Journal Title
XX(X):2–68

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

A survey of neurosymbolic artificial intelligence: foundations, advances, and future trajectories

Otto Mättas[1], Priit Järv[1] and Tanel Tammet[1]

Abstract

Neurosymbolic artificial intelligence seeks to integrate the strengths of learning and symbolic reasoning to deliver systems that are effective, interpretable, reliable, and accountable. This survey compiles advances from 2020-2025, organized into four themes: performance, understandability, reliability, and ethics.

We treat an approach as neurosymbolic only when symbolic structures with well-defined semantics participate directly in learning or inference; retrieval or external tool use without such coupling is treated as adjacent context. Within this scope, we describe recurring interface patterns - in practical terms, the ways neural components consume, produce, or are constrained by symbolic representations and reasoning operators (e.g., programs/queries, constraints, or structured traces). We use these patterns to organize and compare approaches across the functional roles they target in AI systems (perception, knowledge, reasoning, planning/control, and oversight).

Keywords

neurosymbolic artificial intelligence, survey, performance, understandability, reliability, ethics

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

[1]Tallinn University of Technology, Estonia

Corresponding author:
Otto Mättas, Applied Artificial Intelligence Group, Department of Software Science, School of Information Technologies, Ehitajate tee 5, 19086 Tallinn, Estonia
Email: otto.mattas@taltech.ee

Prepared using sagej.cls

Abbreviations

Abbreviation	Expansion
Venues and journals	
AAAI	Association for the Advancement of Artificial Intelligence
ACL	Association for Computational Linguistics (Annual Meeting)
AISTATS	International Conference on Artificial Intelligence and Statistics
EMNLP	Empirical Methods in Natural Language Processing
ICLR	International Conference on Learning Representations
ICML	International Conference on Machine Learning
IJCAI	International Joint Conference on Artificial Intelligence
NAACL	North American Chapter of the Association for Computational Linguistics
NeurIPS	Advances in Neural Information Processing Systems
PMLR	Proceedings of Machine Learning Research
WWW	The Web Conference
Methods and concepts	
AI	Artificial Intelligence
ASP	Answer Set Programming
DNN	Deep Neural Network
GNN	Graph Neural Network
HCI	Human-Computer Interaction
HIL	Human-in-the-Loop
ILP	Inductive Logic Programming
JEPA	Joint Embedding Predictive Architecture
KG	Knowledge Graph
KGQA	Knowledge Graph Question Answering
KR	Knowledge Representation
LLM	Large Language Model
MLN	Markov Logic Network
NAS	Neural Architecture Search
NeSy	Neurosymbolic
NLP	Natural Language Processing
QA	Question Answering
QALD	Question Answering over Linked Data
RAG	Retrieval-Augmented Generation
RL	Reinforcement Learning
SAT	Boolean Satisfiability
SMT	Satisfiability Modulo Theories
TD	Temporal-Difference
XAI	Explainable Artificial Intelligence

Introduction

Neurosymbolic artificial intelligence (NeSy AI) aims to combine the strengths of statistical learning with explicit knowledge and reasoning in order to build systems that are effective, interpretable, reliable, and accountable. We set the stage by briefly revisiting the symbolic and neural traditions and the case for integration, then articulating the problem this survey addresses, its scope and novelty, and the contributions it makes.

Background: Symbolic vs. Neural and the Case for Integration

Artificial intelligence has advanced through two intertwined traditions. In the symbolic lineage, early systems demonstrated task-oriented language interaction and structured manipulation within constrained domains (e.g., *ELIZA* and *SHRDLU*) (Weizenbaum, 1966; Winograd, 1971). This line of work also established long-running debates about representation, symbol grounding (in the classical sense of relating symbols to what they denote in the world; later in the paper, "grounding" also refers to retrieval/citation/tool grounding in modern language-model pipelines, which does not by itself imply semantic grounding or correctness guarantees), and the relationship between symbols and procedures (Moran, 1973; Cohen, 1983; Sloman et al., 1983). Subsequent critiques clarified limitations of purely symbolic accounts and emphasized the role of execution, procedures, and operational semantics (Touretzky & Minton, 1985; Dahlback, 1989; Russell, 1989). Empirical comparisons and early integration attempts in the late 1980s and 1990s explored how symbolic structure and neural learning can be combined in practice (Barnden, 1989; Mooney et al., 1989; Frixione & Spinelli, 1989). Case-based reasoning provided additional evidence that explicit representations and retrieval can support structured reasoning in applied settings (Ashley & Aleven, 1997; Rosa & Franeozo, 1999).

In parallel, the connectionist lineage established core learning principles and scalable representations, from early perceptrons and associative memory through backpropagation and modern deep learning (Rosenblatt, 1958; Hopfield, 1982; Rumelhart et al., 1986; Bengio et al., 2021). Subsequent work extended these capabilities to attention-based architectures and large-scale generative models that support broad task coverage (Vaswani et al., 2017; Ramesh et al., 2021). Neural systems paired with explicit search provide a reference point for how learning and planning can be composed in complex decision-making (Silver et al., 2016). Domain-specialized language models

illustrate both gains and limitations of purely neural approaches in knowledge-intensive settings (Shin et al., 2020; Jurafsky & Martin, 2025).

Viewed through a broader historical lens, the field alternates between symbolic and neural emphases, and position pieces argue for integration as a recurring response to limitations in either approach (Kautz, 2022). Surveys and textbooks motivate neurosymbolic systems as a way to combine learned percepts with explicit, inspectable representations and operators (Garcez et al., 2019; Russell & Norvig, 2020; Garcez & Lamb, 2020). Recent neurosymbolic surveys emphasize practical architectures, evaluation concerns, and the role of explicit KR in modern pipelines (Hitzler et al., 2022, 2024). Cognitive perspectives provide a functional motivation for pairing fast pattern recognition with deliberative reasoning (Kahneman, 2011; Laird et al., 2017). Human-centric perspectives emphasize systems that can explain, align, and collaborate rather than replace (Horvatić & Lipic, 2021).

Experiences from large-scale deployments illustrate both the potential and fragility of purely data-driven methods, reinforcing the need for auditable, knowledge-guided reasoning within AI pipelines (Strickland, 2019). Conceptual roadmaps and position pieces articulate why and how to integrate neural competence with symbolic structure to achieve generality with accountability (Marcus, 2020; Sheth et al., 2023a,b). Complementary perspectives emphasize system-level design choices and integration trade-offs (Sheth & Roy, 2024; Ganguly & Mukherjee, 2025). Together, these lines of evidence motivate a neurosymbolic agenda: retain the strengths of scalable learning, perception, and generation, while introducing explicit knowledge, inference, and control to support interpretation, transfer, and reliable decision-making across complex tasks.

Problem Statement: A Fragmented and Rapidly Evolving Landscape

Foundation models have accelerated rapidly, expanding the scope of tasks that can be addressed by learned systems while exposing new limitations and open questions about evaluation and control. Reports on frontier LLMs describe broad, cross-domain capability alongside uneven reliability and opaque failure modes, underscoring the need for principled assessment beyond benchmark-by-benchmark comparisons (Bubeck et al., 2023). At the same time, progress signals remain difficult to compare across subfields: evaluation measures differ, data and tasks shift, and there is no universally accepted quantitative lens for characterizing improvement. Recent work proposes technology-improvement-rate measurements using patent citation networks to quantify and compare

advancement across AI subdomains, but such instruments are only beginning to connect to practice-level evaluation in research benchmarks (Rezazadegan et al., 2024).

These dynamics also interact with ongoing challenges in reproducibility and reporting. Earlier audits quantified documentation gaps in empirical AI research (Gundersen & Kjensmo, 2018), and while community norms and tooling have improved since then, more recent analyses still find that software and experimental artifacts are provided unevenly and that reproduction can remain non-trivial in practice (Wolter et al., 2025). In neurosymbolic AI, the coupling between learned components and explicit knowledge/reasoning further increases sensitivity to implementation and documentation choices; recent systematic review practice has therefore used code availability as an explicit inclusion criterion (Colelough & Regli, 2025). We therefore pose the problem addressed by this survey as one of synthesis and normalization: to chart what the community is doing across rapidly evolving lines of work, to relate methods to common system functions and evaluation levers, and to consolidate practical, theme-oriented guidance that supports comparability, reproducibility, and decision-making across applications (Bubeck et al., 2023; Rezazadegan et al., 2024).

Novelty of this Synthesis and Scope

This survey emphasizes a goals-first perspective: we evaluate advances by their contribution across four recurring themes - performance, understandability, reliability, and ethics. The temporal scope focuses on 2020–2025 while using foundational anchors that illustrate the historical developments in the domain. We unify cross-domain results (NLP, vision, robotics, knowledge graphs, and autonomous systems) and illustrate end-to-end implications through a consistent mapping to system functions and evaluation measures.

Scope boundary (what counts as neurosymbolic evidence in this survey). We use a strict boundary rule to avoid conflating modern tool augmentation with symbolic reasoning. Throughout, we treat a method as neurosymbolic evidence only when it includes (i) an explicit symbolic representation with defined operators/semantics (e.g., logic/rules, executable programs, KGs with query/entailment operators, planners/controllers, SAT/SMT constraints, proof traces) and (ii) an explicit coupling where those operators constrain, check, or otherwise participate directly in training or inference. In contrast, retrieval-augmented generation, tool calling, and natural-language reasoning traces are treated as adjacent context unless they produce

typed/executable artifacts that are executed and/or checked by explicit operators (e.g., emitting a query/program and validating it against a KG/reasoner, or enforcing constraints via checking or shields).

What this survey adds beyond prior surveys. Compared to method- or domain-centered surveys, we emphasize an interface-centric characterization (Table 1) paired with an explicit evidence protocol (Section 11) so that breadth does not force overgeneralization. In particular, we treat costs and guarantees as first-class comparison dimensions, distinguish tool grounding (retrieval/citations/tool use) from correctness guarantees, and avoid rigid taxonomy assignments in favor of recurring interface patterns that can be instantiated across domains (Hitzler et al., 2022; Hamilton et al., 2024; Michel-Deletie & Sarker, 2025). In practical terms, those patterns are the ways neural components consume, produce, or are constrained by symbolic representations and reasoning operators (e.g., programs/queries, constraints, or structured traces).

Contributions

The paper offers:

(i) A theme-based compilation (2020–2025, with foundational anchors) organizing the literature by performance, understandability, reliability, and ethics.
(ii) A mapping from methods to system functions (perception, knowledge, reasoning, planning/control, oversight) with representative benchmarks and evaluation measures, summarized in Table 2.
(iii) A consolidated view of how papers *evaluate* each theme in practice, summarizing commonly reported evaluation measures, benchmarks, and reproducibility signals (e.g., availability of code/data and ablations when reported).
(iv) A cross-theme analysis of recurring system-design pitfalls when combining learning with symbolic representations and operators (e.g., cost vs. guarantees; grounding vs. correctness), described via interface patterns rather than rigid taxonomies.
(v) Future directions and open challenges, with concrete evaluation criteria and test considerations.

The summary matrix in Table 2 provides a consistent thread for mapping advances to goals, placing results within a practical system setting, and clarifying where knowledge and explanations originate.

Overview of the Paper

This subsection provides a roadmap for what follows. Section 9 details the evidence collection and synthesis protocol (source selection, tagging rules, and comparative synthesis) and introduces the summary matrix (Table 2). Section 14 presents the four core themes with a consistent micro-structure: Section 14 (Performant), Section 21 (Understandable), Section 25 (Reliable), and Section 28 (Ethical). Section 30 integrates these results into a system-oriented view, outlining architecture patterns and design imperatives with an application spotlight. Section 34 situates this work relative to prior surveys and discusses broader implications. Section 36 articulates future directions with evaluation criteria and test considerations. Finally, Section 37 offers a concise recap and outlook.

Methods of the Survey

We adopted a transparent, goals-oriented evidence synthesis to capture recent neurosymbolic advances and relate them to the four themes and system functions. The protocol balances breadth (cross-domain coverage) and depth (clear operationalization of constructs and evaluation).

Source Selection

We focused on peer-reviewed work from 2020–2025, supplemented with foundational anchors when essential for context. Sources included prominent AI venues (e.g., AAAI, IJCAI, NeurIPS, ICML/AISTATS via PMLR, ICLR, ACL/EMNLP/NAACL/Findings, WWW) and journals/publishers relevant to neurosymbolic AI and KR (e.g., Semantic Web (IOS Press), ACM, IEEE venues and journals, Nature and Nature Communications). We also considered relevant Springer publications and reputable preprints from arXiv when influential and cited by peer-reviewed work. Searches were performed over digital libraries and indexing services using combinations of terms referring to neurosymbolic integration (logic, KGs, differentiable reasoning, planning/control, verification, explainability, symbol grounding, tool grounding). Inclusion prioritized works that: (i) integrate neural and symbolic elements; (ii) provide empirical or formal evaluation; and (iii) report sufficient methodological detail for assessment. Exclusions included purely neural or purely symbolic works without a substantive integration point, or position papers lacking concrete contribution. We maintained a transparent accounting of screening and inclusion decisions; selection counts are provided in the Appendix (Section 39). To accelerate screening at scale, we used ASReview for AI-aided prioritization of records during title/abstract screening (Van De Schoot et al., 2021). We recorded sources, query terms, and screening decisions for transparency; summarized query formulations are provided in the Appendix (Section 39), along with selection counts (Section 39).

Thematic Organization

Items were tagged with a primary theme based on the dominant claim (e.g., efficiency, explainability, safety proofs, governance), and optionally one secondary tag for cross-cutting effects. The four themes are a *lens* for synthesis, not a taxonomy. Within each

theme we describe recurring interface patterns and trade-offs, and we use the interface-centric coding dimensions (Table 1) as the stable reference frame for comparing systems across domains.

For the main body, results are organized by themes (Section 14) with a consistent micro-structure (problem framing, representative advances, evaluation/benchmarks, limitations, takeaway). Tool grounding and end-to-end system aspects are revisited in Synthesis (Section 30) to show design trade-offs. Cross-domain works (e.g., KR-centric explainability, KG-grounded LLMs, safe RL with shields) are placed where they most strongly advance a theme and cross-referenced where relevant.

Critical Analysis and Synthesis

For each subsection we extracted: (i) problem abstractions and integration patterns; (ii) evaluation designs, datasets, and reported evaluation measures; and (iii) limitations and threats to validity. We emphasize representative works over exhaustive listings and group closely related contributions to avoid redundancy.

Comparative matrices and summary tables align methods to system functions (perception, KR, reasoning, planning/control, oversight) and to evaluation levers per theme. Reproducibility considerations (dataset/publication clarity, code/data availability, and "remove-one-component" tests (ablations) for robustness) inform the discussion and future directions. Where possible, we connect study claims to practical system components to make implications concrete.

Interface-Centric Coding of Neurosymbolic Systems

To avoid forced taxonomies where systems are assigned to a single "paradigm", we treat neurosymbolic AI as the engineering of interfaces between (i) statistical learners over continuous representations and (ii) explicit symbolic representations with defined operators (logic, programs, graphs, planners, constraint solvers).

This perspective is consistent with earlier structured views of integration that emphasize the learning cycle and the roles of representation and extraction (Bader & Hitzler, 2005; Garcez et al., 2019), and with survey work that argues for characterizing systems along shared dimensions rather than bins (Marra et al., 2024; Raedt et al., 2020; Marra, 2024). It also aligns with architecture-mapping approaches that separate component-coupling patterns (e.g., composite vs. monolithic integration) (Feldstein et al., 2024).

For each representative system, we therefore coded it along a compact set of interface dimensions, summarized in Table 1. This coding supports narrative synthesis without overstating guarantees (e.g., whether outputs are truly symbolic objects or merely natural language). It also makes trade-offs explicit in terms of guarantees and costs, which are central to evaluating whether neurosymbolic methods meet their promises in practice (Hitzler et al., 2022; Hamilton et al., 2024). In addition, it supports more rigorous discussion of trustworthiness and validation by making explicit which components are being explained or verified, and under what assumptions (Renkhoff et al., 2024; Michel-Deletie & Sarker, 2025). Finally, where appropriate, we note reproducibility signals such as public availability of code and materials as part of the evidence context (Colelough & Regli, 2025).

Operationally, we treat an approach as "symbolic" only when it employs an explicit representation with defined operators/semantics (e.g., logic/rules, executable programs, KGs with query/entailment operators, planners/controllers, SAT/SMT constraints, proof traces). In contrast, natural-language rationales (e.g., chain-of-thought prompting) are not treated as symbolic evidence by default unless they are typed, executed, or verified by explicit operators (Qiao et al., 2023; Mialon et al., 2023). Similarly, tool grounding (retrieval/citations) is treated as neuro→tool augmentation unless coupled to typed artifacts and explicit checking/constraint enforcement; we therefore avoid equating "grounded" with "guaranteed" (Gao et al., 2024). Finally, for tighter symbolic→neuro couplings that claim correctness of the encoding itself, we distinguish empirical performance claims from semantic correspondence/encoding conditions (Odense & Garcez, 2022).

Evidence and Citation Protocol

To balance breadth (citing the full bibliography) with defensible claims, we use an explicit evidence protocol that distinguishes *how* a reference is used from *what* it proves. In particular, we avoid treating broad technique families (e.g., "LLM prompting" or "RAG") as neurosymbolic evidence unless an explicit symbolic representation and operator-level coupling is present (e.g., typed/executable artifacts, constraints, or verification).

Citation roles (how we use a reference).

- **Spine (definition/axes authority):** supports definitions, boundary rules, and coding axes (Table 1).

