Symbols and Neurons: A Review of Symbolic XAI in Deep Learning

IONEL EDUARD STAN*, University of Milano-Bicocca, Italy
GUIDO SCIAVICCO, University of Ferrara, Italy
PAOLO NAPOLETANO, University of Milano-Bicocca, Italy

Background: Deep neural networks increasingly power language, vision, and decision systems, yet many deployments require explanations that are faithful, compositional, and governance-ready. Symbolic techniques promise these properties, but the literature mixes post-hoc extraction, knowledge injection, and intrinsically hybrid designs without a unifying view.
Objectives: We provide a systematic review and synthesis of symbolic explainable AI (XAI) for deep learning (January 2017–June 2025), organize the field around a three-part taxonomy—Symbolic Knowledge Extraction (SKE), Symbolic Knowledge Injection (SKI), and Hybrid neurosymbolic architectures—and propose a conceptual framework that clarifies training–inference flows, explanation interfaces, human feedback, and governance touchpoints.
Methods: Beginning from ≈50,000 records, we deduplicated and screened full texts, analyzed 393 PDFs, and included 273 primary studies in the synthesis. We coded each paper for model domain, modality, symbolic formalism, explanation scope and stage, evaluation protocol, and governance alignment. Analyses combine descriptive statistics with stratification by domain and formalism; we qualitatively assess evidence for faithfulness, robustness, data efficiency, and constraint satisfaction.
Results: Research activity accelerates after 2020, with a marked turn toward hybrids. Across the corpus, SKE, SKI, and Hybrid account for approximately 29%, 26%, and 45% of studies, respectively. Rule sets/decision trees remain the dominant explanation artifacts, while logic- and program-based formalisms grow in NLP and planning. SKI most often targets constraint satisfaction and robustness improvements; SKE emphasizes global surrogates and faithfulness auditing; hybrids report gains in sample efficiency and traceable reasoning. However, evaluation practices are heterogeneous, human-subject studies are scarce, and explicit links to policy/risk controls appear in a minority of works.
Conclusions: Our framework unifies how data, priors, and symbolic reasoning interact with neural learners, the explanation interface, human stakeholders, and governance. We distill actionable recommendations: (1) report faithfulness and constraint-satisfaction metrics alongside accuracy; (2) specify symbolic assumptions and training-time injections precisely; (3) include user studies or auditor-centric protocols for high-stakes use; and (4) develop benchmarks that couple tasks with machine-readable knowledge bases. We highlight open problems in scalable formal reasoning with foundation models, verifying generated rationales, and measuring causal faithfulness at scale.

JAIR Track: Integration of Logical Constraints in Deep Learning

JAIR Associate Editor: Matteo Zavatteri

JAIR Reference Format:
Ionel Eduard Stan, Guido Sciavicco, and Paolo Napoletano. 2026. Symbols and Neurons: A Review of Symbolic XAI in Deep Learning. *Journal of Artificial Intelligence Research* 86, Article 36 (July 2026), 60 pages. DOI: 10.1613/jair.1.22096

*Corresponding Author.

Authors' Contact Information: Ionel Eduard Stan, ORCID: 0000-0001-9260-102X, ioneleduard.stan@unimib.it, University of Milano-Bicocca, Milano, Italy; Guido Sciavicco, ORCID: 0000-0002-9221-879X, guido.sciavicco@unife.it, University of Ferrara, Ferrara, Italy; Paolo Napoletano, ORCID: 0000-0001-9112-0574, paolo.napoletano@unimib.it, University of Milano-Bicocca, Milano, Italy.

DOI: 10.1613/jair.1.22096

1 Introduction

A tension between symbolic and sub-symbolic (Minsky 1991) has defined modern artificial intelligence (AI). Early AI systems were largely symbolic, relying on human-readable rules and logic (e.g., expert systems like DENDRAL and MYCIN) to perform reasoning (Russell and Norvig 2020). These approaches offered *transparency*—one could trace a chain of logical inferences—but proved brittle and hard to scale in complex, noisy environments. In contrast, the rapid rise in deep learning has produced neural networks that excel at *pattern recognition from raw data* with unprecedented accuracy in tasks ranging from computer vision to natural language processing. However, this power comes at the cost of interpretability: deep neural networks are often "opaque black-boxes," with millions—even trillions—of parameters whose inner workings are complex for humans to comprehend. This opacity undermines trust and poses risks in safety-critical or decision-critical domains. Concerns range from lack of *reliability*, as deep neural networks can behave unpredictably or exhibit vulnerability to adversarial inputs, to *ethical and legal compliance*, since individuals have a right to understand the reasoning of automated decision-making (as reflected, for example, in the EU GDPR's "right to explanation"). A widely discussed example is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism-risk tool: ProPublica's investigation reported racially disparate error patterns in COMPAS risk scores[1], and subsequent work examined the accuracy, fairness, and limits of such recidivism prediction systems (Dressel and Farid 2018). Another striking example was ALPHAGO's famous 37th move against Lee Sedol, which astonished experts and even its creators—DeepMind's David Silver admitted he had "no insight" into why the AI chose that move until much later analysis.

Such episodes demand for *explainable AI (XAI)* (Gunning and Aha 2019; Gunning, Stefik, et al. 2019) techniques that can illuminate the reasoning behind a model's decisions as AI systems are systematically deployed in sensitive and mission-critical domains. In healthcare, finance, autonomous driving, and other areas, an unexplained algorithmic decision can erode trust or have dire consequences. Stakeholders—from end-users to regulators—now expect AI decisions to be transparent, fair, and accountable. Research in XAI has exploded accordingly, yielding methods like feature attribution, saliency maps, and example-based explanations. Yet many of these popular approaches provide only shallow insights (e.g., highlighting pixels or input features) and may not satisfy a user's need to understand reasoning. By contrast, symbolic explanations (such as logic rules, decision trees, or structured knowledge) align more closely with human reasoning. They can answer the "how" and "why" in a more interpretable, human-understandable form. This realization has rekindled interest in marrying symbolic AI with deep learning to create systems that are both powerful and explainable by-design.

Symbolic XAI in deep learning refers to a broad class of techniques that incorporate symbolic knowledge or produce symbolic artifacts (such as rules, programs, logic statements, ontologies) to explain or constrain neural network behavior. This class includes post-hoc methods that extract human-readable rules from trained networks, as well as integrated neurosymbolic systems that embed logical structures into the model's architecture or training process. For example, researchers have developed algorithms to distill deep neural networks into decision trees or sets of IF–THEN rules, yielding transparent approximations of the model's logic. Others incorporate ontologies and knowledge graphs (KGs) to guide and explain a model's understanding (e.g., mapping latent features to high-level concepts defined by domain knowledge). Program synthesis approaches induce executable programs (e.g., simple code or mathematical expressions) from a model's outputs or internal representations, providing step-by-step explanatory reasoning. Formal methods from software verification have also been applied to reason about neural network decisions in logical terms (e.g., verifying that a network's decisions uphold specific safety rules or using satisfiability solvers to find conditions under which a model changes its prediction).

Research at the intersection of symbolic reasoning and deep learning is not entirely new. The past few years have witnessed a resurgence of these ideas, driven by three key developments. First, the complexity of modern AI

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

(deep CNNs, transformers, large language models) has made explainability a critical bottleneck for deployment. Second, instrumental regulatory frameworks, such as the EU GDPR's 2016 document, have shifted explainability from a desirable feature to a legally mandated requirement in high-stakes scenarios. Third, the community recognizes that achieving higher-level reasoning and trustworthy AI may require integrating the strengths of symbolic and neural approaches. Leading researchers have advocated combining fast, pattern-based "System 1" intelligence with the slower, logical "System 2" reasoning that symbolic methods provide. This vision has led to neurosymbolic AI—what some call the "third wave of AI" (A. d. Garcez and Lamb 2023)—wherein robust neural learning is blended with symbolic reasoning and explanation.

In this paper, we present a systematic literature review of symbolic XAI in deep learning. To our knowledge, this is the first review to focus specifically on symbolic explainability techniques applied to deep neural architectures, following a rigorous PRISMA 2020 methodology (Page et al. 2021) for identifying, screening, and analyzing the literature (the detailed protocol and selection flow are provided in Appendix A). From an initial pool of thousands of papers, we identified 273 peer-reviewed research papers that squarely address symbolic XAI for deep learning, supplemented by relevant survey papers. We intentionally define symbolic XAI broadly to include any approach that provides or leverages symbolic knowledge representations—such as logical rules, KGs, programs, or formal specifications—in explaining or constraining deep models. By doing so, we include:

(1) *rule-based explanation methods* (e.g., extracting decision trees or rule sets from networks),
(2) *ontology- or KB-based explanations* (e.g., mapping network outputs to structured, semantical concepts),
(3) *programmatic and model-based explanations* (e.g., via program synthesis or symbolic policy learning),
(4) *formal logic and verification techniques* applied to neural networks (e.g., using formal methods to explain or constrain network behavior), and
(5) *hybrid neurosymbolic architectures* that embed interpretability by-design.

Moreover, we discuss *neuro-fuzzy hybrids*, which overlap with our themes (see Section 4.3 on hybrid architectures).

Despite this flurry of research, no comprehensive literature review has yet explicitly focused on symbolic XAI in deep learning, spanning the full spectrum of techniques and applications. Existing surveys have only partially covered this area. For example, some authors offered a broad survey on neural network interpretability and proposed a taxonomy along dimensions like passive/active and local/global explanations (Y. Zhang, Tiňo, et al. 2021). While insightful, their work covered a wide range of explainability methods (e.g., feature attributions, saliency maps, example-based explanations) and did not specifically focus on symbolic techniques. Other surveys have reviewed neurosymbolic AI in particular domains—such as on neural-symbolic computing with graph neural networks (Lamb et al. 2020), or more recently, on neurosymbolic reasoning over KGs (Delong et al. 2025). Further reviews have summarized the findings of integrating fuzzy reasoning with neural computation—such as logic-oriented fuzzy neural networks (Alateeq and Pedrycz 2024), or more in general, deep neuro-fuzzy systems (Talpur et al. 2023). Additional review papers have addressed scattered pieces of the puzzle—such as explainable reinforcement learning (XRL) with a focus on integrating explanations into agent behavior (Dazeley et al. 2023), explainable affective computing that gravitates around the importance of end-users' needs (Cortiñas-Lorenzo and Lacey 2024), interpretable-by-design methods for reasoning-based learning (Ignatiev et al. 2021), or neural trees which combine the advantages of both neural networks and decision trees (H. Li, J. Song, et al. 2025).

Broad XAI meta-surveys organize the landscape across many explanation families, where symbolic surrogates and constraint-based mechanisms typically appear as specific sub-threads rather than the primary organizing principle (Arrieta et al. 2020; Saeed and Omlin 2023). In contrast, we focus specifically on the symbolic–deep learning interface and structure the field around *when* symbols intervene and *which* symbolic artifacts they yield: post-hoc extraction (SKE), training-time injection (SKI), and intrinsic mixed designs (Hybrid)—a distinction also adopted in recent systematic work that treats SKE and SKI as complementary activities at scale (Ciatto et al. 2024).

Finally, to connect this explanation-centered view to established neurosymbolic *architecture* perspectives, we include an explicit crosswalk to Kautz's six design patterns (Table 1) (Kautz 2022).

While these works confirm growing interest in combining symbolic techniques (rules, decision trees, logic, etc.) and neural learning, they typically cover a subset of methods or specific applications. Accordingly, our work differs in scope: we consider *symbolic explainability across all deep learning*. Purely symbolic rule-learning also remains an active research line—including recent work on deep/shallow rule learning and FIND-RS/Bayes-point aggregation (Beck and Fürnkranz 2021; Bergamin et al. 2025)—but we treat these methods as adjacent rather than central unless they explain or constrain deep neural predictors. We also explicitly connect to XAI in a broader sense (including regulatory context), which those domain-specific surveys do not. We build on the insights of prior work—for instance, adopting the view that rules and logical explanations are highly desirable for human understanding—and expand the discussion to include ontologies, program synthesis, formal verification, and hybrid architectures under one umbrella. Given the rapid developments from 2017 to mid-2025—including the maturation of techniques and the context of new regulatory frameworks—we argue that summarizing the findings of symbolic XAI in deep learning is both novel and timely. It fills an important gap by consolidating knowledge scattered across thousands of papers and by critically analyzing whether the integration of symbolic AI indeed offers a viable path to more trustworthy AI. By doing so systematically, we intend to map the landscape in a way that is both comprehensive and structured, providing a one-stop reference for researchers, practitioners, and policy-makers interested in symbolic approaches to explaining deep learning. To our knowledge, no published survey or meta-study has synthesized these elements into a unified framework, especially with the breadth of 273 primary research papers and integration of insights from over a dozen earlier reviews.

In light of societal and regulatory pressures, our review is especially pertinent. The EU's AI Act demands transparency and explainability for high-risk AI systems. Compliance will require AI developers to provide clear documentation and understandable rationales for automated decisions. Similar principles are echoed in guidelines worldwide (e.g., the US NIST XAI report and the earlier GDPR's "right to explanation"). By producing human-comprehensible knowledge (like logical rules or causal explanations) alongside or within neural models, symbolic XAI approaches are uniquely well-suited to meet these mandates for qualified transparency. Moreover, society needs—from doctors needing interpretable diagnoses, to loan applicants seeking fair explanations, to autonomous vehicle operators requiring justified decisions—underscore that explainability is not merely a research ideal but a practical necessity. By linking the technical developments in symbolic XAI to these broader needs, our review provides a narrative of how and why the field evolved, and why now "symbol-infused deep learning" is poised to play a critical role in enabling trustworthy and human-aligned AI systems.

Our contributions in this paper are as follows:

(1) **A PRISMA-compliant systematic review.** We systematically review 273 peer-reviewed studies (January 2017–June 2025) that incorporate symbolic structure into deep learning for explainability, complemented by prior surveys for positioning. To our knowledge, this is the first review to focus exclusively on *symbolic* XAI in deep learning at this scale, tracking the full pipeline from identification to inclusion with a PRISMA 2020 (Page et al. 2021) flow diagram (Figure 1) and an auditable methodology (Appendix A). We release a complete reproducibility bundle (code, version-controlled YAML configs, topic/metric tables, and statistical reports).[2]

(2) **A unifying taxonomy that resolves when/how symbols interact with neurons.** We formalize the design space along the *when/how* axis—post-hoc Symbolic Knowledge Extraction (SKE), training-time Symbolic Knowledge Injection (SKI), and Hybrid neurosymbolic architectures—decorated by the symbolic object in use (rules, trees, logic/programs, KGs, fuzzy sets, concepts) (Figure 2; Section 4). This resolves

[2] https://github.com/eduardstan/xai_dl_logic

common ambiguities (e.g., fuzzy methods as SKI vs. as hybrid models; ontology/KG methods as post-hoc mappings vs. embedded reasoning) and provides a consistent legend for comparing methods across modalities and formalisms.

(3) **A conceptual framework linking learning, inference, explanation interfaces, and governance.** Beyond a catalog, we introduce a lifecycle framework (Figure 3; Section 6) that makes explicit where symbolic assets enter training (constraints, priors), where explanations aggregate at inference (rules, clause traces, concept paths), and how governance signals and human feedback update both models and knowledge. This framework turns the taxonomy into deployable patterns by mapping artifacts to concrete touchpoints in development and operations.

(4) **An empirical map of the field with validated composition and temporal trends.** We quantify the corpus composition—79 SKE, 71 SKI, 123 Hybrid (≈29% : 26% : 45%)—and show a post-2020 shift from post-hoc rationalization to *explainability-by-design*, with SKI and Hybrid methods accelerating while SKE plateaus (Table 6; Section 5.1). The analysis couples counts with qualitative trade-offs (faithfulness, constraint satisfaction, intrinsic traces), offering a balanced view of where symbolic methods deliver the most value.

(5) **Governance alignment distilled into a practitioner checklist.** We connect symbolic artifacts to specific legal and assurance duties—EU AI Act transparency and record-keeping, GDPR "meaningful information," credit adverse-action reasons, FDA GMLP/PCCP, and NIST AI RMF—showing what to log and where each artifact fits in the lifecycle (Table 7; Section 5.3). This converts explanations into evidence by specifying minimal records (e.g., rule paths, constraint violation reports, version links) that satisfy oversight and audit requirements.

(6) **Standardized case-study tables enabling like-for-like comparison across domains.** We present harmonized tables for healthcare, finance/economics, NLP, CV, robotics/agentic systems, cybersecurity/privacy, scientific and engineering modeling, and industry, each tying a symbolic pattern to its explanation object and reported impact (Section 7). The result is a transferable "pattern library" that enables practitioners to match domain needs to symbolic structures and expected evidence types.

(7) **Deployment playbooks that translate priorities into designs, artifacts, and controls.** We condense the review into actionable playbooks (Table 19): choose SKE when you need governance-ready reasons without retraining; prefer SKI when properties must hold by construction; adopt Hybrids when intrinsic traces are operationally required. Each playbook specifies what to implement and log (artifacts/metrics), as well as flags typical gotchas (surrogate drift, conflicting priors, loss-weighting, and spec drift), along with suggested controls, aligning directly with regulatory expectations.

(8) **Open, reproducible research.** All parameters (seeds, environments, selection criteria) are documented "as code," and we release machine-readable outputs to enable reproducibility. This level of transparency supports independent verification and facilitates downstream meta-analysis by other researchers.

2 Background

In this section, we provide background on explainable AI, symbolic vs. sub-symbolic AI, and the emergence of neurosymbolic techniques. We define key terms and discuss foundational work that underpins the symbolic XAI approaches reviewed in later sections.

2.1 Explainable AI and Its Evolution

Explainable AI (XAI) refers to methods that make the AI decision-making process understandable to end-users (e.g., policymakers, researchers, practitioners). There are two broad approaches to achieve explainability:

(1) *interpretability by-design*, where the model is inherently understandable (e.g., linear models, decision trees, rule-based systems),[3] and
(2) *post-hoc explainability*, where an external mechanism is used to explain an otherwise opaque model (e.g., explaining a trained deep neural network).

These have been identified as the two major paradigms for bringing interpretability to intelligent systems. Early in the era of expert systems (1970s–1980s), the dominant approach was by-design—AI systems were often rule-based and could expose their inference chains to users. For example, MYCIN (an early medical expert system) could explain which rules led to a diagnosis. However, as the field shifted towards machine learning and sub-symbolic models (e.g., neural networks and ensemble methods) in the 1990s and 2000s, accuracy often came at the cost of interpretability, even though some researchers consider this trade-off a myth (Rudin 2019). Complex models were treated as black boxes, and explanation capabilities were largely set aside.

The resurgence of interest in XAI can be traced to around the mid-2010s. Influential events—such as the adoption of the EU's GDPR, the launch of DARPA's XAI program both in 2016, and IJCAI's first workshop on XAI in 2017—underscored the need for "opening the black box" of deep learning. Researchers began developing post-hoc techniques. For instance, LIME (Ribeiro et al. 2016) and SHAP (Lundberg and S. Lee 2017) approximated a complex model locally with simpler models to explain individual predictions, saliency maps and feature visualizations tried to highlight important input features for neural network decisions, and counterfactual explanations described how input changes would alter the output. These methods were important first steps and are covered extensively in other surveys. However, they typically produce low-level explanations (e.g., pixel importance or abstract feature importances) that may not satisfy a user's need for a higher-level, semantic rationale. It became evident that explanations need to be contextual, contrastive, and tied to concepts that users understand (for example, a doctor might want an explanation in terms of medical conditions or risk factors, not just pixel intensities). This understanding has led researchers to focus on integrating structured knowledge and symbols into AI explanations.

The term neural-symbolic integration—the "third wave of AI" (A. d. Garcez and Lamb 2023)—refers to efforts to combine the strengths of symbolic AI (explicit knowledge and reasoning) with neural networks (learning from data). This idea has a long history in AI: research on neural-symbolic systems can be traced back to the 1990s and early 2000s (e.g., A. S. d. Garcez et al. (2001) provided early theory on knowledge extraction from neural networks). Classic work in the field explored how to translate trained neural networks into logic rules, how to embed logical constraints into networks, and even how to design neural networks that simulate logical operators. However, for many years, these approaches were of niche interest, overshadowed by the escalating performance of pure deep learning.

It is in the context of XAI's rebirth that neural-symbolic ideas found a new and essential application. Many researchers recognized that symbolic representations could be the key to more meaningful explanations. For instance, one could extract a set of human-readable IF–THEN rules that resemble the model's behavior, instead of explaining a model's decision with a list of feature weights. Likewise, one could constrain a model during training to follow logical rules (like conservation laws or domain axioms) so that its internal reasoning aligns with known concepts, making the eventual explanation easier (the model inherently follows a reasoning that can be explained in those terms).

At a high level, symbolic approaches in XAI aim to move explanations from the sub-symbolic level of neurons and weights to the symbolic level of concepts and relations. This can happen either post-hoc (extracting or inferring symbols from the trained model) or by-design (integrating symbols so that the model operates on them). The historical perspective of XAI by Confalonieri, Coba, et al. (2021) highlights that after the era of expert systems,

[3]We adopt the term *interpretability by-design* (or *transparent interpretability*) following Doshi-Velez and B. Kim (2017) and Biran and Cotton (2017), acknowledging ongoing debate in the rigorous science of interpretability. Alternative formulations include "by-design interpretability" and "ante-hoc interpretability," each with slightly different connotations regarding the timing and nature of transparency mechanisms.

explainability research diversified—covering not only expert systems and ML, but also areas like recommender systems, explanations and neural-symbolic reasoning. They conclude that explainability in AI has come full circle: what was once done through explicit knowledge in expert systems is now being revisited through hybrid systems that embed knowledge into learning.

2.2 Deep Learning in a Nutshell and Why Explainability is Harder

Deep learning refers to neural networks with many stacked nonlinear layers trained end-to-end by gradient-based optimization (Goodfellow et al. 2016; LeCun et al. 2015). Compared to classical ML (e.g., linear models, shallow trees, kernel methods), deep models learn *representations*: intermediate feature spaces that are not manually specified but emerge from data. This is a major strength (they learn from raw pixels, text, and signals), but it complicates explanation.

From an explainability viewpoint, deep learning is challenging for three recurring reasons:

- *distributed knowledge:* behavior is encoded across many parameters rather than a small set of explicit rules;
- *nonlinearity and composition:* depth composes many nonlinear transformations, so local changes can interact in complex ways;
- *spurious shortcuts and entanglement:* high-capacity models can rely on correlations that are predictive but not meaningful to stakeholders, producing explanations that look plausible but do not reflect true causal structure.

These properties motivate symbolic XAI: symbolic artifacts (rules, trees, clauses, programs, KG rationales) aim to provide higher-level, structured accounts of behavior that can be inspected and communicated, and—in SKI and hybrids—can be used to constrain learning toward stakeholder-aligned semantics.

2.3 Symbolic vs. Sub-symbolic AI

To fully appreciate symbolic XAI methods, it is instrumental to understand the distinction between symbolic AI and sub-symbolic AI:

- *Symbolic AI* refers to approaches that manipulate *symbols*: discrete, human-interpretable entities that represent knowledge (e.g., logical predicates, graphs). These systems use explicit rules, logic, or structured knowledge bases. Classical AI, often referred to as Good Old-Fashioned AI (GOFAI), was largely symbolic (e.g., rule-based expert systems, logic programming, planning algorithms, ontology-based reasoning). Symbols come with a combinatorial, often explainable structure, allowing users to trace inference steps easily. Additionally, prior knowledge can be encoded directly, and reasoning tends to be transparent, following logical rules. However, pure symbolic systems usually lack learning *from unstructured data*—while rule induction methods (RIPPER (Cohen 1995), ILP systems (Evans and Grefenstette 2018)) can learn from structured symbolic data, they do not perform representation learning from raw sensory inputs (e.g., images, text) as neural networks do. This limitation motivates hybrid approaches.
- *Sub-symbolic AI* refers to approaches like neural networks and other statistical learners that operate on *numeric representations* rather than discrete symbols. Knowledge in a sub-symbolic system is *distributed* across numeric parameters (weights) rather than stored as explicit facts. Deep learning is the prototypical sub-symbolic approach. These models are particularly adept at pattern recognition and can learn from data, effectively handling noise and raw low-level inputs like pixels and waveforms. Their ability to approximate complex functions often results in high accuracy. However, the *knowledge they acquire remains implicit*, leading to opaque models that make it challenging to extract human-readable explanations of their learning processes. Additionally, sub-symbolic AI typically requires large datasets for training and may generalize in unpredictable ways, struggling with systematic generalization and compositionality—a known limitation of these approaches.

This dichotomy suggests a natural synergy: *symbolic systems provide transparency and precise semantics, while sub-symbolic systems provide learning and robustness*. Neural-symbolic AI attempts to integrate the two, so that we can have systems that *learn from data but also reason with knowledge*. In the context of XAI, this integration means the system's decisions can be explained in terms of high-level concepts or logic, not just low-level features.

When are symbolic explanations actually human-comprehensible? A key assumption behind symbolic XAI is that discrete artifacts (rules, trees, clauses, programs) can function as *human-comprehensible knowledge*. We make this assumption explicit and qualify it: symbolic form alone does not guarantee interpretability. Comprehensibility depends on the artifact's *representation* (e.g., DNF vs. CNF vs. decision lists), *size/complexity* (e.g., number of predicates, clause length, tree depth), and *presentation* (grouping, naming, and domain grounding). Empirical studies support this nuance: Booth et al. (2019) analyze how alternative logical representations affect interpretability for humans; Confalonieri, Weyde, et al. (2021) show that grounding post-hoc surrogate explanations in ontologies can improve understandability by mapping low-level features to domain concepts. In practice, this motivates treating symbolic XAI as an *interface* between models and stakeholders: domain experts can supply concepts, constraints, or ontologies, while the system returns artifacts that can be inspected, edited, and logged for auditing.

2.4 Symbolic and Sub-symbolic Integration

One way to visualize this is via a neural-symbolic cycle. In this cycle, we treat a neural network and a symbolic knowledge base as two interacting modules. We can inject symbolic knowledge into the neural network (e.g., by modifying its architecture or training objective to respect certain rules) and then, after training, we extract refined symbolic knowledge back out (e.g., new rules or adjusted rules that the network has learned from data). This iterative loop can continue, with the extracted knowledge potentially being used to update the symbolic module or to inform further training, in a human-in-the-loop fashion. The cycle captures four pillars of neural-symbolic systems:

- *representation* (how symbols and neural representations are mapped),
- *learning* (how the network learns with possibly symbolic input),
- *reasoning* (how symbolic reasoning is performed on the knowledge), and
- *extraction* (how learned knowledge is extracted in symbolic form).

We will see concrete examples of this paradigm in later sections, such as neural networks that learn to obey logical rules and can then output those rules as explanations.

In the context of symbolic XAI in deep learning, we often encounter three broad strategies:

- *Symbolic Knowledge Extraction (SKE):* After training a neural model, the task is to derive a symbolic description of its decision-making behavior (e.g., extracting rules, decision trees, finite state machines, or other structures that mimic the model's input-output mapping). The goal is that these symbolic descriptions are approximate explainers of the model's logic. The extracted symbolic model should faithfully reproduce the original model's decisions (otherwise, the explanation might be misleading). As noted by A. S. d. Garcez et al. (2001), while extraction from large networks may never be 100% complete, it should strive to be sound (i.e., any rule given is truly something the network resembles).
- *Symbolic Knowledge Injection (SKI):* Before or during training a neural model, the strategy is to inject symbolic knowledge to guide or constrain it (e.g., by regularization terms that penalize rule violations, architectural choices like adding a logic layer, or pre-training a network to mimic a symbolic system). The motivation is to embed prior knowledge so that the model not only learns faster or better (data efficiency and generalization can improve), but also follows patterns that are easier to explain (because we have built-in concepts or reasoning steps). By injecting knowledge, we often end up with models that can explain their decisions in terms of the injected knowledge. For instance, a network trained with logical constraints

might output whether those logical conditions were triggered in making a decision, giving a symbolic rationale.

