaches) and their functional role, is provided in Appendix . This appendix also includes a broader set of relevant temporal neuro-symbolic works that were examined during the survey process but not discussed in detail in the main text, in order to provide a more comprehensive overview of the landscape.

Differentiable temporal reasoning and integrated constraints

The architectures presented previously can be understood as particular cases of more general interactive architectures, in which neural learning and logical reasoning are articulated according to increasing degrees of integration. Figure 7 synthesizes the main interaction schemes between neural and symbolic components used in differentiable reasoning.

It highlights three complementary forms of integration. The first corresponds to direct incorporation of logical constraints as differentiable operations within neural computation. The second relies on co-learning using explicit exchange of representations and rules between modules. The third involves closed-loop operation in which verification mechanisms trigger iterative corrections. This synthesis thus illustrates how explicit constraints can be integrated at the very core of inference, rather than being confined to an a posteriori verification role.

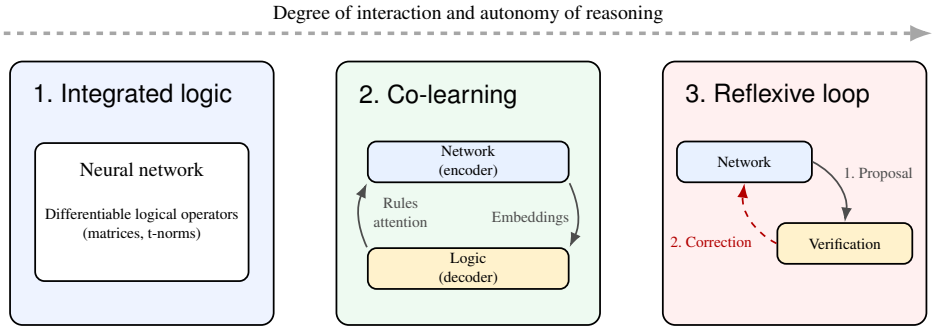


Figure 7. Synthesis of interactive Network ↔ Logic architectures. (1) Logic is integrated into neural operations. (2) Modules co-learn through the exchange of representations and rules. (3) The system operates in a closed loop with verification and dynamic correction.

Temporal neuro-symbolic integration becomes more directly exploitable in architectures where explicit logical constraints can interact with differentiable learning mechanisms. The literature indeed shows that logical statements, including temporal ones, can be translated into continuous constraints integrated into the optimization process, such that these constraints effectively influence gradient descent and the learned representation (Ma et al. 2020; Badreddine et al. 2022; Lorello et al. 2025a; Xiong et al. 2023).

A typical example corresponds to situations in which a neural network predicts the evolution of a continuous signal, such as velocity or position over time, while a temporal specification imposes global requirements on safety thresholds or objective attainment delays. The logical constraint is then no longer reduced to an a posteriori verification mechanism, but acts as a continuous corrective signal, integrated into the learning process, that progressively guides predicted trajectories toward behaviors compatible with formulated temporal requirements.

In this type of configuration, temporal logic can intervene according to several complementary modalities. It can act as a differentiable loss term influencing learning, as an intermediate evaluation module analyzing produced trajectories, or as an explicit correction mechanism projecting predictions toward temporally coherent behaviors.

A set of approaches integrate explicit temporal properties into gradient-optimized architectures, by making certain operations such as satisfaction, violation, or robustness computable within differentiable graphs. These methods are grouped under the term *differentiable temporal reasoning*. Methods such as STLnet or backpropagation through STL specifications illustrate this compatibility by compiling temporal formulas into computation graphs whose outputs can be used as loss terms or as intermediate learning signals (Ma et al. 2020; Leung et al. 2021). Other approaches, such as TILP, learn temporal rules in a differentiable manner in formalisms distinct from STL, while relying on the same principle of gradient-compatible integration of temporal structures (Lorello et al. 2025a; Xiong et al. 2023).

A major technical obstacle to this integration lies however in the very nature of logical operators used to evaluate constraints. Many continuous connectives from fuzzy logic, for example conjunctions defined by a minimum (Gödel logic) or certain classical t-norms, exhibit selective behavior, in which only one sub-formula effectively contributes to the global value of the formula (Klement et al. 2000). During backpropagation, this selection translates into zero gradients on non-selected branches, preventing joint adjustment of different components of the logical constraint. This *vanishing gradient* phenomenon, widely documented in works on differentiable logical reasoning (Krieken et al. 2022), tends to propagate in compound formulas and degrade learning stability. It thus constitutes a structural obstacle that makes necessary the introduction of smooth relaxations of logical operators, an essential condition to enable gradient-based optimization that is effectively exploitable in temporal reasoning approaches.

This difficulty can be illustrated by a simple case of continuous conjunction. Consider a logical constraint of the form $\varphi = \varphi_1 \wedge \varphi_2$, evaluated using the Gödel t-norm, defined by $V(\varphi) = \min\big(V(\varphi_1), V(\varphi_2)\big)$. If, for a given trajectory, we have $V(\varphi_1) = 0.8$ and $V(\varphi_2) = 0.3$, then the global value of the formula is determined only by φ_2. During backpropagation, the gradient of the associated loss is zero with respect to φ_1, while only the branch corresponding to φ_2 is effectively updated. Thus, part of the logical constraint remains invisible to optimization, which prevents joint adjustment of sub-formulas.

To address this limitation, one approach relies on relaxation of temporal logic into continuous operators, to preserve informative gradient propagation. In the context of Signal Temporal Logic, formulas are no longer evaluated in a strictly Boolean manner, but associated with robustness semantics that quantitatively measure the degree of satisfaction or violation of a temporal property (a positive value indicating robust satisfaction, a negative value a violation) (Donzé and Maler 2010). These semantics can be implemented as differentiable computation graphs, allowing backpropagation of gradients through temporal specifications (Figure 8) (Leung et al. 2021).

STL robustness as a signed distance. In quantitative STL semantics, a predicate $\mu(x) > c$ has robustness $\rho(s_t, \mu_c) = \mu(x_t) - c$, which directly encodes the satisfaction margin ($\rho > 0$ satisfied, $\rho < 0$ violated).

Example. For $x(t) \geq 1$ we have $\rho(x(t)) = x(t) - 1$. Consider $x(0) = 1.4$, $x(1) = 0.8$, $x(2) = 1.2$, giving margins $\rho(0) = 0.4$, $\rho(1) = -0.2$, $\rho(2) = 0.2$. For $\varphi = \mathbf{G}_{[0,2]}(x(t) \geq 1)$,

$$\rho(x, \varphi) = \min_{t \in \{0,1,2\}} \rho(x(t)) = \min(0.4, -0.2, 0.2) = -0.2,$$

so φ is violated by a margin of 0.2 at the critical instant. If $x(1)$ is increased to 1.1, then $\rho(x, \varphi) = 0.1$ and the formula becomes satisfied. This signed-distance property makes robustness directly usable as a differentiable training signal for constraint-aware learning.

Figure 8. Example: STL robustness as a signed distance.

Furthermore, several differentiable logical reasoning methods encode rules as real-valued constraints integrated either into the loss function or directly into the architecture,

providing a general mechanism for interaction between symbolic rules and gradient-based learning (Riegel et al. 2020; Huang et al. 2021; Badreddine et al. 2022).

When temporal constraints are integrated into the cost function, verification is no longer limited to a posteriori analysis, but intervenes during learning as a regulation mechanism, with violations generating a continuous corrective signal exploited by optimization. In practice, several architectures explicitly translate temporal or causal rules into differentiable penalties guiding gradient descent (Ma et al. 2020; Liu et al. 2025).

In certain architectures, continuous evaluation of these specifications is not only used as a learning signal, but also serves to explicitly correct produced trajectories. Network predictions can then be adjusted by a projection or optimization module, which modifies predicted values or times to satisfy specified requirements. This projection mechanism transforms temporal logic into an operational constraint acting directly on model outputs, analogously to a constraint solver integrated into the inference process.

A representative scenario is that of a control system where a network predicts a continuous trajectory, for example a velocity $v(t)$ or a position $x(t)$. A specification then imposes global requirements, such as permanent compliance with a safety threshold or reaching an objective before a deadline. Figure 9 summarizes how these requirements are translated into a continuous signal (robustness, loss) reinjected into optimization to progressively adjust predicted trajectories.