Table 1. Interface-centric coding dimensions for describing neurosymbolic systems. Each row is a dimension and the second column states the operational question used when coding a system/paper.

Dimension	Operational question (what we record)
Symbolic representation	What symbolic objects and operators are used (e.g., FOL/DL/ASP, KGs/ontologies, SAT/SMT constraints, planners/controllers, proof traces)?
Interface direction	Does the system map neuro→symbolic (extraction/parsing), symbolic→neuro (knowledge injection/constraints), or alternate bidirectionally (verifier-in-the-loop, iterative refinement)?
Where coupling happens	Is the coupling primarily at training time (loss/regularizers), inference time (constrained decoding, tool-use, checking), or both?
Constraint strictness	Are constraints soft (penalties), enforced (hard decoding/filters), or certified (checking/proofs/guarantees)?
Guarantees	What is guaranteed, if anything (well-formed outputs, constraint satisfaction, safety envelopes, audit trails)?
Cost profile	What are the dominant costs (latency, compute, memory, number of tool calls, search/verification overhead), and where do they arise?
Failure modes	What are the main failure modes (hallucination/invalid outputs, brittleness, distribution shift, combinatorial explosion)?

- **Pattern exemplar:** provides a concrete instance of an interface pattern (e.g., constraints-in-loss; verifier-in-the-loop; neuro→sym extraction).
- **Evidence citation:** supports a measured claim (task/dataset/measure) or a formal guarantee (theorem/checked property), with explicit scope.
- **Context/background:** provides adjacent framing (e.g., LLM or KG surveys) without being treated as evidence of neurosymbolic coupling or guarantees.
- **Position/opinion:** used only as "argues/suggests"; not as evidence of performance or guarantees.

Evidence tags (attached to evaluative statements). When a statement is comparative (e.g., "improves reliability") or appears as a table cell, we tag it as **Measured**, **Claimed**, or **Not evaluated**, and we keep scope explicit (task/domain; dataset/benchmark where applicable). This pre-empts overgeneralization from benchmark results and aligns with critiques that commonsense and reasoning benchmarks can be flawed proxies for the intended capability (Davis, 2023).

Notes. Tags are applied per dimension and should be read with explicit scope (task/domain and benchmark in the cited paper). "Not eval." indicates we do not infer the dimension from that reference. This table illustrates the protocol in Table 3; it does

Table 2. Summary matrix mapping each survey theme to (i) the main system functions it concerns, (ii) typical neurosymbolic interface patterns/methods, and (iii) evaluation levers commonly reported in the literature.

Theme	System functions	Typical methods (examples)	Evaluation levers
Performant	Perception, planning/control	Architecture/compilers; constraints-in-loss (constraints encoded as loss penalties); typed/executable artifacts + execution/checking; constrained decoding; planning+RL hybrids	Accuracy/utility vs. latency/cost (incl. tool/verification calls when present)
Understandable	KR, explanation, oversight	Explicit KR + inspectable operators; neuro→sym extraction; proof/traces when available	Faithfulness/usefulness; provenance (source/derivation trace); human-centered evaluation
Reliable	Reasoning, control, monitors	Constraint enforcement/checking; verifier-in-the-loop (check/accept/reject/repair at inference time); formal specs where applicable; robustness tests	Safety conformance; invalid-output rates; shift/attack tests
Ethical	Policy/governance, HIL	Norm encoding; audits/logging; human-in-the-loop correction; oversight structures	Fairness/audits; accountability evidence; governance readiness

Table 3. Evidence tags used in tables and comparative statements. Each tag specifies what the corresponding citation supports (measured result/assurance claim vs. an asserted claim vs. not evaluated).

Tag	Operational meaning (what it licenses us to claim)
Measured	The paper reports an explicit experimental setup (task/dataset/measure) or a formal statement (e.g., theorem, checked property) that supports the claim.
Claimed	The paper asserts the claim, but does not directly evaluate it with an explicit measure/setting or provide a formal guarantee.
Not evaluated	The paper does not evaluate the dimension in question (e.g., cost, robustness, or whether an explanation is faithful to the system's actual decision process (faithfulness)), so we do not infer it.

not imply that all listed systems are directly comparable or that any dimension transfers across tasks.

Notes. The goal of this table is navigational: it makes it easier to locate papers by (i) interface pattern and (ii) problem setting. Cells are not intended as completeness claims. Citations are ordered chronologically within each cell.

Table 4. Worked cross-table: interface patterns mapped to problem settings (illustrative, not exhaustive). Each cell lists representative exemplars showing the interface pattern instantiated in that setting.

Interface pattern	Knowledge-intensive QA / grounding (KG/evidence)	Planning / control / RL	Verification / robustness	Explainability / oversight
neuro→tool grounding (retrieve/call tools)	(Lewis et al., 2020b; Schick et al., 2023; Gao et al., 2024)	(Besta et al., 2024; Valmeekam et al., 2024)	-	(Weir et al., 2024)
neuro→sym extraction (typed artifacts)	(Chen et al., 2021; Tammet et al., 2024)	(Asai & Muise, 2020)	-	(Mao et al., 2019a)
sym→neuro constraints-in-loss (training-time)	(Hu et al., 2016)	(Achiam et al., 2017)	(Xu et al., 2018; Ahmed et al., 2023)	(Koh et al., 2020)
inference-time enforcement (filters/shields)	-	(Alshiekh et al., 2018)	(Ahmed et al., 2023)	-
verifier-in-the-loop / certified checking	-	-	(Katz et al., 2017a; Elboher et al., 2020; Xie et al., 2022)	(Ignatiev et al., 2021)

Core Themes in Neurosymbolic artificial intelligence

We now organize the surveyed literature around four themes - performance, understandability, reliability, and ethics. These themes are anchored to system functions and evaluation measures summarized in Table 2. Each theme follows a consistent structure to support comparability across domains and to surface design trade-offs that will be synthesized in Section 30.

Performant AI: Efficiency and Capability

We evaluate recent literature on efficiency, capability, and cost trade-offs in neurosymbolic systems, highlighting representative methods, benchmarks, and measurement considerations. Following the interface-centric coding dimensions (Table 1) and the scope boundary in Section 4, we found interface patterns that most directly drive performance *via operator-level symbolic coupling*: (i) symbolic→neuro compilation and constraints-in-loss (training-time coupling), (ii) neuro→symbolic generation of typed/executable artifacts with execution/checking (inference-time

coupling), and (iii) planning/control couplings that combine learned skills with explicit search, planners, or symbolic controllers. We discuss retrieval and tool augmentation as adjacent context and treat it as neurosymbolic evidence only when paired with typed artifacts and explicit checking/constraint enforcement. Throughout, performance claims are interpreted jointly with cost profile (latency, compute, memory, tool calls) and scoped to the reported benchmark setting.

Architectural Paradigms for Efficiency and Scalability We cite a small set of neural architectures as *context* for the neural architectures that hybrid interfaces attach to; these are not treated as neurosymbolic evidence by themselves. Examples include distributed and contextualized representations for language (Mikolov et al., 2013; Peters et al., 2018), attention-based sequence transduction (Bahdanau et al., 2014; Shaw et al., 2018; Vaswani et al., 2023), and long-context sequence models (Dai et al., 2019). We also include representative sequence-to-sequence pretraining and recurrent encoders as common base models that hybrid systems build on (Graves & Schmidhuber, 2005; Lewis et al., 2020a). For relational inputs, graph attention networks are a commonly used neural architecture (Veličković et al., 2018). Scaling discussions and emergent abilities in large language models provide important context for why tool grounding and verifier-in-the-loop designs are increasingly used as *interfaces* in practice, even when they do not imply correctness guarantees (Wei et al., 2022). Graph Transformers generalize attention to arbitrary graphs via connectivity-aware attention and spectral positional encodings, and are reported to improve generalization on relational inputs in the evaluated settings (Dwivedi & Bresson, 2020). Surveys of GNNs within neural-symbolic computing synthesize applications in combinatorial optimization, constraint satisfaction, and relational reasoning, and discuss GNNs as common neural architectures for hybrid systems (Lamb et al., 2020). Domain-specific foundation representations, such as geospatial embedding fields, illustrate how compact, reusable embeddings can ground downstream mapping and analysis at scale (Brown et al., 2025).

Representative hybrid paradigms integrating neural and symbolic components include Logic Tensor Networks and related differentiable semantics (Donadello et al., 2017; Yang et al., 2017), probabilistic logic programming approaches such as DeepProbLog (Manhaeve et al., 2018), and neural logic machines and differentiable rule learning (Dong et al., 2019). Early knowledge-injection and embedding-based integration frameworks provide additional coupling patterns (Hu et al., 2016; Kolb et al., 2018; Chen et al., 2019, 2023a).

A comprehensive integration framework harmonizes symbolic constraints and domain knowledge with deep learning components to improve reasoning, generalization, and transfer (Himabindu et al., 2023).

Compositional integration treats neural and symbolic modules as black boxes with deduction, abduction, and induction interfaces, enabling modular coupling without assuming internal semantics (Tsamoura et al., 2021).

Recent workload characterizations of neurosymbolic systems (runtime profiling across representative operators and hardware) highlight the compute bottlenecks and parallelism profiles that differentiate symbolic reasoning from neural components, informing design trade-offs for scalable hybrid pipelines (Susskind et al., 2021).

On extreme-edge platforms, hardware-aware neurosymbolic architecture search reports joint optimization of symbolic and neural operators under tight memory and latency budgets, generating microcontroller-ready code for multiple NeSy model families in the evaluated settings (Saha et al., 2024).

Neurosymbolic logic programming frameworks based on stochastic derivations, such as DeepStochLog, offer improved scaling for inference and learning compared to neural probabilistic logic programs while retaining end-to-end trainability (Winters et al., 2022). Automated architecture innovation moves beyond classical NAS toward autonomous hypothesis generation and empirical evaluation for model design, suggesting new pathways for scaling hybrid systems (Liu et al., 2025).

Efficient reasoning can be further supported by program-guided perception and learned prioritization of proof paths, which reduce search overhead while retaining interpretability (Mao et al., 2019b; Morris, 2022). Differentiable logic compilation and declarative neurosymbolic languages streamline training and inference by leveraging deep-learning backends for logical queries (Cohen et al., 2017; Yang et al., 2020; Li et al., 2023b). Modular couplings with cognitive architectures can orchestrate hybrid components efficiently at system level, improving throughput and latency via division of labor and tool-use (West et al., 2023; Romero et al., 2024; Liu et al., 2024; Joshi & Ustun, 2024; Thomson & Bastian, 2024; Roy et al., 2025). Self-supervised representation learning (e.g., I-JEPA) provides perceptual base models that can be paired with symbolic interfaces (Assran et al., 2023).

Tool grounding and structured artifacts for factuality and task success Retrieval-augmented generation and tool-augmented inference are *reported* to improve factuality and task success on specific benchmarks and workloads, often at the cost of additional latency, compute, and tool calls (Mialon et al., 2023; Zhao et al., 2023c; Gao et al.,

2024; Annepaka & Pakray, 2025; Li et al., 2025). In line with our evidence protocol (Section 11), we treat tool grounding as an interface pattern rather than a guarantee: tool grounding can reduce error rates but does not, by itself, certify correctness. Under our scope boundary, tool grounding becomes neurosymbolic *evidence* only when it is coupled to explicit symbolic artifacts and operators - for example, when a model emits a typed/executable query or program that is executed and checked by a KG/reasoner, constraint system, or verifier. Retrieval and tool-use provide mechanisms for connecting models to external sources and computations (Lewis et al., 2020b; Mialon et al., 2023; Schick et al., 2023; El-Kishky et al., 2024). For this survey, the key dividing line is whether the interface produces an *explicit, checkable artifact*. When the model emits typed/executable structures (e.g., logical queries/programs) and these are executed/validated by symbolic operators (KG query engines, reasoners, constraint systems, verifiers), the coupling can convert tool grounding into an operator-level neurosymbolic interface (Chen et al., 2021; Weir et al., 2024; Li et al., 2022; Tammet et al., 2024). When such checking is absent, we treat tool grounding as adjacent context and avoid interpreting it as a correctness mechanism (Gao et al., 2024; Sahoo et al., 2024; Huang et al., 2025). Knowledge-intensive pipelines also leverage graph-structured corpora and link structure (document graphs, KG-derived training signals) as *interfaces* that shape retrieval and attribution behavior, with gains and limitations reported in the evaluated settings (Yasunaga et al., 2022; Da et al., 2021). Within NLP, neuro-symbolic reasoning is often framed as bridging neural language models with explicit logic, latent structures, or knowledge bases; applied overviews and reviews provide context on representative integration patterns (Aithal et al., 2022; Keber et al., 2024; Liu et al., 2023). Popular commentary sometimes frames neurosymbolic AI as a remedy for hallucination (Garcez, 2025); in this survey we treat that framing as motivation rather than as technical evidence unless operator-level coupling and scoped evaluations are reported. Practical toolkits and resources support commonsense inference and persona- or task-specific knowledge acquisition that can feed neurosymbolic interfaces (Ismayilzada & Bosselut, 2023; Gao et al., 2023). Solver-in-the-loop training and symbolic feedback loops have been proposed as a way to improve reasoning in math and software engineering tasks without relying solely on scale; such approaches are evaluated and should be interpreted in the scope of their feedback signals and benchmarks (Jana, 2024). In knowledge-intensive QA, coupling LMs to explicit KB/KG artifacts and symbolic teachers has been reported to improve generalization and robustness in the

Table 5. Boundary cases for "tool grounding" in performance-oriented pipelines. Rows distinguish adjacent tool augmentation from operator-level neurosymbolic coupling under the scope boundary in Section 4.

Pattern	Artifact / operator	What is often reduced (scoped)	What is not guaranteed + typical costs
Tool grounding (adjacent context)	retrieval/tool calls; no executable symbolic artifact	factuality errors in reported settings (Lewis et al., 2020b; Gao et al., 2024)	does not certify correctness; costs include tool-call latency and additional context handling (Gao et al., 2024; Huang et al., 2025)
Typed artifacts + execution/checking (NeSy evidence)	typed query/program + KG/reasoner/constraint execution	invalid-output rates; some factuality failures when the checker is sound for the artifact class (Chen et al., 2021; Weir et al., 2024)	well-formed-but-wrong artifacts remain possible; costs include execution and checking overhead
Verifier-in-the-loop (NeSy evidence)	explicit verifier / constraints that accept/reject/repair outputs	constraint violations and some unsafe/invalid behaviors under stated assumptions (Katz et al., 2017a; Alshiekh et al., 2018)	guarantees are property- and assumption-scoped; costs include solver/verifier overhead and potential rejection loops

evaluated settings, while introducing additional failure modes and overhead (Oltramari et al., 2021; Aakur & Sarkar, 2023).

Planning, Control, and Reinforcement Learning Planning and control integrations leverage symbolic models and cognitive frameworks for improved efficiency and decision quality (Clark et al., 2016; Yang et al., 2018). Earlier neurosymbolic planning/control architectures and distributed cognitive robotics systems provide additional coupling patterns (Mastrogiovanni et al., 2007; Belle & Lakemeyer, 2011; de Penning et al., 2011). A neurosymbolic planning architecture, Plan-SOFAI, instantiates dual-process fast/slow reasoning to combine planners across classical scenarios, illustrating modular integration patterns for deliberative control (Fabiano et al., 2023). Structure-of-thought prompting strategies, including chains, trees, and graphs of thought, provide flexible scaffolds for decomposition, search, and evaluation in reasoning and planning with language models; these scaffolds are not treated as symbolic evidence unless intermediate artifacts are typed and executed/checked by explicit operators (Besta et al., 2024).

Neurosymbolic pipelines improve sample efficiency and generalization by inducing symbolic abstractions and models for classical planners (Asai & Muise, 2020; Shah, 2023). Symbolic controllers and differentiable planners elevate decision quality and long-horizon optimization (Zhang & Hannaford, 2020; Jeong et al., 2021; Chatterjee et al., 2023). Language-model interfaces to planning and control include translating domains to plans and shaping RL with structured signals (Karia & Srivastava, 2022; Mitchener et al., 2022; Pallagani et al., 2023). Related approaches couple language models to symbolic planning or control abstractions for decision-making (Kimura et al., 2021; McDonald et al., 2024). Surveys benchmark hybrid methods and outline open challenges for sequential decision-making (Núñez-Molina, 2022; Núñez-Molina et al., 2024; Valmeekam et al., 2024). Explainable AI planning perspectives further contextualize requirements for transparent decision-making in planners (Chakraborti et al., 2020). Classical TD-network formulations illuminate predictive representations useful for control and planning (Sutton & Tanner, 2004). Evaluations of large reasoning models on combinatorial tasks reveal current limitations and the need for symbolic planners and reasoners in the loop (Hazra et al., 2025).

Evaluation and Measures Benchmarking considerations for knowledge-intensive tasks and QA datasets are synthesized in (Rogers et al., 2023), including classic open-book settings that stress retrieval and multi-hop reasoning requirements (Mihaylov et al., 2018). Cost-effectiveness measures for neural models, such as TALES, quantify resource–accuracy trade-offs using FLOPs, parameter counts, and predicted latency to guide model selection on constrained platforms (Zhao et al., 2023b). System-level workload studies categorize NeSy algorithms and profile runtime, memory, sparsity, and operator mixes across CPUs, GPUs, and edge SoCs, informing architecture-aware evaluation (Wan et al., 2024b). A systems perspective extends this profiling toward joint hardware-software design (co-design) for NeSy acceleration (Wan et al., 2024a). General-purpose evaluators and exam-style test suites facilitate comparable assessment of generation quality and task success (Zhong et al., 2022; He et al., 2024; Zhong et al., 2024). Reflections on benchmark validity and historical use inform careful interpretation of reported gains (Orr & Kang, 2024).