- *Hybrid models (neurosymbolic by-design):* Here, the system is architected from the start to integrate neural and symbolic components such that symbolic state and/or symbolic reasoning is part of the decision process. In contrast to SKE (post-hoc) and SKI (training-time injection), Hybrids aim to make explanations *intrinsic* by exposing intermediate symbolic artifacts (e.g., a reasoning trace) as the model runs. In practice, we observe three recurring hybrid patterns:
  - *pipeline hybrids (Neuro | Symbolic):* a neural perception module produces symbols (objects, entities, predicates) that a symbolic reasoner consumes (program execution, rule chaining, constraint solving). Explanations often take the form of *proof/plan steps* or *executed program traces*.
  - *intrinsic differentiable hybrids (Neuro$_{Symbolic}$):* symbolic structure is compiled into differentiable components (e.g., neuro-fuzzy rules, neural decision trees, differentiable logic/rule layers), yielding *by-design* traces such as activated clauses, rule weights, or decision paths.
  - *solver-in-the-loop hybrids (Neuro[Symbolic] / Symbolic[Neuro]):* neural modules invoke symbolic solvers (or are invoked by them) as subroutines (planning, search, theorem proving, KG reasoning). Explanations typically combine *invocation traces* (why a solver was called) with *symbolic solver artifacts* (paths/proofs/plans).

  These patterns are not mutually exclusive, and a single system may blend them (e.g., a differentiable rule layer plus a symbolic consistency checker). For context, Table 1 relates these hybrid patterns to Kautz's architecture-level taxonomy, while our review focuses on *when and how* symbolic artifacts support explanation (SKE/SKI/Hybrid).

Note that these strategies are not mutually exclusive—they can be combined. For example, one might inject knowledge during training and also be able to extract knowledge after training. Indeed, the neural-symbolic cycle mentioned earlier inherently combines injection and extraction in an iterative loop. A recent systematic review by Ciatto et al. (2024) uses the terms SKE and SKI to categorize the literature, and they identified 132 methods of the former and 117 of the latter. They highlight that both activities are complementary means to tackle the opacity of sub-symbolic predictors. Moreover, several forms of neurosymbolic integration exist. Kautz (2022) describes six recurring neurosymbolic system designs that differ by (i) *where* symbolic structures appear (inputs/outputs, intermediate representations, or internal solvers) and (ii) *how tightly* neural and symbolic components are coupled. Table 1 aligns these architecture-level patterns with our explanation-centered taxonomy (SKE/SKI/Hybrid). The key point is that Kautz's taxonomy organizes *system structure*, whereas our taxonomy organizes *when and how symbolic artifacts support explanation*: extraction after training (SKE), injection during learning (SKI), or explicit mixed pipelines (Hybrid). This alignment clarifies how the same neurosymbolic architecture family can host different explanation strategies depending on whether symbolic artifacts are produced at inference time, enforced during training, or both. Such distinctions inspire our taxonomy in Section 4.

Toy examples of symbolic explanation formats. To make the remainder of the survey more self-contained, Table 2 provides small schematic examples of the main symbolic artifacts discussed throughout (rules, tree paths, logical constraints, KG rationales, and program traces). These are illustrative and intentionally simplified; later sections analyze how real systems instantiate these artifacts at scale.

3 Methodology

We conducted a systematic literature review of *symbolic* XAI for *deep learning* following PRISMA 2020 (Page et al. 2021). The full protocol, algorithms, and parameter settings appear in the Extended Methodology (Appendix A); here we summarize the essentials used to generate the evidence base and figures reported in the main text.

Table 1. Comparison between Kautz's six neurosymbolic *architecture* patterns and our *explanation-centered* taxonomy (SKE/SKI/Hybrid). Kautz's notation emphasizes control/coupling: brackets [ ] indicate a subordinate module (subroutine), | indicates a pipeline/coroutine boundary, :→ highlights a symbolic-driven training regime, and subscripts indicate rules compiled into neural templates.

Kautz pattern	Where symbols live / coupling intuition	Canonical instantiations (illustrative)	Closest match in our taxonomy	How explanations typically arise (within our scope)
(1) Symbolic Neuro symbolic	Symbolic I/O with a neural core: symbolic tokens/categories are mapped to vectors, processed by a neural model, then mapped back to symbols.	NLP standard operating procedure; LLM-style pipelines; symbolic I/O with sub-symbolic internals.	**Contextual baseline** (not inherently XAI)	Not explanatory by itself; becomes SKE or SKI only when paired with explicit explanation mechanisms (e.g., post-hoc rule extraction, concept interfaces, constraint training).
(2) Symbolic[Neuro]	A symbolic problem solver calls a neural pattern-recognition module as a *subroutine*. The symbolic layer is "in charge."	MCTS + value/policy nets (e.g., game-playing); robotics stacks with neural perception submodules.	**Hybrid** (symbolic top-level)	Explanations often come from symbolic traces (plans/search trees/rules). Neural submodules may still require SKE-style local explanations if audited independently.
(3) Neuro | Symbolic	Neural front-end converts raw input (e.g., pixels) into a symbolic data structure that is processed by a symbolic reasoner; feedback may train the neural part.	Neural perceptual parser + symbolic reasoning (program execution/logic/constraints).	**Hybrid** (neural → symbolic)	Symbolic explanations arise naturally from the reasoning trace (executed program, proof steps, constraint satisfaction), often complemented by SKE for the perception module.
(4) Neuro: Symbolic→Neuro	Neural model trained with a symbolic-driven regime (rules/solvers generate data, labels, or curricula); the deployed predictor may remain purely neural.	Solver-generated supervision; rule/algorithmic curriculum; synthetic-label training.	**SKI** (training regime/structured supervision)	The symbolic component influences learning, but explanations are not guaranteed at inference unless the supervision includes symbolic rationales or the trained model is paired with SKE.
(5) Neuro$_{\text{Symbolic}}$	Symbolic rules compiled into neural structures (templates/layers) so that parts of the computation reflect symbolic relations "by-design."	Differentiable rule/logic templates; logic-tensor/rule-form layers; constrained decoders.	**SKI** (architecture-level injection)	Explanations can be by-design when the compiled structures remain inspectable (e.g., rule weights/activated clauses), or via constraint-satisfaction traces during training/inference.
(6) Neuro[Symbolic]	A neural system embeds a symbolic reasoning engine as a *subroutine*, invoked on demand (tight coupling; neural decides when to reason).	Agents that call planners/search/solvers; "tool-using" neural controllers with internal symbolic routines.	**Hybrid** (tight coupling)	Explanations may combine (i) invocation traces (why the solver was called), (ii) symbolic solver artifacts (plans/proofs), and (iii) neural state summaries—spanning both SKE and SKI in practice.

Sources and search. We queried Embase, IEEE Xplore, Scopus, and Web of Science (January 2017–June 2025; English only) using a Boolean expression that combines three concept domains: symbolic/knowledge, neural/sub-symbolic, and explainability (complete query in Appendix A). We deduplicated records in Mendeley and applied a venue-quality filter (SCImago Artificial Intelligence Q1 journals and CORE A*/A conferences). The PRISMA flow diagram is shown in Figure 1.

Screening pipeline. We used a three-stage computational pipeline to structure screening and to surface representative studies for close reading:

Table 2. Toy examples of symbolic explanation artifacts used throughout this survey. Examples are schematic and illustrate the *format* (rules, paths, clauses, KG rationales, and programs) rather than any specific dataset.

Symbolic artifact	Toy instance (schematic)	How it explains (in simple terms)
IF–THEN rule	`IF (age > 60) AND (smoker = yes) THEN risk = high`	States a human-readable condition that is sufficient for the predicted label; supports auditing and editing.
Decision-tree path	`(cholesterol > 240)` $\rightarrow$ `(bp > 140)` $\rightarrow$ `leaf: high risk`	Explains a prediction as a sequence of threshold tests; each root-to-leaf path is an interpretable rule.
Logical constraint (SKI)	$\forall x$: `diabetes`$(x) \wedge$ `pregnant`$(x) \Rightarrow$ `high_risk`(x)	Encodes domain knowledge as a condition the model should satisfy; explanations can report satisfied/violated constraints.
Ontology/KG rationale	`patient` $\xrightarrow{\texttt{hasCondition}}$ `diabetes` $\xrightarrow{\texttt{increasesRisk}}$ `complication`	Grounds a prediction in named concepts/relations; explanation is the semantically meaningful path/subgraph.
Program/trace explanation	`score = 0; if fever then score+=2; if cough then score+=1; return (score` $\geq$ `2)`	Explains via an executable sequence of steps; supports replay and "why" questions at the step level.

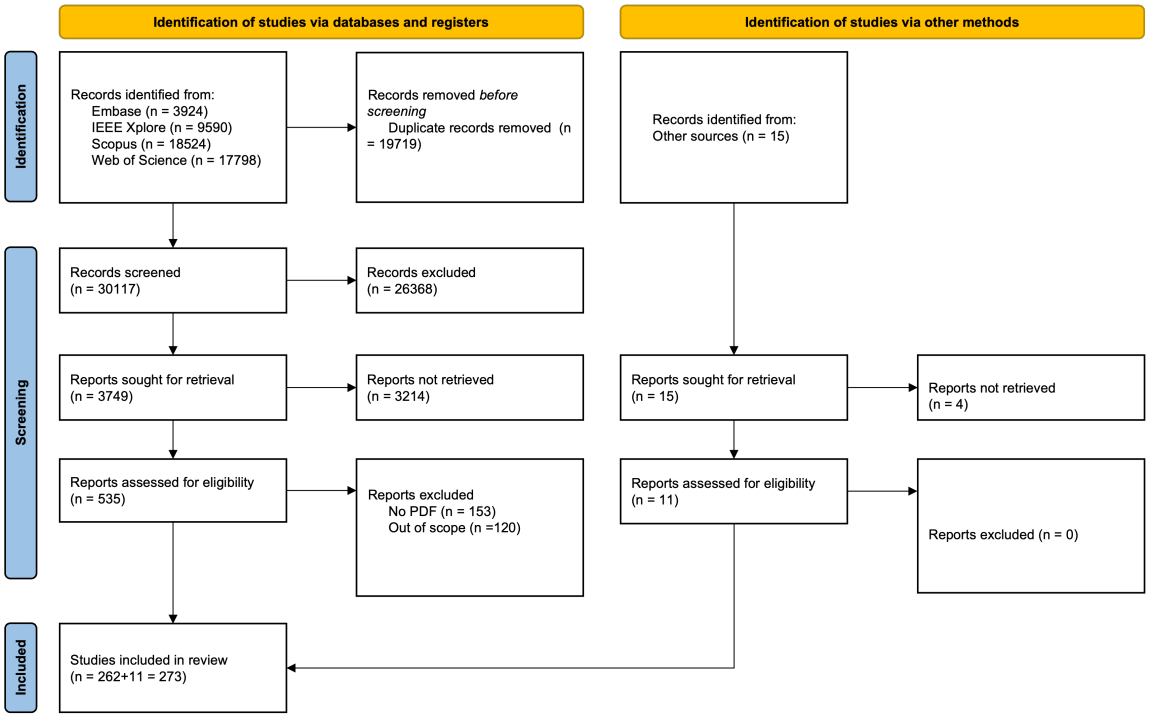

Fig. 1. PRISMA 2020 flow diagram for identification, screening, eligibility, and inclusion.

(1) *Topic induction* with BERTopic (Sentence-BERT `all-MiniLM-L6-v2`, UMAP, HDBSCAN), yielding 58 coherent topics with all items assigned.
(2) *Representative selection* inside each topic using a composite score that balances centrality and diversity. Let topic/cluster T contain papers $i \in T$ with embedding vectors $\mathbf{e}_i$ and centroid $\mathbf{c}_T$. We define the *centrality*

and *diversity* of paper i:

$$\text{centrality}_i = \cos(\mathbf{e}_i, \mathbf{c}_T), \quad \text{diversity}_i = 1 - \max_{j \in T, j \neq i} \cos(\mathbf{e}_i, \mathbf{e}_j),$$

and min-max normalize both metrics to $[0, 1]$ per topic. We combine centrality and diversity with a coverage-first weight to define the *representativeness* score:

$$\text{rep\_score}_i = 0.55 \cdot \text{diversity}_i + 0.45 \cdot \text{centrality}_i.$$

with duplicate avoidance and topic-size–proportional quotas (target ratio 15%). The weight $\alpha = 0.55$ was determined through internal validation.

(3) *Statistical validation* of selection with two-sided Mann–Whitney U tests plus Benjamini–Hochberg correction, Cohen's d (95% bootstrap CIs), and a Kolmogorov–Smirnov test for temporal bias.

From 3749 screenable reports, the procedure identified 535 representatives (14.27%). Validation confirmed significant, *practically meaningful* separation on diversity ($d = 0.573$) and representativeness ($d = 0.636$), no differences on alignment axes, and no temporal bias ($p_{\text{KS}} = 0.206$). Visual summaries appear in Figures 5 and 6.

Inclusion for synthesis. Full-text assessment of the candidate set led to 262 included studies that meet our symbolic/deep-learning criteria. We extracted standardized metadata, methodological descriptors, explanation types (inputs/outputs, local/global, ante-/post-hoc), and key findings.

Quality appraisal. We appraised risk of bias with ROBIS (Whiting and al. 2016), complemented by internal/external validity checks. A sensitivity analysis without the venue filter did not change qualitative conclusions (details in Appendix A).

Complementary literature identification via expert input. To ensure comprehensive coverage while maintaining reproducibility, we extended our systematic search using PRISMA 2020's "Identification of studies via other methods" branch (Page et al. 2021), including 11 papers identified by domain experts during the peer review phase (reviewer-identified gaps). We distinguish between the *reproducible database pipeline* (frozen as described above) and the *expert-augmented corpus*. Thus, we included 273 (262+11) papers (see Figure 1). By doing so, the latter is easily extendible to structured snowballing (Wohlin 2014)—we leave this for future work and identify it as a limitation.

Methodological failure modes and safeguards. The expert-augmented stream also makes explicit several failure modes of the semi-automated pipeline. First, *terminological under-retrieval* can occur when a seminal contribution uses the vocabulary of a neighboring subfield rather than the survey query terms or later taxonomy labels. For example, concept-bottleneck work such as Koh et al. (Koh et al. 2020) is central to concept-based interpretability, but may be weakly connected in bibliographic metadata to terms such as "symbolic", "logic", or "knowledge". This failure mode is therefore not only a retrieval issue, but also a vocabulary-alignment problem in horizontal surveys that span several research communities. Second, *salience flattening* can occur when many papers are presented in compact reference lists: foundational, representative, and peripheral works may appear to carry similar weight. Our centrality/diversity score and representative-paper tables are intended as navigation aids that reduce this effect, not as definitive rankings of scientific importance. Third, *taxonomy-boundary and coding errors* can occur when a paper lies near the SKE/SKI/Hybrid boundary or when a reference is reused in multiple explanatory roles. The original placement of Van Looveren and Klaise's counterfactual-explanation work (Looveren and Klaise 2021) under SKI is one such example: its root cause was a table-level coding/mapping error around the boundary between counterfactual explanation and training-time symbolic injection, not evidence that the method injects symbolic constraints during model training. We corrected such cases through manual audit and treat them as expected failure modes that future replications should actively check for.

Consequently, we interpret the corpus as a reproducible, expert-audited map of the literature rather than as a guarantee of exhaustive coverage for every vertical subfield. The safeguards are: (i) freezing and releasing the database-driven pipeline for auditability; (ii) reporting reviewer-identified additions through PRISMA's separate "other methods" stream; (iii) retaining provenance for expert-added papers; and (iv) manually auditing high-impact omissions and boundary cases before using category-level counts in the synthesis. Broader methodological limitations, including venue, language, time-window, and corpus-bias effects, are discussed in Appendix A, Section A.10.

Reproducibility. All parameters are specified in version-controlled YAML files; we log seeds, package versions, and intermediate artifacts. We release the code, configuration files, topic/metric tables, and statistical reports to enable exact regeneration of Figures 5–6 in the supplementary repository.[2] For full specifications, see Appendix A.

4 Taxonomy

We analyze symbolic XAI for deep learning along a technical axis that reflects *when* and *how* the symbolic structure is utilized. We distinguish:

- *Symbolic knowledge extraction (SKE)*, where a derived symbolic artifact explains a trained (neural) model;
- *Symbolic knowledge injection (SKI)*, where symbolic priors act during learning via losses, constraints, or structured supervision; and
- *Hybrid neurosymbolic architectures*, where the forward pass embeds symbolic computation and explanations are intrinsic.

Within each macro-category, we "decorate" methods by their symbolic structure (rules, trees, programs/logic, KGs, fuzzy rules) and alignment mechanism (logic/physics/monotone/regex constraints; concept interfaces; teacher–student distillation). This resolves common ambiguities. For example, fuzzy methods belong to SKI when used as training-time regularizers and to hybrids when the model is a fuzzy inference system; in the same vein, ontology/KG methods appear as post-hoc concept mappings or integrated reasoning. Figure 2 illustrates the taxonomy. Moreover, benefits, challenges, and limitations of these macro-categories are listed in Tables 3–5 (at the end of each subsection).

4.1 Post-hoc Symbolic Knowledge Extraction

Let f be a trained neural network and $\mathcal{S}$ a space of symbolic hypotheses (e.g., rule sets, decision trees, etc.). A procedure E is SKE if:

(1) it leaves f unchanged,
(2) it returns $s \in \mathcal{S}$ whose predictive behavior approximates f on a target distribution, and
(3) s has explicit, human-interpretable semantics (e.g., truth-conditional, graph-theoretic, etc.).

Evaluation emphasizes fidelity, coverage, size/complexity, and comprehensibility.

SKE methods differ chiefly by the *symbolic structure* used to encode the explanation. Each subcategory below outlines its scope and aim, accompanied by instances from our corpus.

- *Rule extraction:* These techniques produce global or local IF–THEN rules or (ordered) rule lists that emulate the network via surrogate learning, oracle probing, or constrained discretization, supporting audits and policy communication (An et al. 2022; Barbiero, Giannini, et al. 2024; Y. Chen, Yuan, et al. 2023; Gottesman et al. 2020; Honke et al. 2021; Junior 2020; Kasabov et al. 2023; Kozielski et al. 2025; Looveren and Klaise 2021; Y. Ma et al. 2024; Moradi and Samwald 2021; Muyama et al. 2024; Ragab et al. 2019; Ru et al. 2021; Schmaltz 2021; Sen et al. 2020; Thakker et al. 2020; Y. Wang, K. Li, et al. 2024; Watson et al. 2022; J. Yu and G. Liu 2020; Y. Zhang, Xiong, et al. 2022; Y. Zhang, X. Wang, et al. 2024). These rules can be propositional (Boolean conditions on input features) (Junior 2020; Kozielski et al. 2025; Moradi and Samwald 2021; Watson et al.

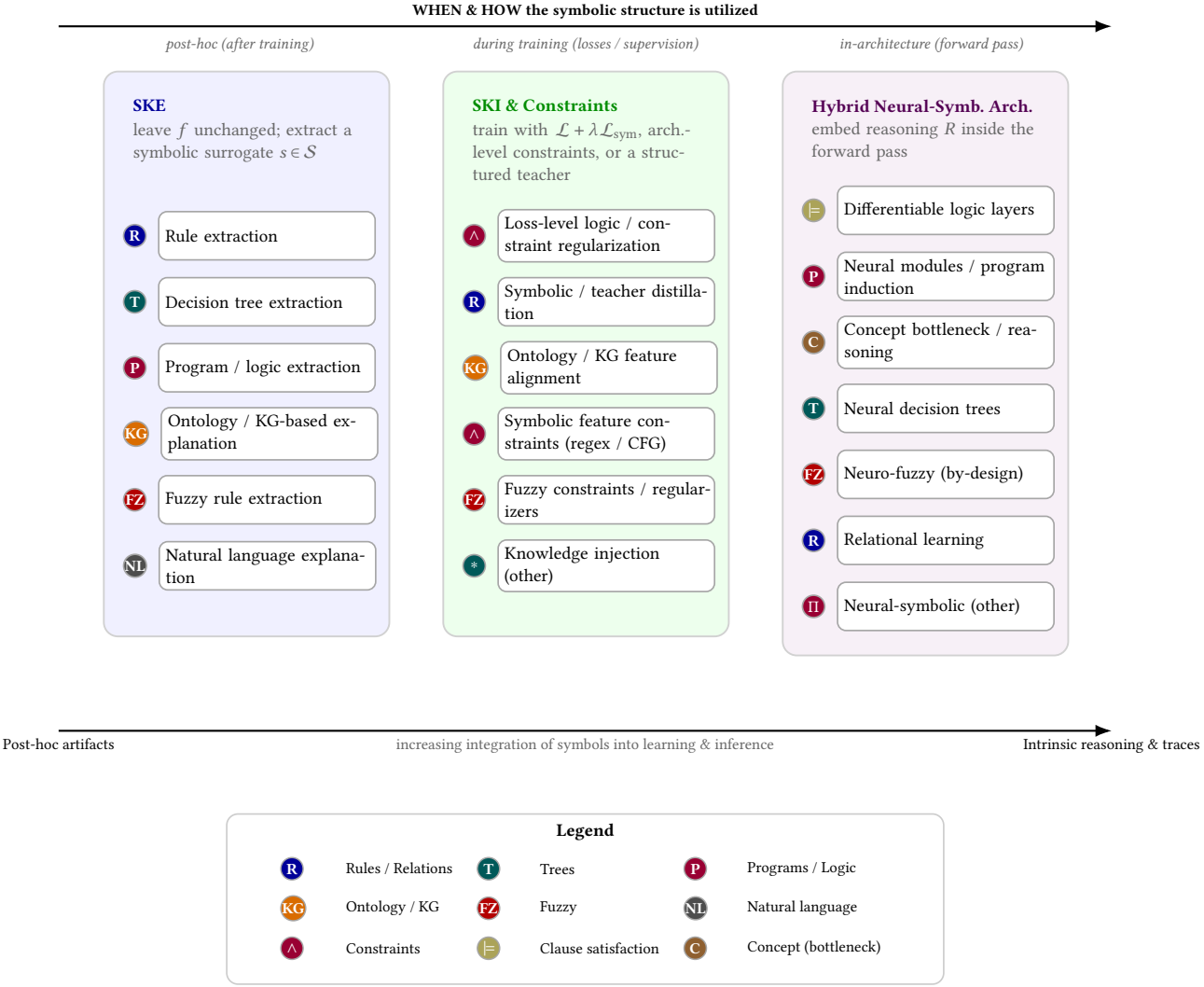

Fig. 2. Taxonomy of symbolic XAI for deep learning along the *when/how* axis: post-hoc Symbolic Knowledge Extraction (SKE), Symbolic Knowledge Injection during learning (SKI), and Hybrid neurosymbolic architectures with intrinsic reasoning. Colored badges denote the symbolic structure used.

2022; J. Yu and G. Liu 2020; Y. Zhang, X. Wang, et al. 2024) or (fragments of) first-order logic rules (which can express relations between entities, often seen with GNNs and relational data) (Barbiero, Giannini, et al. 2024; Y. Chen, Yuan, et al. 2023; Kasabov et al. 2023; Y. Ma et al. 2024; Ru et al. 2021; Sen et al. 2020; Thakker et al. 2020).

- *Decision tree extraction:* This category fits an interpretable decision tree to mimic the network, encoding readable rules that typically trade fidelity for global structure (Bertsimas and Stellato 2021; Bewley and Lawry 2021; Biggs et al. 2021; M. Lee et al. 2022; Marton et al. 2024; Sigut et al. 2023; H. Wu, Ruan, et al.

2023; H. Wu, C. Wang, et al. 2018; Yao et al. 2021; L. Zhang, X. Lu, et al. 2022; Q. Zhang et al. 2019; Zhuo et al. 2024; Zou et al. 2025). In decision trees, each path from root to leaf is essentially a rule.

- *Program/logic extraction:* Here, we derive logical clauses or concise programs (e.g., ILP, ASP, DSP, FSM, etc.) capturing decision boundaries, enabling compositional reasoning and formal inspection (Bauer et al. 2024; Choi et al. 2020; Huang, J. Zhao, et al. 2024; Kambhampati et al. 2022; M. Li et al. 2023; Padalkar et al. 2024; Y. Tang et al. 2021; Y. Wang, Su, et al. 2018; S. Yan et al. 2023).
- *Ontology/KG-based explanation:* These approaches align outputs to named concepts/relations (ontology or KG), yielding explanatory subgraphs or concept attributions grounded in domain semantics (H. Chang et al. 2023; K. Chen and Forbus 2021; Z. Chen, J. Chen, et al. 2024; Chhablani et al. 2024; L. Cui et al. 2020; Fu et al. 2021; Guan et al. 2025; M. Li et al. 2023; Y. Li et al. 2025; Mueller et al. 2023; Peng et al. 2022; Postic and Subasic 2025; Prado-Romero et al. 2024; H. Wang, S. Zhang, et al. 2025; S. Wang, H. Li, et al. 2022; X. Wang, K. Liu, et al. 2022; H. Wu, Fang, et al. 2023; Xian et al. 2021; L. Zhang, L. Jin, et al. 2023; J. Zheng et al. 2022).
- *Fuzzy rule extraction:* The extraction procedure translates decisions into linguistic rules with membership functions, enhancing readability for continuous inputs and noisy regimes (Aghaeipoor et al. 2023; Jarraya et al. 2019; C. Li, Ledo, et al. 2017; Y. Wang, H. Liu, et al. 2022; Zang et al. 2024).
- *Natural language explanation:* A variant worth mentioning is when methods generate controlled rationales in natural language, possibly anchored to symbolic conditions or templates for accessibility and auditability (Arnold et al. 2021; Z. Chen, J. Chen, et al. 2024; Chi and Liao 2022; Kalyanpur et al. 2022; L. Li et al. 2020; P. Lu et al. 2022; Mao et al. 2023; Shrivastava et al. 2023; M. H. Wang et al. 2024; Yen et al. 2021; Yi et al. 2025).

Abstract-interpretation-based techniques are an adjacent thread to our scope that primarily target robustness certification rather than explainability (Mirman et al. 2018). We mention them as they provide symbolic abstractions of network behavior. However, we clarify that our review focuses on symbolic structures used to *explain predictions* or to *constrain predictors in service of explainability*, and thus, we exclude such approaches.

Table 3 lists the benefits, challenges, and limitations of SKE methods.

4.2 Symbolic Knowledge Injection and Constraints

Let θ denote parameters and C symbolic conditions (e.g., logical clauses, ontology relations, etc.). A learner is SKI if:

- its loss is $\mathcal{L}(\theta) + \lambda \mathcal{L}_{\text{sym}}(\theta; C)$;
- it uses architecture-level constraints that structurally restricts computations to respect C; and/or
- it employs structured distillation such that the predictor systematically aligns with C during training.

We write the standard SKI objective as $\mathcal{L}(\theta) + \lambda \mathcal{L}_{\text{sym}}(\theta; C)$, where $\mathcal{L}(\theta)$ is the task loss (e.g., cross-entropy), C is a set of symbolic conditions (rules, constraints, ontology relations), and $\mathcal{L}_{\text{sym}}$ measures violation of C (soft penalties) or distance to satisfying assignments (relaxations). The scalar λ controls the trade-off: larger λ enforces stronger compliance but can reduce accuracy when constraints are incomplete or mis-specified. In later sections, we distinguish soft-penalty SKI from hard-constrained or architecture-level SKI, since only the latter can guarantee satisfaction "by-construction." Explanations follow from constraint satisfaction, aligned concept activations, or teacher rule traces.

SKI methods differ by the *channel* through which the symbolic knowledge enters the objective or supervision channel. We therefore group by constraint/supervision family.

- *Loss-level logic/constraint regularization:* A common approach is to add semantic/logic losses, monotonicity/shape, or physics-informed penalties that bias learning toward structured compliance (Ben-Younes et al. 2017; Bhatt et al. 2023; Böhle et al. 2022; Z. Chen, C. Chen, et al. 2019; Chu et al. 2018; Florensa et al. 2017; Gao et al. 2024; Gruber et al. 2023; He et al. 2024; X. Hu et al. 2023; F. Jiang et al. 2025; T. Jiang et al. 2023;

Table 3. Post-hoc Symbolic Knowledge Extraction (SKE): benefits, challenges, and limitations per subcategory, with representative works supporting each claim.