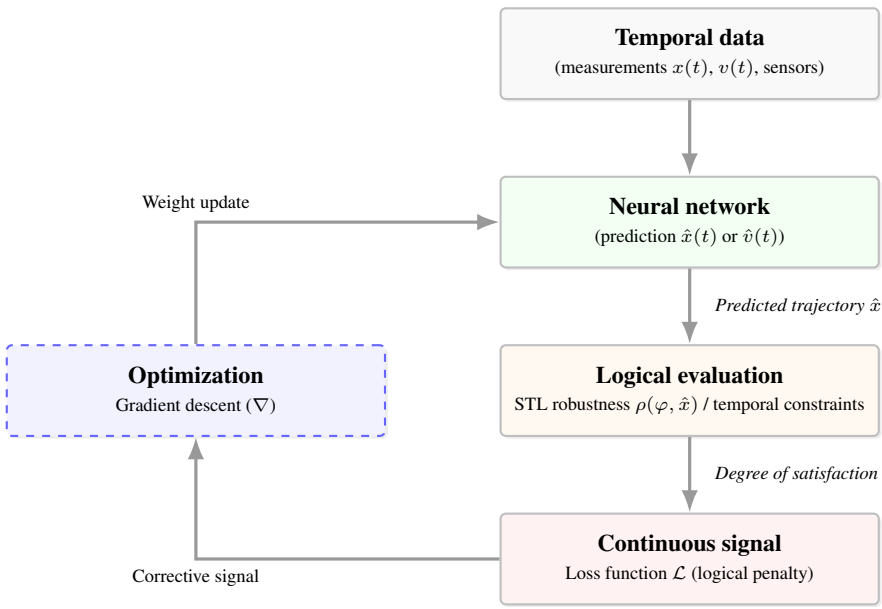


Figure 9. Pipeline for differentiable temporal Reasoning integrating explicit logical constraints.

These mechanisms can also be used to learn temporal rules from data. Certain approaches aim for direct induction of STL formulas from observations, relying on quantitative robustness to handle noisy signals (Bombara and Belta 2021). Other methods, such as TILP or TEILP, learn temporal rules in a differentiable manner in inductive formalisms adapted to event graphs (Xiong et al. 2024, 2023). Rules thus learned are used for classification, analysis, or temporal monitoring tasks, and can contribute to partially structuring inference according to the architecture considered (Bombara and Belta 2021).

These approaches however introduce structural trade-offs. Logical relaxation departs from strict discrete semantics by introducing continuous evaluation of satisfaction (Donzé and Maler 2010). Furthermore, computational costs tend to grow with specification expressiveness and the number of formulas considered, which motivates adapted representation and implementation choices (Ma et al. 2020; Leung et al. 2021). Nevertheless, by making possible direct interaction between explicit temporal specifications and gradient-based optimization, these mechanisms constitute a concrete technical path for integrating formal constraints at the core of differentiable architectures (Ma et al. 2020; Leung et al. 2021; Xiong et al. 2024; Lorello et al. 2025a).

Coherence guarantees, explainability, and controllability

Joint integration of neural mechanisms and explicit temporal logics enables introduction of structural properties that remain difficult to access when temporal reasoning relies exclusively on learned statistical correlations. In particular, several works show that introduction of explicit symbolic constraints makes observable and verifiable temporal commitments that remain implicit in standard neural architectures (Marconato et al. 2024; Marín 2025).

A first property concerns temporal coherence of reasoning. When ordering, duration, or causality constraints are explicitly formulated, trajectory validity can be evaluated independently of local plausibility of produced predictions. This distinction highlights the possible gap between point-wise accuracy and satisfaction of global specifications, empirically emphasized in the analysis of reasoning shortcuts and systematic violations observed in unconstrained models (Marconato et al. 2024; Marín 2025). For example, a locally correct prediction may respect instantaneous observations while violating a global requirement that event A systematically precedes event B over the entire trajectory.

A second property concerns structural explainability. Certain neuro-symbolic architectures make explicitly accessible the rules, constraints, and transitions used during inference, enabling each decision to be linked to interpretable symbolic objects. Systems such as PyReason illustrate this capability by providing complete inference traces, including applied rules, temporal annotations, and deduction steps, which allows detailed inspection of reasoning trajectories over extended horizons (Aditya et al. 2023). This explicitation facilitates analysis of actually operating mechanisms and localization of unexpected or unstable behaviors.

Controllability constitutes a third important property. In certain integrated architectures, violations of temporal specifications can be detected and exploited as

adjustment signals for reasoning. Approaches based on abductive correction show that it is possible to correct or refine erroneous inference through symbolic verification loops, without resorting to complete retraining of the neural model, notably in zero-shot or weakly supervised contexts (Liang et al. 2025).

Thus, a temporal ordering violation detected during inference can lead to local revision of an intermediate hypothesis, without modifying underlying neural parameters. This capability distinguishes architectures explicitly integrating logical reasoning from strategies relying solely on a posteriori analysis or explanation.

These properties also have direct implications for evaluation. By making explicit the temporal commitments of the system, neuro-symbolic integration enables defining evaluation criteria based on structural validity of trajectories rather than solely on accuracy of final outputs. Robust satisfaction metrics from Signal Temporal Logic provide quantitative tools to evaluate the degree of satisfaction or violation of explicit temporal properties, and can be interpreted as reasoning verification instruments beyond a binary predictive verdict (Donzé and Maler 2010).

These benefits however come with well-documented constraints and trade-offs. Introduction of explicit logical structures incurs design costs, notably related to knowledge formalization and the rule acquisition bottleneck (Liu et al. 2022), as well as computational and memory costs associated with grounding and manipulation of expressive temporal specifications (Aditya et al. 2023; Mukherji et al. 2025; Chen et al. 2025a). Converging analyses also emphasize the existence of trade-offs between logical expressiveness, differentiability, and computational efficiency, which impose architectural choices such as approximation of exact inference, limitation of logical exhaustiveness, or adoption of probabilistic mechanisms (Garcez and Lamb 2020; Huang et al. 2021; Mukherji et al. 2025; DeLong et al. 2025).

Temporal neuro-symbolic integration can thus be understood as a methodological approach to analyze, constrain, and regulate temporal reasoning, by articulating statistical learning, symbolic reasoning, and explicit verification mechanisms. Certain contributions position it as a structuring perspective to study the interaction between perception, temporal reasoning, and control, and to equip systems with tools for analysis, correction, and debugging of reasoning in dynamic environments (Garcez and Lamb 2020; Aditya et al. 2023; Lee et al. 2023; Liang et al. 2025).

Illustrative applications of temporal neuro-symbolic reasoning

Temporal neuro-symbolic approaches have been primarily used in contexts where reasoning validity depends on compliance with explicit sequential constraints, rather than point-wise accuracy of isolated predictions. In many sequential reasoning tasks, a trajectory can indeed contain all expected events while becoming invalid as soon as their order or temporal position violates a specified global constraint (Liang et al. 2025; Lorello et al. 2025b). Such cases are characteristic of so-called critical systems, in which events must satisfy formal relations of ordering, duration, or causality over the entire temporal trajectory.

A representative scenario consists of analyzing or producing a continuous temporal trajectory while respecting explicit global requirements. In a control or monitoring situation, a neural model can for example predict the evolution of a signal $x(t)$, corresponding to a position, velocity, or physiological activity, while a specification imposes conditions such as permanent compliance with a safety threshold or reaching a target state within a given time window.

The literature highlights that certain domains, particularly critical and cyber-physical systems, impose explicit requirements regarding safety, reliability, and compliance with formal constraints, which purely statistical models struggle to guarantee. For example, in an autonomous navigation system, a predicted trajectory can locally satisfy kinematic constraints while violating a global safety rule, such as entering a forbidden zone after a given delay, a violation that is only detectable at the scale of the complete trajectory.

In these situations, integration of logical representations is presented as a means to improve reliability and trust in neural systems, by allowing explicit expression of safety properties and temporal specifications, such as compliance with deadlines, execution orders, or safety conditions that must be satisfied over an entire trajectory (Lorello et al. 2025a,b). Several works show that these specifications can be used to verify trajectory coherence or detect violations of expected properties during execution or simulation, notably via temporal monitoring mechanisms based on continuous evaluation of formal temporal formulas (Donzé and Maler 2010). These approaches thus enable continuous evaluation of system behavior compliance with explicit requirements, without assuming automatic trajectory correction, with logic playing here a role of observation and diagnosis rather than direct control. Figure 10 illustrates this mechanism in the case of a continuous trajectory subject to a global specification, highlighting the evaluation window considered as well as local detection of a violation.

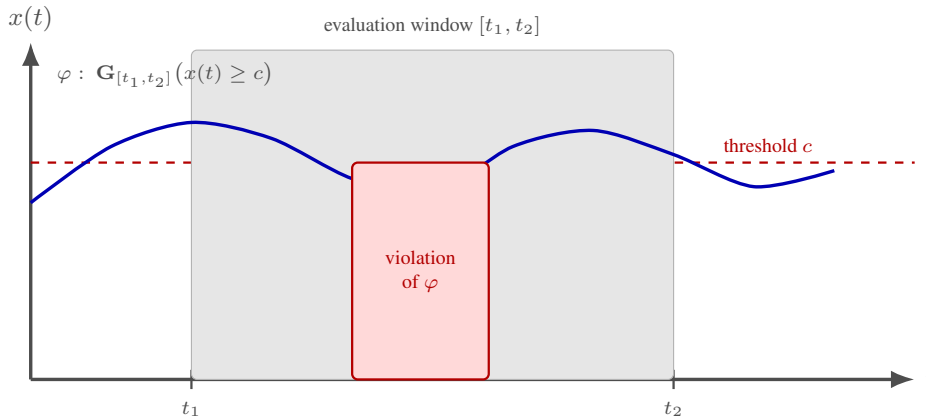


Figure 10. Temporal monitoring of a constraint over a continuous trajectory. A signal $x(t)$ is evaluated with respect to a global temporal specification (e.g., $\mathbf{G}_{[t_1,t_2]}(x(t) \geq c)$) over a given observation window. The shaded area delimits the temporal interval over which the constraint is assessed, while the colored region highlights a local violation of the property, corresponding to a time instant at which the signal does not satisfy the imposed threshold. This type of monitoring makes it possible to detect structural violations independently of the local plausibility of the predictions.