Notes. “Cost” refers to reported latency/compute/memory/tool-call or verification overhead in the cited paper; “Not eval.” indicates cost was not evaluated.

Theme takeaway (interface patterns and trade-offs). Across performant systems, measured gains typically arise from two interface strategies: (i) training-time coupling

Table 6. Representative systems discussed under the *Performant* theme, summarized by interface pattern and scoped evidence. Columns indicate the interface type, the main artifact, the evaluation scope, the evidence tag (Table 3), and whether cost/overhead is measured or only discussed. Tool-augmentation systems are included as adjacent context baselines and are not treated as neurosymbolic evidence unless paired with typed artifacts and explicit checking/constraint enforcement (Section 4).

Reference	Interface	Artifact	Scope	Tag	Cost
(Saha et al., 2024) (TinyNS)	sym→neuro co-design	explicit symbolic operators in pipeline	edge workloads (microcontrollers)	Measured	Measured
(Winters et al., 2022) (DeepStochLog)	sym→neuro (stochastic logic)	logic program; derivations	probabilistic logic programming tasks	Measured	Not eval.
(Li et al., 2023b) (Scallop)	sym→neuro (differentiable queries)	Datalog-style rules; provenance	differentiable query learning; inference	Measured	Not eval.
(Manhaeve et al., 2018) (DeepProbLog)	sym→neuro (differentiable inference)	probabilistic program; proofs	neurosymbolic probabilistic reasoning	Measured	Not eval.
(Schick et al., 2023) (Toolformer)	neuro→tool augmentation	tool calls; structured outputs	tool-augmented tasks	Measured	Claimed
(Lewis et al., 2020b) (RAG)	neuro→tool augmentation	retrieval; generation pipeline	retrieval-augmented tasks	Measured	Claimed
(Asai & Muise, 2020) (neuro→planner)	neuro→sym extraction	symbolic abstractions for planning	planning with learned abstractions	Measured	Not eval.
(Zhao et al., 2023b) (TALES)	evaluation instrumentation	cost/accuracy trade-off measure	model selection under constraints	Measured	Measured

(symbolic constraints or structures shaping learning) and (ii) inference-time coupling (constrained decoding, checking, or search, and tool use when coupled to typed artifacts and explicit operators). These gains are rarely "free": they trade accuracy/utility against latency, compute, memory, and tool-/verification-call overhead (Table 1). We therefore treat performance claims as inseparable from cost profiles and evaluate them under explicitly reported workloads and benchmarks. In the running example, this means reporting end-to-end utility alongside latency and the overhead of retrieval, execution, and checking calls.

Understandable AI: Opening the Black Box

We review works that make AI reasoning more interpretable, focusing on the interface patterns by which systems produce inspectable artifacts. Concretely, we structure the discussion around (i) symbolic representations that support explanation (KR, ontologies, KGs, rules), (ii) neuro→symbolic extraction and trace generation (programs, proofs, paths, structured events), and (iii) concept-level bottlenecks that make intermediate variables human-meaningful. We distinguish explanation *artifacts* from natural-language rationales and treat faithfulness/provenance as first-class evaluation dimensions.

Evaluation pitfalls (what to avoid inferring). Because explanation claims are easy to overstate, we apply two guardrails throughout this theme: (i) *plausibility is not faithfulness* - a coherent narrative can still be an inaccurate account of the decision process (faithfulness asks whether the explanation matches what the system actually used); and (ii) *artifact validity is not task correctness* - a well-formed query, rule, or proof trace can still support a wrong conclusion if the upstream mapping or the knowledge base is incomplete or mis-specified. We therefore treat provenance and faithfulness as distinct evaluation targets, and avoid treating post-hoc natural-language rationales as symbolic evidence unless they are typed and executed/checked by explicit operators (Section 11).

Knowledge Representation as a Foundation Structured KR provides the semantics that make explanations and audits interpretable: it defines *what counts* as an entity, relation, rule, or constraint, and which operators are valid for querying and inference. We therefore treat KR choices as a first-order design decision for understandability. Foundational work and surveys motivate how symbols are represented and queried and why semantics matter for explanation (Brachman et al., 1985; Miller, 1995; Ding, 2007; Donadello et al., 2017; Dumancic et al., 2019; Jaeger, 2023; Raedt et al., 2020; Marra et al., 2024; Marra, 2024; Kramer, 2020).

KGs/ontologies and queryable semantics. When knowledge is stored as a KG or ontology, explanations can be grounded in explicit relations and retrieved via typed queries; provenance can be attached to edges, documents, and entailment steps. This supports explanation artifacts such as query traces, retrieved subgraphs, and entailment chains, but it also introduces maintenance risks (coverage gaps, stale knowledge, inconsistent schemas) that should be reported and versioned. Representative resources and surveys discuss how KGs/ontologies can support human-aligned concepts and interpretable inference across domains (Lecue, 2020; Hogan et al., 2022; Ji et al., 2022; Kau, 2024; Rajabi & Etminani, 2024).

Rules/programs and executable structure. Rule- and program-based representations (logic programs, ASP, ILP, Datalog-style rules) support *executable* explanations: the explanation is the program/rule trace that leads from inputs to outputs. Such artifacts are inspectable and editable, but the central validity question becomes whether the neuro→symbolic mapping is faithful and whether the program semantics match the intended domain assumptions. Representative mechanisms include coupling neural predictions to symbolic constraints, inducing programs/rules under constraints, and using declarative languages that expose provenance (Yang et al., 2020; Cropper & Morel, 2021; Ciatto et al., 2021; Li et al., 2023b). Learning and transferring symbolic representations, and using symbolic priors as soft guides to neural semantic parsing, provide additional context on how executable structure can improve sample efficiency and interpretability in the evaluated settings (James, 2018; Xiao et al., 2017).

Artifact family C: probabilistic and differentiable semantics. When uncertainty is central, probabilistic-symbolic or differentiable semantics can provide explanation artifacts that encode both structure and uncertainty (e.g., weighted rules, probabilistic programs, differentiable query provenance). These can improve interpretability by making uncertainty explicit, but they also complicate evaluation because explanation faithfulness depends on both the symbolic structure and the learned scoring/inference dynamics (Sato & Kameya, 1997; Minato et al., 2007; Cohen et al., 2017; Qu & Tang, 2019; Badreddine et al., 2022; Serafini & Garcez, 2016; Serafini et al., 2017). Grammar-based symbolic structure and refined nonterminals illustrate a classical symbolic representation where learnable components yield explicit, inspectable structure, while retaining strong empirical performance in the evaluated settings (Shindo et al., 2013).

Further reading (KR assets and construction at scale). Human-interpretable KR assets and large ontologies provide reusable concept spaces for explanation (e.g., commonsense and biomedical ontologies), while knowledge construction and integration work addresses coverage and consistency in deployed pipelines (Speer et al., 2018; Mostafazadeh et al., 2020; Hwang et al., 2021; Robinson et al., 2008; Smirnov et al., 2024; Bordes et al., 2013; Han et al., 2018; Schockaert et al., 2021; Odense & Garcez, 2022; Werner, 2024; Järv et al., 2022, 2023; Bosselut et al., 2019; Fang et al., 2021; Wang et al., 2025a; Hitzler et al., 2020; Jain et al., 2025). Formal argumentation further contextualizes symbolic reasoning artifacts (e.g., non-monotonic logics, inconsistency handling, and argumentative traces) that can support auditable reasoning (Ulbricht, 2024).

Intrinsic Explainability and Transparent Decision-Making Intrinsic transparency via rule extraction and differentiable/provable reasoning is exemplified by early rule extraction and neural-symbolic approaches (Craven & Shavlik, 1995; d'Avila Garcez et al., 2002a,b), as well as later differentiable and end-to-end formulations (Rocktäschel & Riedel, 2016, 2017). Program-guided perception and manipulation induce symbolic structure from images, enabling extrapolation and regularity editing within the proposed framework (Mao et al., 2019b). Adaptive proof-path selection policies improve tractability in neural theorem proving by learning to prioritize likely derivations, retaining interpretability while scaling reasoning (Morris, 2022).

Logical reasoners and hybrid provers provide verifiable traces and coherent explanation graphs for domain tasks and QA (Kalyanpur et al., 2022; Tammet et al., 2023; Vakharia et al., 2024; Weir et al., 2024). Symbol-aware pipelines disentangle perception from logic to expose step-by-step reasoning and support contrastive explanations (Yi et al., 2018; Amizadeh et al., 2020; Eiter et al., 2023). Interpretable model classes and compilation to tractable forms support minimal feature-based and rule-level explanations (Shih et al., 2018; Riegel et al., 2020; Ignatiev et al., 2021). Related approaches connect explanation to program induction and end-to-end neurosymbolic learning (Rocktäschel & Riedel, 2017; Evans & Grefenstette, 2018). Domain pipelines (e.g., sentiment) illustrate that conversion to symbolic representations can make decision paths more transparent by exposing explicit intermediate structure (Cambria et al., 2022).

Concept-Based Models for Interpretable Bottlenecks Concept bottlenecks make intermediate predictions about human-understandable concepts that mediate from features to decisions, improving transparency and controllability (Koh et al., 2020). Neurosymbolic concept learners ground visual and relational concepts in language or structured supervision, enabling programmatic reasoning over explicit symbols (Mao et al., 2019a). End-to-end integrations learn latent concepts alongside symbolic rules or ASP programs to support decision pipelines with checkable intermediate structure (Murali et al., 2022; Cunnington et al., 2023). Domain adaptations leverage ontologies to define concept spaces that strengthen interpretability and performance in specialized settings (Glauer et al., 2023). Learning strategies invent new interpretable relational concepts and efficient search heuristics to scale symbolic learning (Daniele et al., 2023; Demir & Ngomo, 2023; Sha et al., 2025).

Evaluation and Measures Evaluation of explainability combines quantitative and qualitative criteria to assess usefulness, faithfulness, and human factors (Islam et al.,

Table 7. Representative systems discussed under the *Understandable* theme. Columns specify the produced explanation artifact (e.g., proofs, programs, concepts), the interface pattern, the evaluation scope, and an evidence tag (Table 3) indicating whether the explainability-related claim is directly evaluated.

Reference	Artifact	Interface	Scope	Tag	Notes
(Koh et al., 2020) (Concept bottlenecks)	concept variables	neuro→sym bottleneck	concept classification	Measured	faithfulness evaluated
(Mao et al., 2019a) (NS-VQA)	programs over symbols	neuro→sym extraction	visual QA with program execution	Measured	artifact-based
(Weir et al., 2024) (NELLIE)	proof trees	verifier-in-the-loop	grounded QA with proof artifacts	Measured	provenance provided
(Vakharia et al., 2024) (ProSLM)	logical traces	sym→neuro / hybrid prover	domain tasks + QA	Measured	trace quality scoped
(Tammet et al., 2023) (reasoner traces)	explanation graphs/traces	sym reasoning + trace output	reasoning with explicit traces	Measured	semantics-dependent
(Ignatiev et al., 2021) (SAT/abduction)	minimal explanations	sym checking/compilation	explanation via tractable forms	Measured	local explanations
(Stammer et al., 2021) (interactive concepts)	editable concepts/rules	human-in-the-loop refinement	interactive correction workflows	Measured	user-study scoped

2024). Interactive interpretability leverages language models to generate or critique explanations, requiring human-centered protocols for validation (Singh et al., 2024). Post-hoc natural language explanations and task-specific settings (e.g., education, grading) illustrate measurement of explanation quality in practice (Tornqvist et al., 2023). Task addresses high-level understanding (e.g., humor comprehension) expose current gaps and help benchmark progress beyond surface correlations (Hessel et al., 2023). Symbolic domains such as music highlight discrete-generation evaluation measures and evaluation needs (Plasser et al., 2023). Multilingual KGQA benchmarks support fairer accessibility in grounded QA evaluation (Perevalov et al., 2022). Comprehensive surveys of XAI and explainability taxonomies inform evaluation protocols for hybrid systems (Zhang & Sheng, 2024; Ullah et al., 2025). Evaluation implications extend to domain-specific symbolic generation tasks, including symbolic music rearrangement (Zhao et al., 2023a).

Notes. "Artifact" is the inspectable object used for explanation/oversight (not a natural-language rationale by default). "Interface" indicates neuro→sym, sym→neuro, or verifier-in-the-loop style coupling.

Theme takeaway (interface patterns and trade-offs). Understandability is strongest when the interface produces inspectable artifacts with stable semantics (rules, programs, proof traces, KG queries) rather than post-hoc narratives. Neuro→symbolic extraction and concept bottlenecks can increase editability and oversight, but they introduce failure modes (spurious rules, brittle concept definitions, dataset leakage) that require faithfulness and provenance evaluation. We therefore prioritize evidence-tagged claims about explainability quality and explicitly separate artifact validity from human-perceived plausibility (Table 3). In the running example, the most defensible explanations are checkable artifacts (queries, rules, traces) paired with provenance and explicit evaluation of faithfulness.

Reliable AI: Robustness and Verifiability

We summarize approaches to robustness and verification in hybrid systems, organized by how reliability is enforced at the interface. We group work into (i) certified or checkable constraints (formal verification, SMT/ASP-style checking, proof-carrying artifacts), (ii) robustness mechanisms that use symbolic knowledge/constraints as invariants or inductive biases, and (iii) safe planning/RL couplings where shields, specifications, and constrained objectives narrow failure modes. Reliability is reported with explicit scope (assumed attacker/shift conditions (threat model), shift type, task) and—when present—what is guaranteed versus only mitigated.

Guarantee scope checklist (what a "guarantee" depends on). Where papers claim formal guarantees, we interpret them as property- and assumption-scoped. Concretely, the meaning of a guarantee depends on at least: (i) **what is specified** (property class and how it is formalized), (ii) **what is modeled** (environment assumptions, input bounds, and the assumed attacker/shift conditions (threat model)), (iii) **what is covered** (which component(s) of the end-to-end pipeline are verified or constrained), and (iv) **what method limits apply** (solver completeness, abstractions, approximations, and any uncertified heuristics). We therefore avoid implying global safety from local checks, and treat "certified" results as scoped to the verified object and stated assumptions.

Formal Verification and Safety Guarantees Formal guarantees in hybrid/multi-agent contexts are explored in (Kouvaros et al., 2018). Neural SAT solvers illustrate learned

propositional reasoning capabilities that can support verification and analysis tasks within hybrid pipelines (Selsam et al., 2018).

Verification of neural components has progressed via SMT-based solvers, abstraction techniques, and neurosymbolic specifications, expanding the space of certifiable properties (Katz et al., 2017a,b; Elboher et al., 2020; Xie et al., 2022). Additional formulations and implementations of Reluplex further characterize the feasibility and limitations of DNN property checking (Katz et al., 2017a). Methodologies and surveys outline processes for certifiable AI in safety-critical domains and for multi-agent settings (Elia et al., 2024; Renkhoff et al., 2024; Kouvaros et al., 2024). These works emphasize testability, evidence-backed assurance cases, and the role of formal safety arguments in deployment (Kouvaros, 2023; Lenat & Marcus, 2023; Lu et al., 2024). Neuro-symbolic theorem proving provides a complementary reliability pattern: neural guidance can prioritize proof search while a symbolic deduction engine produces checkable (and often human-readable) proofs (Trinh et al., 2024).

Robustness to Adversarial and Distributional Shifts Symbolic constraints embedded in objectives improve robustness by enforcing invariants during learning (Xu et al., 2018; Chen et al., 2023b; Ahmed et al., 2023). Encoding contextual knowledge into model parameters and deconfounding strategies mitigate distractors and shifts (Chen et al., 2023c; Wang et al., 2024). Evaluation requires systematic tests across domains and shift conditions to characterize reliability boundaries. Reasoning under uncertainty via probabilistic logics (e.g., Logical Credal Networks) provides calibrated inference with imprecise probabilities, supporting reliability in open settings (Qian et al., 2022). Constraint-guided fine-tuning of generative models has been proposed as another route to reduce forbidden outputs under specified constraint classes, but its reliability should be interpreted as constraint- and evaluation-scoped (Yin et al., 2024).

Constraint Satisfaction and Safe Reinforcement Learning Classical symbolic solvers illustrate constraint handling and approximate reasoning under uncertainty, informing safety mechanisms in hybrid controllers (Sacks, 1989).

Constrained optimization and temporal-logic shields are used to enforce or encourage safety specifications during exploration and execution, within the scope of the stated constraints and environment assumptions (Achiam et al., 2017; Alshiekh et al., 2018; Yang et al., 2023). Neurosymbolic controllers leverage logical constraints to guide policies and (in some settings) enforce checkable safety properties during learning and deployment.

Table 8. Representative systems discussed under the *Reliable* theme. Columns specify the checked object (e.g., properties, shields, constraints), the interface pattern, the evaluation scope (threat model/task), evidence tag (Table 3), and the form of guarantee claimed or measured.

Reference	Object	Interface	Scope	Tag	Guarantee
(Katz et al., 2017a) (Reluplex)	SMT-checked properties	certified checking	DNN property checking	Measured	Measured (scoped)
(Xie et al., 2022) (specifications)	neurosymbolic specs	spec→ verification	certifiable properties	Measured	Claimed; Measured
(Alshiekh et al., 2018) (shields)	safety shields	constraint enforcement	safe RL policies	Measured	Measured (scoped)
(Achiam et al., 2017) (constrained RL)	constrained objective	training-time constraints	safe policy learning	Measured	Claimed
(Xu et al., 2018) (semantic constraints)	invariants; constraints	constraints-in-loss	robustness under shifts	Measured	Not eval.
(Qian et al., 2022) (credal nets)	probabilistic logic model	sym inference	uncertainty-aware reasoning	Measured	Claimed
(Renkhoff et al., 2024) (V&V survey)	V&V process model	evaluation protocol	testing; validation taxonomy	Measured	Not eval.