SKE Category	Benefits	Challenges	Limitations
Rule extraction	• Readable global/local rules for auditing and policy communication; high empirical fidelity when pruned/regularized (Kozielski et al. 2025; Moradi and Samwald 2021; Ragab et al. 2019). • Model-agnostic post-hoc artifacts that can be versioned and shared across stakeholders (Moradi and Samwald 2021; J. Yu and G. Liu 2020). • Supports error analysis and feature attribution at the clause/path level (Y. Chen, Yuan, et al. 2023; Junior 2020; Watson et al. 2022).	• Fidelity–simplicity trade-off: compact rule sets miss edge cases; larger sets become unreadable (Kozielski et al. 2025; Moradi and Samwald 2021). • Scalability and stability issues in high-dimensional data; sensitivity to hyperparameters and discretization choices (Barbiero, Giannini, et al. 2024; Junior 2020; Kozielski et al. 2025). • Capturing high-order interactions of deep models is difficult for flat rules (Barbiero, Giannini, et al. 2024; Y. Chen, Yuan, et al. 2023; J. Yu and G. Liu 2020).	• Post-hoc rationalization risk under distribution shift; artifacts may require re-extraction (Kozielski et al. 2025; Ragab et al. 2019). • Limited causal accountability due to indirect mapping from continuous features to discrete rules (Moradi and Samwald 2021; Watson et al. 2022).
Decision tree extraction	• Provides a single global structure where each path is a human-readable rule (Bewley and Lawry 2021; H. Wu, C. Wang, et al. 2018; Q. Zhang et al. 2019). • Can compress model behavior and enable fast inference/editing of decisions (Bertsimas and Stellato 2021; Biggs et al. 2021).	• Trees may grow deep/wide to match fidelity, reducing interpretability (Marton et al. 2024; Sigut et al. 2023; Q. Zhang et al. 2019). • Instability to small data perturbations and threshold choices (Sigut et al. 2023; H. Wu, C. Wang, et al. 2018; Zou et al. 2025).	• Limited ability to capture complex feature interactions compared to deep nets (Q. Zhang et al. 2019; Zou et al. 2025). • Requires re-fitting when data drifts; global trees can miss local phenomena and long-tail behaviors (Bewley and Lawry 2021; Marton et al. 2024).
Program/logic extraction	• Produces formal clauses/programs amenable to verification, compositional reasoning, and constraint checking (Huang, J. Zhao, et al. 2024; Kambhampati et al. 2022; Padalkar et al. 2024). • Enables modular explanations and "reasoning traces" that can be inspected, compared, or reused as symbolic components (Choi et al. 2020; Y. Wang, Su, et al. 2018; S. Yan et al. 2023).	• Search and optimization over symbolic spaces is combinatorial and compute-intensive (Huang, J. Zhao, et al. 2024; Padalkar et al. 2024; Y. Tang et al. 2021). • Noise and hyperparameter sensitivity (e.g., clause depth, pruning) (Choi et al. 2020; Y. Wang, Su, et al. 2018; S. Yan et al. 2023).	• Rigidity to the chosen language/DSL; domain engineering is often required (Kambhampati et al. 2022; Padalkar et al. 2024). • Indirect mapping from continuous representations to discrete logic may reduce faithfulness (Choi et al. 2020; Y. Wang, Su, et al. 2018; S. Yan et al. 2023).
Ontology/KG-based explanation	• Grounds explanations in domain vocabularies (entities/relations), aiding stakeholder understanding (L. Cui et al. 2020; Fu et al. 2021; S. Wang, H. Li, et al. 2022). • Supports multi-hop reasoning and subgraph rationales (Peng et al. 2022; H. Wu, Fang, et al. 2023; Xian et al. 2021).	• Dependence on KG coverage/quality; misaligned or incomplete schemas cause partial explanations (H. Chang et al. 2023; L. Cui et al. 2020; J. Zheng et al. 2022). • Entity/linking errors and concept drift require maintenance (Fu et al. 2021; H. Wu, Fang, et al. 2023; Xian et al. 2021).	• Out-of-KG phenomena cannot be explained; explanations can mirror KG incompleteness and biases (H. Chang et al. 2023; L. Cui et al. 2020). • Tight coupling to particular ontologies reduces portability across domains (S. Wang, H. Li, et al. 2022; J. Zheng et al. 2022).
Fuzzy rule extraction	• Linguistic IF–THEN rules with memberships handle uncertainty and continuous features gracefully (C. Li, Ledo, et al. 2017; Y. Wang, H. Liu, et al. 2022; Zang et al. 2024). • Often robust to noise and provides human-readable control logic (Aghaeipoor et al. 2023; Jarraya et al. 2019).	• Selecting/learning membership functions and rule bases without explosion (C. Li, Ledo, et al. 2017; Y. Wang, H. Liu, et al. 2022). • Balancing accuracy and interpretability as rules proliferate (Aghaeipoor et al. 2023; Zang et al. 2024).	• May require expert priors/heuristics; scalability issues for very high dimensions (Aghaeipoor et al. 2023; Jarraya et al. 2019). • Faithfulness limited by discretization/approximation of continuous dynamics (C. Li, Ledo, et al. 2017; Y. Wang, H. Liu, et al. 2022).
Natural language explanation	• Accessible free-text rationales/templates improve user communication and audit (Arnold et al. 2021; Chi and Liao 2022; Shrivastava et al. 2023). • Can be anchored to symbolic conditions for clarity (Kalyanpur et al. 2022; P. Lu et al. 2022; Yen et al. 2021).	• Faithfulness vs. plausibility gap; risk of hallucination (Chi and Liao 2022; P. Lu et al. 2022; Mao et al. 2023). • Evaluation of textual explanations remains subjective and task-dependent (Arnold et al. 2021; Peng et al. 2022; Shrivastava et al. 2023).	• May not reflect internal mechanisms; prone to rationalization (Mao et al. 2023; Peng et al. 2022). • Length and readability constraints limit detail; tighter constraints improve usability but reduce completeness (Shrivastava et al. 2023; Yen et al. 2021).

Jiao et al. 2024; Khosravi et al. 2019; W. Kim and Y. Lee 2019; Lawless et al. 2022; S. Lin et al. 2023; H. Liu et al. 2021; Mercea et al. 2022; Pati and Lerch 2021; Quan et al. 2023; Quessard et al. 2020; Roth et al. 2022; Upadhyay et al. 2023; J. Wang and M. Zhang 2022; P. Wang et al. 2024; X. Wang, Dai, et al. 2024; Z. Wu et al. 2024; J. Xu et al. 2018; S. Xu et al. 2024; H. Yang et al. 2024; L. Yang et al. 2022; L. Yu et al. 2022; B. Zhang, W. Zheng, et al. 2023; Z. Zhang et al. 2025; Y. Zhao, H. Wu, et al. 2025; Zhuo et al. 2024). During training, these constraints guide the model to behave in more interpretable ways; at inference, the model can then explain that it satisfies certain properties, which we can articulate symbolically.

- *Symbolic/teacher distillation:* Here, we supervise a student with a symbolic/structured teacher (e.g., rules, trees, etc.), often improving data efficiency and structure adherence (W. Chen et al. 2024; K. Das et al. 2023; Gupta et al. 2023; S. Lee and B. C. Song 2021; Pan et al. 2020; Shang et al. 2021; Q. Wu et al. 2018). The network may then use additional data to fine-tune, but because it was initialized or guided by a symbolic teacher, it tends to respect those semantics.
- *Ontology/KG feature alignment:* Another approach to injection involves aligning latent spaces to named entities/relations with auxiliary heads or graph-consistency losses, so explanations reference domain concepts (C. Chang et al. 2023; Le et al. 2023; B. Lee et al. 2023; S. Li et al. 2025; Marchetti et al. 2024; Meng et al. 2025; Ong et al. 2020; Sufriyana et al. 2023; Vu and Le 2025; S. Wang, B. Xie, et al. 2025; C. Xiao et al. 2017; W. Zhang et al. 2024; X. Zhu et al. 2023). These concepts serve as a built-in explanation vocabulary.
- *Symbolic feature constraints:* A less common approach, but worth mentioning, enforces regular expressions, grammars, or symbolic knowledge (e.g., first-order rules) as hard/soft constraints in information extraction and structured generation (Gan et al. 2021; Lan et al. 2022; Z. Liu, X. Chen, et al. 2024).
- *Fuzzy constraints/regularizers:* This class uses fuzzy memberships/relations as training-time penalties without architectural change, aiding noise robustness and semantic control (Y. Chen, W. Ding, et al. 2025; C. Li, Ledo, et al. 2017).
- *Knowledge injection (other):* Some approaches inject structured priors that do not neatly fit the above (e.g., game-theoretic concepts, causal edges, prototype–concept ties, engineered features with auxiliary losses) (Angelov et al. 2025; Q. Cao et al. 2017; Chouzenoux and Elvira 2024; Feng et al. 2025; Z. Li and X. Si 2022; Trivedi and Zha 2020; W. Wang et al. 2023; K. Zheng et al. 2024).

Table 4 synthesizes the benefits, challenges, and limitations of SKI methods.

4.3 Hybrid Neural-Symbolic Architectures

A hybrid neurosymbolic architecture integrates symbolic computation into the model's forward pass. Let g_θ denote a differentiable mapping whose computation includes a symbolic reasoning operator R (e.g., soft logic evaluation, program execution trace, KG path reasoning, fuzzy inference). A model is hybrid if $R(g_\theta(x))$ (possibly interleaved), such that reasoning steps are part of the predictive computation and produce intrinsic explanations (e.g., activated clauses, proof trees, program/module layouts, concept activations, tree paths, linguistic rules).

We categorize by the core *symbolic mechanism embedded* in the architecture; this determines the form of the intrinsic explanation and the training regime.

- *Differentiable logic layers:* This category encodes clause satisfaction or probabilistic/soft logic as differentiable modules; explanations identify which clauses/predicates fire and with what truth values (Aly and Vlachos 2024; Badreddine et al. 2022; Bereket and Karaletsos 2023; Y. Chen, J. Zhang, et al. 2025; Ciravegna et al. 2023; O. Csiszár, G. Csiszár, et al. 2020; O. Csiszár, Pusztaházi, et al. 2023; Fischer and Vreeken 2021; Hua et al. 2024; Hubin et al. 2021; Jain et al. 2023; Jaques et al. 2020; Jeong et al. 2024; Katzir et al. 2021; Lima and Lorini 2024; L. Ma et al. 2021; Mistry et al. 2022; Petersen et al. 2022; W. Shen et al. 2024; Shi et al. 2021; Z. Tang et al. 2025; Veer et al. 2023; K. Wang et al. 2020; Q. Wang et al. 2023; Z. Wang, W. Yan, et al. 2023; Z. Wang, W. Zhang, et al. 2024, 2021, 2020; Wei et al. 2018; J. Yu and G. Liu 2021; B. Zhang and L. Li 2023; Y. Zhao, Y. Shen, et al. 2023). In simple words, the network is trained to follow rules; at test time, we can show which rules fired.
- *Neural modules/program induction:* These are methods that compose neural modules mirroring a query's structure or induces DSL programs; that module layout/program trace is the explanation (Ai et al. 2023; Araki et al. 2022; Bauer et al. 2024; Ceylan et al. 2020; G. Cui and H. Zhu 2021; Cunnington et al. 2023; Ghosh et al. 2020; Y. Gu et al. 2024; X. Hu et al. 2023; Huang, F. Chen, et al. 2023; Hunter and Potyka 2023; M. Jin et al. 2022; Pettit et al. 2025; Tarzariol et al. 2022; S. Wang, B. Xie, et al. 2025; Y. Yang et al. 2025).

Table 4. Symbolic Knowledge Injection (SKI) and Constraints: benefits, challenges, and limitations per subcategory, with representative works supporting each claim.

SKI Category	Benefits	Challenges	Limitations
Loss-level logic/constraint regularization	• Encodes priors (logic, monotonicity, physics) directly into learning, improving robustness and sample efficiency (Bhatt et al. 2023; Gruber et al. 2023; He et al. 2024; Pati and Lerch 2021; P. Wang et al. 2024; J. Xu et al. 2018). • Yields explanations as satisfied constraints and clause/truth scores (Bhatt et al. 2023; W. Kim and Y. Lee 2019; J. Xu et al. 2018).	• Choosing constraint sets and weights is delicate; mis-specified priors degrade accuracy (Bhatt et al. 2023; Pati and Lerch 2021; J. Xu et al. 2018). • Scaling constraint evaluation is compute-intensive for long sequences/graphs (Bhatt et al. 2023; Gruber et al. 2023; He et al. 2024). • Conflicting/inactive constraints require curricula or annealing (Pati and Lerch 2021; X. Wang, Dai, et al. 2024; J. Xu et al. 2018).	• Certifies compliance rather than revealing internal causal mechanisms (Böhle et al. 2022; He et al. 2024; W. Kim and Y. Lee 2019). • Effectiveness depends on coverage/quality of the symbolic resource (He et al. 2024; Quessard et al. 2020; J. Xu et al. 2018).
Symbolic/teacher distillation	• Transfers structure from symbolic teachers (rules/trees) to flexible students, keeping capacity while respecting priors (Gupta et al. 2023; Pan et al. 2020; Shang et al. 2021; Q. Wu et al. 2018). • Often improves data efficiency and yields faithful teacher traces for explanations (W. Chen et al. 2024; K. Das et al. 2023; Gupta et al. 2023).	• Teacher limitations and bias are inherited by the student (K. Das et al. 2023; Pan et al. 2020). • Balancing teacher imitation with exploration requires careful objectives (W. Chen et al. 2024; Gupta et al. 2023; Shang et al. 2021).	• Over-regularization can reduce expressivity; coverage is limited by teacher rules (Gupta et al. 2023; Shang et al. 2021).
Ontology/KG feature alignment	• Aligns latent features to named entities/relations for concept-level explanations (C. Chang et al. 2023; S. Li et al. 2025; C. Xiao et al. 2017; W. Zhang et al. 2024). • Enables graph-consistency or multi-hop reasoning signals during training (Le et al. 2023; Meng et al. 2025; Vu and Le 2025; S. Wang, B. Xie, et al. 2025).	• Relies on high-quality KGs/ontologies; noise and drift require maintenance (S. Li et al. 2025; Meng et al. 2025; Vu and Le 2025). • Entity linking/disambiguation errors propagate to explanations (C. Chang et al. 2023; W. Zhang et al. 2024).	• Portability is limited by ontology scope and domain specificity (Vu and Le 2025; S. Wang, B. Xie, et al. 2025). • Out-of-KG facts remain unexplained or misaligned (S. Li et al. 2025; Meng et al. 2025).
Symbolic feature constraints	• Hard/soft structural constraints (regex/CFG) guarantee output validity and make checks auditable (Lan et al. 2022; Z. Liu, X. Chen, et al. 2024).	• Constraint design is labor-intensive; rigid patterns can conflict with learning (Lan et al. 2022; Z. Liu, X. Chen, et al. 2024).	• Limited expressivity outside language-like structure; brittle under domain shift (Lan et al. 2022; Z. Liu, X. Chen, et al. 2024).
Fuzzy constraints/regularizers	• Fuzzy memberships as penalties improve noise robustness and controllability of decision boundaries (Y. Chen, W. Ding, et al. 2025).	• Selecting membership shapes and weights requires tuning and may cause rule proliferation if combined (Y. Chen, W. Ding, et al. 2025).	• Scales poorly to very high dimensions without structure; effects can be task-dependent (Y. Chen, W. Ding, et al. 2025).
Knowledge injection (other)	• Flexible channel for injecting prototypes, causal edges, or engineered features via auxiliary losses (Angelov et al. 2025; Z. Li and X. Si 2022; Trivedi and Zha 2020).	• Ambiguous attribution of gains to the injected knowledge vs. generic regularization or data (Angelov et al. 2025; Trivedi and Zha 2020).	• Often lacks formal guarantees and may require heavy domain engineering (Angelov et al. 2025; Z. Li and X. Si 2022).

- *Concept bottleneck/reasoning:* Concept bottleneck models are a recent idea where a neural network is forced to predict intermediate human-meaningful concepts (e.g., attributes, relations, etc.), then symbolic reasoning (e.g., rules, logic, constraints, etc.) uses these concepts to explain decisions (Barbiero, Giannini, et al. 2024; Debot et al. 2024; Hong et al. 2023; L. Hu et al. 2024; Koh et al. 2020; Laurent et al. 2024; H. Li, C. Shen, et al. 2024; Ren et al. 2023; Stammer et al. 2021; Tong et al. 2020; S. Wang, Zeng, et al. 2024; Z. Wang, T. Chen, et al. 2017; Yeh et al. 2017).
- *Neural decision trees:* Here, the techniques build differentiable tree-shaped predictors where paths encode rules with soft thresholds, offering structured global transparency (Barbiero, Ciravegna, et al. 2022; Geyer et al. 2024; Ibrahim et al. 2023; Z. Liu, Y. Zhu, et al. 2025; Nauta et al. 2021; Tilwani et al. 2024; Z. Xu et al. 2022; R. Yan et al. 2022; Zharmagambetov and Carreira-Perpiñán 2022; Y. Zheng et al. 2023).
- *Neuro-fuzzy (by-design):* This category realizes a fuzzy inference system (e.g., TSK, ANFIS, etc.) as a neural network, where explanations are linguistic rule bases with learned memberships and pruning strategies (Bian et al. 2024; J. Cao et al. 2025; O. Csiszár, G. Csiszár, et al. 2020; Ebadzadeh and Salimi-Badr 2018; X. Gu 2022; X. Gu et al. 2024; F. Hu et al. 2024; Iyer et al. 2018; Kridalukmana et al. 2022; Lang et al.

2023; S. Liu et al. 2024; Y. Liu et al. 2025; Sajid et al. 2024; Salimi-Badr 2024; N. Wang et al. 2022; Xue et al. 2025; E. Zhou et al. 2023; H. Zhu et al. 2025).

- *Relational learning:* These are architectures for relational/graph data that learn or execute rules over entities/relations using neural scoring (C. Chen et al. 2024; Y. Chen, Bijral, et al. 2019; Y. Chen, Natarajan, et al. 2022; Christmann et al. 2023; Q. Cui et al. 2024; Defresne et al. 2023; Erion et al. 2021; González-Briones et al. 2025; Gujral et al. 2020; Y. Jiang and Bansal 2019; Kang et al. 2018; J. Li et al. 2025; T. Lin et al. 2025; Z. Liu, L. Wu, et al. 2025; Nezerka et al. 2024; J. Si et al. 2023; Wan et al. 2020; J. Wang, P. Ding, et al. 2025; L. Zhang, Su, et al. 2023; T. Zhang et al. 2025; Zuo et al. 2023). Here, the symbolic trace is a rule body (paths, clauses) annotated with neural confidences, where the model identifies patterns (e.g., multi-hop QA reasoning, graph-based association rules, etc.) and presents them as its reasoning.
- *Neurosymbolic (other):* These approaches integrate symbolic components (e.g., ILP, ASP, theorem prover, etc.) with neural scoring; explanations are symbolic constructs (e.g., proofs, clauses, causality graphs, etc.) witnessing entailment or consistency (Ciocarlan et al. 2024; S. Das et al. 2021; Deng et al. 2025; Z. Hu et al. 2017; S. W. Kim et al. 2020; C. Li, B. Zhang, et al. 2025; P. Lin et al. 2020; J. Ma et al. 2022; Pang et al. 2025; Rymarczyk et al. 2021; Sen et al. 2020; P. Song et al. 2023; H. Wang, Che, et al. 2024; Wen et al. 2024; Xia et al. 2022; T. Xiao et al. 2024; Y. Xie et al. 2021; X. Xu et al. 2023; K. Yan et al. 2021; L. Zhou et al. 2024; Zunic et al. 2022).

Table 5 lists the benefits, challenges, and limitations of hybrid methods.

5 Results and Analysis

Using the taxonomy as a scaffold, we analyze empirical trends across our curated corpus and assess where symbolic methods for deep learning deliver most value. We combine a quantitative view (publication dynamics by year and macro-category) with a qualitative synthesis of effectiveness, trade-offs, and regulatory fit. Our corpus spans 273 peer-reviewed works (79 SKE, 71 SKI, 123 Hybrids) from 2017–2025, screened per the methodology in Section 3 (and its extended version in Appendix A) and restricted to peer-reviewed venues.

5.1 Publication Trends (2017–2025)

Three observations emerge:

(1) *Strong growth, then diversification:* Early activity (2017–2019) is modest—8—9 papers/year—with some emphasis on post-hoc SKE (e.g., rule/tree extraction) and first hybrid exemplars (e.g., neural modules, fuzzy-neural). Over the same horizon, SKE plateaus (37 works in 2020–2022 vs. 35 in 2023–2025), while SKI and Hybrids accelerate (SKI: from 18 to 42; Hybrids: from 46 to 69), consistent with a shift from explaining black-boxes to "explainability by-design." This rebalancing reflects the community's return to neurosymbolic integration as an explicit research focus after 2020 and aligns with taxonomic distinctions, such as SKE vs. SKI, popularized by recent systematic reviews. Table 6 summarizes these trends.

(2) *Domain broadening:* While early works concentrate on vision and tabular settings for rule/tree surrogates, the later years bring (i) NLP/sequence models informed by constraints or concept interfaces, (ii) graph/KG reasoning and KG-aligned explanations, and (iii) RL with temporal/logical structure or policy abstractions. Our corpus also contains applied slices in healthcare and credit/risk, where symbolic rationales double as documentation and reason codes.

(3) *Macro-category composition:* Over the full window SKE : SKI : Hybrids $\approx 79 : 71 : 123$, i.e., $\approx 29\% : 26\% : 45\%$. The hybrid share is largest overall, reflecting the maturation of differentiable logic layers, neural decision trees, fuzzy-neural designs, and modular/program-induction pipelines that produce intrinsic traces (clause activations, tree paths, rule firings, module layouts) rather than post-hoc surrogates.

Table 5. Hybrid Neural-Symbolic Architectures: benefits, challenges, and limitations per subcategory, with representative works supporting each claim.

Hybrid Category	Benefits	Challenges	Limitations
Differentiable logic layers	• Intrinsic explanations via clause/predicate activations and truth values during the forward pass (Aly and Vlachos 2024; O. Csiszár, G. Csiszár, et al. 2020; O. Csiszár, Pusztaházi, et al. 2023; Fischer and Vreeken 2021; Lima and Lorini 2024; Mistry et al. 2022; Veer et al. 2023). • Improves compositional generalization by enforcing rule structure end-to-end (Fischer and Vreeken 2021; Jaques et al. 2020; Mistry et al. 2022; Veer et al. 2023).	• Non-convex relaxations and gradient instability for discrete logic; careful initialization and annealing often needed (O. Csiszár, Pusztaházi, et al. 2023; Fischer and Vreeken 2021; Jaques et al. 2020). • Computational overhead for evaluating many clauses/predicates at scale (Aly and Vlachos 2024; Fischer and Vreeken 2021; Lima and Lorini 2024; Veer et al. 2023).	• Bound to the chosen logic formalism; portability across domains can be limited (Aly and Vlachos 2024; O. Csiszár, G. Csiszár, et al. 2020; O. Csiszár, Pusztaházi, et al. 2023; Lima and Lorini 2024). • Truth value approximations may reduce faithfulness to crisp semantics (O. Csiszár, Pusztaházi, et al. 2023; Fischer and Vreeken 2021; Jaques et al. 2020).
Neural modules/program induction	• Produces explicit module layouts/program traces as step-by-step explanations (Ceylan et al. 2020; G. Cui and H. Zhu 2021; Y. Gu et al. 2024; X. Hu et al. 2023; Huang, F. Chen, et al. 2023; M. Jin et al. 2022). • Supports compositional queries and improves data efficiency when structure matches the task (Ai et al. 2023; Ceylan et al. 2020; G. Cui and H. Zhu 2021; X. Hu et al. 2023; Hunter and Potyka 2023).	• Learning discrete structures/programs is combinatorial; search and supervision can be heavy (Ceylan et al. 2020; G. Cui and H. Zhu 2021; Y. Gu et al. 2024; Huang, F. Chen, et al. 2023; M. Jin et al. 2022). • Brittleness to spurious layouts and exposure bias in sequence-style planners (Ai et al. 2023; Cunnington et al. 2023; X. Hu et al. 2023; Hunter and Potyka 2023).	• Strong priors on DSL/module library may restrict expressivity outside the domain (Ceylan et al. 2020; G. Cui and H. Zhu 2021; Y. Gu et al. 2024; M. Jin et al. 2022). • Inference latency due to program execution or multi-step reasoning (Ai et al. 2023; Araki et al. 2022; Huang, F. Chen, et al. 2023; Hunter and Potyka 2023).
Concept bottleneck/reasoning	• Localizes evidence to human-meaningful concepts; enables intervention and debugging (L. Hu et al. 2024; Stammer et al. 2021; Z. Wang, T. Chen, et al. 2017; Yeh et al. 2017). • Transparent decision paths when rules operate over predicted concepts (Hong et al. 2023; L. Hu et al. 2024; Tong et al. 2020; S. Wang, Zeng, et al. 2024).	• Requires high-quality concept labels/definitions; concept drift and label noise hurt reliability (L. Hu et al. 2024; Stammer et al. 2021; S. Wang, Zeng, et al. 2024). • Trade-off between concept accuracy and end-task performance needs balancing (Hong et al. 2023; Tong et al. 2020; Yeh et al. 2017).	• Coverage limited to the predefined concept set; unseen concepts are invisible (L. Hu et al. 2024; Z. Wang, T. Chen, et al. 2017; Yeh et al. 2017). • Concept leakage and shortcut learning can mislead explanations (Hong et al. 2023; Stammer et al. 2021; Tong et al. 2020).
Neural decision trees	• Global tree structure with soft thresholds yields readable rule paths and smooth training (Nauta et al. 2021; Z. Xu et al. 2022; R. Yan et al. 2022; Y. Zheng et al. 2023). • Supports pruning/regularization to balance accuracy and interpretability (Barbiero, Ciravegna, et al. 2022; Nauta et al. 2021; Tilwani et al. 2024; Z. Xu et al. 2022).	• Depth/width can grow to match complex decision boundaries, hurting clarity (Z. Xu et al. 2022; R. Yan et al. 2022; Zharmagambetov and Carreira-Perpiñán 2022; Y. Zheng et al. 2023). • Split instability and sensitivity to initialization/temperature schedules (Barbiero, Ciravegna, et al. 2022; Nauta et al. 2021; Zharmagambetov and Carreira-Perpiñán 2022).	• Limited modeling of high-order interactions compared to general deep nets (Nauta et al. 2021; Z. Xu et al. 2022; R. Yan et al. 2022). • Retraining is needed to maintain faithful paths under distribution shifts (Geyer et al. 2024; Tilwani et al. 2024; Y. Zheng et al. 2023).
Neuro-fuzzy (by-design)	• Produces linguistic rule bases with learned memberships; naturally handles uncertainty (Bian et al. 2024; Ebadzadeh and Salimi-Badr 2018; Sajid et al. 2024; N. Wang et al. 2022; E. Zhou et al. 2023). • Often robust and interpretable in noisy/continuous domains (J. Cao et al. 2025; X. Gu 2022; Iyer et al. 2018; S. Liu et al. 2024; Salimi-Badr 2024).	• Tuning membership shapes and avoiding rule proliferation require pruning and curricula (J. Cao et al. 2025; Ebadzadeh and Salimi-Badr 2018; S. Liu et al. 2024; N. Wang et al. 2022). • Optimization can be sensitive and slower due to fuzzy inference layers (Bian et al. 2024; Sajid et al. 2024; E. Zhou et al. 2023).	• Scalability to high-dimensional inputs and many rules is limited without structure (J. Cao et al. 2025; X. Gu 2022; Salimi-Badr 2024). • Linguistic labels may require domain expertise to define and maintain (Ebadzadeh and Salimi-Badr 2018; X. Gu 2022; Iyer et al. 2018).
Relational learning	• Learns/executes multi-hop rules and path-based reasoning with intrinsic rule/path explanations (Y. Chen, Natarajan, et al. 2022; Y. Jiang and Bansal 2019; J. Si et al. 2023; Wan et al. 2020). • Naturally fits graph/relational data with interpretable symbolic traces (Christmann et al. 2023; Defresne et al. 2023; Gujral et al. 2020).	• Search space over entities/relations explodes; requires sampling, pruning, or differentiable relaxations (Y. Chen, Natarajan, et al. 2022; Y. Jiang and Bansal 2019; J. Si et al. 2023; Wan et al. 2020). • Noisy/incomplete graphs and link errors degrade reasoning quality (Christmann et al. 2023; Defresne et al. 2023; Gujral et al. 2020).	• Portability depends on the graph schema; out-of-graph facts cannot be explained (Y. Chen, Natarajan, et al. 2022; Gujral et al. 2020). • Inference cost can be high for long paths or large KGs (Y. Jiang and Bansal 2019; J. Si et al. 2023; Wan et al. 2020).
Neurosymbolic (other)	• Integrates powerful symbolic components (provers, planners), giving proofs/derivations as explanations (Ciocarlan et al. 2024; Z. Hu et al. 2017; J. Ma et al. 2022; Xia et al. 2022; T. Xiao et al. 2024). • Can impose strong correctness guarantees on subproblems (Deng et al. 2025; Z. Hu et al. 2017; S. W. Kim et al. 2020).	• Bridging differentiability with discrete solvers requires surrogates or bi-level optimization (J. Ma et al. 2022; Rymarczyk et al. 2021; Xia et al. 2022; Y. Xie et al. 2021). • Engineering and dependency complexity is high for hybrid stacks (S. W. Kim et al. 2020; T. Xiao et al. 2024; X. Xu et al. 2023).	• Tight coupling to specialized symbolic tools reduces portability and increases maintenance costs (Deng et al. 2025; S. W. Kim et al. 2020; T. Xiao et al. 2024). • Latency and memory overhead during inference limit deployment scenarios (J. Ma et al. 2022; Y. Xie et al. 2021; X. Xu et al. 2023).