In the domain of symbolic temporal reasoning, neuro-symbolic integration is also used to structure and constrain the interpretation of complex sequences.

Concretely, a prediction may rely on a sequence of correctly identified individual events, but be invalid if the expected order (for example diagnosis $\rightarrow$ treatment $\rightarrow$ improvement) is reversed or partially violated.

Recent works show that introduction of explicit rules promotes interpretable and coherent forms of reasoning over complete trajectories, by reducing internal inconsistencies and improving robustness of sequential reasoning (Singh et al. 2023; Liang et al. 2025) (Liu et al. 2022).

Furthermore, several approaches based on temporal or causal logical rules show that these symbolic structures can support better generalization to temporal configurations not observed during training, by constraining the space of admissible trajectories independently of observed statistical distributions, particularly when constraints encode stable structural regularities (Garcez and Lamb 2020; Xiong et al. 2024).

More generally, the literature emphasizes that the main contribution of temporal neuro-symbolic approaches lies less in marginal performance improvement than in their capacity to produce evaluable, interpretable, and reliable reasoning over extended horizons, that is, reasoning that can be inspected, confronted with explicit specifications, and analyzed with regard to its internal dependencies. These approaches are thus presented as methodological instruments for analysis, verification, and control of complex systems, particularly in contexts where reasoning reliability and transparency

constitute central requirements (Garcez and Lamb 2020; Aditya et al. 2023; Lorello et al. 2025b).

Limitations and structural tensions

Despite the contributions of temporal neuro-symbolic integration, several structural tensions persist and condition the choice of formalisms and architectures. A first tension concerns the trade-off between logical expressiveness and computational feasibility. Introduction of explicit temporal constraints enables representing non-local dependencies and extended temporal relations, but these capabilities come with increased costs when relations become distant or the temporal horizon extends, leading in certain cases to reasoning that is correct in theory but difficult to exploit in practice when structure size or trajectory duration increases. Thus, a constraint requiring that an initial event conditions a response several hundred instants later can be formally expressible, while making inference prohibitive on long trajectories or dense event graphs.

Recent analyses show that, while symbolic or hybrid methods can capture long-range dependencies, their inference does not scale well on large structures, while increasing temporal depth leads to measurable growth in inference time and computational requirements (Dean 1989; Yu et al. 2024; DeLong et al. 2025).

A second tension concerns the articulation between differentiability and strict logical semantics. Approaches based on quantitative or fuzzy semantics overcome the limitations of Boolean verdicts by making measurable the degree of satisfaction of a property, but they modify the very nature of the satisfaction relation, with a property being able to be weakly satisfied or weakly violated according to a continuous measure rather than classified absolutely. A trajectory can then be considered weakly satisfactory according to a robustness measure, while remaining unacceptable in a context where even marginal violation leads to categorical rejection.

It is now acknowledged that, in these semantics, boundaries between satisfaction and violation become gradual, reflecting the vague nature of many properties derived from real-world data, which makes strictly binary interpretation inappropriate (Donzé and Maler 2010; Frigeri et al. 2012; Conradie et al. 2020). This semantic evolution constitutes a deliberate methodological choice rather than formal weakening, but it implies interpretation precautions according to application requirements, notably when decisions require a categorical verdict based on explicit thresholds.

A third tension concerns architectural design and integration of neural and symbolic modules. Empirical works show that simple modular combination of components that perform well in isolation can lead to learning instabilities, optimization problems, or increased global complexity without proportional benefit, for example when signals from symbolic specifications interfere with neural optimization or when module articulation introduces dependencies that are difficult to calibrate, particularly in relational and temporal contexts (Lorello et al. 2025a; DeLong et al. 2025). These findings have led several recent surveys to emphasize the necessity of explicit architectural choices and clear taxonomies, as well as to identify management of trade-offs between integration,

interpretability, performance, and scalability as a central challenge for the maturation of spatio-temporal neuro-symbolic reasoning (DeLong et al. 2025).

Interpretability and temporal rule Extraction

Why temporal explainability is a condition for reasoning

Application of post-hoc explainability methods to sequential architectures raises specific difficulties related to temporal dependencies, as shown by analyses dedicated to applying attribution methods, which aim to estimate the relative contribution of past inputs or instants to a given prediction, in temporal contexts (Tonekaboni et al. 2020).

Approaches such as LIME or SHAP, initially designed for static data, are frequently applied to time series without explicitly modeling dependencies between successive instants. It has been shown that this application leads to treating time steps as quasi-independent entities, which is problematic when the decision depends on delayed relations (Tonekaboni et al. 2020).

Reported experiments indicate that these methods struggle to correctly locate observations that are truly determinant for the decision, particularly when the influence of an event depends on its relative positioning in the global trajectory. This difficulty is accentuated by data non-stationarity, explicitly identified as a central issue in clinical time series analysis, where the effect of a variable can evolve along with system regime changes (Tonekaboni et al. 2020).

Perturbation strategies used to estimate local importances also artificially modify certain parts of analyzed sequences. Existing contributions show that these perturbations can produce unrealistic counterfactuals and out-of-distribution configurations, which do not correspond to trajectories actually encountered during inference (Tonekaboni et al. 2020). Obtained explanations may then reflect more the model's reaction to artificial configurations than its behavior on plausible sequences.

These findings emphasize that, for temporal data, interpretation of a prediction often depends on global structure rather than isolated instantaneous values. Studies conducted on continuous physiological data notably illustrate that extended dynamics play a central role in the decision, making interpretation based exclusively on point-wise importance scores delicate (Tonekaboni et al. 2020).

Such situations align with more general criticisms addressed to post-hoc explanations that are plausible but potentially misleading. It is indeed well established that an explanation can appear convincing while masking statistical shortcuts or incorrect concepts learned by the model, contributing to an illusion of understanding rather than faithful analysis of underlying reasoning (Rudin 2019; Marconato et al. 2024).

Faced with these limitations, several approaches propose using explicit structures, such as rules, graphs, or inference chains, to represent in a traceable manner the relations used by the system, by making explicit the fact that an event can only be concluded if preceded by a given sequence of conditions satisfied in a precise order.

It appears that these structures make visible the causal and temporal dependencies exploited during inference and facilitate reconstruction of reasoning chains, by

identifying activated rules or relations actually used, as well as verification of conditions that led to a prediction (Aditya et al. 2023; Liang et al. 2025; Chen et al. 2025b; Zhu et al. 2025). In these methods, explanation is no longer limited to local a posteriori attribution, but is part of a reasoning process explicitly structured in time, where intermediate decisions can be linked to identifiable constraints or relations.

Temporal rules and symbolic structures as explanation supports

Temporal rules and symbolic structures enable, in several approaches proposed in the literature, explicit representation of ordering, duration, and causality relations used during inference. Unlike continuous latent representations, these objects offer explicit and inspectable inference chains linking initial observations to a final decision, thus facilitating analysis of reasoning followed by the system (Aditya et al. 2023). These properties contribute to making produced decisions more easily verifiable and discussable from a logical perspective.

Recent contributions have shown that it is possible to extract, from complex neural models, graphs, rules, or traces reproducing certain aspects of observed predictive behavior. However, this algorithmic capacity does not in itself guarantee the explanatory value of produced objects. In particular, so-called spurious rules can emerge through simple statistical coincidence, without coherent inferential trajectory or causal plausibility, notably in knowledge graphs (Chen et al. 2025b). For example, in event analysis contexts, ignoring intermediate historical conditions can make prediction of a future event unexplainable from recent observations alone (Yu et al. 2024).

In contrast to these approaches based on purely correlative extraction, certain works propose structural explainability mechanisms integrated directly into reasoning. The STAR-RAG system, for example, organizes inference as rule graphs aligned with predictions, producing explicit traces that reflect the logical structure actually used by the model (Zhu et al. 2025). These inference graphs enable inspection of causal and temporal dependencies associated with decisions, illustrating a form of explainability obtained by construction.

The explanatory quality of a rule is not reduced to its predictive performance. Several works emphasize the importance of complementary criteria, among which mental simulation capacity occupies a central place. An explanation is called simulable when it allows a user to anticipate system behavior on new cases from the provided explanatory object (Doshi-Velez and Kim 2017; Hoffman et al. 2019). This requirement implies a controlled trade-off between completeness and readability because the explanation must contain sufficient information to be informative, without introducing superfluous details that could harm clarity or user satisfaction (Hoffman et al. 2019).