Evaluation and Measures Foundational theories and empirical studies on assessment, progress quantification, and reproducibility include (Liu et al., 2018; Gundersen & Kjensmo, 2018; Rezazadegan et al., 2024). Measurement models and testing frameworks support standardized reliability evaluation for commercial and research systems (Zhang et al., 2022; Li et al., 2023a).

Notes. "Guarantee" entries should be interpreted as assumption-scoped to the cited work (e.g., property class, environment model, or robustness condition).

Theme takeaway (interface patterns and trade-offs). Reliable systems tend to externalize correctness into checkable interfaces: constraints, verifiers, shields, and specifications that bound outputs and actions. The main synthesis risk is conflating mitigation with guarantees; we therefore keep threat models explicit and treat "guarantee" as property- and assumption-scoped (Table 1). Robustness evidence is strongest when accompanied by standardized test suites, coverage criteria, and transparent reporting of assumptions and "remove-one-component" tests (ablations). In the running example, reliability should be reported as constraint-violation rates and

robustness under stated shifts/threats, with guarantees described only as assumption-scoped.

Ethical AI: Value Alignment and Accountability

We consolidate literature on alignment, fairness, human-in-the-loop refinement, and governance for accountable neurosymbolic systems, emphasizing what is specific to neurosymbolic design: *accountability interfaces*. In practice, accountability is realized by explicit, inspectable artifacts and operators - norms as rules/constraints, compliance checks, audit logs, and revision traces - and by clearly specifying where they act (training-time objectives, inference-time checking, or governance/workflow controls). We therefore treat value alignment and ethics not as abstract principles but as interface properties that can be implemented, evaluated, and revised without retraining an entire model.

Value Alignment and Encoding Ethical Principles Human-compatible AI principles and preference-uncertainty arguments are articulated in (Russell, 2019).

Encoding norms into symbolic components supports auditable reasoning about duties, rights, and constraints. Under our interface-centric lens, the key question is: what is the *executable norm artifact* (rules, constraints, policies) and what operator checks it (reasoner, constraint solver, verifier) at design time or runtime? Neurosymbolic pipelines can translate legal or policy text into logical code for transparent compliance, producing inspectable artifacts that can be reviewed, tested, and updated (Chanin & Hunter, 2023; School, 2024; Kant et al., 2024). Expressivity limits of standard reward formulations motivate explicitly structured objectives for alignment (Abel et al., 2021). Comprehensive assessment frameworks guide ethical integration and evaluation in domain settings such as education (Kılınç, 2024). Alternate discussions of reward expressivity emphasize implications for alignment specifications (Abel et al., 2021).

Algorithmic Fairness and Bias Mitigation Auditing for dataset and model biases requires culturally aware diagnostics and targeted mitigations; surveys highlight Western-centric biases as a persistent risk (Abbas, 2025). From a neurosymbolic perspective, the contribution is to make fairness-related assumptions *explicit and testable* as constraints, documentation artifacts, and audit procedures rather than implicit behavior. Symbolic knowledge and constraints can support debiasing and accountability by (i) exposing which attributes/relations are used, (ii) enabling constraint-based checks, and (iii) providing versioned artifacts that support rollback and redress.

Human-in-the-Loop Learning for Collaborative Refinement Interactive frameworks allow users to correct concepts, rules, and reasoning traces, improving both performance and trust (Kim et al., 2020; Stammer et al., 2021; Crochepierre et al., 2022). Complementary conversational and iterative learning setups emphasize user-driven correction loops and structured feedback signals (Kirk & Laird, 2019; Arabshahi et al., 2021). Interactive reward and policy learning with human feedback complements symbolic refinement pathways for safer, aligned behavior (MacGlashan et al., 2017). For accountability, HIL protocols should capture rationale, provenance, and rollback: which artifact was changed (concept, rule, policy), why it was changed (human justification), and how the system behaved before/after the change under a fixed evaluation protocol.

Governance and Oversight National strategies such as Estonia's Kratt Strategy document concrete policy instruments for accountable AI deployment in public services (of Economic Affairs and Communications, 2022). Strategic and programmatic drivers for explainable and third-wave AI are summarized in (Daws, 2018; Gunning et al., 2021).

Oversight structures for agent-based AI specify roles, audit trails, and escalation protocols in public organizations (Schmitz et al., 2025). Operational governance typically requires logging of data access, tool-use, and explanations to enable audits and redress.

Evaluation and Measures Ethics-focused evaluation spans fairness measures, participatory audits, and governance checklists; documentation of assumptions and failure modes is essential. Systematic reviews of evaluation criteria for trustworthy AI provide broader taxonomies that can be mapped onto neurosymbolic accountability interfaces (McCormack & Bendechache, 2024).

Notes. "Locus" refers to where accountability constraints are applied: training-time objectives, inference-time checks, or governance/workflow controls.

Theme takeaway (interface patterns and trade-offs). Ethical and accountable NeSy systems place normative constraints in explicit, revisable structures (policies, rules, audits/logs) and couple them to learning and decision-making via objectives, checks, and oversight workflows. A central trade-off is between flexibility (updating norms) and assurance (demonstrating compliance under deployment conditions). We therefore treat auditability, traceability, and redress mechanisms as concrete interface properties rather than purely aspirational claims. In the running example, accountability is operationalized by explicit policy constraints, audit logs for data/tool access, and a revision trail for norm updates.

Table 9. Representative systems and workflows discussed under the *Ethical* theme. Columns indicate the explicit accountability interface (e.g., norms as rules, audits/logs), where it is enforced (training/inference/workflow), the evaluation scope, and an evidence tag (Table 3).

Reference	Accountability interface	Locus	Scope	Tag	Notes
(Chanin & Hunter, 2023) (norm encoding)	legal/policy rules as code	inference-time checks	compliance reasoning	Measured	auditability
(Schmitz et al., 2025) (oversight)	roles/audit trails	governance workflow	public org oversight	Measured	process-level
(Stammer et al., 2021) (interactive)	editable concepts/rules	human-in-the-loop	user correction loops	Measured	HCI-scoped
(Russell, 2019) (principles)	preference uncertainty framing	objective design	alignment principles	Claimed	normative
(Abbas, 2025) (bias audits)	explicit audit criteria	evaluation protocol	bias diagnostics	Measured	dataset-sensitive
(of Economic Affairs and Communications, 2022) (strategy)	governance instruments	policy deployment	public-sector AI	Measured	context
(MacGlashan et al., 2017) (interactive RL)	feedback signals + constraints	training-time coupling	safer policy learning	Measured	cost varies

Synthesis and Application Spotlight: The Neurosymbolic System

Building on the thematic synthesis, this section integrates findings into a system-oriented perspective. We relate components and interfaces across perception, knowledge, reasoning, planning/control, and oversight, and examine cross-theme trade-offs among performance, interpretability, reliability, and ethics. We outline architectural patterns and design imperatives, then illustrate implications through a focused application spotlight.

Architectures for Collaborative and Autonomous Systems

At the theoretical level, the Common Model of Cognition can be adapted to coordinate large generative networks via shadow production systems that interface with a central controller, offering a principled scaffold for system design (West et al., 2023). High-level reasoning agendas propose cognitive architectures as orchestrators for integrating symbolic knowledge with learning components, with the goal of improving commonsense and reliable decision-making (Oltramari, 2023b,a, 2024). Complementary integration patterns combine large language models with cognitive architectures,

spanning modular tool-augmented pipelines, agent societies, and neurosymbolic schemes that translate learned representations into explicit control structures (Romero et al., 2024). A survey of fusion strategies between cognitive architectures and generative models maps design options and integration tactics across components (Liu et al., 2024). Concrete augmentation patterns instrument Soar and Sigma with large language models, outlining benefits, limitations, and required extensions for effective coupling (Joshi & Ustun, 2024).

Applied frameworks and case studies describe design patterns with modular memory, tools, and governance hooks in practical domains (Siyaev et al., 2023; Bai et al., 2024; Sumers et al., 2024). Domain-focused surveys and applications (e.g., robotics/embodiment) further motivate interface-level integration patterns for agent-based systems (Basumatari, 2025; Ugur et al., 2025). Human–agent collaboration benefits from logic-guided models of intent and role assignment, enabling safer, more efficient teamwork (Cao et al., 2023; Smirnov et al., 2023). Public-sector agent initiatives highlight requirements for interoperability, auditability, and oversight in at-scale deployments (of Estonia Information System Authority (RIA), 2021; Ilves, 2025).

Design Imperatives in Neurosymbolic Systems

System design must balance efficiency, transparency, safety, and accountability, with component choices tightly coupled across themes. Performance and cost are shaped by workload characteristics and joint hardware-software design (co-design), influencing feasible explainability and verification budgets (Wan et al., 2024b,a). Explainability quality depends on KR choices and user-centered protocols, affecting oversight effectiveness and trust (Lecue, 2020; Hogan et al., 2022; Islam et al., 2024). Reliability hinges on formal methods and training objectives that encode constraints or specifications, with verification scope informed by domain risk and regulatory expectations (Katz et al., 2017a; Elboher et al., 2020; Renkhoff et al., 2024). Governance and oversight frameworks provide constraints and audit requirements that shape tool-use, data access, and logging (Daws, 2018; Russell, 2019; Gunning et al., 2021; Schmitz et al., 2025).

Worked Examples: Interface Patterns in Practice

To make the interface-centric dimensions (Table 1) concrete without implying transfer from any single benchmark, we consolidate the worked example into the same running

scenario used throughout this paper: a manufacturing maintenance copilot (Section 8; application spotlight in Section **??**). The goal is to show, end to end, what is coupled, what is checked, and what should be measured.

Worked example: Manufacturing maintenance copilot (end-to-end interfaces)

Task setting. The system receives streaming sensor anomalies and operator reports, and must (i) diagnose likely fault causes, (ii) propose safe corrective actions, and (iii) produce an auditable trace suitable for oversight. The system therefore couples perception/prediction components with explicit knowledge and constraints (equipment KG/ontology, safety rules, and operating procedures).

Interfaces and artifacts (what becomes explicit).

- **Neuro→tool grounding (adjacent context):** retrieve manuals, maintenance logs, and relevant KG subgraphs to constrain the hypothesis space (Lewis et al., 2020b; Gao et al., 2024; Li et al., 2025; Tilwani et al., 2024).
- **Neuro→symbolic extraction (NeSy evidence):** emit typed artifacts such as (a) a structured diagnostic query/program over the equipment KG (e.g., SPARQL-like query), and (b) a structured work-order or plan skeleton (Chen et al., 2021; Tammet et al., 2024). Adjacent work on structured intermediate representations (e.g., scene graphs, discourse/action graphs) provides additional context on how typed artifacts can improve salience and factuality for downstream reasoning, though these works are not neurosymbolic evidence unless paired with explicit operator-level checking (Benetatos et al., 2023; Chen & Yang, 2021; Tan et al., 2020).
- **Symbolic execution/checking (NeSy evidence):** execute the query/program against the KG/reasoner; reject ill-formed or inconsistent artifacts; optionally attach proof trees or trace artifacts for provenance (Weir et al., 2024).
- **Inference-time safety enforcement (NeSy evidence):** filter or repair proposed actions using explicit safety constraints (rules or temporal-logic style shields) and record interventions (Alshiekh et al., 2018).
- **Optional planning/control coupling:** use explicit planners to improve long-horizon decision quality under constraints (e.g., sequencing maintenance steps) (Fabiano et al., 2023; Chatterjee et al., 2023).

What to measure (tie to the interface axes). In this example, "performance" and "trustworthiness" decompose into measurable interface properties: (i) **task success** (diagnosis accuracy; action utility under stated conditions), (ii) **invalid-artifact**

Table 10. Typical failure modes for the manufacturing maintenance copilot example, mapped to interface components. The purpose is to clarify what checks can and cannot certify under stated assumptions.

Component	Failure mode	What to report / mitigate
Retrieval/tool grounding	correct evidence not retrieved; biased/stale sources	retrieval coverage/latency; source versioning; audit logs of access
Typed artifact generation	well-formed but wrong query/plan; spurious structure	invalid-artifact rate; counterfactual tests; human review pathways
Symbolic execution/checking	checker is sound but incomplete; KB is inconsistent	KB quality/coverage; checker assumptions; trace/proof scope
Safety enforcement	constraint set incomplete; shield over-constrains actions	violation + intervention rates; assumption-scoped guarantees; rollback of constraints

rate (ill-formed queries/plans; rejected outputs), (iii) **safety-spec violation rate** and **intervention frequency** for the safety layer (assumption-scoped) (Achiam et al., 2017; Alshiekh et al., 2018), (iv) **provenance/trace availability and quality** (when proof/traces are produced), and (v) **end-to-end cost profile** (latency, compute, number of retrieval/execution/checking calls) (Wan et al., 2024b,a).

Discussion

This section positions our synthesis relative to prior surveys and theories, clarifies scope and novelty, and discusses broader implications for human-compatible AI. We first compare with related works and taxonomies, then consider impacts, risks, and policy considerations that inform practical deployment and governance.

To make cross-paper synthesis defensible, we avoid three recurring anti-patterns. First, we do not force techniques into rigid bins (e.g., treating prompting or retrieval as intrinsically "symbolic"); instead we describe systems by interface patterns and coding dimensions (Table 1). Second, we do not equate tool grounding (citations, retrieval, tool use) with correctness guarantees; tool grounding is treated as an interface that can reduce error rates but remains subject to dataset bias, benchmark artifacts, and residual hallucination (Davis, 2023). Third, we avoid qualitative scoring of architecture families without an evidence-tagged rubric: comparisons are scoped to specific systems/papers, tasks, and evaluation protocols, and claims are labeled as measured versus claimed (Section 11).

Comparison with Previous Works and Theories

Table 11 summarizes how this survey differs from recent complementary surveys and systematic reviews. We use these works as navigation aids for the reader (rather than as evidence categories), and we avoid implying that one taxonomy or survey scope is universally superior; instead, we make explicit which comparison dimensions (interfaces, guarantees, and costs) we emphasize.

This synthesis complements prior reviews by organizing the field around human-compatibility goals (themes) rather than methods alone, while mapping to system functions and evaluation practices. For NLP and knowledge-graph reasoning, structured reviews and surveys provide domain-specific taxonomies and evaluation emphases that we use for triangulation (Hamilton et al., 2024; DeLong et al., 2025). Trustworthy-AI lenses and governance perspectives supply adjacent evaluation criteria and deployment considerations (Acharya et al., 2024). Systems/workload perspectives motivate cost-aware comparisons that complement algorithm-centric summaries (Wan et al., 2024a). Surveys on RL/planning and cross-domain applications provide additional entry points where neurosymbolic interfaces are instantiated in sequential decision-making and applied domains (Zhang et al., 2021; Yu et al., 2023; Cheng et al., 2024). Meta-analyses and mappings aggregate the literature into architecture- or application-oriented

Table 11. Comparison with representative recent surveys and systematic reviews. Columns summarize each work's primary organizing lens and whether it explicitly separates evidence from context (e.g., measured vs. claimed), treats costs/guarantees as first-class dimensions, and distinguishes tool grounding from symbolic-operator coupling. This table is illustrative (not exhaustive).

Work (representative)	Primary organizing lens	Evidence separation	Costs / guarantees	Tool grounding vs. symbolic coupling
(Bhuyan et al., 2024) (broad survey)	broad taxonomy (representation, learning, reasoning, decision-making)	mostly narrative synthesis	discussed, but not a central comparison axis	may cover both; boundary varies by subtopic
(Renkhoff et al., 2024) (V&V)	verification/validation/testing taxonomy	explicit evaluation framing for V&V settings	central (assurance arguments, testability); guarantee scope emphasized	focuses on assurance; tool grounding treated case-by-case
(DeLong et al., 2025) (KG reasoning)	KG reasoning tasks and method families	task- and benchmark-oriented survey framing	costs/guarantees typically task-scoped	focuses on symbolic structure in KG pipelines; tool-use is peripheral
(Michel-Deletie & Sarker, 2025) (trustworthy NeSy)	systematic review lens focused on interpretability/trustworthiness	systematic review methodology; categorization dimensions	trustworthiness dimensions emphasized; costs variable	emphasizes symbolic structures; boundary depends on inclusion criteria
(Colelough & Regli, 2025) (systematic mapping)	mapping / taxonomy of NeSy research areas	systematic review-style aggregation	costs/guarantees typically not the primary axis	boundary depends on mapping categories
This survey	themes (performance, understandability, reliability, ethics) + interface patterns	explicit citation roles + evidence tags (Measured/Claimed/Not evaluated)	first-class comparison dimensions (cost profiles + guarantee scope)	strict boundary: tool grounding is context unless paired with typed artifacts + explicit operators

taxonomies; we use these primarily as navigational aids rather than as evidence categories (Bouneffouf & Aggarwal, 2022; Gibaut et al., 2023; Feldstein et al., 2024; Colelough & Regli, 2024, 2025). Compendia and broad surveys provide additional breadth and terminology alignment across subcommunities (Hitzler & Sarker, 2022; Hitzler et al., 2023; Wang et al., 2025b). Finally, surveys and position papers on neurosymbolic agents and system-level integration motivate how the same interface patterns recur in agentic pipelines (Yu et al., 2021; Belle et al., 2024; Bhuyan et al., 2024; Kishor, 2022; Bougzime et al., 2025; Silver & Mitchell, 2023). Related works in neural architecture optimization provide useful contrasts to hybrid approaches (Wang et al., 2021). Domain and scope-specific reviews situate advances in healthcare and visual reasoning (Frisoni et al., 2021; Hossain & Chen, 2025; Khan et al., 2025). Robotics and embodied settings add distinct integration constraints and evaluation practices (Basumatari, 2025; Ugur et al., 2025). Foundational and historical perspectives anchor the trajectory of neural-symbolic computing up to the present (Garcez et al., 2015; Besold et al., 2017; Garcez & Lamb,

2023). Complementary reviews emphasize the evolution of architectures, datasets, and evaluation practices (Sarker et al., 2022).