All numbers derive from our screened corpus (peer-reviewed only; 2017–2025 inclusive). They complement, and in places update, broader neurosymbolic/XAI surveys by focusing specifically on symbolic explainability in deep learning rather than on accuracy-oriented neurosymbolic systems in general.

Table 6. Publication trends by year and category. Reported numbers are counts and percentages for a specific year. The last row is the total sum of the columns.

Year	SKE	SKI	Hybrids
2017	1 (12.5%)	4 (50.0%)	3 (37.5%)
2018	2 (22.2%)	4 (44.5%)	3 (33.3%)
2019	4 (44.4%)	3 (33.3%)	2 (22.2%)
2020	8 (30.8%)	5 (19.2%)	13 (50.0%)
2021	16 (44.4%)	5 (13.9%)	15 (41.7%)
2022	13 (33.3%)	8 (20.5%)	18 (46.2%)
2023	13 (22.8%)	19 (33.3%)	25 (43.9%)
2024	14 (25.9%)	12 (22.2%)	28 (51.9%)
2025 (June)	8 (22.9%)	11 (31.4%)	16 (45.7%)
Total	**79 (28.9%)**	**71 (26.0%)**	**123 (45.1%)**

5.2 Efficiency of Symbolic XAI Approaches

Post-hoc SKE. For tabular and structured tasks, global rule lists and compact trees often achieve high fidelity with tractable size after pruning/regularization, enabling audits and model governance. For high-dimensional vision or highly entangled representations, surrogates either balloon (hurting readability) or restrict scope (local rules), foregrounding the fidelity–simplicity tension. Program/logic extraction yields formally interpretable artifacts (clauses/programs) that support compositional checks, but their search/regularization hyperparameters can affect stability and scalability.

SKI and constraints. Injecting logic/semantic losses or monotone/shape/physics constraints can enforce properties users expect (e.g., monotonic ranking) with small accuracy deltas relative to unconstrained baselines, while symbolic/teacher distillation transfers structure into high-capacity students without freezing design space. KG/ontology alignment readily equips models with a named concept vocabulary, improving controllability and explanatory granularity (e.g., attributing to relations, not just raw features).

Hybrids (explanations by-design). Differentiable logic layers, neural decision trees, neuro-fuzzy systems, and neural module/program architectures render reasoning steps intrinsic to the forward pass. On synthetic or compositional benchmarks (e.g., visual reasoning), hybrids often close the performance gap while preserving legible traces; on unconstrained real-world data, some variants still trade a few points of top-line accuracy for inspectability and intervention (e.g., edit a concept, inspect which clause fired), a trade-off increasingly acceptable in regulated domains.

5.3 Alignment with Regulatory and Governance Requirements

Symbolic XAI aligns naturally with evolving law- and risk-based expectations:

- *EU AI Act (Reg. 2024/1689):*[4] High-risk systems must meet transparency/documentation duties; providers must maintain technical documentation and information enabling users to interpret outputs. Symbolic artifacts (rules, clauses, KG rationales) are concrete, versionable units that can be surfaced in user information and technical files, and constraint satisfaction offers auditable evidence of property compliance.
- *GDPR (Art. 22; Recital 71—Profiling):*[5] Where automated decisions produce legal/similar effects, controllers must provide meaningful information about the logic involved and allow for human contestation. Explicit decision rules or concept-logic chains meet the spirit of "meaningful information" more readily than opaque attributions.

[4] https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
[5] https://gdpr-info.eu/art-22-gdpr/

Table 7. Checklist that ties symbolic artifacts to legal/assurance duties and the associated logging.

Requirement	Legal source	Symbolic artifact(s) to provide	Minimal evidence to retain	Where it fits
Transparency and information to deployers	EU AI Act Art. 13	Rule lists/decision paths; clause truth-value tables; KG path rationales; constraint compliance summaries	Model and data version IDs; extraction timestamp; dataset snapshot hash; fidelity/coverage metrics; explanation scope; linkage from decision to artifact	Technical documentation (Art. 11); user information package; conformity/assurance file
Record–keeping (event logging)	EU AI Act Art. 12	Per-decision traces (activated rules, tree paths, module layouts); constraint satisfaction/violation records	Automatic event logs (timestamp, input/output hashes, chosen branch/module, violated/active constraints); retention schedule; signer	Logging subsystem; post-market monitoring; quality management system
Automated decisions: "meaningful information about the logic involved"	GDPR Art. 22 (and Rec. 71)	Human-readable rule summaries; monotonicity statements; concept/ontology flow graphs; rationale templates	Per-decision reason string referencing rule(s)/constraint(s); mapping from model features to named concepts; DPIA cross-reference	Data protection impact assessment (DPIA); user-facing notices; model cards
Adverse action reasons in credit decisions	ECOA / Reg B (12 CFR 1002.9)	Rule-based "reason codes" mapped to standardized categories; monotone/eligibility constraints for fairness explanations	Stored reasons per decision; model version; score factors; audit IDs; retention for regulatory lookback	Adverse action notice (AAN) pipeline; compliance reporting
Good ML Practice/transparency for SaMD	FDA GMLP; agency transparency guidance	Explainability specification; constraint definitions tied to hazards; proof/trace exemplars (e.g., clause activations)	Design history links from requirement to test to artifact; verification logs; change control (e.g., PCCP)	Design history file; quality system records; post-market surveillance
Risk evidence and continuous monitoring	NIST AI RMF 1.0	Risk evidence register: explanation coverage and stability metrics; constraint dashboards; rule drift reports	Continuous monitoring logs; metric thresholds and alerts; reviews and sign-offs; incident tickets	RMF Govern–Map–Measure–Manage artifacts; model registry

- *Credit decisioning (ECOA/Reg B, CFPB):*[6] Adverse-action notices must state specific reasons, including when decisions rely on complex algorithms—symbolic rule extraction readily maps to standardized reason codes.
- *Medical devices (FDA/HC/MHRA GMLP; PCCP):*[7] Good ML practice stresses transparency, traceability, and data/model management, and the FDA's PCCP guidance underscores change control. Symbolic constraints and intrinsic traces (e.g., rule firings) support traceable behavior and post-market monitoring.
- *Risk management (NIST AI RMF 1.0):*[8] The RMF's GOVERN/MAP/MEASURE/MANAGE cycle treats explainability and transparency as cross-cutting properties; symbolic artifacts act as risk evidence—testable, monitorable, and communicable across stakeholders.

Table 7 synthesizes the evidence-to-duty checklist that ties symbolic artifacts to specific duties/requirements and how to log them. Historically, the renewed emphasis on explainability traces to DARPA's XAI program (2016–)[9] and the EU's Trustworthy AI guidelines (2019)[10], both of which positioned understandability as a design goal rather than a post-hoc patch—an impetus reflected in the shift from SKE to SKI/Hybrids in our counts.

[6] https://www.consumerfinance.gov/rules-policy/regulations/1002/9/; https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/
[7] https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles; https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device
[8] https://www.nist.gov/itl/ai-risk-management-framework
[9] https://www.darpa.mil/research/programs/explainable-artificial-intelligence
[10] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Table 8. Three prioritized patterns for integrating symbolic XAI in deep learning (distilled from the survey). Each pattern specifies when to use it, the symbolic artifacts it yields, what to validate, and common failure modes.

Pattern	When to use (primary driver)	Symbolic artifacts produced	Validate & log (minimum)	Common failure modes
P1: SKE-first	You need governance-ready explanations *without* changing the deployed predictor; retrofitting and auditing are priorities.	Rules/trees/programs fitted to model behavior; local rationales; symbolic surrogates.	Fidelity to the predictor (global/local); stability across perturbations; artifact complexity (depth/length); drift monitoring of extracted artifacts.	Surrogate drift (faithfulness gap); over-complex artifacts; brittle local explanations; dataset shift breaks explanation validity.
P2: SKI-first	You must enforce or encourage domain constraints *during learning* (safety, monotonicity, consistency, concept interfaces).	Constraint satisfaction reports; clause activations; concept bottlenecks; aligned representations.	Constraint violation rates (train/test); calibration of constraint trade-off; coverage of constraints (where they apply); provenance of C and updates.	Mis-specified/incomplete constraints; "gaming" soft penalties; reduced accuracy under partial knowledge; hidden violations outside constraint scope.
P3: Hybrid-by-design	You need intrinsic, inspectable reasoning traces (planning/proofs/KG paths) as part of the product behavior; structured decision support.	Reasoning traces (plans/proofs); KG paths/subgraphs; program traces; decision policies over symbolic state.	Trace validity (soundness wrt rules/KG); end-to-end robustness of perception→symbols; trace coverage; audit logs linking decisions to trace steps.	Symbol grounding errors (perception→symbols); brittle symbolic pipelines; trace incompleteness; cascading errors across modules.

5.4 Challenges and Open Directions

Fidelity vs. simplicity (SKE). Global surrogates risk over-smoothing minority phenomena; local rules can overgeneralize without coverage control. Extracted programs/clauses require careful search/pruning to remain stable across re-trains.

Specification and scaling (SKI). Constraint design and weighting are delicate. Multiple priors can conflict, and symbolic losses may be compute-intensive—necessitating curricula/annealing and violation diagnostics to sustain training stability.

Optimization and coupling (hybrids). Integrating discrete reasoning incurs optimization overhead and may couple models to particular symbolic formalisms (e.g., logic, DSL, KG, etc.), raising portability and latency concerns without specialized engineering.

Evaluation and faithfulness. Beyond accuracy, communities need fidelity, coverage, and stability metrics for symbolic artifacts, plus human-factors evaluations (usability, contestability). With the rise of LLM-assisted rationales, faithfulness checks (e.g., constraint audits, proof witnesses) become central—an arena where symbolic representations provide verifiable anchors.

6 Conceptual Framework for Symbolic Integration in Deep Learning

Building on the survey findings, this section distills an evidence-grounded *integration playbook* for symbolic XAI in deep learning. Rather than enumerating all possible system design variants, we prioritize three recurring integration patterns aligned with our taxonomy—SKE-first, SKI-first, and Hybrid-by-design—and specify for each (i) when it is appropriate, (ii) which symbolic artifacts it produces, (iii) what should be evaluated and logged, and (iv) the most common failure modes. Figure 3 summarizes the shared system components.

6.1 Prioritized Integration Patterns

Table 8 summarizes the three patterns we observe repeatedly across domains (see Section 7) and that align with the trade-offs reported in our synthesis tables (Section 4).

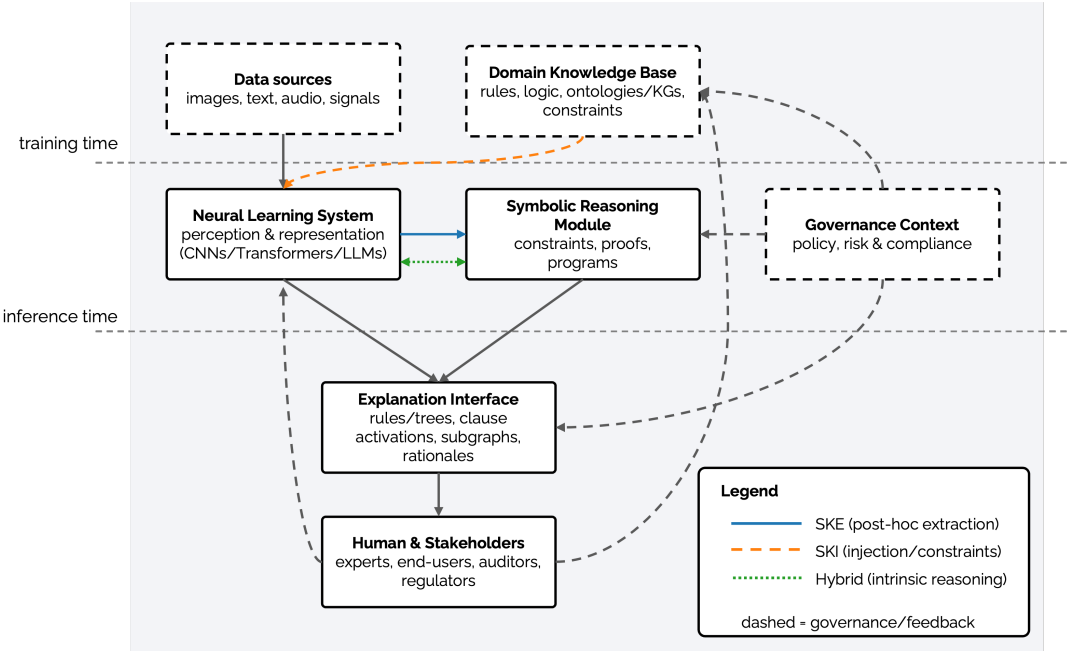

Fig. 3. Conceptual playbook for symbolic integration in deep learning. *Data* and the *Domain Knowledge Base* feed neural and symbolic cores. SKI injects priors during training; SKE extracts symbolic artifacts post-hoc; Hybrid designs exchange symbolic state and reasoning traces intrinsically. Explanations are consolidated at an interface layer for stakeholders; governance constraints and human feedback induce updates to knowledge, monitoring, and models.

6.2 Components (Operational View)

We keep the same five components as in Figure 3, but state them operationally in terms of what must be *represented, constrained, and audited.*

Domain Knowledge Base (symbolic memory). A versioned repository of symbolic information (rules, ontologies/KGs, constraints, concept vocabularies). Operationally, it provides (i) training-time priors (SKI), (ii) an explanation vocabulary, and (iii) audit provenance (what knowledge was assumed at decision time).

Neural Learning System. The deep model(s) that map raw inputs to predictions and/or intermediate representations. Operationally, this component must expose either (i) query access for extraction (SKE), (ii) constraint hooks for training (SKI), or (iii) symbol interfaces for hybrid reasoning.

Symbolic Reasoning Module. A mechanism that applies rules/constraints/programs to symbolic state (possibly inferred by the neural system). Operationally, it should emit *checkable artifacts*: proofs, traces, constraint logs, or path rationales.

Explanation Interface. A presentation layer that selects the explanation object (rule list, path, proof, KG rationale, trace) and adapts granularity to the stakeholder. Operationally, it should also report *uncertainty and scope* (what the explanation does and does not cover).

Human Feedback Loop. A structured mechanism for curation and correction: updating the KB (exceptions, new concepts), correcting labels or concept annotations, and triggering monitoring actions (e.g., drift alerts). Operationally, it defines who can edit knowledge, under what review process, and how changes propagate.

6.3 Governance, Monitoring, and "What to Record"

A consistent theme in real deployments is that explanations are not only user-facing text: they are also *records* that support auditing, risk assessment, and continuous monitoring. Risk-management guidance emphasizes that limited explainability and poor documentation increase operational risk; therefore, symbolic XAI systems should treat explanation artifacts as first-class monitoring signals (e.g., rule drift, constraint violation dashboards, trace coverage). This perspective aligns naturally with lifecycle-oriented risk frameworks (e.g., govern–map–measure–manage) and motivates logging the following minimal evidence:

- *provenance:* KB version, constraint set C, and any exceptions active at inference time;
- *artifact logs:* extracted rules/trees (SKE), constraint satisfaction traces (SKI), or proof/plan/KG-path traces (Hybrid);
- *stability signals:* artifact drift over time; sensitivity of explanations to small input perturbations; changes after model updates;
- *scope statements:* whether an explanation is local vs. global, and whether it approximates or guarantees behavior (soft vs. hard constraints).

6.4 A Practical Instantiation Guideline (and A Key Caveat)

A practical way to instantiate the playbook is to start from the stakeholder requirement and select the minimal pattern in Table 8. For example, if a domain requires intrinsically checkable decision traces, Hybrid-by-design is typically unavoidable; if retrofitting is required, SKE-first is the minimal intervention; if systematic compliance must be enforced, SKI-first is preferred.

Caveat: concept alignment is often the bottleneck. We emphasize (and do not assume away) that aligning KB concepts with neural representations is a delicate, often architecture-defining challenge. In the surveyed literature, successful systems typically (i) restrict to a small, task-relevant concept vocabulary, (ii) explicitly measure artifact complexity and constraint violations, and (iii) treat mismatches as actionable monitoring signals (rather than hiding them in a single aggregate score). Case studies in Section 7 illustrate these choices across domains.

7 Case Studies and Applications

Symbolic XAI methods have evolved from conceptual proposals to domain-anchored solutions that generate verifiable artifacts of reasoning—rules, logical clauses, KG paths, fuzzy memberships, and prototype traces—usable by domain experts and auditors. The patterns identified in our taxonomy (Section 4) repeatedly support three practical objectives:

(1) making model behavior legible in the units of the domain,
(2) enabling testing and monitoring of properties such as monotonicity, safety, or physics consistency, and
(3) furnishing governance-ready evidence such as rule lists, constraint-satisfaction reports, or concept-level rationales.

7.1 Healthcare

Clinical decision support and biomedical modeling demand explanations that connect neural evidence to established medical semantics or measurement protocols. The corpus contains hybrid pipelines that align perception or graph encodings with structured medical knowledge, as well as SKI formulations that inject semantic or physical

Table 9. Healthcare: corpus examples illustrating symbolic designs, explanation objects, and reported impacts.

Example	Task / Domain	Symbolic pattern and structure	Explanation object	Reported evidence / impact
(J. Wang, P. Ding, et al. 2025)	Drug–target affinity	Hybrid GNN with interpretable multi-scale modules	Feature and path rationales	Interpretable binding factors with competitive accuracy
(K. Wang et al. 2020)	Emergency diagnosis	Hybrid decoding with knowledge-based tree	Rule paths and thresholds	Transparent triage logic consistent with expert criteria
(Ong et al. 2020)	Biomarker discovery	SKE associative rule mining	Top-k class rules	Actionable markers aligned to clinical factors
(Muyama et al. 2024)	Personalized pathways	Hybrid reinforcement learning with clinical structure	Policy traces and state features	Patient-specific decision pathways with legible transitions

Table 10. Finance and economics: symbolic explanations and governance-oriented evidence.

Example	Task / Domain	Symbolic pattern and structure	Explanation object	Reported evidence / impact
(Ghosh et al. 2020)	Credit and risk classification	SKE relaxed logical rules	Compact rule lists	Human-readable reason sets compatible with decision rationale
(J. Xu et al. 2018)	Tasks with domain priors	SKI semantic loss with logical constraints	Constraint satisfaction scores	Enforced logical and monotone properties at training time
(Fischer and Vreeken 2021)	Pattern mining	SKE differentiable pattern set mining	Pattern sets and weights	Concise descriptors suited for audits
(Y. Ma et al. 2024)	Finance signals in crises	SKI or hybrid with model checking	Temporal logic verdicts	Policy-auditable alarms and interpretable conditions

priors into the learning objective. Representative examples (Table 9) include interpretable drug–target affinity modeling with multi-scale graph reasoning, knowledge-based tree decoding for emergency diagnosis, and the discovery of associative rules for biomarker identification, supplemented by work on personalized diagnostic pathways. Together, they illustrate how constraint or rule traces become auditable artifacts aligned to clinical guidelines and device documentation.

Takeaway. Healthcare applications benefit from concept- and rule-level explanations that map directly to clinical artifacts. Hybrid and SKI designs surface constraint-satisfaction reports and rule firings, which are traceable for audits and post-market monitoring.

7.2 Finance and Economics

Financial decisioning emphasizes adverse-action reasons, monotonic behaviors, and policy alignment. The corpus documents SKE rule extraction that produces compact reason lists, SKI training that enforces logical or monotone properties, and pattern-set mining that yields concise descriptors. Works on model checking for finance signals further illustrate how temporal logic verdicts translate into audit-ready conditions during crises. Table 10 summarizes some examples.

Takeaway. Symbolic artifacts directly support obligations such as adverse-action notices and internal model governance. Encoding properties during learning reduces reliance on post-hoc rationalization and yields verifiable evidence for regulators.

Table 11. Natural language processing: rule-, prototype-, and graph-based explanatory signals.

Example	Task / Domain	Symbolic pattern and structure	Explanation object	Reported evidence / impact
(Sen et al. 2020)	Sentence classification	Hybrid neural inductive logic programming	Linguistic rules and clauses	Human-readable expressions triggering decisions
(Z. Chen, J. Chen, et al. 2024)	LLM explanations	SKE with knowledge-augmented supervision	Grounded explanation labels	Reliability gains through grounding
(Christmann et al. 2023)	Conversational QA	Hybrid iterative GNN over heterogeneous sources	Knowledge-graph paths	Source-attributed chains improving interpretability
(Z. Liu, X. Chen, et al. 2024)	Relation extraction	SKI with regular expressions as constraints	Regex-conditioned checks	Auditable structure constraints during training

Table 12. Computer vision: intrinsically interpretable architectures and structured rationales.

Example	Task / Domain	Symbolic pattern and structure	Explanation object	Reported evidence / impact
(Nauta et al. 2021)	Fine-grained recognition	Hybrid prototype tree	Prototype images and decision path	Part-level evidence with explicit paths
(Böhle et al. 2022)	Interpretable conv nets	Hybrid alignment objective	Concept alignments	Intrinsic interpretability at inference
(Padalkar et al. 2024)	Image classification	Hybrid neurosymbolic pipeline	Symbolic factors in forward pass	High fidelity with symbolic traces
(Mao et al. 2023)	Visual recognition rationales	SKE or hybrid rationale generation	Structured why-prompts	More faithful justifications than gradient maps

7.3 Natural Language Processing

In NLP, symbolic approaches lift explanations from token attributions to rule sets, prototype trajectories, and KG paths. The corpus features neural inductive logic programming for linguistic expressions, knowledge-augmented explanation benchmarking for LLMs, and iterative graph-based reasoning for conversational question answering, as well as methods that integrate regular expressions as structural constraints for relation extraction. These designs produce clause-level and path-level evidence that can be adjudicated against linguistic theory and sources. Table 11 synthesizes some examples.

Takeaway. Symbolic NLP methods supply explanations at the level where users reason (e.g., rules, prototypes, paths), supporting testing of linguistic hypotheses and stronger provenance than token saliency alone.

7.4 Computer Vision

Symbolic XAI in vision emphasizes intrinsically interpretable features and programmatic rationales. Prototype trees and alignment-oriented networks create by-design legibility, while hybrids translate visual evidence into structured rationales. Table 12 illustrates examples that include neural prototype trees for fine-grained recognition, alignment-focused convolutional networks, a neurosymbolic framework with explicit symbolic factors, and a why-prompt formulation that yields structured visual reasons.

Takeaway. Prototype- and alignment-based hybrids deliver explanations that are stable to distribution shifts and remain legible without additional post-hoc procedures.

Table 13. Robotics and autonomous systems: verifiable policies, monitors, and motion primitives.

Example	Task / Domain	Symbolic pattern and structure	Explanation object	Reported evidence / impact
(L. Yang et al. 2022)	Safe reinforcement learning	SKI with constrained updates	Constraint feasibility traces	Safer policy updates with explicit constraints
(X. Xu et al. 2023)	Movement primitives	SKE editable primitives	Program-like segments	Debuggable behaviors and human-readable edits
(Pang et al. 2025)	Handover reconstruction	SKI geometry and physics priors	Structured shape and pose reasoning	Robustness with interpretable intermediates
(P. Wang et al. 2024)	Transparent pose estimation	SKI with language and implicit physics	Constraint reports	Generalization on transparent objects with auditable checks

Table 14. Recommender systems: knowledge-graph paths and structured reasoning traces.

Example	Task / Domain	Symbolic pattern and structure	Explanation object	Reported evidence / impact
(C. Chang et al. 2023)	Multi-category recommendation	Hybrid knowledge-graph transformer	Path rationales	Path-grounded explanations alongside accuracy
(S. Wang, H. Li, et al. 2022)	KG-aware recommendation	Hybrid bidirectional propagation	Path segments and rules	Interpretable chains supporting decisions
(X. Zhu et al. 2023)	Robust recommendation	SKI or SKE with knowledge refinement	Refined features and paths	Stability improvements with transparent adjustments
(Meng et al. 2025)	Social and knowledge signals	Hybrid contrastive with KG	Contrastive factors and paths	Robustness with interpretable factors

7.5 Robotics and Autonomous Systems

Safety-critical perception-to-control stacks profit from explicitly representable policies and verifiable constraints. The corpus covers safe policy optimization via constrained updates, editable program-like motion primitives, human-to-robot handover with geometry-aware reasoning, and pose estimation that integrates language and implicit physical constraints. These artifacts—constraint logs, decision nodes, and structured motion segments—facilitate incident analysis and debugging. Table 13 tabulates some of these contributions.

Takeaway. Embedding rules and constraints into control yields legible logs and explanations that are useful for debugging and post-incident analysis, and that align with safety assurance practices.

7.6 Recommender Systems

Recommendation often requires user-facing justifications and internal provenance. KG reasoning and logic-aware propagation provide path-level explanations; contrastive objectives over social and knowledge signals yield interpretable factors. Table 14 highlights KG transformers, bidirectional embedding propagation, knowledge-refined robustness, and contrastive graph structure learning for explainable recommendation.

Takeaway. KG-grounded rationales satisfy user-facing justification requirements and provide internal chains for debugging and monitoring.

7.7 Cybersecurity and Privacy

Security-relevant classifiers must justify alerts while respecting privacy. Table 15 includes argumentation frameworks for misinformation detection, post-hoc itemset explanations that provide sparse reasons, differentially

Table 15. Cybersecurity and privacy: argumentation, itemsets, and privacy-aware explainability.

Example	Task / Domain	Symbolic pattern and structure	Explanation object	Reported evidence / impact
(Chi and Liao 2022)	Fake-news detection	Hybrid quantitative argumentation	Attack and support graphs	Debatable reasoning trails aligned with claims
(Moradi and Samwald 2021)	Post-hoc explainability	SKE confident itemsets	Itemset rules	Clear and sparse reasons for classification
(Mueller et al. 2023)	Privacy-preserving GNN	SKI with differential privacy mechanisms	Privacy and utility reports	Whole-graph DP with competitive accuracy
(Zhuo et al. 2024)	Adversarial robustness	Hybrid symbolic tree	Tree paths and margins	Robustness with interpretable splits

Table 16. Scientific and engineering modelling: symbolic constraints, physics-informed terms, and fuzzy rules.