Stability constitutes another criterion frequently used to analyze explanation quality. It has been shown that high sensitivity of explanations to slight variations in inputs or formulation can reveal fragility of underlying reasoning. In particular, marked instability has been interpreted as a sign of dependence on surface regularities rather than robust reasoning (Marín 2025). Furthermore, certain approaches based on quantitative or fuzzy semantics propose explicitly measuring spatio-temporal robustness of expressed

properties, by evaluating their degree of satisfaction in the face of signal perturbations (Bombara and Belta 2021).

Temporal neuro-symbolic approaches integrate interpretable logical languages at the very core of reasoning. Learning rules or Signal Temporal Logic (STL) formulas from traces enables formally explicating validity conditions associated with observed sequences, making violations themselves detectable and verifiable (Bombara and Belta 2021). By projecting the model's implicit choices into explicit symbolic structures, these approaches make visible certain fragile assumptions, causal dependencies, or reasoning shortcuts, opening the way to critical analysis and argued discussion of produced decisions (Marconato et al. 2024).

Thus, temporal rules and symbolic structures are not limited to descriptive a posteriori artifacts. A set of contributions emphasize the intelligibility and analytical capacity of symbolic structures, rather than faithful and exhaustive imitation of model behavior, in contexts of verification, inspection, or decision discussion objectives (Aditya et al. 2023; Liang et al. 2025; Zhu et al. 2025). Their interest lies less in their capacity to faithfully imitate model behavior than in their ability to make this behavior intelligible, simulable, and analyzable over time, by enabling examination of decision coherence over complete trajectories, anticipation of effects of local modifications, and identification of fragile or persistent dependencies over time.

Concretely, symbolic structures used for explanation purposes can take several formats, corresponding to distinct levels of abstraction and uses (Figure 11). At the most global level, temporal automata provide a comprehensive view of reasoning by making explicit possible system states and transition conditions between them, which makes them particularly suited to verification of process or protocol compliance. Causal graphs operate at a similar level of abstraction but focus instead on highlighting directed influence relations and their associated delays, in the form of explicit dependencies of the type $A_t \rightarrow B_{t+\tau}$, thereby facilitating analysis of underlying causal mechanisms. Moving towards more instance-based approaches, prototypes explain decisions by similarity with complete reference trajectories, offering intuitive interpretive support through comparison of global behaviors. At the finest granularity, shapelets identify discriminant local patterns, characteristic subsequences responsible for classification, thus enabling explanations focused on specific temporal segments.

The choice of one or another of these formats directly conditions the nature of interaction with the end user, whether it involves verifying a global property, analyzing a causal relation, or visually recognizing typical situations.

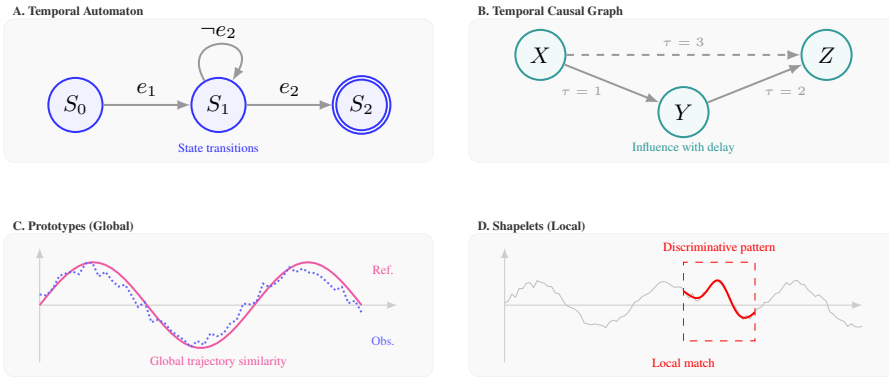


Figure 11. The four symbolic formats for temporal explanation. (A) Automata make the state dynamics explicit. (B) Causal graphs represent influence delays. (C) Prototypes explain through global similarity. (D) Shapelets isolate local discriminative patterns.

Traceability, human control, and trust

It is well established, in interpretability as in social sciences of explanation, that reasoning traceability plays a central role in humans' capacity to evaluate, control, and trust complex artificial intelligence systems. In particular, when formal or exhaustive verification of a system is impractical, access to explicit reasoning traces constitutes a privileged means to enable informed human inspection and critical evaluation of system behavior (Doshi-Velez and Kim 2017; Miller 2018).

Understanding an isolated prediction often appears insufficient. Users expect explanations that enable reconstructing mechanisms or reasoning steps that led to a decision, in order to assess its coherence, limitations, and application conditions (Miller 2018). Symbolic and neuro-symbolic approaches precisely propose representations in which rules, constraints, and dependencies used during inference are made explicit and inspectable (Aditya et al. 2023).

Recent analyses however highlight that systems can produce locally plausible explanations while relying on fragile or unstable inferential mechanisms. Such explanatory gaps have notably been observed in evaluation of temporal reasoning systems, where apparently coherent justifications mask high sensitivity to formulation or surface regularities (Marín 2025). These situations illustrate the risk that explanation, when not solidly grounded in actual reasoning, contributes to a false impression of control or robustness.

Traceability then constitutes a means of making these gaps observable. By making explicit inference chains, causal dependencies, and commitments used by the system, it provides support for human inspection, identification of structural fragilities, and reasoned discussion of produced decisions (causal narrative). It appears that provision of explicit symbolic structures facilitates reasoning audit and detection of undesirable or unexpected behaviors (Marconato et al. 2024).

Thus, several lines of work converge toward close articulation between reasoning traceability, human control, and trust building. Temporal neuro-symbolic approaches, by making explicit rules, constraints, and inference trajectories, offer concrete mechanisms to support inspection, contestation, and critical analysis of decisions produced by complex systems, without being limited to local or purely performative justification (Doshi-Velez and Kim 2017; Aditya et al. 2023; Zhu et al. 2025).

These different mechanisms can be interpreted as successive steps of an abstraction chain, ranging from raw continuous signals to high-level explanations usable by a human. Figure 12 synthesizes this progression, by distinguishing levels of signal, event, symbolic structure, logical rule, and explanatory causal narrative.

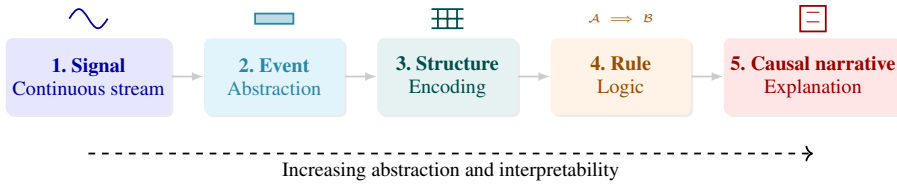


Figure 12. Temporal abstraction chain: from raw signals to an explanatory causal narrative.

Evaluation of temporal reasoning

Evaluation of temporal reasoning constitutes a central and still largely open problem in artificial intelligence. Unlike static tasks, where model quality can be estimated from correctness of point-wise predictions, temporal reasoning involves extended inference trajectories, non-local dependencies, and evolving constraints. A correct prediction at a given instant can thus mask global inconsistencies, causal violations, or reasoning relying on statistical shortcuts rather than valid structures (Chen et al. 2024; Bazaga et al. 2025).

This specificity challenges evaluation practices from classical supervised learning. Usual metrics, such as accuracy or Hits@k, certify local agreement but remain blind to global coherence of the decision trajectory over time (Marín 2025). Moreover, by treating time as a simple ranking problem, these measures fail to capture essential metric properties such as temporal distance or compliance with explicit deadlines (Xiong et al. 2024). In these contexts, these metrics can thus validate interpolation behaviors over observed periods, while masking the model's inability to generalize to future distributions or unseen scenarios (Lorello et al. 2025b).

A central challenge in temporal reasoning evaluation thus lies in shifting from result to trajectory. A model can correctly predict a final event while systematically violating the causal order of intermediate events that led to this result, a failure that remains invisible when only the final output is evaluated. Evaluating a system therefore no longer consists only of measuring the quality of a final output, but of analyzing the chain of intermediate inferences that lead to it. This perspective highlights reasoning shortcut phenomena, in which models exploit superficial correlations or unstable statistical regularities to produce locally plausible but globally inconsistent predictions (Marín

2025). The absence of mechanisms to inspect these trajectories makes these failures difficult to detect with classical metrics.

Recent works show that certain causal dependencies can extend over very long horizons, far exceeding contextual windows used by classical sequential architectures. For example, in ONSEP, apparently innocuous events occurring several months before the target event are identified as determinant causes, illustrating the failure of local attention mechanisms and the necessity of explicitly non-local reasoning (Yu et al. 2024).