Systematic reviews of trustworthy neurosymbolic AI foreground interpretability-focused taxonomies and open questions (Michel-Deletie & Sarker, 2025).

Implications for Human-Compatible AI

Framing progress by themes emphasizes co-design of performance, understandability, reliability, and ethics for deployable systems. Policy-driven and principled approaches highlight value alignment, oversight structures, and operational accountability (Russell, 2019; Schmitz et al., 2025). National and sector programs illustrate governance requirements for agent-based AI in practice (of Economic Affairs and Communications, 2022; Ilves, 2025). Evaluation against structured levels of capability and autonomy facilitates clearer communication of risk and fitness for purpose (Morris et al., 2024). Beyond applications, AI is transforming the scientific process itself, with implications for governance, reproducibility, and collaboration (Roded & Slattery, 2025).

Future Directions and Open Challenges

Drawing from the preceding analysis, we propose concrete directions for advancing neurosymbolic AI. We group open problems into four areas: scalability and deeper integration; comprehensive trustworthiness and meta-cognition; research-to-engineering design principles; and meaningful evaluation, outlining evaluation criteria and test considerations for each.

Scalability and Deep Integration of Hybrid Architectures

Scalable NeSy systems require workload-aware acceleration (optimizing for the actual runtime/operator mix), memory-efficient KR integration, and differentiable reasoning backends (Susskind et al., 2021; Wan et al., 2024b,a). Promising directions include hardware-aware architecture search, stochastic logic programming, and declarative differentiable languages (Winters et al., 2022; Li et al., 2023b; Saha et al., 2024). Long-term visions for modular systems motivate deeper fusion of perception, KR, planning, and self-supervised objectives (LeCun, 2022).

Advancing Towards Comprehensive Trustworthiness and Meta-Cognition

Reliability at scale will combine formal verification of neural components with symbolic specifications and constraint-aware training objectives (Katz et al., 2017a; Elboher et al., 2020; Xie et al., 2022). Meta-cognitive strategies (self-monitoring that chooses among reasoning modes) can arbitrate between reasoning modes via conflict detection and self-monitoring, improving adaptation in novel contexts (Raja et al., 2024; Hu et al., 2025). Stress-testing for hallucination, compositional complexity, and shift remains essential to validate claims (Huang et al., 2025; Saxena et al., 2025; Shojaee et al., 2025).

From Research to Engineering: Design Principles

Methodological guidance should cover dataset governance, KR curation, tool-use auditing, and end-to-end testability. Standardized testing and measurement models can improve comparability and deployment readiness (Gundersen & Kjensmo, 2018; Zhang et al., 2022; Li et al., 2023a).

The Challenge of Meaningful Evaluation and Benchmarking

Evaluation must move beyond benchmark chasing to measure reasoning, tool grounding (retrieval/citations/tool use), and robustness with higher-quality datasets and protocols (Davis, 2023; Rogers et al., 2023; Orr & Kang, 2024). Multi-dimensional evaluators and exam-style test suites offer practical instruments for comparable assessment (Zhong et al., 2022; He et al., 2024; Zhong et al., 2024).

Conclusion

Performant neurosymbolic AI balances capability with efficiency. Progress stems from architecture and compiler choices, workload-aware design, and the strategic use of tool grounding, tools, and structured knowledge that is *reported* to improve factuality and task success in specific settings, while increasing measurable costs (latency, compute, tool calls). Hybrid planning and control can further improve long-horizon performance by pairing learning with search and symbolic structure. Going forward, reporting should pair accuracy with resource and latency measures, and evaluate system-level throughput under realistic constraints.

Understandable neurosymbolic AI puts knowledge representations and explicit reasoning at the center of explanation. KR-centric pipelines, intrinsic proof traces, and concept bottlenecks make model behavior inspectable and editable, enabling human oversight and targeted error analysis. However, evaluation must move beyond plausibility to measure faithfulness and usefulness with human-centered protocols, provenance, and versioned assets. Investments in KR quality, coverage, and maintenance are a recurring requirement for systems that rely on explicit KR.

Reliable neurosymbolic AI combines formal specifications, verifiable components, and robustness objectives to maintain integrity under shift and attack. Constraint satisfaction, shields, and differentiable planners can provide enforceable guarantees for specified properties under stated assumptions, while uncertainty-aware reasoning offers calibrated behavior in open settings. Reliability claims should be supported by standardized test suites, coverage criteria, and clearly scoped guarantees aligned to domain risk. Reproducible robustness requires transparent reporting of data, assumptions, and ablations.

Ethical neurosymbolic AI encodes values, rights, and duties into explicit structures that can be audited and refined. Human-in-the-loop protocols enable correction of concepts, rules, and policies; fairness assessments and mitigation plans should be integrated into the development lifecycle. Operational governance (roles, logging, and escalation) turns principles into accountable practice, especially for public-sector and high-stakes deployments. Documentation of limitations and routes to redress is essential for trust.

What we do not claim. Because neurosymbolic results are often setting-dependent, we do not claim that neurosymbolic methods are universally superior to purely neural or purely symbolic baselines, nor that tool grounding or tool use implies correctness. We also do not claim global safety or reliability from local checks: guarantees are interpreted as property- and assumption-scoped to the verified object and deployment conditions. Finally, we do not infer benefits (e.g., interpretability, robustness, fairness) unless they are explicitly evaluated in the cited work or clearly labeled as claimed rather than measured.

Outlook. One direction is deeper integration: scalable differentiable reasoning and KR backends, platform-aware co-design, and meta-cognitive control that arbitrates among reasoning modes. Methodological guidance and open assets (datasets, KR resources, evaluators, and reference implementations) can raise comparability and deployment readiness. By co-designing performance, interpretability, reliability, and ethics around concrete system functions and evaluation levers, the field can move toward systems that are understandable, dependable, and accountable.

Search Strategy and Information Sources

We queried leading AI venues and digital libraries covering learning, knowledge representation, reasoning, and systems between 2020 and 2025. Sources included prominent conference proceedings and journals, publisher portals, and indexing services. Representative query terms combined neurosymbolic and theme-specific keywords (e.g., logic, ILP, differentiable reasoning; KGs, symbol grounding, tool grounding, retrieval/tool use; planning/control, safe RL; verification, SMT; explainability, bottlenecks; governance). Searches were oriented toward identifying systems with an explicit symbolic representation and operator-level coupling, consistent with our interface-centric coding and evidence protocol (Sections 10 -).

Query log summary (representative). Table 12 summarizes representative query formulations used across sources.

Screening workflow. Records from multiple sources were consolidated and screened with AI-aided assistance using ASReview (Van De Schoot et al., 2021) to prioritize likely-relevant items, with final inclusion decisions made by the authors.
Notes. Dates are month-level to indicate when queries were last run.

Study Selection Flow and Counts

Screening and selection were performed with AI-aided assistance (ASReview) (Van De Schoot et al., 2021) over an initial candidate set of **912** consolidated records aggregated across sources. Table 13 reports the resulting stage counts as summarized for this manuscript. **Included technical studies in synthesis: 319.**
Notes. "Reports excluded" refers to records removed during title/abstract screening due to lack of explicit neurosymbolic coupling, insufficient evaluative detail, or out-of-scope framing.

Data Extraction Dimensions and Materials

For each included study we extracted: (i) problem abstraction and integration pattern; (ii) interface-centric coding dimensions (Table 1); (iii) data, tasks, and benchmarks; (iv) evaluation design and reported measures; (v) limitations and threats to validity; and (vi) alignment to the four themes and system functions. For evaluative statements and summary table cells, we tracked an evidence tag (**Measured**, **Claimed**, **Not evaluated**)

Table 12. Illustrative query-log summary showing information sources used during evidence collection, example query formulations, typical filters, and the approximate date last searched. This table is a reporting aid and does not imply equal coverage across sources.

Source	Query string (example)	Typical filters	Date
Semantic Scholar / OpenAlex	(neurosymbolic OR "neural symbolic") AND (logic OR "knowledge graph" OR "differentiable reasoning" OR ASP OR ILP OR SMT)	2020-2025; CS/AI; English	2026-01
arXiv	(neurosymbolic OR "neurosymbolic" OR "neural-symbolic") AND (logic OR "knowledge graph" OR ILP OR ASP OR SMT OR verifier OR "constrained decoding")	2020-2025; cs.AI; cs.CL; cs.LG	2026-01
ACL Anthology	(neurosymbolic) AND (reasoning OR "knowledge graph" OR "tool use" OR retrieval OR verification)	2020-2025; ACL; EMNLP; NAACL; Findings	2026-01
IEEE Xplore	("neuro symbolic" OR neurosymbolic) AND (verification OR safety OR "formal methods" OR "constraint" OR SMT)	2020-2025; journals; conferences	2026-01
ACM Digital Library	(neurosymbolic) AND (survey OR review OR evaluation OR explainability OR governance)	2020-2025; surveys; journals	2026-01
SpringerLink / IOS Press portals	(neurosymbolic) AND (knowledge representation OR ontology OR "semantic web" OR explainability)	2020-2025; AI; KR venues	2026-01

Table 13. Record screening and inclusion counts used for this survey. Stages reflect a typical identification → screening → eligibility workflow.

Stage	Records (n)
Records identified (all sources)	912
Records screened (title/abstract)	912
Reports excluded (title/abstract screening)	593
Reports assessed for eligibility (full text)	319
Studies included in synthesis	319

and explicit scope (task/domain and dataset/benchmark when available), consistent with Section 11 and Table 3. Where available, we noted code/data availability and license.

Acknowledgements

We used generative AI tools for manuscript preparation: an LLM-based writing assistant in the Cursor editor (Cursor, 2026), powered by OpenAI GPT models (OpenAI, 2026), was used to improve the syntax and grammar of several paragraphs and to draft early versions of some figure

captions and discussion text. All such edits were reviewed and revised by the authors, who take full responsibility for the final manuscript.

Declaration of conflicting interests

The authors declare no potential conflicts of interest.

Funding

This research was supported by the European Union and the Estonian Research Council through project TEM-TA141.

References

Aakur, S. N. & Sarkar, S. (2023). Leveraging Symbolic Knowledge Bases for Commonsense Natural Language Inference using Pattern Theory. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (pp. 1–18).

Abbas, D. A. (2025). Western Bias in AI: Why Global Perspectives Are Missing.

Abel, D., Dabney, W., Harutyunyan, A., Ho, M. K., Littman, M. L., Precup, D. & Singh, S. (2021). On the Expressivity of Markov Reward. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Volume 34 of *NIPS '21* (pp. 7799–7812). Red Hook, NY, USA: Curran Associates, Inc.

Acharya, K., Raza, W., Dourado, C., Velasquez, A. & Song, H. H. (2024). Neurosymbolic Reinforcement Learning and Planning: A Survey. *IEEE Transactions on Artificial Intelligence*, *5(5)*, 1939–1953.

Achiam, J., Held, D., Tamar, A. & Abbeel, P. (2017). Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 22–31). PMLR.

Ahmed, K., Chang, K.-W. & Broeck, G. V. d. (2023). A Pseudo-Semantic Loss for Autoregressive Models with Logical Constraints. Volume 36 of *NeurIPS 2023* (pp. 18325–18340). Curran Associates, Inc.

Aithal, S. G., Rao, A. B., B, C. C. & Singh, S. (2022). Application of Neuro-Symbolic Reasoning in Natural Language Processing. In *2022 IEEE 6th Conference on Information and Communication Technology (CICT)* (pp. 1–5). Gwalior, India: IEEE.

Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S. & Topcu, U. (2018). Safe reinforcement learning via shielding. In *Proceedings of the Thirty-Second*

AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18 (pp. 2669–2678). New Orleans, Louisiana, USA: AAAI Press.

Amizadeh, S., Palangi, H., Polozov, A., Huang, Y. & Koishida, K. (2020). Neuro-Symbolic Visual Reasoning: Disentangling. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 279–290). PMLR.

Annepaka, Y. & Pakray, P. (2025). Large language models: a survey of their development, capabilities, and applications. *Knowledge and Information Systems*, *67(3)*, 2967–3022.

Arabshahi, F., Lee, J., Gawarecki, M., Mazaitis, K., Azaria, A. & Mitchell, T. (2021). Conversational Neuro-Symbolic Commonsense Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35(6)*, 4902–4911.

Asai, M. & Muise, C. (2020). Learning Neural-Symbolic Descriptive Planning Models via Cube-Space Priors: The Voyage Home (to STRIPS). Volume 3 (pp. 2676–2682).

Ashley, K. D. & Aleven, V. (1997). Reasoning Symbolically About Partially Matched Cases. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (I)* (pp. 335–341). Nagoya.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y. & Ballas, N. (2023). Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 15619–15629). Vancouver, BC, Canada: IEEE.

Bader, S. & Hitzler, P. (2005). Dimensions of Neural-symbolic Integration - A Structured Survey.

Badreddine, S., Garcez, A. d., Serafini, L. & Spranger, M. (2022). Logic Tensor Networks. *Artificial Intelligence*, *303*, 103649.

Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*.

Bai, J., Mosbach, S., Taylor, C. J., Karan, D., Lee, K. F., Rihm, S. D., Akroyd, J., Lapkin, A. A. & Kraft, M. (2024). A dynamic knowledge graph approach to distributed self-driving laboratories. *Nature Communications*, *15(1)*, 462.

Barnden, J. A. (1989). Neural-Net Implementation of Complex Symbol-Processing in a Mental Model Approach to Syllogistic Reasoning. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (I)* (pp. 568–573). Detroit.

Basumatari, H. (2025). Neuro-Symbolic AI in Robotics: A State-of-the-Art Overview.

Belle, V., Fisher, M., Russo, A., Komendantskaya, E. & Nottle, A. (2024). Neuro-Symbolic AI + Agent Systems: A First Reflection on Trends, Opportunities and Challenges. In Amigoni, F. & Sinha, A. (Eds.), *Autonomous Agents and Multiagent Systems. Best and Visionary Papers* (pp. 180–200). Cham: Springer Nature Switzerland.

Belle, V. & Lakemeyer, G. (2011). On Progression and Query Evaluation in First-Order Knowledge Bases with Function Symbols. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* (pp. 744–749). Barcelona.

Benetatos, A., Diomataris, M., Pitsikalis, V. & Maragos, P. (2023). Generating Salient Scene Graphs with Weak Language Supervision. In *2023 31st European Signal Processing Conference (EUSIPCO)* (pp. 526–530). Helsinki, Finland: IEEE.

Bengio, Y., Lecun, Y. & Hinton, G. (2021). Deep learning for AI. *Commun. ACM*, *64(7)*, 58–65.

Besold, T. R., Garcez, A. d., Bader, S., Bowman, H., Domingos, P., Hitzler, P., Kuehnberger, K.-U., Lamb, L. C., Lowd, D., Lima, P. M. V., de Penning, H. L. H. L., Pinkas, G., Poon, H. & Zaverucha, G. (2017). Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. arXiv:1711.03902 [cs].

Besta, M., Memedi, F., Zhang, Z., Gerstenberger, R., Piao, G., Blach, N., Nyczyk, P., Copik, M., Kwaśniewski, G., Müller, J., Gianinazzi, L., Kubicek, A., Niewiadomski, H., O'Mahony, A., Mutlu, O. & Hoefler, T. (2024). Demystifying Chains, Trees, and Graphs of Thoughts.

Bhuyan, B. P., Ramdane-Cherif, A., Tomar, R. & Singh, T. P. (2024). Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications*, *36(21)*, 12809–12844.

Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J. & Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-relational Data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, Volume 26 of *NIPS'13* (pp. 2787–2795). Red Hook, NY, USA: Curran Associates, Inc.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A. & Choi, Y. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In Korhonen, A., Traum, D. & Màrquez, L. (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4762–4779). Florence, Italy: Association for Computational Linguistics.

Bougzime, O., Jabbar, S., Cruz, C. & Demoly, F. (2025). Unlocking the Potential of Generative AI through Neuro-Symbolic Architectures: Benefits and Limitations. arXiv:2502.11269 [cs] version: 1.

Bouneffouf, D. & Aggarwal, C. C. (2022). Survey on Applications of Neurosymbolic Artificial Intelligence.

Brachman, R. J., Gilbert, V. P. & Levesque, H. J. (1985). An Essential Hybrid Reasoning System: Knowledge and Symbol Level Accounts of KRYPTON. In *Readings in Artificial Intelligence and Databases* (pp. 532–540).

Brown, C. F., Kazmierski, M. R., Pasquarella, V. J., Rucklidge, W. J., Samsikova, M., Zhang, C., Shelhamer, E., Lahera, E., Wiles, O., Ilyushchenko, S., Gorelick, N., Zhang, L. L., Alj, S., Schechter, E., Askay, S., Guinan, O., Moore, R., Boukouvalas, A. & Kohli, P. (2025). AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T. & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712.

Cambria, E., Liu, Q., Decherchi, S., Xing, F. & Kwok, K. (2022). SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J. & Piperidis, S. (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3829–3839). Marseille, France: European Language Resources Association.