Example	Task / Domain	Symbolic pattern and structure	Explanation object	Reported evidence / impact
(J. Xu et al. 2018)	Deep learning with priors	SKI semantic loss with logic	Constraint satisfaction and violations	Domain laws encoded during training
(Y. Chen, J. Zhang, et al. 2025)	Manufacturing multi-physics	Hybrid or SKI physics-informed network	Term-level contributions	Accurate multi-field modeling with traceable terms
(Geyer et al. 2024)	Building design	Hybrid systems-engineering with deep learning	Component-level rationale	Explainable trade-offs for energy-efficient design
(Xue et al. 2025)	Trustworthy regression	Hybrid interval type-2 neuro-fuzzy inference	Linguistic rules and memberships	Stable, uncertainty-aware regressors

private graph neural networks for whole-graph classification, and adversarial-resilient symbolic trees. These methods supply forensic-grade artifacts without exposing sensitive data.

Takeaway. Argumentation, itemsets, and privacy-aware training provide actionable forensics without compromising sensitive data. They also create durable artifacts for incident response.

7.8 Scientific and Engineering Modelling

Scientists seek mechanism-level insight rather than opaque predictions. Table 16 documents SKI formulations that encode logical or physics priors, physics-informed multi-field modeling with explicit term-level attributions, interpretable systems-engineering approaches for design, and interval type-2 neuro-fuzzy regressors. These methods output testable scientific objects such as equation terms, satisfaction reports, and linguistic rules.

Takeaway. By converting model behavior into equations, constraints, and interpretable rules, symbolic XAI supports hypothesis refinement, ablation studies, and peer review in science and engineering.

7.9 Industry and Manufacturing

Industrial analytics benefits from explanations that align with operator know-how and process physics. Table 17 reports interpretable feature-based classification in waste management, and physics-informed modeling of press-fitting with explicit term attributions; both produce artifacts that are straightforward to validate on the shop floor.

Table 17. Industry and manufacturing: process-aligned explanations and physical term attributions.

Example	Task / Domain	Symbolic pattern and structure	Explanation object	Reported evidence / impact
(Nezerka et al. 2024)	Waste fragment classification	SKE feature extraction compared with CNNs	Feature-based rules	Comparable accuracy with higher transparency
(Y. Chen, J. Zhang, et al. 2025)	Press-fitting modelling	Hybrid or SKI physics-informed network	Physical term scores	Process insight with verifiable physical contributions

Table 18. Agentic systems: decision trees, natural-language plans, and constraint logs.

Example	Task / Domain	Symbolic pattern and structure	Explanation object	Reported evidence / impact
(Z. Liu, Y. Zhu, et al. 2025)	Multi-agent control	Hybrid soft decision trees	Path-level policy justifications	Interpretable cooperative strategies
(Peng et al. 2022)	Language-mediated RL	Hybrid with program-like chains	Natural-language plan traces	Human-auditable agent intentions
(L. Yang et al. 2022)	Safe policy improvement	SKI with constrained updates	Constraint satisfaction logs	Verifiable safety during learning

Takeaway. Symbolic explanations help operators and engineers validate and refine processes, and they integrate naturally into quality and safety documentation.

7.10 Agentic Systems

Last but not least, agentic pipelines combine neural perception with symbolic control to produce inspectable policies and plans. Table 18 includes multi-agent reinforcement learning with soft decision trees, inherently explainable reinforcement learning in natural language, and safe policy optimization using constrained updates. These methods produce node-level decision paths, plan traces, and constraint logs that are suitable for human inspection and incident reconstruction.

Takeaway. Symbolic traces make agentic systems scrutable at deployment time, elevating them from opaque reward maximizers to accountable components within a larger socio-technical process.

Summary. Across domains, we observe a consistent division of labor among SKE, SKI, and Hybrid designs. SKE is effective when the goal is to furnish governance-friendly reason sets without retraining; SKI is preferable when critical properties must hold by construction; hybrids provide intrinsic traces that reduce the fidelity gap between the model and its explanation. These observations align with the taxonomy and governance analysis presented in the main paper and suggest a practical approach: matching the symbolic artifact to the assurance requirement and reporting satisfaction of the assurance requirement and constraint alongside rule activation and accuracy.

8 Discussion

In this section, we step back to reflect on the broader implications of our review, discuss the challenges that remain for symbolic XAI in deep learning, and outline promising directions for future research. We also connect our findings to regulatory and ethical considerations, reinforcing how symbolic XAI can contribute to the development of trustworthy AI. Moreover, we discuss deployment playbooks for shipping deep learning-based symbolic XAI systems. Finally, we present a reader's guide on how to use this review.

8.1 Benefits of Symbolic XAI in Deep Learning

Our review demonstrates clear benefits of integrating symbolic methods with deep learning.

Improved interpretability. By design, symbolic XAI yields explanations in forms (e.g., rules, logic, concepts) that humans readily understand. This addresses the opacity of deep models, turning some of the invisible reasoning visible. For example, instead of a cryptic vector output, a user sees a rule, which is much easier to grasp.

Knowledge integration. Symbolic XAI allows injection of domain expertise, leading to models that behave more in line with reality (and thus more robust to edge cases). This can reduce training data requirements and prevent nonsensical outputs because a priori knowledge guides the model. For instance, a physics-informed neural network that respects conservation laws (enforced symbolically) will not output energy-violating predictions, improving both trust and plausibility of results.

User trust and adoption. When users (be it a doctor, customer, or engineer) can get an explanation, they are more likely to trust and adopt AI decisions. However, this relationship is nuanced: Poursabzi-Sangdeh et al. (2021) demonstrate that transparency alone does not guarantee better outcomes—in their experiments with 3,800 participants, users shown transparent models were paradoxically *less* able to detect and correct for the model's sizable mistakes, seemingly due to information overload and anchoring effects. This suggests that the psychological aspect of explanation is double-edged: while the mere fact that a system can explain itself fosters trust (Kliegr et al. 2021), such trust may be misplaced if explanations trigger cognitive biases. Indeed, Kliegr et al. (2021) review twenty cognitive biases that can distort the interpretation of rule-based models, from anchoring on specific feature values to confirmation bias in rule selection. Their work emphasizes that interpretability should not be equated with mere syntactic comprehensibility, but must account for the cognitive plausibility of explanations—how reasonable they seem to human judgment. Thus, aligning explanations with users' mental models and providing debiasing interventions (such as outlier-focus messages) is crucial for fostering well-calibrated trust rather than blind trust.

Error analysis and debugging. Symbolic explanations act as a diagnostic tool. By inspecting extracted rules or logic traces, developers can identify why a model made an error. Perhaps a rule is over-generalizing (covering cases it should not)—developers can refine features or add exceptions. Traditional neural nets often require blind parameter tweaking or dataset augmentation to fix issues. In contrast, symbolic insights provide targeted cues (e.g., *"model thinks all users under 25 are risky because training data was biased"*—then one can fix that bias).

Compliance and accountability. We have stressed how symbolic XAI aligns with legal requirements for explainability. In case of adverse events (like algorithmic bias allegations), having symbolic logic to show can demonstrate due diligence or reveal where the flaw was. It also makes it easier to implement governance policies. For example, a bank might have a policy *"do not use ZIP code in lending decisions to avoid redlining"*. However, models may exploit correlated features (e.g., property values, local business types) to implicitly reconstruct ZIP code information (Kallus et al. 2022). This illustrates that symbolic constraints at the input level are insufficient—fairness requires causal reasoning about proxy variables.

8.2 Trade-offs of Symbolic XAI in Deep Learning

However, there are trade-offs.

Performance vs. interpretability. While not universal, some tasks show a trade-off where the most interpretable models (e.g., a small rule set) are slightly less accurate than black-box models. We saw cases where integrated approaches narrowed this gap, but it can persist. The question is whether the small loss in accuracy is acceptable

given the gain in transparency. In high-stakes scenarios, many argue yes, preferring a model that is maybe 1–2% less accurate but much more trustworthy and auditable.

Complexity of explanations. Paradoxically, a very complex symbolic explanation (like a huge rule list or deeply nested logic) can itself become hard to understand. There is a sweet spot of abstraction needed. Being too fine-grained (explaining every neuron's role) is overwhelming, while being too high-level (just a generic statement) is not informative. Hence, human-friendly summarization and interactive drill-down are needed so users can manage complexity. This trade-off is being addressed by research into selective or contrastive explanations, which present only the salient logic relevant to the query or case.

Development effort. It can be more involved to develop a neurosymbolic system than a pure deep learning pipeline. One might need to encode knowledge bases, ensure data aligns with them, and maintain that knowledge as things change. Some industries lack readily available ontologies or experts to craft rules. This trade-off is between initial effort vs. runtime benefit. AutoML tools and emerging libraries aim to reduce this barrier by auto-extracting candidate symbols or by providing templates for common knowledge (e.g., fairness constraints).

Symbolic imperfections. Symbolic knowledge itself can be incomplete or incorrect. If we rely on it too much, we might constrain the model in suboptimal ways. For example, early expert systems sometimes failed because the rules did not cover edge cases that data-driven learning could discover. A neural net might find a valid pattern that is not in our knowledge base, and a strict symbolic overlay could suppress it. The hybrid approach mitigates this by balancing data and knowledge, but finding that balance (e.g., weight of rule regularization) is non-trivial. Essentially, we trade off some flexibility for interpretability, which usually is positive, but we must be cautious not to hold models to possibly outdated or heuristic rules rigidly.

8.3 Challenges and Limitations

Despite progress, several challenges remain for symbolic XAI.

Scalability of symbolic reasoning. While neural nets scale to millions of parameters and huge data, symbolic reasoning (especially logic inference) can become a bottleneck for large knowledge bases or complex rule sets (logic inference is often NP-hard). In integrated systems, one has to ensure the symbolic module can handle the throughput. One workaround is using differentiable approximations (as in Logic Tensor Networks), but that sometimes sacrifices the crisp semantics of symbolic logic. Another solution is hierarchical or modular approaches—break knowledge into smaller chunks contextually invoked. Research is needed in scaling up reasoning or combining symbolic with vectorized reasoning (like neural theorem provers that use embeddings to guide proof search).

Learning symbolic representations. Not all tasks have clear symbolic counterparts. Identifying what concepts or rules are relevant can be hard. Unsupervised discovery of concepts from deep nets is an open problem (e.g., can we automatically label neurons with semantic concepts?) It is a "chicken and egg": we want symbols to explain the net, but we need to know what symbols to use. Techniques such as disentanglement—which aim to separate distinct factors of variation into independent latent variables—have been actively studied since at least the early 2010s (Bengio et al. 2013), with roots in independent component analysis and slow feature analysis. However, reliably extracting symbolic concepts from neural representations remains challenging. Future directions combining disentanglement with concept embeddings and causal representation learning might help neural networks naturally align with symbolic factors.

Dynamic and continuous domains. Some domains involve continuous control or diffuse concepts that do not lend themselves to crisp rules (e.g., what is the precise rule for distinguishing a casual conversation from hate speech? It may require nuanced understanding beyond simple logic). Fuzzy logic can model gradations, but

even so, symbolic logic might struggle with subtleties. Neural nets excel at these fuzzy pattern recognitions, but explaining them in crisp terms is tough. More advanced forms of explanation (like exemplars or analogies, which are somewhat symbolic) might be needed for those cases.

User understanding and cognitive load. Just because something is symbolic does not guarantee a non-expert will understand it. A decision tree with depth 10 is technically symbolic and transparent, but a layperson might still be overwhelmed. This observation aligns with Miller's seminal finding that human working memory capacity is limited to approximately 7 ± 2 chunks of information (Miller 1956). When the number of independent items exceeds this threshold, processing efficiency degrades, and errors increase—a phenomenon that applies directly to complex symbolic explanations. So, human factors research is needed on how to present symbolic explanations effectively (graphical formats? step-by-step narratives?). Miller's concept of *chunking*—organizing information into meaningful, coherent units—offers one path forward: rather than presenting raw rules, explanation systems might group related conditions into higher-level concepts that users can grasp as single chunks. Explanation systems may also need to be interactive: symbolically explain a bit, gauge user understanding, and refine, thereby respecting the severe constraints on human information processing capacity.

Integration with ML workflows. Many existing AI pipelines and frameworks are built for pure numeric optimization. Incorporating symbolic knowledge often feels like "hacking around" the usual training. For widespread adoption, we need better integration—perhaps deep learning libraries that natively support constraints or have an API for knowledge injection. There's movement in this direction (e.g., PyTorch has some constraint libraries, TensorFlow has Neural Structured Learning, etc.), but it is early.

Verification of explanations. Ironically, we also need to ensure the explanations themselves are correct (faithful). There is a risk of explanation techniques that produce plausible but inaccurate explanations (the "rationalization" problem, akin to how an LLM might give a confident-sounding but wrong chain of thought). For trust, we need guarantees that the symbolic explanation aligns with the model's actual logic. Some approaches enforce by-design faithfulness (like intrinsic symbols rather than post-hoc), which is good. Another idea is verifying explanations: given an extracted rule set, one can test it on data or use it to probe the model further to ensure it truly reflects model behavior. This meta verification is an area for development.

8.4 Alignment with Regulations and Societal Expectations

Our review underscores that symbolic XAI is not just a technical endeavor but a response to broader societal needs.

Regulatory alignment. The EU AI Act specifically mentions transparency, and though it does not prescribe technical means, symbolic explanations fit naturally into its mandates. Article 14, for example, requires human oversight—an AI system that can explain its decisions in symbols (rules, reasons) can be more effectively overseen by a human who can understand those reasons. We foresee that compliance checklists might include "Does the system provide understandable logic for its outputs?"—symbolic XAI would check that box. Similarly, sector-specific regulations (like FDA's SaMD guidance) emphasize "clinically meaningful explanations"—again pointing towards symbolic or at least semantically rich outputs. Our survey, by cataloging how such outputs can be produced, supports practitioners aiming for compliance.

Ethical AI and bias mitigation. There is an ethical dimension: if an AI's decision-making is transparent, it is easier to spot and correct biases. For instance, if a lending model's rule *"IF resident_of_zipcode = 12345 THEN deny"* emerged, one can immediately flag that as potentially discriminatory (redlining). In a black box, such a pattern might go unnoticed, manifesting as correlated inputs but not explicitly seen. Symbolic extraction thus serves fairness auditing. Moreover, one can proactively inject fairness constraints symbolically (like adding rules

to ensure equal treatment of sensitive groups). This shows that symbolic XAI is aligned with the goal of AI that respects values and rights.

Public perception and acceptance: Society's trust in AI has been shaken at times (e.g., fatal Tesla Autopilot accidents, biased COMPAS recidivism predictions). A key reason is the "black box" nature leading to fear and mystique around AI decisions. Symbolic XAI by making AI "speak human language" can demystify AI. This is important for public acceptance, especially as AI decisions start affecting everyday life (loans, medical diagnoses, driving, etc.). It gives people a sense of control—if they get an explanation, they can contest it or understand it. Several surveys on public opinion show people are more willing to use or be subjected to AI decisions if an explanation is provided, even if they do not fully agree with the decision. Essentially, explanation is seen as a sign of respect for the user's right to know.

Education and literacy. There is also a beneficial side-effect: symbolic XAI can educate users about the domain or the AI's functioning. If a credit model explains *"your debt ratio is too high relative to income,"* the applicant learns about creditworthiness factors. If a health AI explains *"we found these symptoms suggest this condition,"* patients can better understand their health. Over time, this could raise general literacy about decision factors in those domains (though care must be taken to provide correct advice).

8.5 Deployment Playbooks

Symbolic XAI offers multiple routes from insight to decision; which route to take depends on deployment priorities and constraints observed in practice (cf. Sections 4–7). In particular, Table 7 already ties symbolic artifacts to legal/assurance duties and logging requirements; the playbooks in Table 19 translate that checklist into concrete design choices and monitoring patterns suitable for shipping systems.

Regulatory alignment note. The artifacts and logs in Table 19 directly support transparency (e.g., EU AI Act Article 13) and record-keeping (Article 12) for high-risk systems. Similarly, for medical AI, the FDA's transparency and GMLP principles emphasize audience-appropriate explanation and lifecycle evidence. Moreover, for enterprise risk governance, the NIST AI RMF's Govern–Map–Measure–Manage functions motivate coverage/stability dashboards and incident-ready traces.

8.6 Future Outlook

Given the trajectory observed, we anticipate several future developments.

Standardization of XAI practices. It is likely that industries will develop standard XAI frameworks (perhaps influenced by our conceptual framework and others). For example, a "model factsheet" might become as standard as a nutrition label, containing sections like "Key decision rules of this model" or "Knowledge used by model," which are produced by symbolic analysis. Regulators might enforce such documentation. Tools to automatically generate these from trained models will be in demand.

Hybrid AI systems as norm. The distinction between "symbolic AI" and "neural AI" will blur further. We foresee the emergence of additional platforms and languages for hybrid AI. For instance, future AI IDEs could allow writing parts of the model in Python (neural layers) and parts in a logic DSL, then train them jointly. This seamless integration will lower the barrier to implementing symbolic XAI.

AI self-explanation and dialogue. With the rise of large language models (LLMs), a notable emerging concept is AI capable of engaging in dialogue to explain itself. An LLM, guided by a symbolic reasoning chain, could answer follow-up questions about a decision. That would fully realize an interactive explanation, as in our framework. Ensuring those answers are truthful to the model's process likely requires an approach like ours (rather than

Table 19. Deployment playbooks for matching priorities to designs, artifacts, and controls.

Deployment priority/context	Recommended symbolic pattern	What to implement and log (artifacts & metrics)	Typical gotchas & controls
Auditability with minimal retraining (stable model already in production; need reasons per decision, rapid compliance response)	**SKE** (post-hoc rule/tree extraction; KG path rationales)	• **Artifacts:** decision-linked reason sets; rule/tree paths; KG path traces • **Metrics:** fidelity to base model; coverage; stability across time/windows; rule size/complexity caps • **Ops:** snapshot model/data hashes; extraction timestamps; retention schedule aligned to policy	• **Surrogate drift** (fidelity decay): schedule re-extraction; drift alarms • **Minority over-smoothing:** stratified fidelity/coverage; per-group checks; reason diversity tests • **Scope creep:** clearly label explanation scope and limitations (local/global)
Properties must hold by construction (safety, fairness, physics, and eligibility constraints are non-negotiable)	**SKI** (logic, constraints, and priors in training; constrained decoding/control)	• **Artifacts:** constraint suites; satisfaction/violation reports; proof/trace exemplars • **Metrics:** per-constraint violation rates; training stability (loss curves w/ constraint terms); robustness under perturbations • **Ops:** governance linkage from requirement to test to artifact (traceability)	• **Soft-constraint pitfall:** $\lambda \mathcal{L}_{\text{sym}}$ reduces violations but does not guarantee their absence; "by construction" is false unless enforcement is hard • **Conflicting priors:** conflict analysis; disable or relax per validated policy • **Loss weighting:** ablation to calibrate constraint weights; monitor accuracy–satisfaction trade-offs • **Spec drift:** version constraint sets; change-control gates
Explanations by-design (intrinsic traces needed for ops, incident response, or user-facing justification)	**Hybrids** (differentiable logic layers; neural decision trees; fuzzy/program modules)	• **Artifacts:** clause activations; tree paths; module layouts; natural-language plan traces (for agentic systems) • **Metrics:** trace completeness/coverage; latency budget; explanation stability across runs • **Ops:** default trace logging; selective sampling for deep traces; privacy filters for logs	• **Latency and engineering overhead:** micro-benchmarking; cache/compile traces; route heavy traces to async pipelines • **Tool portability:** containerize symbolic engines; fallback modes when engines unavailable

simply letting the LLM hallucinate an explanation). Therefore, the integration of LLMs (which are neural) with KGs and formal logic is a rapidly evolving area.

Causal and counterfactual explanations. Future XAI will likely emphasize why and what-if (counterfactuals) more, as these are very intuitive forms of explanation for humans. Symbolic models are well-suited to generate counterfactual statements (e.g., *"If X had not been true, the decision would change"*) because one can manipulate rules to see outcome differences. We expect to see more methods that explicitly combine causal inference (often using symbolic graphs) with deep learning. This would not only explain but also help models make decisions that generalize under interventions, improving robustness.

Ethical reasoning in AI. As AI moves into spaces that require ethical judgments (such as autonomous weapons and content moderation), symbolic encodings of ethical principles (e.g., Asimov's laws or rules regarding hate speech) will likely be incorporated to ensure compliance with human values. This is an area where pure learning is risky (we do not want a model to learn its own, possibly flawed, ethical guidelines from data—it is better to encode basic principles and let it learn specifics around them). Therefore, neurosymbolic ethics may become a distinct field, bridging AI with philosophy and law.

8.7 How to Use This Review: A Reader's Guide

This survey can be navigated selectively depending on the specific role.

Practitioners. Start with the taxonomy to understand the design space (Section 4), then use the framework (Section 6) to see where symbolic assets enter training and inference and how explanation interfaces and human feedback are wired. For evidence and trade-offs by macro-category, consult results (Section 5, especially Section 5.1 for composition and trends) and discussion (Sections 8.1–8.3). Finally, jump to deployment playbooks (Section 8.5) for concrete choices under common constraints, and to case studies (Section 7) for domain patterns.

Policy-makers. Begin with regulatory and societal alignment (Section 8.4) to understand how symbolic artifacts relate to oversight, documentation, and contestability. Cross-reference results on alignment with regulatory and governance requirements (Section 5.3) and the framework touchpoints (Section 6) to locate auditable artifacts (e.g., rule sets, constraint reports, clause traces) in the lifecycle. For process assurance, refer to Appendix A (reproducible pipeline, configuration-as-code, and artifacts logged), which aligns with governance expectations, such as those outlined in the EU AI Act (human oversight and transparency) and AI risk frameworks (NIST AI RMF; ISO/IEC 42001).

Researchers. Use background (Section 2) for positioning, then the taxonomy (Section 4) to locate related work and gaps. Consult results (Section 5) for empirical scope and modality/formalism trends; discussion on challenges/limitations (Section 8.3) and future outlook (Section 8.6) summarize open problems (e.g., scaling formal reasoning, faithfulness verification, human-factor evidence). The conclusion (Section 9) consolidates method-selection guidance and calls for benchmarks coupling tasks to machine-readable knowledge bases.

Summary. Overall, our survey suggests that for many critical applications, the trade-offs favor using symbolic XAI. Gains in trust, safety, and compliance outweigh the slight loss in raw performance or added development complexity. A novel and timely contribution of our work is showing that with today's techniques (2017–2025), these trade-offs are smaller than historically assumed—we are approaching a point where we can achieve high accuracy and interpretability (especially as we design better hybrid models and leverage powerful neural nets in tandem with symbolic structure).

Moreover, our systematic review has shown that the foundation for these futures is being laid. Symbolic XAI approaches from 2017 to 2025 are prototypes that will evolve into mature systems of the next decade. Continued interdisciplinary collaboration (ML experts, knowledge engineers, HCI specialists, domain experts) is crucial to address the challenges identified.

In conclusion, the synthesis provided by this paper demonstrates that summarizing and conceptualizing symbolic XAI in deep learning is indeed a novel and timely contribution. It brings together threads from disparate research communities into a coherent narrative and framework, aligned with current needs for explainability, accountability, and trust in AI. We have shown that symbolic XAI is not a step backward to old AI, but rather a step forward to a new kind of AI that is explainable by design —an AI that does not just produce intelligent decisions, but can also engage in an intelligent conversation about how and why those decisions were made. This is ultimately what is required for AI systems to be fully integrated into human society.

9 Conclusion

This review aims to clarify the meaning of "symbolic explainability" in the context of modern deep learning and to organize the rapidly growing body of work into a coherent picture. We proposed a three-part taxonomy—Symbolic Knowledge Extraction (SKE), Symbolic Knowledge Injection (SKI), and Hybrid neurosymbolic architectures—and introduced a conceptual framework that makes explicit the training–inference flows, explanation interfaces, human feedback, and governance touchpoints. In doing so, we provided a unified vocabulary and set of design patterns for relating symbols and neurons that apply across modalities and domains.

Our synthesis of 273 peer-reviewed studies (from 2017 to mid-2025) shows a clear shift from post-hoc explanations toward explainability by-design. While SKE remains valuable, activity after 2020 tilts toward SKI and,

especially, Hybrids (overall composition ≈ 29% SKE, 26% SKI, 45% Hybrids). In parallel, the field has broadened beyond tabular data and vision to encompass NLP, graphs/KGs, RL, and multi-modal pipelines. Across categories, symbolic artifacts—rules, trees, clauses, programs, KG paths, fuzzy rule bases—enable audits, constraint checking, and human-interpretable rationales, but evaluation practices remain heterogeneous and human-subject studies comparatively rare.

A central message of this review is that symbolic artifacts are governance-ready evidence. In high-risk or regulated settings, symbolic rules and clauses, as well as constraint-satisfaction reports, are more straightforward to version, test, and communicate than gradient maps or opaque embeddings; they integrate naturally into documentation, post-market monitoring, adverse-action reasons, and auditor workflows. Building on the corpus and our framework, we distilled actionable practice recommendations:

(1) report faithfulness and constraint-satisfaction metrics alongside accuracy;
(2) specify symbolic assumptions and training-time injections precisely;
(3) include user studies or auditor-centric protocols in high-stakes use; and
(4) develop benchmarks that couple tasks with machine-readable knowledge bases.

We also offered concrete guidance on method selection. Use SKE when the priority is to surface governance-friendly reason sets without touching the production model; prefer SKI when domain properties must hold by construction (monotonicity, physics, or policy constraints); and adopt Hybrid designs when intrinsic traces (clause activations, concept bottlenecks, tree paths, module/program layouts) are needed during inference to support inspection and intervention. Regardless of category, package explanations with explicit scopes (local/global), fidelity/coverage summaries, stability analyses, and, where relevant, user-study evidence of comprehensibility.

At the same time, we identified open problems. Scaling formal reasoning and constraint evaluation to foundation-model regimes, verifying the truthfulness of generated rationales (closing the plausibility–faithfulness gap), measuring causal faithfulness at scale, and standardizing evaluation protocols remain pressing challenges. Methodologically, constraint specification and weighting in SKI can be brittle. Hybrid stacks may couple strongly to particular logics/DSLs, and post-hoc surrogates face fidelity–simplicity trade-offs that call for principled complexity controls and coverage diagnostics. Finally, the evidence base would benefit from more controlled human-factor studies and auditor-centric evaluations.

A research agenda emerges:

(1) scalable, verifiable reasoning layers for large models (including differentiable–discrete hybrids with certifiable traces);
(2) standardized, governance-aligned reporting bundles that pair performance with faithfulness, constraint satisfaction, and stability;
(3) community benchmarks that bind tasks to machine-readable knowledge (ontologies/KGs) and include auditor/user tasks;
(4) tooling that lowers integration costs—APIs for constraints, concept interfaces, and rule extraction within mainstream ML stacks; and
(5) interdisciplinary studies that evaluate how symbolic explanations actually support decisions, contestability, and oversight in real organizations.

In conclusion, symbol-infused deep learning is not a retreat to GOFAI but a pragmatic path toward trustworthy, human-aligned AI. Our taxonomy, framework, and synthesis demonstrate how symbolic representations can render neural systems legible and controllable without compromising modern performance, and how they can translate model behavior into testable, communicable units that align with stakeholder needs. We hope this review provides both a map of the terrain and a practical playbook for researchers and practitioners aiming to build AI systems that not only predict but also explain, justify, and are governed.

Acknowledgments

Ionel Eduard Stan and Paolo Napoletano acknowledge that this work was funded by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them. This research was also partially funded by the INdAM - GNCS Project "Neurosymbolic semantic contract for rigorous reachability and certificates in cyber-physical systems (CPS)" (CUP: E53C25002010001); Ionel Eduard Stan and Guido Sciavicco are INdAM members.

References

F. Aghaeipoor, M. Sabokrou, and A. Fernández. 2023. "Fuzzy Rule-Based Explainer Systems for Deep Neural Networks: From Local Explainability to Global Understanding." *IEEE Trans. Fuzzy Syst.*, 31, 9, 3069–3080. doi:10.1109/TFUZZ.2023.3243935.

L. Ai, J. Langer, S. H. Muggleton, and U. Schmid. 2023. "Explanatory machine learning for sequential human teaching." *Mach. Learn.*, 112, 10, 3591–3632. doi:10.1007/S10994-023-06351-8.