Several complementary dimensions are necessary to characterize the validity of temporal reasoning. Accuracy remains a minimal condition, but it must be complemented by coherence criteria, evaluating compliance with explicit orderings, delays, and causal dependencies (Yu 2025). Generalization constitutes another critical dimension, particularly in non-stationary environments, where distributions evolve over time and render interpolation performance uninformative (Lorello et al. 2025b). Robustness, finally, aims to measure reasoning stability in the face of marginal perturbations of events or signals, and to detect situations close to critical violations (Donzé and Maler 2010).

Existing works propose various metrics to apprehend these dimensions, but none emerges as a unified standard. Logical metrics evaluate satisfaction or violation of formal properties, while quantitative metrics introduce degrees of satisfaction enabling ordering behaviors according to their proximity to violation (Donzé and Maler 2010). Other approaches, based on self-reflection, constrain the model to make explicit an internal chronology then confront its reasoning trace with this chronology (Bazaga et al. 2025). This confrontation reveals internal contradictions, such as confusions between simultaneity and succession, that remain invisible when only the final output is evaluated (Bazaga et al. 2025). These verification mechanisms highlight fragilities that remain undetectable when evaluation is limited to final output, by revealing gaps between statistical regularities learned by the model and the logic actually required by the problem.

Evaluation of temporal reasoning also raises explainability challenges. An explanation can be formally correct while remaining difficult to interpret when the inference trajectory becomes too long or too abstract. Empirical studies show that automatic metrics tend to overestimate explanation quality compared to human judgments, particularly when formal coherence masks reasoning that is implausible or difficult to follow (Hoffman et al. 2019; Huang et al. 2025). This observation emphasizes the necessity of integrating user-centered criteria into evaluation of temporal reasoning systems.

To go beyond binary ranking metrics, continuous measures specific to intervals have been introduced. The aeIOU (*average error Intersection over Union*) quantifies physical overlap between predicted interval and ground truth, while TAC (*Temporal Accuracy*) specifically penalizes errors on start and end boundaries (Figure 13) (Xiong et al. 2024). Furthermore, global evaluation is often formalized as multi-objective optimization, seeking a Pareto front between predictive performance and logical coherence, to identify models that do not sacrifice validity for local accuracy (Figure 14) (Marconato et al. 2024).

In this context, evaluation can no longer be considered a secondary or purely quantitative step. It becomes a normative instrument implicitly defining what it means to reason correctly in time.

1. Context and problem statement

This example, taken from the study of the TEILP model (Xiong et al. 2024), aims to predict the validity period of a political relation, rather than a single point in time.

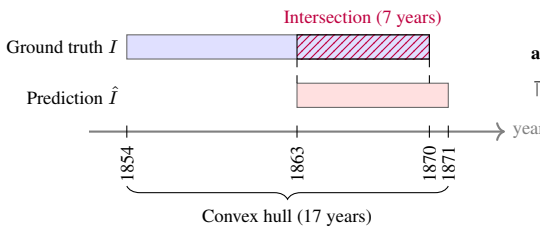
- Query: $(\textit{David Davis}, \textit{isAffiliatedTo}, \textit{Republican party}, ?)$
- Ground truth (I): $[1854, 1870]$ (actual duration of the affiliation).
- Prediction ($\hat{I}$): $[1863, 1871]$ (generated by the model).

A classical ranking-based evaluation (timestamp ranking) would fail to capture the partial quality of this answer. Although the boundaries are not exact (which would be heavily penalized by a binary ranking), the prediction overlaps with a significant portion of the ground truth.

2. Analysis using continuous metrics

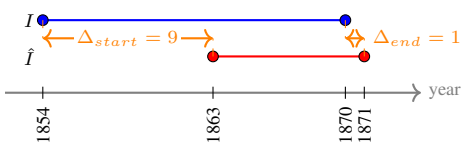
Continuous metrics make it possible to quantify the structural proximity between intervals.

A. Overlap (aeIOU)



aeIOU: Volume ratio $\frac{|I \cap \hat{I}|}{|ConvHull|} = \frac{7}{17} \approx \mathbf{0.41}$

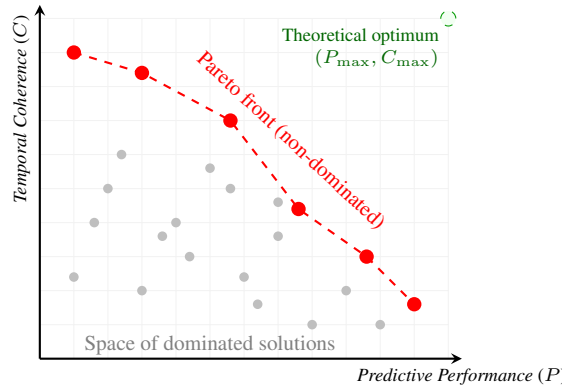
B. Boundary Accuracy (TAC)



TAC: Tolerance to delays $\frac{1}{2}\left(\frac{1}{1+9} + \frac{1}{1+1}\right) = \frac{1}{2}(0.1 + 0.5) = \mathbf{0.30}$

Conclusion: While aeIOU (0.41) emphasizes the strong central overlap, TAC (0.30) heavily penalizes the large delay at the start (9 years), illustrating the complementarity of these two measures for characterizing the validity of temporal reasoning.

Figure 13. Illustration of continuous evaluation of temporal predictions: aeIOU vs. TAC comparison



Analysis of the multi-objective trade-off.
The evaluation of temporal neuro-symbolic models involves a joint consideration of predictive performance (horizontal axis) and temporal coherence, understood as the semantic and logical consistency of the inferred temporal structures (vertical axis). The plot above represents different model configurations in this two-dimensional evaluation space.
Each point corresponds to a candidate model, with gray points denoting dominated configurations that can be simultaneously improved along both dimensions. In contrast, the red points form the Pareto front, characterizing the set of non-dominated solutions. Along this frontier, any further gain in predictive performance necessarily induces a degradation of temporal coherence, and vice versa, making explicit the intrinsic trade-off between statistical accuracy and symbolic consistency.
The theoretical optimum $(P_{\max}, C_{\max})$, indicated in green, represents an ideal but generally unattainable target. In practice, this limitation is often explained by *reasoning shortcuts*, whereby models maximize predictive scores by exploiting spurious statistical correlations at the expense of coherent temporal reasoning (Marconato et al. 2024). Evaluation should therefore focus not on identifying a single optimal point, but on characterizing admissible trade-offs that satisfy application-specific constraints.

Figure 14. Characterizing the trade-off between predictive performance and temporal coherence via the Pareto front.

Synthesis and roadmap for temporal neuro-symbolic reasoning

Analysis of works presented in this survey highlights a cross-cutting finding. Difficulties of temporal reasoning do not solely stem from lack of computational capacity or optimization, but from structural constraints linked to the cumulative and non-local nature of time. Combinatorial explosion induced by explicit instantiation of dependencies limits scaling of symbolic approaches (Mukherji et al. 2025), while probabilistic or differentiable methods face proliferation of proof or explanation trajectories that makes exact computation rapidly intractable and imposes approximations at the cost of completeness loss (Huang et al. 2021). These limitations are accentuated by difficulties of robust integration between discrete rules and continuous learning, which favor

appearance of reasoning shortcuts and explanatory gaps between invoked rules and actually operating mechanisms (Marconato et al. 2024; Marín 2025).

This global diagnosis motivates the necessity of going beyond fragmented reading of existing contributions and adopting explicit structuring of neuro-symbolic reasoning, not in terms of isolated techniques, but in the form of a progressive roadmap linking algorithmic capabilities, evaluation criteria, and control mechanisms.

On the methodological level, the survey highlights persistent fragmentation of formalisms and experimental practices, which complicates neuro-symbolic integration by imposing ad hoc translations between heterogeneous representations (Yu 2025). More generally, the absence of a fully stabilized unifying formalism limits comparability of approaches and the possibility of formulating general theoretical guarantees concerning validity, robustness, or explainability of produced reasoning (Yu 2025). These difficulties appear particularly clearly in evaluation. Existing benchmarks predominantly favor interpolation scenarios over observed periods, which tends to undersample regimes where long-term inconsistencies and extrapolation degradations emerge (Chen et al. 2024). Studies on explainability moreover show that formal correctness of an explanation guarantees neither its readability nor its cognitive relevance, emphasizing the necessity of going beyond purely automatic metrics (Hoffman et al. 2019; Huang et al. 2025).

These observations lead to questioning usual criteria of reasoning validity. Point-wise predictive performance often appears insufficient to qualify reasoning as correct when explicit constraints can be violated despite good final answers (Marín 2025). Several recent works indeed show that evaluation based solely on final output can mask inconsistent or opportunistic inferential trajectories. In response, certain approaches explicitly analyze intermediate reasoning traces and confront produced chronologies to verify coherence of steps leading to the decision (Marconato et al. 2024; Bazaga et al. 2025).

This perspective fits within approaches that recommend explicitly evaluating coherence (order, distance, delays) rather than reducing time to a simple ranking problem (Xiong et al. 2024), and within protocols that make inconsistencies observable through construction and confrontation of intermediate chronologies (Bazaga et al. 2025). It unfolds in increasing maturity levels, linking algorithmic capabilities, evaluation requirements, and control mechanisms, synthesized in Table 2.