Cao, C., Fu, Y., Xu, S., Zhang, R. & Li, S. (2023). Enhancing Human-AI Collaboration Through Logic-Guided Reasoning.

Chakraborti, T., Sreedharan, S. & Kambhampati, S. (2020). The Emerging Landscape of Explainable Automated Planning & Decision Making. Volume 5 (pp. 4803–4811).

Chanin, D. & Hunter, A. (2023). Neuro-symbolic Commonsense Social Reasoning. arXiv:2303.08264 [cs].

Chatterjee, P., Chapagain, A., Chen, W. & Khardon, R. (2023). DiSProD: Differentiable Symbolic Propagation of Distributions for Planning. Volume 5 (pp. 5324–5332).

Chen, B., Hao, Z., Cai, X., Cai, R., Wen, W., Zhu, J. & Xie, G. (2019). Embedding Logic Rules Into Recurrent Neural Networks. *IEEE Access*, *7*, 14938–14946.

Chen, J. & Yang, D. (2021). Structure-Aware Abstractive Conversation Summarization via Discourse and Action Graphs. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T. & Zhou, Y. (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1380–1391). Online: Association for Computational Linguistics.

Chen, Q., Lamoreaux, A., Wang, X., Durrett, G., Bastani, O. & Dillig, I. (2021). Web question answering with neurosymbolic program synthesis. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, PLDI 2021 (pp. 328–343). New York, NY, USA: Association for Computing Machinery.

Chen, Y., Guo, L. & Yu, S. (2023a). Emergence of Symbols in Neural Networks for Semantic Understanding and Communication. arXiv:2304.06377 [cs].

Chen, Z., Sun, H., He, H. & Chen, P. (2023b). Learning from Noisy Crowd Labels with Logics. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)* (pp. 41–52). Anaheim, CA, USA: IEEE.

Chen, Z., Weiss, G., Mitchell, E., Celikyilmaz, A. & Bosselut, A. (2023c). RECKONING: reasoning through dynamic knowledge encoding. In *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23 (pp. 62579 – 62600). Red Hook, NY, USA: Curran Associates Inc.

Cheng, K., Ahmed, N. K., Rossi, R. A., Willke, T. & Sun, Y. (2024). Neural-Symbolic Methods for Knowledge Graph Reasoning: A Survey. *ACM Trans. Knowl. Discov. Data*, *18(9)*, 225:1–225:44.

Ciatto, G., Calegari, R. & Omicini, A. (2021). 2P-Kt: A logic-based ecosystem for symbolic AI. *SoftwareX*, *16*, 100817.

Clark, K., Hengst, B., Pagnucco, M., Rajaratnam, D., Robinson, P., Sammut, C. & Thielscher, M. (2016). A Framework for Integrating Symbolic and Sub-symbolic Representations. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 2486–2492). New York, NY, USA.

Cohen, D. (1983). Symbolic Execution of the Gist Specification Language. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (I)* (pp. 17–21).

Cohen, W. W., Yang, F. & Mazaitis, K. R. (2017). TensorLog: Deep Learning Meets Probabilistic DBs. arXiv:1707.05390 [cs].

Colelough, B. C. & Regli, W. (2024). Neuro-Symbolic AI in 2024: A Systematic Review. In *Proceedings of the First International Workshop on Logical Foundations of Neuro-Symbolic AI (LNSAI 2024) co-located with the 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024)*, LNSAI 2024.

Colelough, B. C. & Regli, W. (2025). Neuro-Symbolic AI in 2024: A Systematic Review. arXiv:2501.05435 [cs].

Craven, M. W. & Shavlik, J. W. (1995). Extracting Tree-Structured Representations of Trained Networks. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, Volume 8 of *NIPS'95* (pp. 24–30). Cambridge, MA, USA: MIT Press.

Crochepierre, L., Boudjeloud-Assala, L. & Barbesant, V. (2022). Interactive Reinforcement Learning for Symbolic Regression from Multi-Format Human-Preference Feedbacks. Volume 6 (pp. 5900–5903).

Cropper, A. & Morel, R. (2021). Learning programs by learning from failures. *Machine Learning*, *110(4)*, 801–856.

Cunnington, D., Law, M., Lobo, J. & Russo, A. (2023). Neuro-Symbolic Learning of Answer Set Programs from Raw Data. Volume 4 (pp. 3586–3596).

Cursor (2026). Cursor: The ai code editor. Accessed 2026-01-30.

Da, J., Bras, R. L., Lu, X., Choi, Y. & Bosselut, A. (2021). Analyzing Commonsense Emergence in Few-shot Knowledge Models. Automated Knowledge Base Construction (AKBC).

Dahlback, N. (1989). A Symbol Is Not A Symbol. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (I)* (pp. 8–14). Detroit.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. & Salakhutdinov, R. (2019). Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2978–2988). Florence, Italy: Association for Computational Linguistics.

Daniele, A., Campari, T., Malhotra, S. & Serafini, L. (2023). Deep Symbolic Learning: Discovering Symbols and Rules from Perceptions. Volume 4 (pp. 3597–3605).

Davis, E. (2023). Benchmarks for Automated Commonsense Reasoning: A Survey. *ACM Comput. Surv.*, *56(4)*, 81:1–81:41.

Daws, R. (2018). DARPA introduces 'third wave' of artificial intelligence.

DeLong, L. N., Mir, R. F. & Fleuriot, J. D. (2025). Neurosymbolic AI for Reasoning Over Knowledge Graphs: A Survey. *IEEE Transactions on Neural Networks and*

Learning Systems, *36(5)*, 7822–7842.

Demir, C. & Ngomo, A.-C. N. (2023). Neuro-Symbolic Class Expression Learning. Volume 4 (pp. 3624–3632).

Ding, L. (2007). A Model of Hierarchical Knowledge Representation – Toward Knowware for Intelligent Systems. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *11(10)*, 1232–1240.

Donadello, I., Serafini, L. & Garcez, A. d. (2017). Logic Tensor Networks for Semantic Image Interpretation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 1596–1602).

Dong, H., Mao, J., Lin, T., Wang, C., Li, L. & Zhou, D. (2019). Neural Logic Machines. ICLR '19.

Dumancic, S., Garcia-Duran, A. & Niepert, M. (2019). A Comparative Study of Distributional and Symbolic Paradigms for Relational Learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 6088–6094). Macao.

Dwivedi, V. P. & Bresson, X. (2020). A Generalization of Transformer Networks to Graphs.

of Economic Affairs and Communications, M. (2022). Estonia's National Artificial Intelligence Strategy (Kratt Strategy) for 2022–2023 | Digital Watch Observatory.

Eiter, T., Geibinger, T., Higuera, N. & Oetsch, J. (2023). A Logic-based Approach to Contrastive Explainability for Neurosymbolic Visual Question Answering. Volume 4 (pp. 3668–3676).

El-Kishky, A., Selsam, D., Song, F., Parascandolo, G., Ren, H., Lightman, H., Chung, H. W., Akkaya, I., Sutskever, I., Wei, J., Gordon, J., Cobbe, K., Yu, K., Kondraciuk, L., Schwarzer, M., Rohaninejad, M., Brown, N., Zhao, S., Bansal, T., Kosaraju, V. & Zhou, W. (2024). Learning to reason with LLMs.

Elboher, Y. Y., Gottschlich, J. & Katz, G. (2020). An Abstraction-Based Framework for Neural Network Verification. In Lahiri, S. K. & Wang, C. (Eds.), *Computer Aided Verification* (pp. 43–65). Cham: Springer International Publishing.

Elia, M., Stieler, F., Ripke, F., Nann, M., Dopfer, S. & Bauer, B. (2024). Towards Certifiable AI in Medicine: Illustrated for Multi-label ECG Classification Performance Metrics. In *2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)* (pp. 1–8). Madrid, Spain: IEEE.

of Estonia Information System Authority (RIA), R. (2021). Bürokratt – a single chatbot for Estonia | Interoperable Europe Portal.

Evans, R. & Grefenstette, E. (2018). Learning Explanatory Rules from Noisy Data. *Journal of Artificial Intelligence Research*, *61*, 1–64.

Fabiano, F., Pallagani, V., Ganapini, M. B., Horesh, L., Loreggia, A., Murugesan, K., Rossi, F. & Srivastava, B. (2023). Plan-SOFAI: A Neuro-Symbolic Planning Architecture.

Fang, T., Zhang, H., Wang, W., Song, Y. & He, B. (2021). DISCOS: Bridging the Gap between Discourse Knowledge and Commonsense Knowledge. In *Proceedings of the Web Conference 2021*, WWW '21 (pp. 2648–2659). New York, NY, USA: Association for Computing Machinery.

Feldstein, J., Dilkas, P., Belle, V. & Tsamoura, E. (2024). Mapping the Neuro-Symbolic AI Landscape by Architectures: A Handbook on Augmenting Deep Learning Through Symbolic Reasoning.

Frisoni, G., Moro, G. & Carbonaro, A. (2021). A Survey on Event Extraction for Natural Language Understanding: Riding the Biomedical Literature Wave. *IEEE Access*, *9*, 160721–160757.

Frixione, M. & Spinelli, G. (1989). Symbols and subsymbols for representing knowledge: a catalogue raisonne. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (I)* (pp. 3–7). Detroit.

Ganguly, P. & Mukherjee, I. (2025). Bridging the Gap: The Rise of Neurosymbolic Artificial Intelligence in Advanced Computing. *IT Professional*, *27(2)*, 48–53.

Gao, S., Borges, B., Oh, S., Bayazit, D., Kanno, S., Wakaki, H., Mitsufuji, Y. & Bosselut, A. (2023). PeaCoK: Persona Commonsense Knowledge for Consistent and Engaging Narratives. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6569–6591). Toronto, Canada: Association for Computational Linguistics.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M. & Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs].

Garcez, A. (2025). Neurosymbolic AI Could Be the Answer to Hallucination in Large Language Models.

Garcez, A. d., Gori, M., Lamb, L. C., Serafini, L., Spranger, M. & Tran, S. N. (2019). Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning. arXiv:1905.06088.

Garcez, A. d. & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd Wave. arXiv:2012.05876.

Garcez, A. d. & Lamb, L. C. (2023). Neurosymbolic AI: the 3rd wave. *Artificial Intelligence Review*, *56(11)*, 12387–12406.

Garcez, A. D. A., Besold, T. R., Raedt, L. D., Földiak, P., Hitzler, P., Icard, T., Kai-Uwe Kühnberger, Lamb, L. C., Miikkulainen, R. & Silver, D. L. (2015). Neural-Symbolic Learning and Reasoning: Contributions and Challenges. In *Papers from the 2015 AAAI Spring Symposium*, number No. 3: Knowledge Representation and Reasoning: Integrating Symbolic and Neural Approaches in AAAI '15. AAAI Press.

d'Avila Garcez, A. S., Broda, K. B. & Gabbay, D. M. (2002a). Neural-Symbolic Integration: The Road Ahead. In A. S. d'Avila Garcez, K. B. Broda & D. M. Gabbay (Eds.), *Neural-Symbolic Learning Systems: Foundations and Applications* (pp. 235–252). London: Springer.

d'Avila Garcez, A. S., Broda, K. B. & Gabbay, D. M. (2002b). *Neural-Symbolic Learning Systems*. Perspectives in Neural Computing. London: Springer.

Gibaut, W., Pereira, L., Grassiotto, F., Osorio, A., Gadioli, E., Munoz, A., Gomes, S. & Santos, C. d. (2023). Neurosymbolic AI and its Taxonomy: a survey.

Glauer, M., Mossakowski, T., Neuhaus, F., Memariani, A. & Hastings, J. (2023). Chapter 21. Neuro-Symbolic Semantic Learning for Chemistry. In P. Hitzler, M. K. Sarker & A. Eberhart (Eds.), *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Graves, A. & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, *18(5-6)*, 602–610.

Gundersen, O. E. & Kjensmo, S. (2018). State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32(1)*.

Gunning, D., Vorm, E., Wang, J. Y. & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, *2(4)*, e61.

Hamilton, K., Nayak, A., Božić, B. & Longo, L. (2024). Is neuro-symbolic AI meeting its promises in natural language processing? A structured review. *Semantic Web*, *15(4)*, 1265–1306.

Han, X., Cao, S., Lv, X., Lin, Y., Liu, Z., Sun, M. & Li, J. (2018). OpenKE: An Open Toolkit for Knowledge Embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 139–144). Brussels, Belgium: Association for Computational Linguistics.

Hazra, R., Venturato, G., Martires, P. Z. D. & Raedt, L. D. (2025). Can Large Language Models Reason? A Characterization via 3-SAT.

He, C., Luo, R., Hu, S., Zhao, R., Zhou, J., Wu, H., Zhang, J., Han, X., Liu, Z. & Sun, M. (2024). UltraEval: A Lightweight Platform for Flexible and Comprehensive Evaluation for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)* (pp. 247–257). Bangkok, Thailand: Association for Computational Linguistics.

Hessel, J., Marasovic, A., Hwang, J. D., Lee, L., Da, J., Zellers, R., Mankoff, R. & Choi, Y. (2023). Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 688–714). Toronto, Canada: Association for Computational Linguistics.

Himabindu, M., V, R., Gupta, M., Rana, A., Chandra, P. K. & Abdulaali, H. S. (2023). Neuro-Symbolic AI: Integrating Symbolic Reasoning with Deep Learning. In *2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Volume 10 (pp. 1587–1592).

Hitzler, P., Bianchi, F., Ebrahimi, M. & Sarker, M. K. (2020). Neural-symbolic integration and the Semantic Web. *Semantic Web*, *11(1)*, 3–11.

Hitzler, P., Eberhart, A., Ebrahimi, M., Sarker, M. K. & Zhou, L. (2022). Neuro-symbolic approaches in artificial intelligence. *National Science Review*, *9(6)*, nwac035.

Hitzler, P., Ebrahimi, M., Sarker, M. K. & Stepanova, D. (2024). Neuro-symbolic AI and the semantic web. *Semantic Web*, *15(4)*, 1261–1263.

Hitzler, P. & Sarker, M. K. (2022). *Neuro-Symbolic Artificial Intelligence: The State of the Art*. IOS Press. Google-Books-ID: uFtcEAAAQBAJ.

Hitzler, P., Sarker, M. K. & Eberhart, A. (Eds.). (2023). *Compendium of Neurosymbolic Artificial Intelligence*, Volume 369 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S. & Zimmermann, A. (2022). Knowledge Graphs. *ACM Computing Surveys*, *54(4)*, 1–37.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79(8)*, 2554–2558.

Horvatić, D. & Lipic, T. (2021). Human-Centric AI: The Symbiosis of Human and Artificial Intelligence. *Entropy*, *23(3)*, 332.

Hossain, D. & Chen, J. Y. (2025). A Study on Neuro-Symbolic Artificial Intelligence: Healthcare Perspectives.

Hu, W.-C., Dai, W.-Z., Jiang, Y. & Zhou, Z.-H. (2025). Efficient Rectification of Neuro-Symbolic Reasoning Inconsistencies by Abductive Reflection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *39(16)*, 17333–17341.

Hu, Z., Ma, X., Liu, Z., Hovy, E. & Xing, E. (2016). Harnessing Deep Neural Networks with Logic Rules. In Erk, K. & Smith, N. A. (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2410–2420). Berlin, Germany: Association for Computational Linguistics.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. & Liu, T. (2025). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, *43(2)*, 42:1–42:55.

Hwang, J. D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A. & Choi, Y. (2021). (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35(7)*, 6384–6392.

Ignatiev, A., Marques-Silva, J., Narodytska, N. & Stuckey, P. J. (2021). Reasoning-Based Learning of Interpretable ML Models. Volume 5 (pp. 4458–4465).

Ilves, L. (2025). The Agentic State: How Agentic AI Will Revamp 10 Functional Layers of Public Administration.

Islam, M. A., Mridha, M. F., Jahin, M. A. & Dey, N. (2024). A Unified Framework for Evaluating the Effectiveness and Enhancing the Transparency of Explainable AI Methods in Real-World Applications. arXiv:2412.03884 [cs].

Ismayilzada, M. & Bosselut, A. (2023). kogito: A Commonsense Knowledge Inference Toolkit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 96–104). Dubrovnik, Croatia: Association for Computational Linguistics.

Jaeger, M. (2023). Learning and reasoning with graph data. *Frontiers in Artificial Intelligence*, *6*, 1124718.

Jain, N., Domingues, A., Baokar, A., Penuela, A. M. & Simperl, E. (2025). Towards Interpretable Embeddings: Aligning Representations with Semantic Aspects. *Neurosymbolic Artificial Intelligence*.

James, S. (2018). Learning Portable Symbolic Representations. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*

(pp. 5765–5766). Stockholm, Sweden.

Jana, P. (2024). NeuroSymbolic LLM for Mathematical Reasoning and Software Engineering. Volume 9 (pp. 8492–8493).

Jeong, J., Jaggi, P. & Sanner, S. (2021). Symbolic Dynamic Programming for Continuous State MDPs with Linear Program Transitions. Volume 4 (pp. 4083–4089).

Ji, S., Pan, S., Cambria, E., Marttinen, P. & Yu, P. S. (2022). A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, *33(2)*, 494–514.

Joshi, H. & Ustun, V. (2024). Augmenting Cognitive Architectures with Large Language Models. *Proceedings of the AAAI Symposium Series*, *2(1)*, 281–285.

Jurafsky, D. & Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*.