M. M. Alateeq and W. Pedrycz. 2024. "Logic-oriented fuzzy neural networks: A survey." *Expert Syst. Appl.*, 257, 125120. doi:10.1016/J.ESWA.2024.125120.

R. Aly and A. Vlachos. 2024. "TabVer: Tabular Fact Verification with Natural Logic." *Trans. Assoc. Comput. Linguistics*, 12, 1648–1671. doi:10.1162/TACL_A_00722.

J. An, Y. Lai, and Y. Han. 2022. "Logic Rule Guided Attribution with Dynamic Ablation." In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022.* AAAI Press, 77–85. doi:10.1609/AAAI.V36I1.19881.

P. Angelov, D. Kangin, and Z. Zhang. 2025. "IDEAL: Interpretable-by-Design ALgorithms for learning from foundation feature spaces." *Neurocomputing*, 626, 129464. doi:10.1016/J.NEUCOM.2025.129464.

B. Araki, J. Choi, L. Chin, X. Li, and D. Rus. 2022. "Learning Policies by Learning Rules." *IEEE Robotics Autom. Lett.*, 7, 2, 1284–1291. doi:10.1109/LRA.2021.3139380.

T. Arnold, D. Kasenberg, and M. Scheutz. 2021. "Explaining in Time: Meeting Interactive Standards of Explanation for Robotic Systems." *ACM Trans. Hum. Robot Interact.*, 10, 3, 25:1–25:23. doi:10.1145/3457183.

A. B. Arrieta et al.. 2020. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Inf. Fusion*, 58, 82–115. doi:10.1016/J.INFFUS.2019.12.012.

S. Badreddine, A. S. d'Avila Garcez, L. Serafini, and M. Spranger. 2022. "Logic Tensor Networks." *Artif. Intell.*, 303, 103649. doi:10.1016/J.ARTINT.2021.103649.

P. Barbiero, G. Ciravegna, F. Giannini, P. Lió, M. Gori, and S. Melacci. 2022. "Entropy-Based Logic Explanations of Neural Networks." In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022.* AAAI Press, 6046–6054. doi:10.1609/AAAI.V36I6.20551.

P. Barbiero, F. Giannini, G. Ciravegna, M. Diligenti, and G. Marra. 2024. "Relational Concept Bottleneck Models." In: *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024.* Ed. by A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang. http://papers.nips.cc/paper_files/paper/2024/hash/8deff86036ee6a73cda883dc7d2d6b07-Abstract-Conference.html.

J. J. Bauer, T. Eiter, N. H. Ruiz, and J. Oetsch. Feb. 2024. "Visual Graph Question Answering with ASP and LLMs for Language Parsing." In: *Proceedings 40th International Conference on Logic Programming, ICLP 2024* (EPTCS). Ed. by P. Cabalar, F. Fabiano, M. Gebser, G. Gupta, and T. Swift. Vol. 416. (Feb. 2024), 15–28. doi:10.4204/EPTCS.416.2.

F. Beck and J. Fürnkranz. 2021. "An Empirical Investigation Into Deep and Shallow Rule Learning." *Front. Artif. Intell.*, 4, 689398. doi:10.3389/frai.2021.689398.

H. Ben-Younes, R. Cadène, M. Cord, and N. Thome. 2017. "MUTAN: Multimodal Tucker Fusion for Visual Question Answering." In: *IEEE International Conference on Computer Vision, ICCV 2017.* IEEE Computer Society, 2631–2639. doi:10.1109/ICCV.2017.285.

Y. Bengio, A. C. Courville, and P. Vincent. 2013. "Representation Learning: A Review and New Perspectives." *IEEE Trans. Pattern Anal. Mach. Intell.*, 35, 8, 1798–1828. doi:10.1109/TPAMI.2013.50.

M. Bereket and T. Karaletsos. 2023. "Modelling Cellular Perturbations with the Sparse Additive Mechanism Shift Variational Autoencoder." In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023.* Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. http://papers.nips.cc/paper_files/paper/2023/hash/0001ca33ba34ce0351e4612b744b3936-Abstract-Conference.html.

L. Bergamin, M. Polato, and F. Aiolli. 2025. "Improving rule-based classifiers by Bayes point aggregation." *Neurocomputing*, 613, 128699. doi:10.1016/j.neucom.2024.128699.

D. Bertsimas and B. Stellato. 2021. "The voice of optimization." *Mach. Learn.*, 110, 2, 249–277. doi:10.1007/S10994-020-05893-5.

T. Bewley and J. Lawry. 2021. "TripleTree: A Versatile Interpretable Representation of Black Box Agents and their Environments." In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI Press, 11415–11422. doi:10.1609/AAAI.V35I13.17360.

G. Bhatt, D. Das, L. Sigal, and V. N. Balasubramanian. 2023. "Mitigating the Effect of Incidental Correlations on Part-based Learning." In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. http://papers.nips.cc/paper_files/paper/2023/hash/c9493f7cb0d1ec4ae5fc6e0c1a5aca63-Abstract-Conference.html.

Z. Bian, J. Zhang, F. Chung, and S. Wang. 2024. "Residual Sketch Learning for a Feature-Importance-Based and Linguistically Interpretable Ensemble Classifier." *IEEE Trans. Neural Networks Learn. Syst.*, 35, 8, 10461–10474. doi:10.1109/TNNLS.2023.3242049.

M. Biggs, W. Sun, and M. Ettl. 2021. "Model Distillation for Revenue Optimization: Interpretable Personalized Pricing." In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021* (Proceedings of Machine Learning Research). Ed. by M. Meila and T. Zhang. Vol. 139. PMLR, 946–956. http://proceedings.mlr.press/v139/biggs21a.html.

O. Biran and C. Cotton. 2017. "Explanation and justification in machine learning: A survey." In: *IJCAI-17 workshop on explainable AI (XAI)* 1. Vol. 8, 8–13.

M. Böhle, M. Fritz, and B. Schiele. 2022. "B-cos Networks: Alignment is All We Need for Interpretability." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*. IEEE, 10319–10328. doi:10.1109/CVPR52688.2022.01008.

S. Booth, C. Muise, and J. Shah. 2019. "Evaluating the Interpretability of the Knowledge Compilation Map: Communicating Logical Statements Effectively." In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. Ed. by S. Kraus. ijcai.org, 5801–5807. doi:10.24963/IJCAI.2019/804.

J. Cao, T. Zhou, S. Lam, Y. Zhang, J. Zhang, X. Fan, D. Sun, and J. Cai. 2025. "A Lightweight TSK Fuzzy Classifier With Quantitative Equivalent Fuzzy Rules via Adaptive Weighting." *IEEE Trans. Fuzzy Syst.*, 33, 8, 2805–2817. doi:10.1109/TFUZZ.2025.3572937.

Q. Cao, H. Shen, K. Cen, W. Ouyang, and X. Cheng. 2017. "DeepHawkes: Bridging the Gap between Prediction and Understanding of Information Cascades." In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017*. Ed. by E. Lim et al. ACM, 1149–1158. doi:10.1145/3132847.3132973.

İ. İ. Ceylan, T. Lukasiewicz, E. Malizia, C. Molinaro, and A. Vaicenavicius. 2020. "Explanations for Negative Query Answers under Existential Rules." In: *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020*. Ed. by D. Calvanese, E. Erdem, and M. Thielscher, 223–232. doi:10.24963/KR.2020/23.

C. Chang, J. Zhou, Y. Weng, X. Zeng, Z. Wu, C. Wang, and Y. Tang. 2023. "KGTN: Knowledge Graph Transformer Network for explainable multi-category item recommendation." *Knowl. Based Syst.*, 278, 110854. doi:10.1016/J.KNOSYS.2023.110854.

H. Chang, J. Cai, and J. Li. 2023. "Knowledge Graph Completion with Counterfactual Augmentation." In: *Proceedings of the ACM Web Conference 2023, WWW 2023*. Ed. by Y. Ding, J. Tang, J. F. Sequeda, L. Aroyo, C. Castillo, and G. Houben. ACM, 2611–2620. doi:10.1145/3543507.3583401.

C. Chen, H. Geng, N. Yang, X. Yang, and J. Yan. 2024. "EasyDGL: Encode, Train and Interpret for Continuous-Time Dynamic Graph Learning." *IEEE Trans. Pattern Anal. Mach. Intell.*, 46, 12, 10845–10862. doi:10.1109/TPAMI.2024.3443110.

K. Chen and K. D. Forbus. 2021. "Visual Relation Detection using Hybrid Analogical Learning." In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI Press, 801–808. doi:10.1609/AAAI.V35I1.16162.

W. Chen, Y. Hu, M. Peng, and B. Zhu. 2024. "Positional normalization-based mixed-image data augmentation and ensemble self-distillation algorithm." *Expert Syst. Appl.*, 252, 124140. doi:10.1016/J.ESWA.2024.124140.

Y. Chen, J. Zhang, P. Feng, Z. Zheng, X. Zhang, and J. Wang. 2025. "Development of a cascaded multitask physics-informed neural network (CM-PINN) to construct the muti-physical field model of rubber bushing press fitting." *J. Intell. Manuf.*, 36, 5, 3607–3624. doi:10.1007/S10845-024-02427-X.

Y. Chen, A. S. Bijral, and J. L. Ferres. 2019. "On Dynamic Network Models and Application to Causal Impact." In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019*. Ed. by A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis. ACM, 1194–1204. doi:10.1145/3292500.3330990.

Y. Chen, W. Ding, J. Huang, W. Zhang, and T. Zhou. 2025. "Multigranularity Fuzzy Autoencoder for Discriminative Feature Selection in High-Dimensional Data." *IEEE Trans. Neural Netw. Learn. Syst.*, 1–15. doi:10.1109/TNNLS.2025.3569893.

Y. Chen, B. Yuan, B. Liao, and D. M. Gabbay. 2023. "A self-explanatory contrastive logical knowledge learning method for sentiment analysis." *Knowl. Based Syst.*, 278, 110863. doi:10.1016/J.KNOSYS.2023.110863.

Y. Chen, S. Natarajan, and N. Ruozzi. 2022. "Relational Neural Markov Random Fields." In: *International Conference on Artificial Intelligence and Statistics, AISTATS 2022* (Proceedings of Machine Learning Research). Ed. by G. Camps-Valls, F. J. R. Ruiz, and I. Valera. Vol. 151. PMLR, 8260–8269. https://proceedings.mlr.press/v151/chen22f.html.

Z. Chen, J. Chen, A. K. Singh, and M. Sra. 2024. "XplainLLM: A Knowledge-Augmented Dataset for Reliable Grounded Explanations in LLMs." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*. Ed. by Y. Al-Onaizan, M. Bansal, and Y. Chen. Association for Computational Linguistics, 7578–7596. doi:10.18653/V1/2024.EMNLP-MAIN.432.

Z. Chen, C. Chen, Z. Zhang, Z. Zheng, and Q. Zou. 2019. "Variational Graph Embedding and Clustering with Laplacian Eigenmaps." In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*. Ed. by S. Kraus. ijcai.org, 2144–2150. doi:10.24963/IJCAI.2019/297.

C. Chhablani, S. Jain, A. Channesh, I. A. Kash, and S. Medya. 2024. "Game-theoretic Counterfactual Explanation for Graph Neural Networks." In: *Proceedings of the ACM on Web Conference 2024, WWW 2024*. Ed. by T. Chua, C. Ngo, R. Kumar, H. W. Lauw, and R. K. Lee. ACM, 503–514. doi:10.1145/3589334.3645419.

H. Chi and B. Liao. 2022. "A quantitative argumentation-based Automated eXplainable Decision System for fake news detection on social media." *Knowl. Based Syst.*, 242, 108378. doi:10.1016/J.KNOSYS.2022.108378.

S. Choi, H. Park, J. Yeo, and S. Hwang. 2020. "Less is More: Attention Supervision with Counterfactuals for Text Classification." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Association for Computational Linguistics, 6695–6704. doi:10.18653/V1/2020.EMNLP-MAIN.543.

E. Chouzenoux and V. Elvira. 2024. "Sparse Graphical Linear Dynamical Systems." *J. Mach. Learn. Res.*, 25, 223:1–223:53. https://jmlr.org/papers/v25/23-0878.html.

P. Christmann, R. S. Roy, and G. Weikum. 2023. "Explainable Conversational Question Answering over Heterogeneous Sources via Iterative Graph Neural Networks." In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023*. Ed. by H. Chen, W. (Duh, H. Huang, M. P. Kato, J. Mothe, and B. Poblete. ACM, 643–653. doi:10.1145/3539618.3591682.

L. Chu, X. Hu, J. Hu, L. Wang, and J. Pei. 2018. "Exact and Consistent Interpretation for Piecewise Linear Neural Networks: A Closed Form Solution." In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018*. Ed. by Y. Guo and F. Farooq. ACM, 1244–1253. doi:10.1145/3219819.3220063.

G. Ciatto, F. Sabbatini, A. Agiollo, M. Magnini, and A. Omicini. 2024. "Symbolic Knowledge Extraction and Injection with Sub-symbolic Predictors: A Systematic Literature Review." *ACM Comput. Surv.*, 56, 6, 161:1–161:35. doi:10.1145/3645103.

A. Ciocarlan, S. L. Hégarat-Mascle, S. Lefebvre, and A. Woiselle. 2024. "Deep-NFA: A deep a contrario framework for tiny object detection." *Pattern Recognit.*, 150, 110312. doi:10.1016/J.PATCOG.2024.110312.

G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Liò, M. Maggini, and S. Melacci. 2023. "Logic Explained Networks." *Artif. Intell.*, 314, 103822. doi:10.1016/J.ARTINT.2022.103822.

W. W. Cohen. 1995. "Fast Effective Rule Induction." In: *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*. Ed. by A. Prieditis and S. Russell. Morgan Kaufmann, 115–123. doi:10.1016/B978-1-55860-377-6.50023-2.

R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold. 2021. "A historical perspective of explainable Artificial Intelligence." *WIREs Data Mining Knowl. Discov.*, 11, 1. doi:10.1002/WIDM.1391.

R. Confalonieri, T. Weyde, T. R. Besold, and F. M. del Prado Martín. 2021. "Using ontologies to enhance human understandability of global post-hoc explanations of black-box models." *Artif. Intell.*, 296, 103471. doi:10.1016/J.ARTINT.2021.103471.

K. Cortiñas-Lorenzo and G. Lacey. 2024. "Toward Explainable Affective Computing: A Review." *IEEE Trans. Neural Networks Learn. Syst.*, 35, 10, 13101–13121. doi:10.1109/TNNLS.2023.3270027.

O. Csiszár, G. Csiszár, and J. Dombi. 2020. "How to implement MCDM tools and continuous logic into neural computation?: Towards better interpretability of neural networks." *Knowl. Based Syst.*, 210, 106530. doi:10.1016/J.KNOSYS.2020.106530.

O. Csiszár, L. S. Pusztaházi, L. Dénes-Fazakas, M. S. Gashler, V. Kreinovich, and G. Csiszár. 2023. "Uninorm-like parametric activation functions for human-understandable neural models." *Knowl. Based Syst.*, 260, 110095. doi:10.1016/J.KNOSYS.2022.110095.

G. Cui and H. Zhu. 2021. "Differentiable Synthesis of Program Architectures." In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*. Ed. by M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, 11123–11135. https://proceedings.neurips.cc/paper/2021/hash/5c5a93a042235058b1ef7b0ac1e11b67-Abstract.html.

L. Cui, H. Seo, M. Tabar, F. Ma, S. Wang, and D. Lee. 2020. "DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation." In: *The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2020*. Ed. by R. Gupta, Y. Liu, J. Tang, and B. A. Prakash. ACM, 492–502. doi:10.1145/3394486.3403092.

Q. Cui, R. Meng, C. Xu, and C. Chang. 2024. "Steganographic Passport: An Owner and User Verifiable Credential for Deep Model IP Protection Without Retraining." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*. IEEE, 12302–12311. doi:10.1109/CVPR52733.2024.01169.

D. Cunnington, M. Law, J. Lobo, and A. Russo. 2023. "FFNSL: Feed-Forward Neural-Symbolic Learner." *Mach. Learn.*, 112, 2, 515–569. doi:10.1007/S10994-022-06278-6.

K. Das, B. Samanta, P. Goyal, S. Lee, S. Bhattacharjee, and N. Ganguly. 2023. "CrysGNN: Distilling Pre-trained Knowledge to Enhance Property Prediction for Crystalline Materials." In: *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*. Ed. by B. Williams, Y. Chen, and J. Neville. AAAI Press, 7323–7331. doi:10.1609/AAAI.V37I6.25892.

S. Das, A. Singh, S. Chatterjee, S. Bhattacharya, and S. Bhattacharya. 2021. "Finding High-Value Training Data Subset Through Differentiable Convex Programming." In: *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD*

2021 (Lecture Notes in Computer Science). Ed. by N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, and J. A. Lozano. Vol. 12976. Springer, 666–681. doi:10.1007/978-3-030-86520-7_41.

R. Dazeley, P. Vamplew, and F. Cruz. 2023. "Explainable reinforcement learning for broad-XAI: a conceptual framework and survey." *Neural Comput. Appl.*, 35, 23, 16893–16916. doi:10.1007/S00521-023-08423-1.

D. Debot, P. Barbiero, F. Giannini, G. Ciravegna, M. Diligenti, and G. Marra. 2024. "Interpretable Concept-Based Memory Reasoning." In: *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*. Ed. by A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang. http://papers.nips.cc/paper_files/paper/2024/hash/223df9b793a8b384bf38b963e35e2cac-Abstract-Conference.html.

M. Defresne, S. Barbe, and T. Schiex. 2023. "Scalable Coupling of Deep Learning with Logical Reasoning." In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023*. ijcai.org, 3615–3623. doi:10.24963/IJCAI.2023/402.

L. N. Delong, R. F. Mir, and J. D. Fleuriot. 2025. "Neurosymbolic AI for Reasoning Over Knowledge Graphs: A Survey." *IEEE Trans. Neural Networks Learn. Syst.*, 36, 5, 7822–7842. doi:10.1109/TNNLS.2024.3420218.

X. Deng, C. Zhang, L. Jiang, J. Xia, and M. Xu. 2025. "DeepSN-Net: Deep Semi-Smooth Newton Driven Network for Blind Image Restoration." *IEEE Trans. Pattern Anal. Mach. Intell.*, 47, 4, 2632–2646. doi:10.1109/TPAMI.2024.3525089.

F. Doshi-Velez and B. Kim. 2017. *Towards A Rigorous Science of Interpretable Machine Learning*. (2017). https://arxiv.org/abs/1702.08608 arXiv: 1702.08608 [stat.ML].

J. Dressel and H. Farid. 2018. "The accuracy, fairness, and limits of predicting recidivism." *Sci. Adv.*, 4, 1, eaao5580. doi:10.1126/sciadv.aao5580.

M. M. Ebadzadeh and A. Salimi-Badr. 2018. "IC-FNN: A Novel Fuzzy Neural Network With Interpretable, Intuitive, and Correlated-Contours Fuzzy Rules for Function Approximation." *IEEE Trans. Fuzzy Syst.*, 26, 3, 1288–1302. doi:10.1109/TFUZZ.2017.2718497.

G. G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, and S. Lee. 2021. "Improving performance of deep learning models with axiomatic attribution priors and expected gradients." *Nat. Mach. Intell.*, 3, 7, 620–631. doi:10.1038/S42256-021-00343-W.

R. Evans and E. Grefenstette. 2018. "Learning Explanatory Rules from Noisy Data." *J. Artif. Intell. Res.*, 61, 1–64. doi:10.1613/JAIR.5714.

Y. Feng, A. Carballo, and K. Takeda. 2025. "Sparse Prototype Network for Explainable Pedestrian Behavior Prediction." *IEEE Robotics Autom. Lett.*, 10, 5, 4196–4203. doi:10.1109/LRA.2025.3550728.

J. Fischer and J. Vreeken. 2021. "Differentiable Pattern Set Mining." In: *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2021*. Ed. by F. Zhu, B. C. Ooi, and C. Miao. ACM, 383–392. doi:10.1145/3447548.3467348.

C. Florensa, Y. Duan, and P. Abbeel. 2017. "Stochastic Neural Networks for Hierarchical Reinforcement Learning." In: *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net. https://openreview.net/forum?id=B1oK8aoxe.

Z. Fu, Y. Xian, Y. Zhu, S. Xu, Z. Li, G. de Melo, and Y. Zhang. 2021. "HOOPS: Human-in-the-Loop Graph Reasoning for Conversational Recommendation." In: *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021*. Ed. by F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, and T. Sakai. ACM, 2415–2421. doi:10.1145/3404835.3463247.

L. Gan, K. Kuang, Y. Yang, and F. Wu. 2021. "Judgment Prediction via Injecting Legal Knowledge into Neural Networks." In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI Press, 12866–12874. doi:10.1609/AAAI.V35I14.17522.

Z. Gao, S. Cheng, E. Wittrup, J. Gryak, and K. Najarian. 2024. "Learning using privileged information with logistic regression on acute respiratory distress syndrome detection." *Artif. Intell. Medicine*, 156, 102947. doi:10.1016/J.ARTMED.2024.102947.

A. d'Avila Garcez and L. C. Lamb. 2023. "Neurosymbolic AI: the 3rd wave." *Artif. Intell. Rev.*, 56, 11, 12387–12406. doi:10.1007/S10462-023-10448-W.

A. S. d'Avila Garcez, K. Broda, and D. M. Gabbay. 2001. "Symbolic knowledge extraction from trained neural networks: A sound approach." *Artif. Intell.*, 125, 1-2, 155–207. doi:10.1016/S0004-3702(00)00077-1.

P. Geyer, M. M. Singh, and X. Chen. 2024. "Explainable AI for engineering design: A unified approach of systems engineering and component-based deep learning demonstrated by energy-efficient building design." *Adv. Eng. Informatics*, 62, 102843. doi:10.1016/J.AEI.2024.102843.

B. Ghosh, D. Malioutov, and K. S. Meel. 2020. "Classification Rules in Relaxed Logical Form." In: *24th European Conference on Artificial Intelligence, ECAI 2020* (Frontiers in Artificial Intelligence and Applications). Ed. by G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, and J. Lang. Vol. 325. IOS Press, 2489–2496. doi:10.3233/FAIA200382.

A. González-Briones, J. Palomino-Sánchez, Z. Vale, C. Ramos, and J. M. Corchado. 2025. "Evolution of Building Energy Management Systems for greater sustainability through explainable artificial intelligence models." *Eng. Appl. Artif. Intell.*, 147, 110324. doi:10.1016/J.ENGAPPAI.2025.110324.

I. J. Goodfellow, Y. Bengio, and A. C. Courville. 2016. *Deep Learning*. Adaptive computation and machine learning. MIT Press. ISBN: 978-0-262-03561-3. http://www.deeplearningbook.org/.

O. Gottesman, J. Futoma, Y. Liu, S. Parbhoo, L. A. Celi, E. Brunskill, and F. Doshi-Velez. 2020. "Interpretable Off-Policy Evaluation in Reinforcement Learning by Highlighting Influential Transitions." In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020* (Proceedings of Machine Learning Research). Vol. 119. PMLR, 3658–3667. http://proceedings.mlr.press/v119/gottesman20a.html.

A. Gruber, K. Lee, and N. Trask. 2023. "Reversible and irreversible bracket-based dynamics for deep graph neural networks." In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*. Ed. by A. Oh,

T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. http://papers.nips.cc/paper_files/paper/2023/hash/7903af0a1cffb43dbb2f8160d110a5f3-Abstract-Conference.html.

X. Gu. 2022. "An explainable semi-supervised self-organizing fuzzy inference system for streaming data classification." *Inf. Sci.*, 583, 364–385. doi:10.1016/J.INS.2021.11.047.

X. Gu, G. D. Howells, and H. Yuan. 2024. "A soft prototype-based autonomous fuzzy inference system for network intrusion detection." *Inf. Sci.*, 677, 120964. doi:10.1016/J.INS.2024.120964.

Y. Gu, K. Zhang, Q. Liu, W. Gao, L. Li, and J. Zhou. 2024. "π-Light: Programmatic Interpretable Reinforcement Learning for Resource-Limited Traffic Signal Control." In: *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*. Ed. by M. J. Wooldridge, J. G. Dy, and S. Natarajan. AAAI Press, 21107–21115. doi:10.1609/AAAI.V38I19.30103.

Q. Guan, X. Cheng, F. Xiao, Z. Li, C. He, L. Fang, G. Chen, Z. Gong, and W. Luo. 2025. "Explainable exercise recommendation with knowledge graph." *Neural Networks*, 183, 106954. doi:10.1016/J.NEUNET.2024.106954.

E. Gujral, R. Pasricha, and E. E. Papalexakis. 2020. "Beyond Rank-1: Discovering Rich Community Structure in Multi-Aspect Graphs." In: *WWW '20: The Web Conference 2020*. Ed. by Y. Huang, I. King, T. Liu, and M. van Steen. ACM / IW3C2, 452–462. doi:10.1145/3366423.3380129.

D. Gunning and D. W. Aha. 2019. "DARPA's Explainable Artificial Intelligence (XAI) Program." *AI Mag.*, 40, 2, 44–58. doi:10.1609/AIMAG.V40I2.2850.

D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Yang. 2019. "XAI - Explainable artificial intelligence." *Sci. Robotics*, 4, 37. doi:10.1126/SCIROBOTICS.AAY7120.

A. Gupta, S. Saini, and P. J. Narayanan. 2023. "Concept Distillation: Leveraging Human-Centered Explanations for Model Improvement." In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. http://papers.nips.cc/paper_files/paper/2023/hash/c9450295fd667740a39a68148fc17f6e-Abstract-Conference.html.

C. He, H. Shi, X. Liu, and J. Li. 2024. "Interpretable physics-informed domain adaptation paradigm for cross-machine transfer diagnosis." *Knowl. Based Syst.*, 288, 111499. doi:10.1016/J.KNOSYS.2024.111499.

D. Hong, T. Wang, and S. Baek. 2023. "ProtoryNet - Interpretable Text Classification Via Prototype Trajectories." *J. Mach. Learn. Res.*, 24, 264:1–264:39. http://jmlr.org/papers/v24/21-0899.html.

G. Honke, I. Higgins, N. Thigpen, V. Miskovic, K. Link, S. Duan, P. Gupta, J. Klawohn, and G. Hajcak. 2021. "Representation learning for improved interpretability and classification accuracy of clinical factors from EEG." In: *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net. https://openreview.net/forum?id=TVjLza1t4hI.

F. Hu, Z. Deng, Z. Xie, T. Zhang, K. Choi, F. Zhang, and S. Wang. 2024. "GFS-Node: Graph Fuzzy Systems for Node Prediction." *IEEE Trans. Fuzzy Syst.*, 32, 12, 6822–6834. doi:10.1109/TFUZZ.2024.3465557.

L. Hu, S. Lai, W. Chen, H. Xiao, H. Lin, L. Yu, J. Zhang, and D. Wang. 2024. "Towards Multi-dimensional Explanation Alignment for Medical Classification." In: *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*. Ed. by A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang. http://papers.nips.cc/paper_files/paper/2024/hash/ea370419760b421ce12e3082eb2ae1a8-Abstract-Conference.html.

X. Hu, X. Kong, and K. Tu. 2023. "A Multi-Grained Self-Interpretable Symbolic-Neural Model For Single/Multi-Labeled Text Classification." In: *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net. https://openreview.net/forum?id=MLJ5TF5FtXH.

Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. 2017. "Toward Controlled Generation of Text." In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017* (Proceedings of Machine Learning Research). Ed. by D. Precup and Y. W. Teh. Vol. 70. PMLR, 1587–1596. http://proceedings.mlr.press/v70/hu17e.html.

Y. Hua, Y. Todo, S. Tao, Z. Tang, T. Cheng, and Z. Qiu. 2024. "Bio-inspired computational model for direction and speed detection." *Knowl. Based Syst.*, 300, 112195. doi:10.1016/J.KNOSYS.2024.112195.

Z. Huang, J. Zhao, and Q. Jin. 2024. "ECR-Chain: Advancing Generative Language Models to Better Emotion-Cause Reasoners through Reasoning Chains." In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*. ijcai.org, 6288–6296. https://www.ijcai.org/proceedings/2024/695.