In the short term, the central challenge is to make reasoning observable and auditable, by relying on interpretable intermediate representations and quantitative robustness measures (Donzé and Maler 2010). In the medium term, structural integration of constraints and abductive correction enable shifting from a posteriori verification to active reasoning control (Liang et al. 2025). In the long term, adaptation under non-stationarity and incremental rule revision constitute key directions to support sustainability of deployed systems, particularly when regularities evolve over time (Garcez and Lamb 2020; Lorello et al. 2025b).

This synthesis positions temporal neuro-symbolic reasoning as a design object in its own right, situated at the intersection of logic, learning, and governance (Garcez and Lamb 2020). It thus prepares the general conclusion, in which implications of

this roadmap for reliability, auditability, and accountability of autonomous systems are discussed.

Table 2. Roadmap for temporal neuro-symbolic reasoning: co-evolution of evaluation criteria, benchmarks, architectures, and control capabilities as temporal reasoning becomes a central design object.

Horizon	What to measure	Required benchmarks	What to build	What this enables
Short term *Foundations*	Verify existence of actual temporal reasoning (not simple exploitation of static correlations). Elementary counterfactual analyses.	Controlled, synthetic or semi-real datasets, integrating explicit temporal perturbations (deletion, permutation, event delays).	Separable chain: data $\rightarrow$ events $\rightarrow$ rules. Isolated and a posteriori inspectable temporal reasoning.	Understand and audit a decision. Distinguish perception error from reasoning error.
Medium term *Integration*	Measure satisfaction or violation of temporal constraints (orderings, delays, causal dependencies).	Constraint-oriented benchmarks, annotated with explicit temporal rules, including deliberately inconsistent cases.	Architectures directly integrating temporal and logical constraints into inference. Detection and abductive correction.	Actively prevent inconsistent or dangerous decisions. Human supervision based on explicit rules.
Long term *Autonomy*	Evaluate reasoning stability under non-stationarity (drift of data, rules, and practices).	Dynamic and evolving benchmarks, integrating rule changes and controlled temporal breaks.	Self-regulated systems capable of revising, creating, or abandoning temporal rules without knowledge loss.	Deploy sustainable systems, evolutionary yet auditable, explainable, and accountable.

This structuring highlights a conditional progression, in which each maturity level constitutes a prerequisite for the next. It formalizes the transition from a posteriori auditability toward active control, then toward regulated autonomy, by aligning architectures, benchmarks, and governance requirements.

Conclusion

This survey has analyzed approaches dedicated to temporal neuro-symbolic reasoning, from logical formalisms to learning architectures explicitly integrating constraints. All examined works show that accounting for time cannot be reduced to simple sequential data processing, but requires reasoning about event trajectories, non-local dependencies, and evolving constraints.

The study highlights a progressive shift in objectives assigned to temporal reasoning systems. Point-wise predictive performance proves insufficient when not accompanied by explicit guarantees concerning global reasoning coherence. Similarly, explainability cannot be limited to local a posteriori analyses, but must enable inspection and verification of inference trajectories over extended horizons. What fundamentally distinguishes temporal neuro-symbolic approaches thus lies not only in constraint

incorporation, but in explicit representation and inspection of inference trajectories, enabling reasoning verifiability and auditability by construction.

By articulating statistical learning, symbolic representations, and explicit temporal constraints, these temporal neuro-symbolic approaches offer a particularly suited perspective for designing systems whose reasoning is controllable, verifiable, and auditable by construction. However, existing works indicate that these guarantees can only be obtained at the cost of explicit methodological structuring, both at the architecture and evaluation protocol levels. This notably requires defining shared criteria to characterize and evaluate temporal reasoning validity beyond classical predictive metrics. Rather than proposing new metrics, this survey identifies operationalizable evaluation dimensions such as trajectory coherence, robustness to perturbations, or explicit specification satisfaction that remain currently underrepresented in existing benchmarks.

The proposed roadmap synthesizes this structuring by identifying successive maturity levels, linking algorithmic capabilities, evaluation criteria, and control mechanisms. It formalizes a conditional progression, from reasoning auditability to controlled adaptation in dynamic environments, and provides a framework to guide future developments in the field. Without claiming exhaustiveness or optimality, this roadmap seeks to offer a possible structuring of temporal neuro-symbolic reasoning challenges, based on regularities and recurrent limitations highlighted in the literature.

Ultimately, temporal neuro-symbolic reasoning asserts itself as a structuring methodological path for designing artificial intelligence systems adapted to dynamic contexts. By placing temporal coherence, verifiability, auditability, and governance at the core of design, it opens concrete perspectives to reconcile adaptability, reliability, and trust requirements in domains where time constitutes a determinant dimension.

Appendix: mapping of temporal neuro-symbolic models

This appendix proposes a structured synthesis of main neuro-symbolic models using temporal reasoning, organized according to the integration mode between neural and formal components. The tables gather and position approaches cited in the body of the text by distinguishing (i) architectures where logic directly structures or constrains neural learning (Logic $\rightarrow$ Network), (ii) those that extract a posteriori rules, automata, or symbolic representations from neural models (Network $\rightarrow$ Logic), and (iii) bidirectional or post-hoc approaches combining iterative learning, explainability, and reliability guarantees.

This synthesis does not claim to be exhaustive. It rather aims to provide a representative sampling of paradigms, mechanisms, and uses identified in the literature, selected for their illustrative or structuring character with regard to axes analyzed in the manuscript. The chronological organization adopted within each category highlights the progressive evolution of neuro-symbolic mechanisms, from weak or differentiable integration forms toward explicit symbolic structures, tighter feedback loops, and mechanisms dedicated to explainability and validation. This appendix thus constitutes a cross-cutting reading tool, intended to facilitate comparison of paradigms, clarify their

respective functional role, and complement the conceptual analysis developed in the main body of the manuscript.

Table 3. Logic → Network for temporal neuro-symbolic reasoning

Model (Ref)	Year	Temporal Neuro-symbolic Mechanism	Main Application
STLnet (Ma et al. 2020)	2020	Teacher–Student architecture. The *teacher* projects predictions into a subspace satisfying STL constraints; the student learns through this distillation process.	Time-series prediction (Smart City)
T-LEAF (Xie et al. 2021)	2021	LTLf formulas are compiled into finite automata that constrain and guide the learning of sequential neural models.	Sequence learning (Robotics)
DynTKG (Liu et al. 2025)	2025	Causal and temporal constraints encoded as soft regularization and rule-guided	distillation mechanisms
NeSyA (Manginas et al. 2025)	2025	Symbolic Finite Automata (SFA) integrated into a neural network. Differentiable inference computes the acceptance probability of a sequence.	Sequence classification (Events)

Table 4. Network → Logic approaches for temporal neuro-symbolic reasoning: rule induction, extraction, and interpretability.

Model (Ref)	Year	Extraction / Induction Mechanism	Main Application
		Rule Induction & Neuro-symbolic Pipelines	
TILP Xiong et al. (2023)	2023	Induction of temporal rules constrained by random walks (constrained random walks).	Explainable temporal forecasting
TECHS (Lin et al. 2023)	2023	Differentiable temporal reasoning framework combining propositional and first-order reasoning, with post-hoc induction of temporal logical rules via attention-based mechanisms.	Time-series forecasting
TEILP (Xiong et al. 2024)	2024	Extension of TILP with conditional probability density modeling to predict the exact *time*.	Temporal prediction (Intervals)
LLM-DA (Wang et al. 2024)	2024	Extraction of temporal rules using LLMs and dynamic adaptation to new data.	Explainable TKG reasoning
MVFF Xu et al. (2024)	2024	Multi-view fusion framework combining temporal rule patterns with embedding-based representations for temporal knowledge graph reasoning.	TKG completion
ONSEP (Yu et al. 2024)	2024	Dynamic causal rule mining via LLMs over real-time data streams.	Online event prediction
PRlogic (Zeng et al. 2025)	2025	Context-aware logical association network for multi-hop temporal and event-based path inference.	Multi-hop temporal and event-based path reasoning
T-CPDL (Yu 2025)	2025	Probabilistic causal temporal description logic to structure agent reasoning (Logic-RAG).	Structured reasoning agents
		Mechanistic Interpretability & Automata Extraction	
DT-STL (Bombara et al. 2016)	2016	Extraction of Signal Temporal Logic (STL) formulas via decision trees to classify time series.	Explainable signal classification
Extract RNN (Weiss et al. 2018)	2018	Active extraction of finite automata (DFA) from RNN hidden-state dynamics using the L* algorithm.	Formal verification of RNNs

Table 5. Bidirectional and **Post-hoc** approaches for temporal neuro-symbolic reasoning, explainability, and reliability.