Järv, P., Tammet, T., Verrev, M. & Draheim, D. (2022). Knowledge Integration for Commonsense Reasoning with Default Logic:. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* (pp. 148–155). Valletta, Malta: SCITEPRESS - Science and Technology Publications.

Järv, P., Tammet, T., Verrev, M. & Draheim, D. (2023). Large-Scale Commonsense Knowledge for Default Logic Reasoning. *SN Computer Science*, *4(5)*, 550.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux. Google-Books-ID: ZuKTvERuPG8C.

Kalyanpur, A., Breloff, T. & Ferrucci, D. A. (2022). Braid: Weaving Symbolic and Neural Knowledge into Coherent Logical Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36(10)*, 10867–10874.

Kant, M., Nabi, M., Kant, M., Carlson, P. & Ma, M. (2024). Equitable Access to Justice: Logical LLMs Show Promise.

Karia, R. & Srivastava, S. (2022). Relational Abstractions for Generalized Reinforcement Learning on Symbolic Problems. Volume 4 (pp. 3135–3142).

Katz, G., Barrett, C., Dill, D. L., Julian, K. & Kochenderfer, M. J. (2017a). Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In Majumdar, R. & Kunčak, V. (Eds.), *Computer Aided Verification* (pp. 97–117). Cham: Springer International Publishing.

Katz, G., Barrett, C., Dill, D. L., Julian, K. & Kochenderfer, M. J. (2017b). Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In R. Majumdar &

V. Kunčak (Eds.), *Computer Aided Verification*, Volume 10426 (pp. 97–117). Cham: Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Kau, A. (2024). Combining Knowledge Graphs With Language Models for Interpretability.

Kautz, H. A. (2022). The third AI summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine*, *43(1)*, 105–125.

Keber, M., Grubišić, I., Barešić, A. & Jović, A. (2024). A Review on Neuro-symbolic AI Improvements to Natural Language Processing. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)* (pp. 66–72). Opatija, Croatia: IEEE.

Khan, M. J., Ilievski, F., Breslin, J. G. & Curry, E. (2025). A survey of neurosymbolic visual reasoning with scene graphs and common sense knowledge. *Neurosymbolic Artificial Intelligence*, *1*, NAI–240719.

Kim, J. T., Kim, S. & Petersen, B. K. (2020). An Interactive Visualization Platform for Deep Symbolic Regression. Volume 5 (pp. 5261–5263).

Kimura, D., Ono, M., Chaudhury, S., Kohita, R., Wachi, A., Agravante, D. J., Tatsubori, M., Munawar, A. & Gray, A. (2021). Neuro-Symbolic Reinforcement Learning with First-Order Logic. In Moens, M.-F., Huang, X., Specia, L. & Yih, S. W.-t. (Eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3505–3511). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Kirk, J. R. & Laird, J. E. (2019). Learning Hierarchical Symbolic Representations to Support Interactive Task Learning and Knowledge Transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 6095–6102). Macao.

Kishor, R. (2022). Neuro-Symbolic AI: Bringing a new era of Machine Learning. *International Journal of Research Publication and Reviews*, *03(12)*, 2326–2336.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B. & Liang, P. (2020). Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 5338–5348). PMLR.

Kolb, S., Mladenov, M., Sanner, S., Belle, V. & Kersting, K. (2018). Efficient Symbolic Integration for Probabilistic Inference. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (pp. 5031–5037). Stockholm, Sweden.

Kouvaros, P. (2023). Towards Formal Verification of Neuro-symbolic Multi-agent Systems. Volume 6 (pp. 7014–7019).

Kouvaros, P., Botoeva, E. & Bonis-Campbell, C. D. (2024). Formal Verification of Parameterised Neural-symbolic Multi-agent Systems. Volume 1 (pp. 103–110).

Kouvaros, P., Lomuscio, A. & Pirovano, E. (2018). Symbolic Synthesis of Fault-Tolerance Ratios in Parameterised Multi-Agent Systems (pp. 324–330).

Kramer, S. (2020). A Brief History of Learning Symbolic Higher-Level Representations from Data (And a Curious Look Forward). Volume 5 (pp. 4868–4876).

Kılınç, S. (2024). Comprehensive AI assessment framework: Enhancing educational evaluation with ethical AI integration. *Journal of Educational Technology and Online Learning*, *7(4 - ICETOL 2024 Special Issue)*, 521–540.

Laird, J. E., Lebiere, C. & Rosenbloom, P. S. (2017). A Standard Model of the Mind: Toward a Common Computational Framework Across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine*, *38(4)*, 13–26.

Lamb, L. C., Garcez, A. d., Gori, M., Prates, M. O. R., Avelar, P. H. C. & Vardi, M. Y. (2020). Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective. Volume 5 (pp. 4877–4884).

Lecue, F. (2020). On the role of knowledge graphs in explainable AI. *Semantic Web*, *11(1)*, 41–51.

LeCun, Y. (2022). A Path Towards Autonomous Machine Intelligence.

Lenat, D. & Marcus, G. (2023). Getting from Generative AI to Trustworthy AI: What LLMs might learn from Cyc. arXiv:2308.04445 [cs].

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L. (2020a). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). Online: Association for Computational Linguistics.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S. & Kiela, D. (2020b). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F. & Lin, H. (Eds.), *Advances in Neural Information Processing Systems*, Volume 33 (pp. 9459–9474). Curran Associates, Inc.

Li, F., Zhu, J., Yan, H. & Zhang, Z. (2022). Grammatically Derived Factual Relation Augmented Neural Machine Translation. *Applied Sciences*, *12(13)*, 6518.

Li, L., Huang, Y., Cui, X., Cheng, X. & Liu, X. (2023a). On Testing and Evaluation of Artificial Intelligence Models. In *2023 IEEE International Conference on Sensors,*

Electronics and Computer Engineering (ICSECE) (pp. 92–97). Jinzhou, China: IEEE.

Li, X., Jin, J., Zhou, Y., Zhang, Y., Zhang, P., Zhu, Y. & Dou, Z. (2025). From Matching to Generation: A Survey on Generative Information Retrieval. arXiv:2404.14851 [cs].

Li, Z., Huang, J. & Naik, M. (2023b). Scallop: A Language for Neurosymbolic Programming. *Proceedings of the ACM on Programming Languages*, *7(PLDI)*, 1463–1487.

Liu, J., Pan, Z., Xu, J., Liang, B., Chen, Y. & Ji, W. (2018). Quality-time-complexity universal intelligence measurement. *International Journal of Crowd Science*, *2(2)*, 99–107.

Liu, X., Lu, Z. & Mou, L. (2023). Chapter 30. Weakly Supervised Reasoning by Neuro-Symbolic Approaches. In P. Hitzler, M. K. Sarker & A. Eberhart (Eds.), *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Liu, Y., Liu, Y. & Shen, C. (2024). Combining Minds and Machines: Investigating the Fusion of Cognitive Architectures and Generative Models for General Embodied Intelligence. *Proceedings of the AAAI Symposium Series*, *2(1)*, 307–314.

Liu, Y., Nan, Y., Xu, W., Hu, X., Ye, L., Qin, Z. & Liu, P. (2025). AlphaGo Moment for Model Architecture Discovery. arXiv:2507.18074 [cs].

Lu, Z., Afridi, I., Kang, H. J., Ruchkin, I. & Zheng, X. (2024). Surveying neuro-symbolic approaches for reliable artificial intelligence of things. *Journal of Reliable Intelligent Environments*, *10(3)*, 257–279.

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Wang, G., Roberts, D. L., Taylor, M. E. & Littman, M. L. (2017). Interactive Learning from Policy-Dependent Human Feedback. In *Proceedings of the 34th International Conference on Machine Learning*, Volume 70 of *Proceedings of Machine Learning Research* (pp. 2285–2294). PMLR.

Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T. & De Raedt, L. (2018). DeepProbLog: Neural Probabilistic Logic Programming. In *Advances in Neural Information Processing Systems*, Volume 31. Curran Associates, Inc.

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B. & Wu, J. (2019a). The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision.

Mao, J., Zhang, X., Li, Y., Freeman, W. T., Tenenbaum, J. B. & Wu, J. (2019b). Program-Guided Image Manipulators (pp. 4030–4039).

Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. arXiv:2002.06177.

Marra, G. (2024). From Statistical Relational to Neuro-Symbolic Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38(20)*, 22678–22678.

Marra, G., Dumančić, S., Manhaeve, R. & De Raedt, L. (2024). From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence*, *328*, 104062.

Mastrogiovanni, F., Sgorbissa, A. & Zaccaria, R. (2007). A Distributed Architecture for Symbolic Data Fusion. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence* (pp. 2153–2158).

McCormack, L. & Bendechache, M. (2024). A comprehensive survey and classification of evaluation criteria for trustworthy artificial intelligence. *AI and Ethics*.

McDonald, C., Malloy, T., Nguyen, T. N. & Gonzalez, C. (2024). Exploring the Path from Instructions to Rewards with Large Language Models in Instance-Based Learning. *Proceedings of the AAAI Symposium Series*, *2(1)*, 334–339.

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y. & Scialom, T. (2023). Augmented Language Models: a Survey. arXiv:2302.07842.

Michel-Deletie, C. & Sarker, M. K. (2025). Neuro-Symbolic methods for Trustworthy AI: a systematic review with a focus on interpretability | Neurosymbolic Artificial Intelligence.

Mihaylov, T., Clark, P., Khot, T. & Sabharwal, A. (2018). Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2381–2391). Brussels, Belgium: Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs].

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38(11)*, 39–41.

Minato, S.-i., Satoh, K. & Sato, T. (2007). Compiling Bayesian Networks by Symbolic Probability Calculation Based on Zero-suppressed BDDs. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence* (pp. 2550–2555).

Mitchener, L., Tuckey, D., Crosby, M. & Russo, A. (2022). Detect, Understand, Act: A Neuro-Symbolic Hierarchical Reinforcement Learning Framework (Extended Abstract). Volume 6 (pp. 5314–5318).

Mooney, R., Shavlik, J., Towell, G. & Gove, A. (1989). An Experimental Comparison of Symbolic and Connectionist Learning Algorithms. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (I)* (pp. 775–780). Detroit.

Moran, T. P. (1973). The Symbolic Nature Of Visual Imagery. In *Proceedings of the Third International Joint Conference on Artificial Intelligence*. Stanford University, California.

Morris, M. (2022). Learning Proof Path Selection Policies in Neural Theorem Proving. *4th Conference on Automated Knowledge Base Construction (AKBC)*.

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C. & Legg, S. (2024). Position: Levels of AGI for Operationalizing Progress on the Path to AGI. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24 (pp. 36308–36321). Vienna, Austria: PMLR.

Mostafazadeh, N., Kalyanpur, A., Moon, L., Buchanan, D., Berkowitz, L., Biran, O. & Chu-Carroll, J. (2020). GLUCOSE: GeneraLized and COntextualized Story Explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4569–4586). Online: Association for Computational Linguistics.

Murali, A., Sehgal, A., Krogmeier, P. & Madhusudan, P. (2022). Composing Neural Learning and Symbolic Reasoning with an Application to Visual Discrimination. Volume 4 (pp. 3358–3365).

Núñez-Molina, C. (2022). Application of Neurosymbolic AI to Sequential Decision Making. Volume 6 (pp. 5863–5864).

Núñez-Molina, C., Mesejo, P. & Fernández-Olivares, J. (2024). A Review of Symbolic, Subsymbolic and Hybrid Methods for Sequential Decision Making. *ACM Computing Surveys*, *56(11)*, 1–36.

Odense, S. & Garcez, A. d. (2022). A Semantic Framework for Neuro-Symbolic Computing.

Oltramari, A. (2023a). Cognitive Neuro-Symbolic Reasoning Systems. In *Proceedings of the AAAI Symposium Series*, AAAI-SS '23. AAAI Press.

Oltramari, A. (2023b). A Path Towards High-Level Reasoning Through Cognitive Neuro-Symbolic Systems. *Neurosymbolic Artificial Intelligence*.

Oltramari, A. (2024). Enabling High-Level Machine Reasoning with Cognitive Neuro-Symbolic Systems. *Proceedings of the AAAI Symposium Series*, *2(1)*, 360–368.

Oltramari, A., Francis, J., Ilievski, F., Ma, K. & Mirzaee, R. (2021). Chapter 13. Generalizable Neuro-Symbolic Systems for Commonsense Question Answering. In

P. Hitzler & M. K. Sarker (Eds.), *Frontiers in Artificial Intelligence and Applications*. IOS Press.

OpenAI (2026). Openai. Accessed 2026-01-30.

Orr, W. & Kang, E. B. (2024). AI as a Sport: On the Competitive Epistemologies of Benchmarking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1875–1884). Rio de Janeiro Brazil: ACM.

Pallagani, V., Muppasani, B., Srivastava, B., Rossi, F., Horesh, L., Murugesan, K., Loreggia, A., Fabiano, F., Joseph, R. & Kethepalli, Y. (2023). Plansformer Tool: Demonstrating Generation of Symbolic Plans Using Transformers. Volume 6 (pp. 7158–7162).

de Penning, H. L. H. L., Garcez, A. d., Lamb, L. C. & Meyer, J.-J. C. (2011). A Neural-Symbolic Cognitive Agent for Online Learning and Reasoning. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* (pp. 1653–1658). Barcelona.

Perevalov, A., Diefenbach, D., Usbeck, R. & Both, A. (2022). QALD-9-plus: A Multilingual Dataset for Question Answering over DBpedia and Wikidata Translated by Native Speakers.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 2227–2237). New Orleans, Louisiana: Association for Computational Linguistics.

Plasser, M., Peter, S. & Widmer, G. (2023). Discrete Diffusion Probabilistic Models for Symbolic Music Generation. Volume 6 (pp. 5842–5850).

Qian, H., Marinescu, R., Gray, A., Bhattacharjya, D., Barahona, F., Gao, T., Riegel, R. & Sahu, P. (2022). Logical Credal Networks. In *Proceedings of the 2022 Conference on Advances in Neural Information Processing Systems*, NeurIPS 2022 (pp. 15325–15337). Curran Associates, Inc.

Qiao, S., Ou, Y., Zhang, N., Chen, X., Yao, Y., Deng, S., Tan, C., Huang, F. & Chen, H. (2023). Reasoning with Language Model Prompting: A Survey. In Rogers, A., Boyd-Graber, J. & Okazaki, N. (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5368–5393). Toronto, Canada: Association for Computational Linguistics.

Qu, M. & Tang, J. (2019). Probabilistic logic neural networks for reasoning. In *Proceedings of the 33rd International Conference on Neural Information Processing*

Systems, number 693 (pp. 7712–7722). Red Hook, NY, USA: Curran Associates Inc.

Raedt, L. d., Dumančić, S., Manhaeve, R. & Marra, G. (2020). From Statistical Relational to Neuro-Symbolic Artificial Intelligence. Volume 5 (pp. 4943–4950).

Raja, A., Leshchenko, A. & Kim, J. (2024). Leveraging Conflict to Bridge Cognitive Reasoning and Generative Algorithms. *Proceedings of the AAAI Symposium Series*, *2(1)*, 391–395.

Rajabi, E. & Etminani, K. (2024). Knowledge-graph-based explainable AI: A systematic review. *Journal of Information Science*, *50(4)*, 1019–1029.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M. & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation.

Renkhoff, J., Feng, K., Meier-Doernberg, M., Velasquez, A. & Song, H. H. (2024). A Survey on Verification and Validation, Testing and Evaluations of Neurosymbolic Artificial Intelligence. *IEEE Transactions on Artificial Intelligence*, *5(8)*, 3765–3779.

Rezazadegan, R., Sharifzadeh, M. & Magee, C. L. (2024). Quantifying the progress of artificial intelligence subdomains using the patent citation network. *Scientometrics*, *129(5)*, 2559–2581.

Riegel, R., Gray, A., Luus, F., Khan, N., Makondo, N., Akhalwaya, I. Y., Qian, H., Fagin, R., Barahona, F., Sharma, U., Ikbal, S., Karanam, H., Neelam, S., Likhyani, A. & Srivastava, S. (2020). Logical Neural Networks. arXiv:2006.13155 [cs].

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D. & Mundlos, S. (2008). The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics*, *83(5)*, 610–615.

Rocktäschel, T. & Riedel, S. (2016). Learning Knowledge Base Inference with Neural Theorem Provers. In Pujara, J., Rocktaschel, T., Chen, D. & Singh, S. (Eds.), *Proceedings of the 5th Workshop on Automated Knowledge Base Construction* (pp. 45–50). San Diego, CA: Association for Computational Linguistics.

Rocktäschel, T. & Riedel, S. (2017). End-to-end Differentiable Proving. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Volume 30 of *NIPS'17* (pp. 3791–3803). Red Hook, NY, USA: Curran Associates, Inc.

Roded, T. & Slattery, P. (2025). AI and the Future of Scientific Discovery.

Rogers, A., Gardner, M. & Augenstein, I. (2023). QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Computing Surveys*, *55(10)*, 1–45.

Romero, O. J., Zimmerman, J., Steinfeld, A. & Tomasic, A. (2024). Synergistic Integration of Large Language Models and Cognitive Architectures for Robust AI: An Exploratory Analysis. *Proceedings of the AAAI Symposium Series*, *2(1)*, 396–405.

Rosa, J. L. G. & Franeozo, E. (1999). Hybrid Thematic Role Processor: Symbolic Linguistic Relations Revised by Connectionist Learning. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (II)* (pp. 852–857).

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65(6)*, 386–408. Place: US.

Roy, K., Wu, S. & Oltramari, A. (2025). Enhancing Foundation Model-Based Reasoning with Neuro-Symbolic Cognitive Methods. In *Handbook on Neurosymbolic AI and Knowledge Graphs* (pp. 712–739). IOS Press.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323(6088)*, 533–536.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group. Google-Books-ID: M1eFDwAAQBAJ.