Z. Huang, F. Chen, Y. Pu, C. Lin, H. Su, and C. Gan. 2023. "DiffVL: Scaling Up Soft Body Manipulation using Vision-Language Driven Differentiable Physics." In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. http://papers.nips.cc/paper_files/paper/2023/hash/5f5f7b6080dcadced61cf5d96f7c6dde-Abstract-Conference.html.

A. Hubin, G. Storvik, and F. Frommlet. 2021. "Flexible Bayesian Nonlinear Model Configuration." *J. Artif. Intell. Res.*, 72, 901–942. doi:10.1613/JAIR.1.13047.

A. Hunter and N. Potyka. 2023. "Syntactic reasoning with conditional probabilities in deductive argumentation." *Artif. Intell.*, 321, 103934. doi:10.1016/J.ARTINT.2023.103934.

S. Ibrahim, G. Afriat, K. Behdin, and R. Mazumder. 2023. "GRAND-SLAMIN' Interpretable Additive Modeling with Structural Constraints." In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. http://papers.nips.cc/paper_files/paper/2023/hash/c057cb81b8d3c67093427bf1c16a4e9f-Abstract-Conference.html.

A. Ignatiev, J. Marques-Silva, N. Narodytska, and P. J. Stuckey. 2021. "Reasoning-Based Learning of Interpretable ML Models." In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021*. Ed. by Z. Zhou, 4458–4465. doi:10.24963/IJCAI.2021/608.

A. R. Iyer, D. K. Prasad, and H. C. Quek. 2018. "PIE-RSPOP: A brain-inspired pseudo-incremental ensemble rough set pseudo-outer product fuzzy neural network." *Expert Syst. Appl.*, 95, 172–189. doi:10.1016/J.ESWA.2017.11.027.

R. Jain, M. Hemani, D. Ceylan, K. K. Singh, J. Lu, M. Sarkar, and B. Krishnamurthy. 2023. "UMFuse: Unified Multi View Fusion for Human Editing applications." In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023*. IEEE, 7148–7157. doi:10.1109/ICCV51070.2023.00660.

M. Jaques, M. Burke, and T. M. Hospedales. 2020. "Physics-as-Inverse-Graphics: Unsupervised Physical Parameter Estimation from Video." In: *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net. https://openreview.net/forum?id=BJeKwTNFvB.

Y. Jarraya, S. Bouaziz, H. Hagras, and A. M. Alimi. 2019. "A Multi-Agent Architecture for the Design of Hierarchical Interval Type-2 Beta Fuzzy System." *IEEE Trans. Fuzzy Syst.*, 27, 6, 1174–1188. doi:10.1109/TFUZZ.2018.2871800.

Y. H. Jeong, S. Hwang, and D. Chae. 2024. "HiLite: Hierarchical Level-implemented Architecture Attaining Part-Whole Interpretability." In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024*. Ed. by E. Serra and F. Spezzano. ACM, 983–993. doi:10.1145/3627673.3679538.

F. Jiang, X. Hou, and M. Xia. 2025. "Spatio-temporal attention-based hidden physics-informed neural network for remaining useful life prediction." *Adv. Eng. Informatics*, 63, 102958. doi:10.1016/J.AEI.2024.102958.

T. Jiang, F. Sohrabi, and W. Yu. 2023. "Active Sensing for Two-Sided Beam Alignment and Reflection Design Using Ping-Pong Pilots." *IEEE J. Sel. Areas Inf. Theory*, 4, 24–39. doi:10.1109/JSAIT.2023.3276296.

Y. Jiang and M. Bansal. 2019. "Self-Assembling Modular Networks for Interpretable Multi-Hop Reasoning." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP 2019*. Ed. by K. Inui, J. Jiang, V. Ng, and X. Wan. Association for Computational Linguistics, 4473–4483. doi:10.18653/V1/D19-1455.

Y. Jiao et al.. 2024. "NAISR: A 3D Neural Additive Model for Interpretable Shape Representation." In: *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net. https://openreview.net/forum?id=wg8NPfeMF9.

M. Jin, Z. Ma, K. Jin, H. H. Zhuo, C. Chen, and C. Yu. 2022. "Creativity of AI: Automatic Symbolic Option Discovery for Facilitating Deep Reinforcement Learning." In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*. AAAI Press, 7042–7050. doi:10.1609/AAAI.V36I6.20663.

J. R. B. Junior. 2020. "Graph embedded rules for explainable predictions in data streams." *Neural Networks*, 129, 174–192. doi:10.1016/J.NEUNET.2020.05.035.

N. Kallus, X. Mao, and A. Zhou. 2022. "Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination." *Manag. Sci.*, 68, 3, 1959–1981. doi:10.1287/MNSC.2020.3850.

A. Kalyanpur, T. Breloff, and D. A. Ferrucci. 2022. "Braid: Weaving Symbolic and Neural Knowledge into Coherent Logical Explanations." In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*. AAAI Press, 10867–10874. doi:10.1609/AAAI.V36I10.21333.

S. Kambhampati, S. Sreedharan, M. Verma, Y. Zha, and L. Guan. 2022. "Symbols as a Lingua Franca for Bridging Human-AI Chasm for Explainable and Advisable AI Systems." In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*. AAAI Press, 12262–12267. doi:10.1609/AAAI.V36I11.21488.

T. Kang, K. Zarringhalam, M. L. Kuijjer, P. Chen, J. Quackenbush, and W. Ding. 2018. "Clustering on Sparse Data in Non-overlapping Feature Space with Applications to Cancer Subtyping." In: *IEEE International Conference on Data Mining, ICDM 2018*. IEEE Computer Society, 1079–1084. doi:10.1109/ICDM.2018.00138.

N. K. Kasabov, Y. Tan, M. Doborjeh, E. Tu, J. Yang, W. W. B. Goh, and J. Lee. 2023. "Transfer Learning of Fuzzy Spatio-Temporal Rules in a Brain-Inspired Spiking Neural Network Architecture: A Case Study on Spatio-Temporal Brain Data." *IEEE Trans. Fuzzy Syst.*, 31, 12, 4542–4552. doi:10.1109/TFUZZ.2023.3292802.

L. Katzir, G. Elidan, and R. El-Yaniv. 2021. "Net-DNF: Effective Deep Modeling of Tabular Data." In: *Proceedings of the 9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net. https://openreview.net/forum?id=73WTGs96kho.

H. A. Kautz. 2022. "The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture." *AI Mag.*, 43, 1, 93–104. doi:10.1609/AIMAG.V43I1.19122.

P. Khosravi, Y. Choi, Y. Liang, A. Vergari, and G. V. den Broeck. 2019. "On Tractable Computation of Expected Predictions." In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*. Ed. by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, 11167–11178. https://proceedings.neurips.cc/paper/2019/hash/fccc64972a9468a11f125cadb090e89e-Abstract.html.

S. W. Kim, Y. Zhou, J. Philion, A. Torralba, and S. Fidler. 2020. "Learning to Simulate Dynamic Environments With GameGAN." In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. Computer Vision Foundation / IEEE, 1228–1237. doi:10.1109/CVPR42600.2020.00131.

W. Kim and Y. Lee. 2019. "Learning Dynamics of Attention: Human Prior for Interpretable Machine Reasoning." In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*. Ed. by H. M. Wallach,

H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, 6019–6030. https://proceedings.neurips.cc/paper/2019/hash/ae3539867aaeec609a4260c6feb725f4-Abstract.html.

T. Kliegr, S. Bahník, and J. Fürnkranz. 2021. "A review of possible effects of cognitive biases on interpretation of rule-based machine learning models." *Artif. Intell.*, 295, 103458. doi:10.1016/J.ARTINT.2021.103458.

P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. 2020. "Concept Bottleneck Models." In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020* (Proceedings of Machine Learning Research). Vol. 119. PMLR, 5338–5348. http://proceedings.mlr.press/v119/koh20a.html.

M. Kozielski, M. Sikora, and L. Wawrowski. 2025. "Towards consistency of rule-based explainer and black box model - Fusion of rule induction and XAI-based feature importance." *Knowl. Based Syst.*, 311, 113092. doi:10.1016/J.KNOSYS.2025.113092.

R. Kridalukmana, H. Lu, and M. Naderpour. 2022. "Self-Explaining Abilities of an Intelligent Agent for Transparency in a Collaborative Driving Context." *IEEE Trans. Hum. Mach. Syst.*, 52, 6, 1155–1165. doi:10.1109/THMS.2022.3202900.

L. C. Lamb, A. S. d'Avila Garcez, M. Gori, M. O. R. Prates, P. H. C. Avelar, and M. Y. Vardi. 2020. "Graph Neural Networks Meet Neural-Symbolic Computing: A Survey and Perspective." In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. Ed. by C. Bessiere, 4877–4884. doi:10.24963/IJCAI.2020/679.

N. G. Lan, M. Geyer, E. Chemla, and R. Katzir. 2022. "Minimum Description Length Recurrent Neural Networks." *Trans. Assoc. Comput. Linguistics*, 10, 785–799. doi:10.1162/TACL_A_00489.

Q. Lang, X. Liu, and W. Jia. 2023. "AFS Graph: Multidimensional Axiomatic Fuzzy Set Knowledge Graph for Open-Domain Question Answering." *IEEE Trans. Neural Networks Learn. Syst.*, 34, 12, 10904–10918. doi:10.1109/TNNLS.2022.3171677.

T. Laurent, J. von Brecht, and X. Bresson. 2024. "Feature Collapse." In: *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net. https://openreview.net/forum?id=gctmyMiPHH.

C. Lawless, J. Kalagnanam, L. M. Nguyen, D. T. Phan, and C. Reddy. 2022. "Interpretable Clustering via Multi-Polytope Machines." In: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*. AAAI Press, 7309–7316. doi:10.1609/AAAI.V36I7.20693.

T. Le, H. Tran, and B. Le. 2023. "Knowledge graph embedding with the special orthogonal group in quaternion space for link prediction." *Knowl. Based Syst.*, 266, 110400. doi:10.1016/J.KNOSYS.2023.110400.

Y. LeCun, Y. Bengio, and G. E. Hinton. 2015. "Deep learning." *Nat.*, 521, 7553, 436–444. doi:10.1038/NATURE14539.

B. Lee, J. Ahn, and C. Park. 2023. "Minimax Risks and Optimal Procedures for Estimation under Functional Local Differential Privacy." In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. http://papers.nips.cc/paper_files/paper/2023/hash/b4dde7f1bc45bf9c0fda8db8f272b758-Abstract-Conference.html.

M. Lee, J. Jeon, and H. Lee. 2022. "Explainable AI for domain experts: a post Hoc analysis of deep learning for defect classification of TFT-LCD panels." *J. Intell. Manuf.*, 33, 6, 1747–1759. doi:10.1007/S10845-021-01758-3.

S. Lee and B. C. Song. 2021. "Interpretable Embedding Procedure Knowledge Transfer via Stacked Principal Component Analysis and Graph Neural Network." In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI Press, 8297–8305. doi:10.1609/AAAI.V35I9.17009.

C. Li, B. Zhang, D. Hong, X. Jia, A. Plaza, and J. Chanussot. 2025. "Learning Disentangled Priors for Hyperspectral Anomaly Detection: A Coupling Model-Driven and Data-Driven Paradigm." *IEEE Trans. Neural Networks Learn. Syst.*, 36, 4, 6883–6896. doi:10.1109/TNNLS.2024.3401589.

C. Li, L. Ledo, M. R. Delgado, M. Cerrada, F. Pacheco, D. Cabrera, R. Sánchez, and J. V. de Oliveira. 2017. "A Bayesian approach to consequent parameter estimation in probabilistic fuzzy systems and its application to bearing fault classification." *Knowl. Based Syst.*, 129, 39–60. doi:10.1016/J.KNOSYS.2017.05.007.

H. Li, C. Shen, P. Torr, V. Tresp, and J. Gu. 2024. "Self-Discovering Interpretable Diffusion Latent Directions for Responsible Text-to-Image Generation." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 12006–12016. doi:10.1109/CVPR52733.2024.01141.

H. Li, J. Song, M. Xue, H. Zhang, and M. Song. 2025. "A Survey of Neural Trees: Co-Evolving Neural Networks and Decision Trees." *IEEE Trans. Neural Networks Learn. Syst.*, 36, 7, 11718–11737. doi:10.1109/TNNLS.2024.3446891.

J. Li, S. Song, S. Yan, G. Hu, C. Lai, and Y. Zhou. 2025. "Advanced dialog state tracking with noetic graphs for complex human-machine interactions." *Pattern Recognit.*, 168, 111842. doi:10.1016/J.PATCOG.2025.111842.

L. Li, Y. Zhang, and L. Chen. 2020. "Generate Neural Template Explanations for Recommendation." In: *The 29th ACM International Conference on Information and Knowledge Management, CIKM 2020*. Ed. by M. d'Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudré-Mauroux. ACM, 755–764. doi:10.1145/3340531.3411992.

M. Li et al.. 2023. "Multi-modal Action Chain Abductive Reasoning." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*. Ed. by A. Rogers, J. L. Boyd-Graber, and N. Okazaki. Association for Computational Linguistics, 4617–4628. doi:10.18653/V1/2023.ACL-LONG.254.

S. Li, F. Dong, R. Li, and W. Li. 2025. "A dual medical ontology and relational graph framework with neural ordinary differential equations for diagnostic prediction." *Neurocomputing*, 647, 130519. doi:10.1016/J.NEUCOM.2025.130519.

Y. Li, M. Li, J. Ding, and Y. Bai. 2025. "Two-layer knowledge graph transformer network-based question and answer explainable recommendation." *Eng. Appl. Artif. Intell.*, 149, 110542. doi:10.1016/J.ENGAPPAI.2025.110542.

Z. Li and X. Si. 2022. "NSNet: A General Neural Probabilistic Framework for Satisfiability Problems." In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. http://papers.nips.cc/paper_files/paper/2022/hash/a40462acc6959034c6aa6dfb8e696415-Abstract-Conference.html.

T. de Lima and E. Lorini. 2024. "Model Checking Causality." In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*. ijcai.org, 3324–3332. https://www.ijcai.org/proceedings/2024/368.

P. Lin, P. Sun, G. Cheng, S. Xie, X. Li, and J. Shi. 2020. "Graph-Guided Architecture Search for Real-Time Semantic Segmentation." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. Computer Vision Foundation / IEEE, 4202–4211. doi:10.1109/CVPR42600.2020.00426.

S. Lin, X. Zhang, C. Chen, X. Chen, and W. Susilo. 2023. "ERM-KTP: Knowledge-Level Machine Unlearning via Knowledge Transfer." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*. IEEE, 20147–20155. doi:10.1109/CVPR52729.2023.01929.

T. Lin et al.. 2025. "Preserving high-order ego-centric topological patterns in node representation in heterogeneous graph." *Knowl. Based Syst.*, 311, 113067. doi:10.1016/J.KNOSYS.2025.113067.

H. Liu, R. Wang, S. Shan, and X. Chen. 2021. "What is a Tabby? Interpretable Model Decisions by Learning Attribute-Based Classification Criteria." *IEEE Trans. Pattern Anal. Mach. Intell.*, 43, 5, 1791–1807. doi:10.1109/TPAMI.2019.2954501.

S. Liu, S. Oh, W. Pedrycz, B. Yang, L. Wang, and Z. Peng. 2024. "SCINN: Semantic Concept-Based Inference Neural Networks With Explainable and Deep Fuzzy Structure." *IEEE Trans. Fuzzy Syst.*, 32, 7, 4133–4147. doi:10.1109/TFUZZ.2024.3398719.

Y. Liu, X. Lu, H. Wang, J. Yi, and C. Li. 2025. "Convolutional fuzzy modules stacked deep residual system with application to classification problems." *Expert Syst. Appl.*, 288, 128282. doi:10.1016/J.ESWA.2025.128282.

Z. Liu, X. Chen, H. Wang, and X. Liu. 2024. "Integrating regular expressions into neural networks for relation extraction." *Expert Syst. Appl.*, 252, 124252. doi:10.1016/J.ESWA.2024.124252.

Z. Liu, Y. Zhu, Z. Wang, Y. Gao, and C. Chen. 2025. "MIXRTs: Toward Interpretable Multi-Agent Reinforcement Learning via Mixing Recurrent Soft Decision Trees." *IEEE Trans. Pattern Anal. Mach. Intell.*, 47, 5, 4090–4107. doi:10.1109/TPAMI.2025.3540467.

Z. Liu, L. Wu, M. He, Z. Guan, H. Zhao, and N. Feng. 2025. "Multi-View Empowered Structural Graph Wordification for Language Models." In: *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*. Ed. by T. Walsh, J. Shah, and Z. Kolter. AAAI Press, 24714–24722. doi:10.1609/AAAI.V39I23.34652.

A. V. Looveren and J. Klaise. 2021. "Interpretable Counterfactual Explanations Guided by Prototypes." In: *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021* (Lecture Notes in Computer Science). Ed. by N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, and J. A. Lozano. Vol. 12976. Springer, 650–665. doi:10.1007/978-3-030-86520-7_40.

P. Lu, S. Mishra, T. Xia, L. Qiu, K. Chang, S. Zhu, O. Tafjord, P. Clark, and A. Kalyan. 2022. "Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering." In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

S. M. Lundberg and S. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NeurIPS 2017*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

J. Ma, R. Guo, S. Mishra, A. Zhang, and J. Li. 2022. "CLEAR: Generative Counterfactual Explanations on Graphs." In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. http://papers.nips.cc/paper_files/paper/2022/hash/a69d7f3a1340d55c720e572742439eaf-Abstract-Conference.html.

L. Ma, Y. Zhang, T. Wang, X. Liu, W. Ye, C. Sun, and S. Zhang. 2021. "Legal Judgment Prediction with Multi-Stage Case Representation Learning in the Real Court Setting." In: *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021*. Ed. by F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, and T. Sakai. ACM, 993–1002. doi:10.1145/3404835.3462945.

Y. Ma, P. Wu, C. Ling, and S. Ding. 2024. "Research on public opinion effecting on stock price during crises based on model checking." *Expert Syst. Appl.*, 249, 123442. doi:10.1016/J.ESWA.2024.123442.

C. Mao, R. Teotia, A. Sundar, S. Menon, J. Yang, X. Wang, and C. Vondrick. 2023. "Doubly Right Object Recognition: A Why Prompt for Visual Rationales." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*. IEEE, 2722–2732. doi:10.1109/CVPR52729.2023.00267.

F. Marchetti, F. Becattini, L. Seidenari, and A. D. Bimbo. 2024. "SMEMO: Social Memory for Trajectory Forecasting." *IEEE Trans. Pattern Anal. Mach. Intell.*, 46, 6, 4410–4425. doi:10.1109/TPAMI.2024.3356755.

S. Marton, S. Lüdtke, C. Bartelt, A. Tschalzev, and H. Stuckenschmidt. 2024. "Explaining neural networks without access to training data." *Mach. Learn.*, 113, 6, 3633–3652. doi:10.1007/S10994-023-06428-4.

S. Meng, X. Zhang, N. Liu, L. Tu, and Q. Li. 2025. "Knowledge&Social-based collaborative method with contrastive graph structure learning for explainable recommendation." *Inf. Sci.*, 709, 122077. doi:10.1016/J.INS.2025.122077.

O.-B. Mercea, L. Riesch, A. S. Koepke, and Z. Akata. 2022. "Audiovisual Generalised Zero-shot Learning with Cross-modal Attention and Language." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*. Computer Vision Foundation / IEEE, 10543–10553. doi:10.1109/CVPR52688.2022.01030.

G. A. Miller. 1956. "The magical number seven, plus or minus two: Some limits on our capacity for processing information." *Psychological Review*, 63, 2, 81–97. doi:10.1037/h0043158.

M. Minsky. 1991. "Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy." *AI Mag.*, 12, 2, 34–51. doi:10.1609/AIMAG.V12I2.894.

M. Mirman, T. Gehr, and M. T. Vechev. 2018. "Differentiable Abstract Interpretation for Provably Robust Neural Networks." In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018* (Proceedings of Machine Learning Research). Ed. by J. G. Dy and A. Krause. Vol. 80. PMLR, 3575–3583. http://proceedings.mlr.press/v80/mirman18b.html.

B. Mistry, K. Farrahi, and J. S. Hare. 2022. "A Primer for Neural Arithmetic Logic Modules." *J. Mach. Learn. Res.*, 23, 185:1–185:58. https://jmlr.org/papers/v23/21-0211.html.

M. Moradi and M. Samwald. 2021. "Post-hoc explanation of black-box classifiers using confident itemsets." *Expert Syst. Appl.*, 165, 113941. doi:10.1016/J.ESWA.2020.113941.

T. T. Mueller, J. C. Paetzold, C. Prabhakar, D. Usynin, D. Rueckert, and G. Kaissis. 2023. "Differentially Private Graph Neural Networks for Whole-Graph Classification." *IEEE Trans. Pattern Anal. Mach. Intell.*, 45, 6, 7308–7318. doi:10.1109/TPAMI.2022.3228315.

L. Muyama, A. Neuraz, and A. Coulet. 2024. "Deep Reinforcement Learning for personalized diagnostic decision pathways using Electronic Health Records: A comparative study on anemia and Systemic Lupus Erythematosus." *Artif. Intell. Medicine*, 157, 102994. doi:10.1016/J.ARTMED.2024.102994.

M. Nauta, R. van Bree, and C. Seifert. 2021. "Neural Prototype Trees for Interpretable Fine-Grained Image Recognition." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*. Computer Vision Foundation / IEEE, 14933–14943. doi:10.1109/CVPR46437.2021.01469.

V. Nezerka, T. Zbíral, and J. Trejbal. 2024. "Machine-learning-assisted classification of construction and demolition waste fragments using computer vision: Convolution versus extraction of selected features." *Expert Syst. Appl.*, 238, Part A, 121568. doi:10.1016/J.ESWA.2023.121568.

H. F. Ong, N. Mustapha, H. Hamdan, R. Rosli, and A. Mustapha. 2020. "Informative top-k class associative rule for cancer biomarker discovery on microarray data." *Expert Syst. Appl.*, 146, 113169. doi:10.1016/J.ESWA.2019.113169.

P. Padalkar, H. Wang, and G. Gupta. 2024. "NeSyFOLD: A Framework for Interpretable Image Classification." In: *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*. Ed. by M. J. Wooldridge, J. G. Dy, and S. Natarajan. AAAI Press, 4378–4387. doi:10.1609/AAAI.V38I5.28235.

M. J. Page et al.. 2021. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." *bmj*, 372. doi:10.1136/bmj.n71.

B. Pan, H. Cai, D. Huang, K. Lee, A. Gaidon, E. Adeli, and J. C. Niebles. 2020. "Spatio-Temporal Graph for Video Captioning With Knowledge Distillation." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. Computer Vision Foundation / IEEE, 10867–10876. doi:10.1109/CVPR42600.2020.01088.

Y. L. Pang, A. Xompero, C. Oh, and A. Cavallaro. 2025. "Stereo Hand-Object Reconstruction for Human-to-Robot Handover." *IEEE Robotics Autom. Lett.*, 10, 6, 5761–5768. doi:10.1109/LRA.2025.3562790.

A. Pati and A. Lerch. 2021. "Attribute-based regularization of latent spaces for variational auto-encoders." *Neural Comput. Appl.*, 33, 9, 4429–4444. doi:10.1007/S00521-020-05270-2.

X. Peng, M. O. Riedl, and P. Ammanabrolu. 2022. "Inherently Explainable Reinforcement Learning in Natural Language." In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. http://papers.nips.cc/paper_files/paper/2022/hash/672e44a114a41d5f34b97459877c083d-Abstract-Conference.html.

F. Petersen, C. Borgelt, H. Kuehne, and O. Deussen. 2022. "Deep Differentiable Logic Gate Networks." In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. http://papers.nips.cc/paper_files/paper/2022/hash/0d3496dd0cec77a999c98d35003203ca-Abstract-Conference.html.

J. F. Pettit, C. S. Lee, J. Yang, A. Ho, D. M. Faissol, B. K. Petersen, and M. Landajuela. 2025. "DisCo-DSO: Coupling Discrete and Continuous Optimization for Efficient Generative Design in Hybrid Spaces." In: *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*. Ed. by T. Walsh, J. Shah, and Z. Kolter. AAAI Press, 27117–27125. doi:10.1609/AAAI.V39I25.34919.

S. Postic and M. Subasic. 2025. "Single-input interpretation of a deep classification neural network." *Knowl. Based Syst.*, 309, 112903. doi:10.1016/J.KNOSYS.2024.112903.

F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. M. Wallach. 2021. "Manipulating and Measuring Model Interpretability." In: *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*. Ed. by Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. M. Drucker. ACM, 237:1–237:52. doi:10.1145/3411764.3445315.

M. A. Prado-Romero, B. Prenkaj, and G. Stilo. 2024. "GRETEL 2.0: Generation and Evaluation of Graph Counterfactual Explanations Evolved." In: *Machine Learning and Knowledge Discovery in Databases. Research Track and Demo Track - European Conference, ECML PKDD 2024* (Lecture Notes in Computer Science). Ed. by A. Bifet, P. Daniusis, J. Davis, T. Krilavicius, M. Kull, E. Ntoutsi, K. Puolamäki, and I. Zliobaite. Vol. 14948. Springer, 363–367. doi:10.1007/978-3-031-70371-3_21.

Y. Quan, X. Yao, and H. Ji. 2023. "Single Image Defocus Deblurring via Implicit Neural Inverse Kernels." In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023*. IEEE, 12566–12576. doi:10.1109/ICCV51070.2023.01158.

R. Quessard, T. D. Barrett, and W. R. Clements. 2020. "Learning Disentangled Representations and Group Structure of Dynamical Environments." In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. https://proceedings.neurips.cc/paper/2020/hash/e449b9317dad920c0dd5ad0a2a2d5e49-Abstract.html.

A. Ragab, M. El-Koujok, H. Ghezzaz, M. Amazouz, M. Ouali, and S. Yacout. 2019. "Deep understanding in industrial processes by complementing human expertise with interpretable patterns of machine learning." *Expert Syst. Appl.*, 122, 388–405. doi:10.1016/J.ESWA.2019.01.011.

J. Ren, M. Li, Q. Chen, H. Deng, and Q. Zhang. 2023. "Defining and Quantifying the Emergence of Sparse Concepts in DNNs." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*. IEEE, 20280–20289. doi:10.1109/CVPR52729.2023.01942.

M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, KDD 2016*. Ed. by B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi. ACM, 1135–1144. doi:10.1145/2939672.2939778.

K. Roth, O. Vinyals, and Z. Akata. 2022. "Integrating Language Guidance into Vision-based Deep Metric Learning." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*. IEEE. doi:10.1109/CVPR52688.2022.01570.

D. Ru, C. Sun, J. Feng, L. Qiu, H. Zhou, W. Zhang, Y. Yu, and L. Li. 2021. "Learning Logic Rules for Document-Level Relation Extraction." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*. Ed. by M. Moens, X. Huang, L. Specia, and S. W. Yih. Association for Computational Linguistics, 1239–1250. doi:10.18653/V1/2021.EMNLP-MAIN.95.

C. Rudin. 2019. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nat. Mach. Intell.*, 1, 5, 206–215. doi:10.1038/S42256-019-0048-X.

S. Russell and P. Norvig. 2020. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson. ISBN: 9780134610993. http://aima.cs.berkeley.edu/.

D. Rymarczyk, L. Struski, J. Tabor, and B. Zielinski. 2021. "ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification." In: *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2021*. Ed. by F. Zhu, B. C. Ooi, and C. Miao. ACM, 1420–1430. doi:10.1145/3447548.3467245.

W. Saeed and C. W. Omlin. 2023. "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities." *Knowl. Based Syst.*, 263, 110273. doi:10.1016/J.KNOSYS.2023.110273.

M. Sajid, A. K. Malik, M. Tanveer, and P. N. Suganthan. 2024. "Neuro-Fuzzy Random Vector Functional Link Neural Network for Classification and Regression Problems." *IEEE Trans. Fuzzy Syst.*, 32, 5, 2738–2749. doi:10.1109/TFUZZ.2024.3359652.