Model (Ref)	Year	Temporal Neuro-symbolic Mechanism	Main Application
		Bidirectional Learning Loops (Iterative)	
TLogic (Liu et al. 2022)	2022	Extraction of temporal logical rules based on *temporal random walks* over the graph.	Explainable temporal forecasting
ILR-IR (Mei et al. 2024)	2024	Iterative extraction of interpretable temporal logical rules from relation paths, combined with embedding-based scoring to support inductive reasoning over temporal knowledge graphs.	Adaptive temporal learning
		Explainability & Reliability (Temporal NeSy)	
TMLN (David et al. 2022)	2022	Temporal extension of Markov Logic Networks to model uncertainty and dynamics through weighted temporal formulas.	Temporal probabilistic inference
LTLZinc (Lorello et al. 2025b)	2025	Data generation and evaluation framework based on Linear Temporal Logic (LTL) specifications for neuro-symbolic learning.	Benchmarking and validation
TISER (Bazaga et al. 2025)	2025	Explicit construction and verification of temporal timelines by LLMs to correct and refine reasoning at test time.	Temporal reasoning in LLMs

References

Aditya D, Mukherji K, Balasubramanian S, Chaudhary A and Shakarian P (2023) PyReason: Software for open world temporal logic. DOI:10.48550/arXiv.2302.13482.

Badreddine S, Garcez Ad, Serafini L and Spranger M (2022) Logic tensor networks. *Elsevier* 303: 103649. DOI:10.1016/j.artint.2021.103649.

Barceló P (2013) Querying graph databases. In: *Proceedings of the 32nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '13. New York, NY, USA: Association for Computing Machinery, pp. 175–188. DOI:10.1145/2463664.2465216.

Bazaga A, Blloshmi R, Byrne B and Gispert Ad (2025) Learning to reason over time: Timeline self-reflection for improved temporal reasoning in language models. DOI:10.48550/arXiv.2504.05258.

Bombara G and Belta C (2021) Offline and online learning of signal temporal logic formulae using decision trees. *ACM Transactions on Cyber-Physical Systems* 5(3): 1–23. DOI:10.1145/3433994.

Bombara G, Vasile CI, Penedo F, Yasuoka H and Belta C (2016) A Decision Tree Approach to Data Classification using Signal Temporal Logic. In: *Proceedings of the 19th International Conference on Hybrid Systems: Computation and Control*. Vienna Austria: ACM. ISBN 978-1-4503-3955-1, pp. 1–10. DOI:10.1145/2883817.2883843. URL https://dl.acm.org/doi/10.1145/2883817.2883843.

Chen J, Ren J, Ding W, Ouyang H, Hu W and Qu Y (2025a) Conflict detection for temporal knowledge graphs:a fast constraint mining algorithm and new benchmarks. DOI:10.48550/arXiv.2312.11053.

Chen K, Song X, Wang Y, Gao L, Li A, Zhao X, Zhou B and Xie Y (2025b) LLM-DR: A novel LLM-aided diffusion model for rule generation on temporal knowledge graphs. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39. AAAI Press, pp. 11481–11489. DOI:10.1609/aaai.v39i11.33249.

Chen K, Wang Y, Li Y, Li A, Yu H and Song X (2024) A unified temporal knowledge graph reasoning model towards interpolation and extrapolation. DOI:10.48550/arXiv.2405.18106.

Clarke E, Grumberg O and Peled D (2001) Model checking. *Springer* .

Conradie W, Della Monica D, Muñoz-Velasco E and Sciavicco G (2020) An approach to fuzzy modal logic of time intervals. In: *ECAI 2020: 24th European Conference on Artificial Intelligence*, *Frontiers in Artificial Intelligence and Applications*, volume 325. IOS Press, pp. 678–685. DOI:10.3233/FAIA200156.

David V, Fournier-S'niehotta R and Travers N (2022) Parameterisation of Reasoning on Temporal Markov Logic Networks. DOI:10.48550/arXiv.2211.16414. URL http://arxiv.org/abs/2211.16414. ArXiv:2211.16414 [cs].

Dean T (1989) Using temporal hierarchies to efficiently maintain large temporal databases. *Journal of the ACM (JACM)* 36(4): 687–718. DOI:10.1145/76359.76360.

Dechter R, Meiri I and Pearl J (1990) Temporal constraint networks. *Elsevier* .

DeLong LN, Mir RF and Fleuriot JD (2025) Neurosymbolic AI for reasoning over knowledge graphs: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 36(5): 7822–7842. DOI:10.1109/TNNLS.2024.3420218.

Dixon C (2007) Tractable temporal reasoning. *IJCAI* .

Donzé A and Maler O (2010) Robust satisfaction of temporal logic over real-valued signals. In: Chatterjee K and Henzinger TA (eds.) *Formal Modeling and Analysis of Timed Systems*, volume 6246. Springer Berlin Heidelberg. ISBN 978-3-642-15296-2 978-3-642-15297-9, pp. 92–106. DOI:10.1007/978-3-642-15297-9_9. Series Title: Lecture Notes in Computer Science.

Doshi-Velez F and Kim B (2017) Towards a rigorous science of interpretable machine learning. DOI:10.48550/arXiv.1702.08608.

Frigeri A, Pasquale L and Spoletini P (2012) Fuzzy time in LTL. DOI:10.48550/arXiv.1203.6278.

Garcez Ad and Lamb LC (2020) Neurosymbolic AI: The 3rd wave. DOI:10.48550/arXiv.2012.05876.

Halpern JY and Shoham Y (1991) A propositional modal logic of time intervals. *Journal of the ACM (JACM)* 38(4): 935–962. DOI:10.1145/115234.115351.

Han L and Srivastava MB (2024) An empirical evaluation of neural and neuro-symbolic approaches to real-time multimodal complex event detection. DOI:10.48550/arXiv.2402.11403.

Hoffman RR, Mueller ST, Klein G and Litman J (2019) Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* .

Hou P (2012) Quantified differential temporal dynamic logic for verifying properties of distributed hybrid systems. DOI:10.48550/arXiv.1207.2531.

Huang J, Li Z, Chen B, Samel K, Naik M, Song L and Si X (2021) Scallop: From probabilistic deductive databases to scalable differentiable reasoning. In: *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., pp. 25134–25145.

Huang S, Wang H, Li P and Chen Z (2025) Document-level future event prediction integrating event knowledge graph and LLM temporal reasoning. *Electronics* 14(19): 3827. DOI: 10.3390/electronics14193827.

Jensen CS, Dyreson CE, Böhlen M, Clifford J, Elmasri R, Gadia SK, Grandi F, Hayes P, Jajodia S, KÄfer W, Kline N, Lorentzos N, Mitsopoulos Y, Montanari A, Nonen D, Peressi E, Pernici B, Roddick JF, Sarda NL, Scalas MR, Segev A, Snodgrass RT, Soo MD, Tansel A, Tiberio P and Wiederhold G (1998) The consensus glossary of temporal database concepts — february 1998 version. In: Etzion O, Jajodia S and Sripada S (eds.) *Temporal Databases: Research and Practice*, volume 1399. Springer Berlin Heidelberg. ISBN 978-3-540-64519-1 978-3-540-69799-2, pp. 367–405. DOI:10.1007/BFb0053710. Series Title: Lecture Notes in Computer Science.

Kamide N (2015) Inconsistency and sequentiality in LTL:. In: *Proceedings of the International Conference on Agents and Artificial Intelligence*. SCITEPRESS - Science and and Technology Publications. ISBN 978-989-758-073-4 978-989-758-074-1, pp. 46–54. DOI:10.5220/0005180800460054.

Kamide N (2017) Paraconsistent sequential linear-time temporal logic: Combining paraconsistency and sequentiality in temporal reasoning. *Reports on Mathematical Logic* 52. DOI:10.4467/20842589RM.17.001.7139.

Klement EP, Mesiar R and Pap E (2000) *Triangular Norms*, *Trends in Logic*, volume 8. Dordrecht: Springer Netherlands. ISBN 978-90-481-5507-1 978-94-015-9540-7. DOI:10.1007/978-94-015-9540-7. URL http://link.springer.com/10.1007/978-94-015-9540-7.

Knez T and Žitnik S (2023) Event-centric temporal knowledge graph construction: A survey. *Mathematics* 11(23): 4852. DOI:10.3390/math11234852.

Krieken Ev, Acar E and Harmelen Fv (2022) Analyzing Differentiable Fuzzy Logic Operators. *Artificial Intelligence* 302: 103602. DOI:10.1016/j.artint.2021.103602. URL http://arxiv.org/abs/2002.06100. ArXiv:2002.06100 [cs].

Kummari DN, Challa SR, Pamisetty V, Motamary S and Meda R (2025) Unifying temporal reasoning and agentic machine learning: A framework for proactive fault detection in dynamic, data-intensive environments. *Metallurgical and Materials Engineering* 31.

Lee JH, Sioutis M, Ahrens K, Alirezaie M, Kerzel M and Wermter S (2023) Neuro-symbolic spatio-temporal reasoning. DOI:10.48550/arXiv.2211.15566.