Russell, S. J. (1989). Execution architectures and compilation. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (I)* (pp. 15–20). Detroit.

Russell, S. J. & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. Pearson. Google-Books-ID: koFptAEACAAJ.

Sacks, E. (1989). An Approximate Solver for Symbolic Equations. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (I)* (pp. 431–434). Detroit.

Saha, S. S., Sandha, S. S., Aggarwal, M., Wang, B., Han, L., Briseno, J. D. G. & Srivastava, M. (2024). TinyNS: Platform-aware Neurosymbolic Auto Tiny Machine Learning. *ACM Trans. Embed. Comput. Syst.*, *23(3)*, 43:1–43:48.

Sahoo, P., Meharia, P., Ghosh, A., Saha, S., Jain, V. & Chadha, A. (2024). A Comprehensive Survey of Hallucination in Large Language, Image, Video and Audio Foundation Models. In Al-Onaizan, Y., Bansal, M. & Chen, Y.-N. (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 11709–11724). Miami, Florida, USA: Association for Computational Linguistics.

Sarker, M. K., Zhou, L., Eberhart, A. & Hitzler, P. (2022). Neuro-symbolic artificial intelligence: Current trends. *AI Communications*, *34(3)*, 197–209.

Sato, T. & Kameya, Y. (1997). PRISM : A Language for Symbolic-Statistical Modeling. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (II)* (pp. 1330–1335). Nagoya.

Saxena, V., Sathe, A. & Sandosh, S. (2025). Mitigating Hallucinations in Large Language Models: A Comprehensive Survey on Detection and Reduction Strategies. In Bansal, J. C., Jamwal, P. K. & Hussain, S. (Eds.), *Sustainable Computing and Intelligent Systems* (pp. 39–52). Singapore: Springer Nature.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N. & Scialom, T. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools.

Schmitz, C., Rystrøm, J. & Batzner, J. (2025). Oversight Structures for Agentic AI in Public-Sector Organizations. In Kamalloo, E., Gontier, N., Lu, X. H., Dziri, N., Murty, S. & Lacoste, A. (Eds.), *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)* (pp. 298–308). Vienna, Austria: Association for Computational Linguistics.

Schockaert, S., Ibanez-Garcia, Y. & Gutierrez-Basulto, V. (2021). A Description Logic for Analogical Reasoning. Volume 2 (pp. 2040–2046).

School, S. L. (2024). Breakthroughs in LLM Reasoning Show a Path Forward for Neuro-symbolic Legal AI.

Selsam, D., Lamm, M., B\"{u}nz, B., Liang, P., Moura, L. d. & Dill, D. L. (2018). Learning a SAT Solver from Single-Bit Supervision.

Serafini, L., Donadello, I. & Garcez, A. d. (2017). Learning and reasoning in logic tensor networks: theory and application to semantic image interpretation. In *Proceedings of the Symposium on Applied Computing*, SAC '17 (pp. 125–130). New York, NY, USA: Association for Computing Machinery.

Serafini, L. & Garcez, A. (2016). Logic Tensor Networks: Deep Learning and Logical Reasoning from Data and Knowledge.

Sha, J., Shindo, H., Kersting, K. & Dhami, D. S. (2025). Neuro-symbolic Predicate Invention: Learning relational concepts from visual scenes. *Neurosymbolic Artificial Intelligence*, *1*, NAI–240712.

Shah, N. (2023). Reliable Neuro-Symbolic Abstractions for Planning and Learning. Volume 6 (pp. 7093–7094).

Shaw, P., Uszkoreit, J. & Vaswani, A. (2018). Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 2 (Short Papers) (pp. 464–468). New Orleans, Louisiana: Association for Computational Linguistics.

Sheth, A. & Roy, K. (2024). Neurosymbolic Value-Inspired Artificial Intelligence (Why, What, and How). *IEEE Intelligent Systems*, *39(1)*, 5–11.

Sheth, A., Roy, K. & Gaur, M. (2023a). Neurosymbolic AI – Why, What, and How. arXiv:2305.00813 [cs].

Sheth, A., Roy, K. & Gaur, M. (2023b). Neurosymbolic Artificial Intelligence (Why, What, and How). *IEEE Intelligent Systems*, *38(3)*, 56–62.

Shih, A., Choi, A. & Darwiche, A. (2018). A Symbolic Approach to Explaining Bayesian Network Classifiers. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)* (pp. 5103–5111). Stockholm, Sweden.

Shin, H.-C., Zhang, Y., Bakhturina, E., Puri, R., Patwary, M., Shoeybi, M. & Mani, R. (2020). BioMegatron: Larger Biomedical Domain Language Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4700–4706). Online: Association for Computational Linguistics.

Shindo, H., Miyao, Y., Fujino, A. & Nagata, M. (2013). Statistical Parsing with Probabilistic Symbol-Refined Tree Substitution Grammars. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (pp. 3082–3086). Beijing.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S. & Farajtabar, M. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529(7587)*, 484–489.

Silver, D. L. & Mitchell, T. M. (2023). The Roles of Symbols in Neural-based AI: They are Not What You Think! In *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning*, CEUR '23 (pp. 420–421). La Certosa di Pontignano, Siena, Italy.

Singh, C., Inala, J. P., Galley, M., Caruana, R. & Gao, J. (2024). Rethinking Interpretability in the Era of Large Language Models.

Siyaev, A., Valiev, D. & Jo, G.-S. (2023). Interaction with Industrial Digital Twin Using Neuro-Symbolic Reasoning. *Sensors*, *23(3)*, 1729.

Sloman, A., McDermott, D. & Woods, W. A. (1983). Under What Conditions Can a Machine Attribute Meanings to Symbols. In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (I)* (pp. 44–45).

Smirnov, A., Ponomarev, A. & Agafonov, A. (2024). Ontology-Based Neuro-Symbolic AI: Effects on Prediction Quality and Explainability. *IEEE Access*, *12*, 156609–156626.

Smirnov, A., Ponomarev, A. & Shilov, N. (2023). Collaborative Decision Support with Ontology-Based Neuro-Symbolic Artificial Intelligence: Challenges and Conceptual Model. In Kovalev, S., Sukhanov, A., Akperov, I. & Ozdemir, S. (Eds.), *Proceedings of the Sixth International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'22)* (pp. 51–59). Cham: Springer International Publishing.

Speer, R., Chin, J. & Havasi, C. (2018). ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. arXiv:1612.03975 [cs].

Stammer, W., Schramowski, P. & Kersting, K. (2021). Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting with their Explanations. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3618–3628). Nashville, TN, USA: IEEE.

Strickland, E. (2019). How IBM Watson Overpromised and Underdelivered on AI Health Care.

Sumers, T. R., Yao, S., Narasimhan, K. & Griffiths, T. L. (2024). Cognitive Architectures for Language Agents. *Transactions on Machine Learning Research*. arXiv:2309.02427 [cs].

Susskind, Z., Arden, B., John, L. K., Stockton, P. & John, E. B. (2021). Neuro-Symbolic AI: An Emerging Class of AI Workloads and their Characterization. arXiv:2109.06133.

Sutton, R. S. & Tanner, B. (2004). Temporal-Difference Networks. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS '04 (pp. 1377–1384). Cambridge, MA, USA: MIT Press.

Tammet, T., Järv, P., Verrev, M. & Draheim, D. (2023). An Experimental Pipeline for Automated Reasoning in Natural Language (Short Paper). In Pientka, B. & Tinelli, C. (Eds.), *Automated Deduction – CADE 29* (pp. 509–521). Cham: Springer Nature Switzerland.

Tammet, T., Järv, P., Verrev, M. & Draheim, D. (2024). Experiments with LLMs for Converting Language to Logic. In Besold, T. R., d'Avila Garcez, A., Jimenez-Ruiz, E., Confalonieri, R., Madhyastha, P. & Wagner, B. (Eds.), *Neural-Symbolic Learning and Reasoning* (pp. 305–314). Cham: Springer Nature Switzerland.

Tan, B., Qin, L., Xing, E. & Hu, Z. (2020). Summarizing Text on Any Aspects: A Knowledge-Informed Weakly-Supervised Approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6301–6309). Online: Association for Computational Linguistics.

Thomson, R. H. & Bastian, N. D. (2024). Integrating Cognitive Architectures with Foundation Models: Cognitively-Guided Few-Shot Learning to Support Trusted Artificial Intelligence. *Proceedings of the AAAI Symposium Series*, *2(1)*, 409–414.

Tilwani, D., Venkataramanan, R. & Sheth, A. P. (2024). Neurosymbolic AI Approach to Attribution in Large Language Models. *IEEE Intelligent Systems*, *39(6)*, 10–17.

Tornqvist, M., Mahamud, M., Mendez Guzman, E. & Farazouli, A. (2023). ExASAG: Explainable Framework for Automatic Short Answer Grading. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 361–371). Toronto, Canada: Association for Computational Linguistics.

Touretzky, D. S. & Minton, G. E. (1985). Symbols Among the Neurons: Details of a Connectionist Inference Architecture (pp. 238–244).

Trinh, T. H., Wu, Y., Le, Q. V., He, H. & Luong, T. (2024). Solving olympiad geometry without human demonstrations. *Nature*, *625(7995)*, 476–482.

Tsamoura, E., Hospedales, T. & Michael, L. (2021). Neural-Symbolic Integration: A Compositional Perspective. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35(6)*, 5051–5060.

Ugur, E., Ahmetoglu, A., Nagai, Y., Taniguchi, T., Saveriano, M. & Oztop, E. (2025). Neuro-Symbolic Robotics.

Ulbricht, M. (2024). Formal Argumentation in Symbolic AI. Volume 9 (pp. 8577–8582).

Ullah, N., Khan, J. A., De Falco, I. & Sannino, G. (2025). Explainable Artificial Intelligence: Importance, Use Domains, Stages, Output Shapes, and Challenges. *ACM Computing Surveys*, *57(4)*, 1–36.

Vakharia, P., Kufeldt, A., Meyers, M., Lane, I. & Gilpin, L. (2024). ProSLM : A Prolog Synergized Language Model for explainable Domain Specific Knowledge Based Question Answering. Volume 14980 (pp. 291–304). arXiv:2409.11589 [cs].

Valmeekam, K., Stechly, K. & Kambhampati, S. (2024). LLMs Still Can't Plan; Can LRMs? A Preliminary Evaluation of OpenAI's o1 on PlanBench.

Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L. & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3(2)*, 125–133.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (pp. 6000–6010). Red Hook, NY, USA: Curran Associates Inc.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2023). Attention Is All You Need. arXiv:1706.03762 [cs].

Veličković, P., Casanova, A., Liò, P., Cucurull, G., Romero, A. & Bengio, Y. (2018). Graph attention networks.

Wan, Z., Liu, C.-K., Yang, H., Raj, R., Li, C., You, H., Fu, Y., Wan, C., Li, S., Kim, Y., Samajdar, A., Lin, Y., Ibrahim, M., Rabaey, J. M., Krishna, T. & Raychowdhury, A. (2024a). Towards Efficient Neuro-Symbolic AI: From Workload Characterization to Hardware Architecture. *IEEE Transactions on Circuits and Systems for Artificial Intelligence*, *1(1)*, 53–68.

Wan, Z., Liu, C.-K., Yang, H., Raj, R., Li, C., You, H., Fu, Y., Wan, C., Samajdar, A., Lin, Y. C., Krishna, T. & Raychowdhury, A. (2024b). Towards Cognitive AI Systems: Workload and Characterization of Neuro-Symbolic AI. In *2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)* (pp. 268–279). Indianapolis, IN, USA: IEEE.

Wang, C., Li, J., Chen, Y., Liu, K. & Zhao, J. (2025a). A Survey of Recent Advances in Commonsense Knowledge Acquisition: Methods and Resources. *Machine Intelligence Research*, *22(2)*, 201–218.

Wang, J., Jiang, Y., Long, Y., Sun, X., Pagnucco, M. & Song, Y. (2024). Deconfounding Causal Inference for Zero-Shot Action Recognition. *IEEE Transactions on Multimedia*, *26*, 3976–3986.

Wang, W., Yang, Y. & Wu, F. (2025b). Towards Data-And Knowledge-Driven AI: A Survey on Neuro-Symbolic Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47(2)*, 878–899.

Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F. & Tu, K. (2021). Automated Concatenation of Embeddings for Structured Prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 2643–2660). Online: Association for Computational Linguistics.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J. & Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.

Weir, N., Clark, P. & Durme, B. V. (2024). NELLIE: A Neuro-Symbolic Inference Engine for Grounded, Compositional, and Explainable Reasoning. Volume 4 (pp. 3602–3612).

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. *Commun. ACM*, *9(1)*, 36–45.

Werner, L. (2024). Neuro-Symbolic Integration for Reasoning and Learning on Knowledge Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38(21)*, 23429–23430.

West, R. L., Eckler, S., Conway-Smith, B., Turcas, N., Tomkins-Flanagan, E. & Kelly, M. A. (2023). Bridging Generative Networks with the Common Model of Cognition. *Proceedings of the AAAI Symposium Series*, *2(1)*, 415–421.

Winograd, T. (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. Accepted: 2004-10-20T20:29:48Z.

Winters, T., Marra, G., Manhaeve, R. & Raedt, L. D. (2022). DeepStochLog: Neural Stochastic Logic Programming. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36(9)*, 10090–10100.

Wolter, M., Veeramacheneni, L. & Hoyt, C. T. (2025). More Rigorous Software Engineering Would Improve Reproducibility in Machine Learning Research. arXiv:2502.00902 [cs].

Xiao, C., Dymetman, M. & Gardent, C. (2017). Symbolic Priors for RNN-based Semantic Parsing. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4186–4192). Melbourne, Australia.

Xie, X., Kersting, K. & Neider, D. (2022). Neuro-Symbolic Verification of Deep Neural Networks. Volume 4 (pp. 3622–3628).

Xu, J., Zhang, Z., Friedman, T., Liang, Y. & Broeck, G. (2018). A Semantic Loss Function for Deep Learning with Symbolic Knowledge. In *Proceedings of the 35th*

International Conference on Machine Learning (pp. 5502–5511). PMLR.

Yang, F., Lyu, D., Liu, B. & Gustafson, S. (2018). PEORL: Integrating Symbolic Planning and Hierarchical Reinforcement Learning for Robust Decision-Making. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence* (pp. 4860–4866). Stockholm, Sweden.

Yang, F., Yang, Z. & Cohen, W. W. (2017). Differentiable learning of logical rules for knowledge base reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17 (pp. 2316–2325). Red Hook, NY, USA: Curran Associates Inc.

Yang, W.-C., Marra, G., Rens, G. & Raedt, L. D. (2023). Safe Reinforcement Learning via Probabilistic Logic Shields. Volume 5 (pp. 5739–5749).

Yang, Z., Ishay, A. & Lee, J. (2020). NeurASP: Embracing Neural Networks into Answer Set Programming. Volume 2 (pp. 1755–1762).

Yasunaga, M., Leskovec, J. & Liang, P. (2022). LinkBERT: Pretraining Language Models with Document Links. In Muresan, S., Nakov, P. & Villavicencio, A. (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8003–8016). Dublin, Ireland: Association for Computational Linguistics.

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P. & Tenenbaum, J. B. (2018). Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Montreal, Canada.

Yin, C., Cappart, Q. & Pesant, G. (2024). An Improved Neuro-Symbolic Architecture to Fine-Tune Generative AI Systems. In Dilkina, B. (Ed.), *Integration of Constraint Programming, Artificial Intelligence, and Operations Research* (pp. 279–288). Cham: Springer Nature Switzerland.

Yu, D., Yang, B., Liu, D., Wang, H. & Pan, S. (2021). A Survey on Neural-symbolic Learning Systems.

Yu, D., Yang, B., Liu, D., Wang, H. & Pan, S. (2023). A survey on neural-symbolic learning systems. *Neural Networks*, *166*, 105–126.

Zhang, D. & Hannaford, B. (2020). IKBT: Solving Symbolic Inverse Kinematics with Behavior Tree (Extended Abstract). Volume 5 (pp. 5145–5148).

Zhang, J., Chen, B., Zhang, L., Ke, X. & Ding, H. (2021). Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open*, *2*, 14–35.

Zhang, X. & Sheng, V. S. (2024). Neuro-Symbolic AI: Explainability, Challenges, and Future Trends.

Zhang, X., Sun, J., Cheng, Z. & Chen, H. (2022). Research on the Embedded Mathematical Model of Artificial Intelligence Measurement. In *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)* (pp. 362–366). Wuhan, China: IEEE.

Zhao, J., Xia, G. & Wang, Y. (2023a). Q&A: Query-Based Representation Learning for Multi-Track Symbolic Music re-Arrangement. Volume 6 (pp. 5878–5886).

Zhao, J., Zhao, Z., Shi, L., Kuang, Z., Wang, R. & Li, H. (2023b). Deep Learning Cost-Effective Evaluation Metrics. In *2023 China Automation Congress (CAC)* (pp. 7190–7195). Chongqing, China: IEEE.

Zhao, W., Zhou, K., Junyi, L., Tianyi, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z. & Wen, J.-R. (2023c). A Survey of Large Language Models.

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H. & Han, J. (2022). Towards a Unified Multi-Dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2023–2038). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Zhong, W., Cui, R., Guo, Y., Liang, Y., Lu, S., Wang, Y., Saied, A., Chen, W. & Duan, N. (2024). AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2299–2314). Mexico City, Mexico: Association for Computational Linguistics.