Leung K, Aréchiga N and Pavone M (2021) Backpropagation through signal temporal logic specifications: Infusing logical structure into gradient-based methods. DOI:10.48550/arXiv.2008.00097.

Liang F, Zeng W, Zhao R and Zhao X (2025) NeSTR: A neuro-symbolic abductive framework for temporal reasoning in large language models. DOI:10.48550/arXiv.2512.07218.

Lichtenstein O, Pnueli A and Zuck L (1985) The glory of the past. In: Parikh R (ed.) *Logics of Programs*, volume 193. Springer Berlin Heidelberg. ISBN 978-3-540-15648-2 978-3-540-39527-0, pp. 196–218. DOI:10.1007/3-540-15648-8_16. Series Title: Lecture Notes in Computer Science.

Lin Q, Liu J, Mao R, Xu F and Cambria E (2023) TECHS: Temporal Logical Graph Networks for Explainable Extrapolation Reasoning. In: *Proceedings of the 61st Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, pp. 1281–1293. DOI:10.18653/v1/2023.acl-long.71. URL https://aclanthology.org/2023.acl-long.71.

Liu Q, Feng S and Huang M (2025) Dynamic subgraph pruning and causal-aware knowledge distillation for temporal knowledge graphs. *J. King Saud Univ. Comput. Inf. Sci.* 37(5): 96. DOI:10.1007/s44443-025-00105-3.

Liu Y, Ma Y, Hildebrandt M, Joblin M and Tresp V (2022) TLogic: Temporal logical rules for explainable link forecasting on temporal knowledge graphs. DOI:10.48550/arXiv.2112.08025.

Lorello LS, Lippi M and Melacci S (2025a) A neuro-symbolic framework for sequence classification with relational and temporal knowledge. DOI:10.48550/arXiv.2505.05106.

Lorello LS, Manginas N, Lippi M and Melacci S (2025b) LTLZinc: a benchmarking framework for continual learning and neuro-symbolic temporal reasoning. DOI:10.48550/arXiv.2507.17482.

Lu Z, Liu J, Augusto JC and Wang H (2010) A linguistic truth-valued temporal reasoning formalism and its implementation. In: Bramer M, Ellis R and Petridis M (eds.) *Research and Development in Intelligent Systems XXVI*. Springer London. ISBN 978-1-84882-982-4 978-1-84882-983-1, pp. 305–310. DOI:10.1007/978-1-84882-983-1_23.

Ma M, Gao J, Feng L and Stankovic J (2020) STLnet: Signal temporal logic enforced multivariate recurrent neural networks. In: *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., pp. 1046–1056.

Manginas N, Paliouras G and Raedt LD (2025) NeSyA: Neurosymbolic Automata. DOI:10.48550/arXiv.2412.07331. URL http://arxiv.org/abs/2412.07331. ArXiv:2412.07331 [cs].

Marconato E, Bortolotti S, Krieken Ev, Vergari A, Passerini A and Teso S (2024) BEARS make neuro-symbolic models aware of their reasoning shortcuts. DOI:10.48550/arXiv.2402.12240.

Maria ED, Despeyroux J and Felty A (2014) A logical framework for systems biology. DOI:10.48550/arXiv.1404.5439.

Marín J (2025) Empirical characterization of temporal constraint processing in LLMs. DOI:10.48550/arXiv.2511.10654.

Mei X, Yang L, Jiang Z, Cai X, Gao D, Han J and Pan S (2024) An Inductive Reasoning Model based on Interpretable Logical Rules over temporal knowledge graph. *Neural Networks* 174: 106219. DOI:10.1016/j.neunet.2024.106219. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608024001436.

Miller T (2018) Explanation in artificial intelligence: Insights from the social sciences. DOI:10.48550/arXiv.1706.07269.

Mukherji K, Patil JM, Aditya D, Shakarian P, Parkar D, Pokala L, Dorman C and Simari GI (2025) Lattice annotated temporal (LAT) logic for non-markovian reasoning. DOI:10.48550/arXiv.2509.02958.

Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE et al. (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372: n71. DOI:10.1136/bmj.n71.

Panayiotopoulos T and Gergatsoulis M (2002) Intelligent Information Processing using TRLi. *Expert Systems* 19(4): 192–209. DOI:10.1111/1468-0394.00205.

Pustejovsky J, Ingria R, Sauri´ R, O JEC, Littman J, Gaizauskas R, Setzer A, Katz G and Mani I (2005) The specification language TimeML. In: Mani I, Pustejovsky J and Gaizauskas R (eds.) *The Language Of Time*. Oxford University PressOxford. ISBN 978-0-19-926853-5 978-1-383-04117-0, pp. 545–558. DOI:10.1093/oso/9780199268535.003.0031.

Riegel R, Gray A, Luus F, Khan N, Makondo N, Akhalwaya IY, Qian H, Fagin R, Barahona F, Sharma U, Ikbal S, Karanam H, Neelam S, Likhyani A and Srivastava S (2020) Logical neural networks. DOI:10.48550/arXiv.2006.13155.

Roşu G and Bensalem S (2006) Allen linear (interval) temporal logic – translation to LTL and monitor synthesis. In: Ball T and Jones RB (eds.) *Computer Aided Verification*, volume 4144. Springer Berlin Heidelberg. ISBN 978-3-540-37406-0 978-3-540-37411-4, pp. 263–277. DOI:10.1007/11817963_25. Series Title: Lecture Notes in Computer Science.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. DOI:10.48550/arXiv.1811.10154.

Shoaff W (1993) Path consistency in a network of non-convex intervals. *IJCAI* .

Singh I, Kaur N, Gaur G and Mausam (2023) NeuSTIP: A neuro-symbolic model for link and time prediction in temporal knowledge graphs. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 4497–4516. DOI:10.18653/v1/2023.emnlp-main.274.

Sistla AP and Clarke EM (1985) The complexity of propositional linear temporal logics. *Journal of the ACM* 32(3): 733–749. DOI:10.1145/3828.3831.

Tang W, Nuamah K and Belle V (2026) LTLBench: Towards benchmarks for evaluating temporal reasoning in large language models. DOI:10.48550/arXiv.2407.05434.

Tonekaboni S, Joshi S, Campbell K, Duvenaud D and Goldenberg A (2020) What went wrong and when? instance-wise feature importance for time-series models. DOI:10.48550/arXiv.2003.02821.

Wang J, Sun K, Luo L, Wei W, Hu Y, Liew AWC, Pan S and Yin B (2024) Large Language Models-guided Dynamic Adaptation for Temporal Knowledge Graph Reasoning. DOI:10.48550/arXiv.2405.14170. URL http://arxiv.org/abs/2405.14170. ArXiv:2405.14170 [cs].

Weiss G, Goldberg Y and Yahav E (2018) Extracting Automata from Recurrent Neural Networks Using Queries and Counterexamples. *Proceedings of the 35 th International Conference on Machine Learning* .

Wolter F and Zakharyaschev M (2000) Spatio-temporal representation and reasoning based on RCC-8. In: *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR 2000)*. San Francisco, CA, USA: Morgan Kaufmann, pp. 3–14.

Xie Y, Zhou F and Soh H (2021) Embedding Symbolic Temporal Knowledge into Deep Sequential Models. DOI:10.48550/arXiv.2101.11981. URL http://arxiv.org/abs/2101.11981. ArXiv:2101.11981 [cs].

Xiong S, Yang Y, Fekri F and Kerce JC (2023) TILP: Differentiable learning of temporal logical rules on knowledge graphs. DOI:10.48550/arXiv.2402.12309.

Xiong S, Yang Y, Payani A, Kerce JC and Fekri F (2024) TEILP: Time prediction over knowledge graphs via logical reasoning. DOI:10.48550/arXiv.2312.15816.

Xu H, Bao J, Li H, He C and Chen F (2024) A Multi-View Temporal Knowledge Graph Reasoning Framework with Interpretable Logic Rules and Feature Fusion. *Electronics* 13(4): 742. DOI:10.3390/electronics13040742. URL https://www.mdpi.com/2079-9292/13/4/742.

Yu HQ (2025) T-CPDL: A temporal causal probabilistic description logic for developing logic-RAG agent. DOI:10.48550/arXiv.2506.18559.

Yu X, Sun W, Li J, Liu K, Liu C and Tan J (2024) ONSEP: A novel online neural-symbolic framework for event prediction based on large language model. In: *Findings of the Association for Computational Linguistics ACL 2024*. Association for Computational Linguistics, pp. 6335–6350. DOI:10.18653/v1/2024.findings-acl.378.

Zeng Y, Hou X, Wang X and Li J (2025) Towards a Unified Temporal and Event Logic Paradigm for Multi-Hop Path Reasoning in Knowledge Graphs. *Electronics* 14(3): 516. DOI:10.3390/electronics14030516. URL https://www.mdpi.com/2079-9292/14/3/516.

Zhu Z, Liu H, He M and Luo S (2025) Right answer at the right time - temporal retrieval-augmented generation via graph summarization. DOI:10.48550/arXiv.2510.16715.