A. Salimi-Badr. 2024. "UNFIS: A Novel Neuro-Fuzzy Inference System with Unstructured fuzzy rules." *Neurocomputing*, 579, 127437. doi:10.1016/J.NEUCOM.2024.127437.

A. Schmaltz. 2021. "Detecting Local Insights from Global Labels: Supervised and Zero-Shot Sequence Labeling via a Convolutional Decomposition." *Comput. Linguistics*, 47, 4, 729–773. doi:10.1162/COLI_A_00416.

P. Sen, M. Danilevsky, Y. Li, S. Brahma, M. Boehm, L. Chiticariu, and R. Krishnamurthy. 2020. "Learning Explainable Linguistic Expressions with Neural Inductive Logic Programming for Sentence Classification." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Association for Computational Linguistics, 4211–4221. doi:10.18653/V1/2020.EMNLP-MAIN.345.

Y. Shang, B. Duan, Z. Zong, L. Nie, and Y. Yan. 2021. "Lipschitz Continuity Guided Knowledge Distillation." In: *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 10655–10664. doi:10.1109/ICCV48922.2021.01050.

W. Shen, Z. Wei, Q. Ren, B. Zhang, S. Huang, J. Fan, and Q. Zhang. 2024. "Interpretable Rotation-Equivariant Quaternion Neural Networks for 3D Point Cloud Processing." *IEEE Trans. Pattern Anal. Mach. Intell.*, 46, 5, 3290–3304. doi:10.1109/TPAMI.2023.3346383.

F. Shi, B. Li, and X. Xue. 2021. "Raven's Progressive Matrices Completion with Latent Gaussian Process Priors." In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*. AAAI Press, 9612–9620. doi:10.1609/AAAI.V35I11.17157.

A. Shrivastava, S. Desai, A. Gupta, A. Elkahky, A. Livshits, A. Zotov, and A. Aly. 2023. "Retrieve-and-Fill for Scenario-based Task-Oriented Semantic Parsing." In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023*. Ed. by A. Vlachos and I. Augenstein. Association for Computational Linguistics, 430–447. doi:10.18653/V1/2023.EACL-MAIN.32.

J. Si, Y. Zhu, and D. Zhou. 2023. "Exploring Faithful Rationale for Multi-Hop Fact Verification via Salience-Aware Graph Learning." In: *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*. Ed. by B. Williams, Y. Chen, and J. Neville. AAAI Press, 13573–13581. doi:10.1609/AAAI.V37I11.26591.

J. F. Sigut, F. Fumero, R. Arnay, J. Estévez, and T. Díaz-Alemán. 2023. "Interpretable surrogate models to approximate the predictions of convolutional neural networks in glaucoma diagnosis." *Mach. Learn. Sci. Technol.*, 4, 4, 45024. doi:10.1088/2632-2153/AD0798.

P. Song, C. Zhao, and B. Huang. 2023. "MPGE and RootRank: A sufficient root cause characterization and quantification framework for industrial process faults." *Neural Networks*, 161, 397–417. doi:10.1016/J.NEUNET.2023.01.030.

W. Stammer, P. Schramowski, and K. Kersting. 2021. "Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting With Their Explanations." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*. Computer Vision Foundation / IEEE, 3619–3629. doi:10.1109/CVPR46437.2021.00362.

H. Sufriyana, Y. Wu, and E. C. Su. 2023. "Human-guided deep learning with ante-hoc explainability by convolutional network from non-image data for pregnancy prognostication." *Neural Networks*, 162, 99–116. doi:10.1016/J.NEUNET.2023.02.020.

N. Talpur, S. J. Abdulkadir, H. Alhussian, M. H. Hasan, N. Aziz, and A. M. Bamhdi. 2023. "Deep Neuro-Fuzzy System application trends, challenges, and future perspectives: a systematic survey." *Artif. Intell. Rev.*, 56, 2, 865–913. doi:10.1007/S10462-022-10188-3.

Y. Tang, H. Lu, X. Li, L. Chen, M. Yuan, and J. Zeng. 2021. "Learning-Aided Heuristics Design for Storage System." In: *International Conference on Management of Data, SIGMOD 2021*. Ed. by G. Li, Z. Li, S. Idreos, and D. Srivastava. ACM, 2597–2601. doi:10.1145/3448016.3457554.

Z. Tang, B. Zhu, Y. Hao, C. Ngo, and R. Hong. 2025. "RAGG: Retrieval-Augmented Grasp Generation Model." In: *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*. Ed. by T. Walsh, J. Shah, and Z. Kolter. AAAI Press, 7311–7319. doi:10.1609/AAAI.V39I7.32786.

A. Tarzariol, M. Gebser, and K. Schekotihin. 2022. "Lifting symmetry breaking constraints with inductive logic programming." *Mach. Learn.*, 111, 4, 1303–1326. doi:10.1007/S10994-022-06146-3.

D. Thakker, B. K. Mishra, A. Abdullatif, S. Mazumdar, and S. Simpson. 2020. "Explainable Artificial Intelligence for Developing Smart Cities Solutions." *Smart Cities*, 3, 4, 1353–1382. doi:10.3390/smartcities3040065.

D. Tilwani, R. Venkataramanan, and A. P. Sheth. 2024. "Neurosymbolic AI Approach to Attribution in Large Language Models." *IEEE Intell. Syst.*, 39, 6, 10–17. doi:10.1109/MIS.2024.3477108.

S. Tong, Q. Liu, W. Huang, Z. Huang, E. Chen, C. Liu, H. Ma, and S. Wang. 2020. "Structure-based Knowledge Tracing: An Influence Propagation View." In: *20th IEEE International Conference on Data Mining, ICDM 2020*. Ed. by C. Plant, H. Wang, A. Cuzzocrea, C. Zaniolo, and X. Wu. IEEE, 541–550. doi:10.1109/ICDM50108.2020.00063.

R. Trivedi and H. Zha. 2020. "Learning Strategic Network Emergence Games." In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. https://proceedings.neurips.cc/paper/2020/hash/4bb236de7787ceedafdff83bb8ea4710-Abstract.html.

U. Upadhyay, S. Karthik, M. Mancini, and Z. Akata. 2023. "ProbVLM: Probabilistic Adapter for Frozen Vison-Language Models." In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023*. IEEE, 1899–1910. doi:10.1109/ICCV51070.2023.00182.

S. Veer, K. Leung, R. K. Cosner, Y. Chen, P. Karkus, and M. Pavone. 2023. "Receding Horizon Planning with Rule Hierarchies for Autonomous Vehicles." In: *IEEE International Conference on Robotics and Automation, ICRA 2023*. IEEE, 1507–1513. doi:10.1109/ICRA48891.2023.10160622.

T. Vu and T. Le. 2025. "TCrossE: Cross-space interaction of bicomplex and quaternion embeddings for temporal knowledge graph completion." *Knowl. Based Syst.*, 315, 113321. doi:10.1016/J.KNOSYS.2025.113321.

G. Wan, B. Du, S. Pan, and G. Haffari. 2020. "Reinforcement Learning Based Meta-Path Discovery in Large-Scale Heterogeneous Information Networks." In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*. AAAI Press, 6094–6101. doi:10.1609/AAAI.V34I04.6073.

H. Wang, S. Zhang, S. Wang, T. Jiang, and Y. Ge. 2025. "Double-Feedback: Enhancing Large Language Models Reasoning in Robotic Tasks by Knowledge Graphs." *IEEE Robotics Autom. Lett.*, 10, 6, 5951–5958. doi:10.1109/LRA.2025.3562776.

H. Wang, Z. Che, Y. Yang, M. Wang, Z. Xu, X. Qiao, M. Qi, F. Feng, and J. Tang. 2024. "RDFC-GAN: RGB-Depth Fusion CycleGAN for Indoor Depth Completion." *IEEE Trans. Pattern Anal. Mach. Intell.*, 46, 11, 7088–7101. doi:10.1109/TPAMI.2024.3388004.

J. Wang and M. Zhang. 2022. "Geo-SIC: Learning Deformable Geometric Shapes in Deep Image Classifiers." In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. http://papers.nips.cc/paper_files/paper/2022/hash/b328c5bd9ff8e3a5e1be74baf4a7a456-Abstract-Conference.html.

J. Wang, P. Ding, Y. Zhu, X. Gao, X. Yu, and B. Yu. 2025. "MSN-DTA: A multi-scale node adaptive graph neural network for interpretable drug-target binding affinity prediction." *Knowl. Based Syst.*, 320, 113699. doi:10.1016/J.KNOSYS.2025.113699.

K. Wang, X. Chen, N. Chen, and T. Chen. 2020. "Automatic Emergency Diagnosis with Knowledge-Based Tree Decoding." In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. Ed. by C. Bessiere. ijcai.org, 3407–3414. doi:10.24963/IJCAI.2020/471.

M. H. Wang et al.. 2024. "AI-Based Advanced Approaches and Dry Eye Disease Detection Based on Multi-Source Evidence: Cases, Applications, Issues, and Future Directions." *Big Data Min. Anal.*, 7, 2, 445–484. doi:10.26599/BDMA.2023.9020024.

N. Wang, W. Pedrycz, W. Yao, X. Chen, and Y. Zhao. 2022. "Disjunctive Fuzzy Neural Networks: A New Splitting-Based Approach to Designing a T-S Fuzzy Model." *IEEE Trans. Fuzzy Syst.*, 30, 2, 370–381. doi:10.1109/TFUZZ.2020.3039371.

P. Wang, L. Garattoni, S. Meier, N. Navab, and B. Busam. 2024. "Improving Self-Supervised Learning of Transparent Category Poses With Language Guidance and Implicit Physical Constraints." *IEEE Robotics Autom. Lett.*, 9, 9, 8114–8121. doi:10.1109/LRA.2024.3440732.

Q. Wang, J. Liu, W. Guo, and X. Wang. 2023. "Evolving stochastic configure network: A more compact model with interpretability." *Inf. Sci.*, 639, 119006. doi:10.1016/J.INS.2023.119006.

S. Wang, Z. Zeng, X. Yang, K. Xu, and X. Zhang. 2024. "Boosting Neural Cognitive Diagnosis with Student's Affective State Modeling." In: *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*. Ed. by M. J. Wooldridge, J. G. Dy, and S. Natarajan. AAAI Press, 620–627. doi:10.1609/AAAI.V38I1.27818.

S. Wang, H. Li, et al.. 2022. "Tower Bridge Net (TB-Net): Bidirectional Knowledge Graph Aware Embedding Propagation for Explainable Recommender Systems." In: *38th IEEE International Conference on Data Engineering, ICDE 2022*. IEEE, 3268–3279. doi:10.1109/ICDE53745.2022.00309.

S. Wang, B. Xie, L. Ding, J. Chen, and Y. Xiang. 2025. "Reinforced logical reasoning over KGs for interpretable recommendation system." *Mach. Learn.*, 114, 4, 95. doi:10.1007/S10994-024-06646-4.

W. Wang, C. Han, T. Zhou, and D. Liu. 2023. "Visual Recognition with Deep Nearest Centroids." In: *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net. https://openreview.net/forum?id=CsKwavjr7A.

X. Wang, K. Liu, D. Wang, L. Wu, Y. Fu, and X. Xie. 2022. "Multi-level Recommendation Reasoning over Knowledge Graphs with Reinforcement Learning." In: *The ACM Web Conference, WWW 2022*. Ed. by F. Laforest, R. Troncy, E. Simperl, D. Agarwal, A. Gionis, I. Herman, and L. Médini. ACM, 2098–2108. doi:10.1145/3485447.3512083.

X. Wang, B. Dai, Q. Miao, Y. Nie, and E. Shang. 2024. "An Efficient Deep Reinforcement Learning-Based Car-Following Method via Rule-Constrained Data Augmentation." *IEEE Trans. Intell. Veh.*, 9, 10, 6381–6393. doi:10.1109/TIV.2024.3367720.

Y. Wang, K. Li, Y. Luo, G. Li, Y. Guo, and Z. Wang. 2024. "Fast, Robust and Interpretable Participant Contribution Estimation for Federated Learning." In: *40th IEEE International Conference on Data Engineering, ICDE 2024*. IEEE, 2298–2311. doi:10.1109/ICDE60146.2024.00182.

Y. Wang, H. Liu, W. Jia, S. Guan, X. Liu, and X. Duan. 2022. "Deep Fuzzy Rule-Based Classification System With Improved Wang-Mendel Method." *IEEE Trans. Fuzzy Syst.*, 30, 8, 2957–2970. doi:10.1109/TFUZZ.2021.3098339.

Y. Wang, H. Su, B. Zhang, and X. Hu. 2018. "Interpret Neural Networks by Identifying Critical Data Routing Paths." In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*. Computer Vision Foundation / IEEE Computer Society, 8906–8914. doi:10.1109/CVPR.2018.00928.

Z. Wang, W. Yan, C. Zeng, Y. Tian, and S. Dong. 2023. "A Unified Interpretable Intelligent Learning Diagnosis Framework for Learning Performance Prediction in Intelligent Tutoring Systems." *Int. J. Intell. Syst.*, 2023, 1–20. doi:10.1155/2023/4468025.

Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin. 2017. "Multi-label Image Recognition by Recurrently Discovering Attentional Regions." In: *IEEE International Conference on Computer Vision, ICCV 2017*. IEEE Computer Society, 464–472. doi:10.1109/ICCV.2017.58.

Z. Wang, W. Zhang, N. Liu, and J. Wang. 2024. "Learning Interpretable Rules for Scalable Data Representation and Classification." *IEEE Trans. Pattern Anal. Mach. Intell.*, 46, 2, 1121–1133. doi:10.1109/TPAMI.2023.3328881.

Z. Wang, W. Zhang, N. Liu, and J. Wang. 2021. "Scalable Rule-Based Representation Learning for Interpretable Classification." In: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*. Ed. by M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, 30479–30491. https://proceedings.neurips.cc/paper/2021/hash/ffbd6cbb019a1413183c8d08f2929307-Abstract.html.

Z. Wang, W. Zhang, N. Liu, and J. Wang. 2020. "Transparent Classification with Multilayer Logical Perceptrons and Random Binarization." In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*. AAAI Press, 6331–6339. doi:10.1609/AAAI.V34I04.6102.

D. S. Watson, L. Gultchin, A. Taly, and L. Floridi. 2022. "Local Explanations via Necessity and Sufficiency: Unifying Theory and Practice." *Minds Mach.*, 32, 1, 185–218. doi:10.1007/S11023-022-09598-7.

C. Wei, S. Luo, L. Pan, Z. Wu, J. Zhang, and Q. G. K. Safi. 2018. "Locally weighted embedding topic modeling by markov random walk structure approximation and sparse regularization." *Neurocomputing*, 285, 35–50. doi:10.1016/J.NEUCOM.2018.01.030.

H. Wen, S. Sun, T. Huang, and D. Xiao. 2024. "An intrinsic integrity-driven rating model for a sustainable reputation system." *Expert Syst. Appl.*, 249, 123804. doi:10.1016/J.ESWA.2024.123804.

P. Whiting and et al.. 2016. "ROBIS: a new tool to assess risk of bias in systematic reviews was developed." *J. Clin. Epidemiol.*, 69, 225–234.

C. Wohlin. 2014. "Guidelines for snowballing in systematic literature studies and a replication in software engineering." In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, EASE 2014*. Ed. by M. J. Shepperd, T. Hall, and I. Myrtveit. ACM, 38:1–38:10. doi:10.1145/2601248.2601268.

H. Wu, W. Ruan, et al.. 2023. "Interpretable Machine Learning for COVID-19: An Empirical Study on Severity Prediction Task." *IEEE Trans. Artif. Intell.*, 4, 4, 764–777. doi:10.1109/TAI.2021.3092698.

H. Wu, C. Wang, J. Yin, K. Lu, and L. Zhu. 2018. "Sharing Deep Neural Network Models with Interpretation." In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018*. Ed. by P. Champin, F. Gandon, M. Lalmas, and P. G. Ipeirotis. ACM, 177–186. doi:10.1145/3178876.3185995.

H. Wu, H. Fang, Z. Sun, C. Geng, X. Kong, and Y. Ong. 2023. "A Generic Reinforced Explainable Framework with Knowledge Graph for Session-based Recommendation." In: *39th IEEE International Conference on Data Engineering, ICDE 2023*. IEEE, 1260–1272. doi:10.1109/ICDE55515.2023.00101.

Q. Wu, P. Wang, C. Shen, I. D. Reid, and A. van den Hengel. 2018. "Are You Talking to Me? Reasoned Visual Dialog Generation Through Adversarial Learning." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*. Computer Vision Foundation / IEEE Computer Society, 6106–6115. doi:10.1109/CVPR.2018.00639.

Z. Wu, A. Arora, Z. Wang, A. Geiger, D. Jurafsky, C. D. Manning, and C. Potts. 2024. "ReFT: Representation Finetuning for Language Models." In: *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*. Ed. by A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang. http://papers.nips.cc/paper_files/paper/2024/hash/75008a0fba53bf13b0bb3b7bff986e0e-Abstract-Conference.html.

X. Xia, W. Yang, J. Ren, Y. Li, Y. Zhan, B. Han, and T. Liu. 2022. "Pluralistic Image Completion with Gaussian Mixture Models." In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. http://papers.nips.cc/paper_files/paper/2022/hash/987913de7a2963359196d4491d0fd4e7-Abstract-Conference.html.

Y. Xian, H. Zhao, T. Y. Lee, S. Kim, R. A. Rossi, Z. Fu, G. de Melo, and S. Muthukrishnan. 2021. "EXACTA: Explainable Column Annotation." In: *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2021*. Ed. by F. Zhu, B. C. Ooi, and C. Miao. ACM, 3775–3785. doi:10.1145/3447548.3467211.

C. Xiao, P. Zhang, W. A. Chaovalitwongse, J. Hu, and F. Wang. 2017. "Adverse Drug Reaction Prediction with Symbolic Latent Dirichlet Allocation." In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI 2017*. Ed. by S. Singh and S. Markovitch. AAAI Press, 1590–1596. doi:10.1609/AAAI.V31I1.10717.

T. Xiao, W. Zhang, Y. Cheng, and J. Suo. 2024. "HOPE: High-Order Polynomial Expansion of Black-Box Neural Networks." *IEEE Trans. Pattern Anal. Mach. Intell.*, 46, 12, 7924–7939. doi:10.1109/TPAMI.2024.3399197.

Y. Xie, Z. Wang, Y. Li, B. Ding, N. M. Gürel, C. Zhang, M. Huang, W. Lin, and J. Zhou. 2021. "FIVES: Feature Interaction Via Edge Search for Large-Scale Tabular Data." In: *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2021*. Ed. by F. Zhu, B. C. Ooi, and C. Miao. ACM, 3795–3805. doi:10.1145/3447548.3467066.

J. Xu, Z. Zhang, T. Friedman, Y. Liang, and G. V. den Broeck. 2018. "A Semantic Loss Function for Deep Learning with Symbolic Knowledge." In: *Proceedings of the 35th International Conference on Machine Learning, ICML 2018* (Proceedings of Machine Learning Research). Ed. by J. G. Dy and A. Krause. Vol. 80. PMLR, 5498–5507. http://proceedings.mlr.press/v80/xu18h.html.

S. Xu, J. Liu, and J. Zhang. 2024. "scACT: Accurate Cross-modality Translation via Cycle-consistent Training from Unpaired Single-cell Data." In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024*. Ed. by E. Serra and F. Spezzano. ACM, 2722–2731. doi:10.1145/3627673.3679576.

X. Xu, M. You, H. Zhou, Z. Qian, W. Xu, and B. He. 2023. "GAN-Based Editable Movement Primitive From High-Variance Demonstrations." *IEEE Robotics Autom. Lett.*, 8, 8, 4593–4600. doi:10.1109/LRA.2023.3285473.

Z. Xu, G. Zhu, C. Yuan, and Y. Huang. 2022. "One-Stage Tree: end-to-end tree builder and pruner." *Mach. Learn.*, 111, 5, 1959–1985. doi:10.1007/S10994-021-06094-4.

C. Xue, J. Gao, and Z. Gu. 2025. "IT2-ENFIS: Interval Type-2 Exclusionary Neuro-Fuzzy Inference System, An Attempt Towards Trustworthy Regression Learning." *IEEE Trans. Artif. Intell.*, 1–15. doi:10.1109/TAI.2025.3574299.

K. Yan, L. Ji, H. Luo, M. Zhou, N. Duan, and S. Ma. 2021. "Control Image Captioning Spatially and Temporally." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Association for Computational Linguistics, 2014–2025. doi:10.18653/V1/2021.ACL-LONG.157.

R. Yan, T. Ma, A. Fokoue, M. Chang, and A. Julius. 2022. "Neuro-symbolic Models for Interpretable Time Series Classification using Temporal Logic Description." In: *IEEE International Conference on Data Mining, ICDM 2022*. Ed. by X. Zhu, S. Ranka, M. T. Thai, T. Washio, and X. Wu. IEEE, 618–627. doi:10.1109/ICDM54844.2022.00072.

S. Yan, Z. Yu, X. Zhang, D. Mahapatra, S. S. Chandra, M. Janda, H. P. Soyer, and Z. Ge. 2023. "Towards Trustable Skin Cancer Diagnosis via Rewriting Model's Decision." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*. IEEE, 11568–11577. doi:10.1109/CVPR52729.2023.01113.

H. Yang, J. Gee, and J. Shi. 2024. "Brain Decodes Deep Nets." In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*. IEEE, 23030–23040. doi:10.1109/CVPR52733.2024.02173.

L. Yang, J. Ji, J. Dai, L. Zhang, B. Zhou, P. Li, Y. Yang, and G. Pan. 2022. "Constrained Update Projection Approach to Safe Policy Optimization." In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. http://papers.nips.cc/paper_files/paper/2022/hash/3ba7560b4c3e66d760fbdd472cf4a5a9-Abstract-Conference.html.

Y. Yang, T. Wang, and W. Xiang. 2025. "Neural transition system abstraction for neural network dynamical system models and its application to Computational Tree Logic verification." *Neural Networks*, 186, 107261. doi:10.1016/J.NEUNET.2025.107261.

Z. Yao, C. Li, T. Dong, X. Lv, J. Yu, L. Hou, J. Li, Y. Zhang, and Z. Dai. 2021. "Interpretable and Low-Resource Entity Matching via Decoupling Feature Learning from Decision Making." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and*

the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Association for Computational Linguistics, 2770–2781. doi:10.18653/V1/2021.ACL-LONG.215.

R. A. Yeh, J. Xiong, W. W. Hwu, M. N. Do, and A. G. Schwing. 2017. "Interpretable and Globally Optimal Prediction for Textual Grounding using Image Concepts." In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NIPS 2017*. Ed. by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, 1912–1922. https://proceedings.neurips.cc/paper/2017/hash/52292e0c763fd027c6eba6b8f494d2eb-Abstract.html.

A. Yen, H. Huang, and H. Chen. 2021. "Unanswerable Question Correction in Question Answering over Personal Knowledge Base." In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, 2021*. AAAI Press, 14266–14275. doi:10.1609/AAAI.V35I16.17678.

P. Yi, Y. Xia, and Y. Long. 2025. "Irony Detection, Reasoning and Understanding in Zero-shot Learning." *IEEE Transactions on Artificial Intelligence*, 1–14. doi:10.1109/TAI.2025.3579452.

J. Yu and G. Liu. 2021. "Extracting and inserting knowledge into stacked denoising auto-encoders." *Neural Networks*, 137, 31–42. doi:10.1016/J.NEUNET.2021.01.010.

J. Yu and G. Liu. 2020. "Knowledge extraction and insertion to deep belief network for gearbox fault diagnosis." *Knowl. Based Syst.*, 197, 105883. doi:10.1016/J.KNOSYS.2020.105883.

L. Yu, W. Xiang, J. Fang, Y. P. Chen, and R. Zhu. 2022. "A novel explainable neural network for Alzheimer's disease diagnosis." *Pattern Recognit.*, 131, 108876. doi:10.1016/J.PATCOG.2022.108876.

Z. Zang, R. Yin, W. Lu, W. Pedrycz, and L. Zhang. 2024. "A Linguistically Interpretable Deep Fuzzy Classification System With Feature Transformation and Reconstruction." *IEEE Trans. Fuzzy Syst.*, 32, 8, 4297–4311. doi:10.1109/TFUZZ.2024.3394897.

B. Zhang and L. Li. 2023. "PIPER: A logic-driven deep contrastive optimization pipeline for event temporal reasoning." *Neural Networks*, 164, 186–202. doi:10.1016/J.NEUNET.2023.04.020.

B. Zhang, W. Zheng, J. Zhou, and J. Lu. 2023. "Bort: Towards Explainable Neural Networks with Bounded Orthogonal Constraint." In: *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net. https://openreview.net/forum?id=My57qBufZWs.

L. Zhang, H. Su, E. Zio, L. Jiang, L. Fan, and J. Zhang. 2023. "A graph structure feature-based framework for the pattern recognition of the operational states of integrated energy systems." *Expert Syst. Appl.*, 213, Part, 119039. doi:10.1016/J.ESWA.2022.119039.

L. Zhang, X. Lu, Z. Chen, T. Liu, Q. Chen, and Z. Li. 2022. "Adaptive deep learning for network intrusion detection by risk analysis." *Neurocomputing*, 493, 46–58. doi:10.1016/J.NEUCOM.2022.04.061.

L. Zhang, L. Jin, X. Sun, G. Xu, Z. Zhang, X. Li, N. Liu, Q. Liu, and S. Yan. 2023. "TOT : Topology-Aware Optimal Transport for Multimodal Hate Detection." In: *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*. Ed. by B. Williams, Y. Chen, and J. Neville. AAAI Press, 4884–4892. doi:10.1609/AAAI.V37I4.25614.

Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu. 2019. "Interpreting CNNs via Decision Trees." In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*. Computer Vision Foundation / IEEE, 6261–6270. doi:10.1109/CVPR.2019.00642.

T. Zhang, J. Fang, H. Jiang, B. Bi, X. Wang, and X. He. 2025. "Explainable and Efficient Editing for Large Language Models." In: *Proceedings of the ACM on Web Conference 2025, WWW 2025*. Ed. by G. Long, M. Blumestein, Y. Chang, L. Lewin-Eytan, Z. H. Huang, and E. Yom-Tov. ACM. doi:10.1145/3696410.3714835.

W. Zhang, L. Kong, S. Lee, Y. Chen, G. Zhang, H. Wang, and M. Song. 2024. "Detecting mental and physical disorders using multi-task learning equipped with knowledge graph attention network." *Artif. Intell. Medicine*, 149, 102812. doi:10.1016/J.ARTMED.2024.102812.

Y. Zhang, Y. Xiong, Y. Sun, C. Shan, T. Lu, H. Song, and Y. Zhu. 2022. "RuDi: Explaining Behavior Sequence Models by Automatic Statistics Generation and Rule Distillation." In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, CIKM 2022*. Ed. by M. A. Hasan and L. Xiong. ACM, 2651–2660. doi:10.1145/3511808.3557441.

Y. Zhang, X. Wang, H. Chen, J. Fan, W. Wen, H. Xue, H. Mei, and W. Zhu. 2024. "Large Language Model with Curriculum Reasoning for Visual Concept Recognition." In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024*. Ed. by R. Baeza-Yates and F. Bonchi. ACM, 6269–6280. doi:10.1145/3637528.3671653.

Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang. 2021. "A Survey on Neural Network Interpretability." *IEEE Trans. Emerg. Top. Comput. Intell.*, 5, 5, 726–742. doi:10.1109/TETCI.2021.3100641.

Z. Zhang, R. Yang, K. Wu, Z. Xu, J. Liu, L. Mu, Z. Gan, and W. Ding. 2025. "Drive in Corridors: Enhancing the Safety of End-to-End Autonomous Driving via Corridor Learning and Planning." *IEEE Robotics Autom. Lett.*, 10, 7, 6728–6735. doi:10.1109/LRA.2025.3570201.

Y. Zhao, H. Wu, and Z. Chen. 2025. "Self-supervised network compression based on quantum mutual information for fast change detection." *Knowl. Based Syst.*, 323, 113691. doi:10.1016/J.KNOSYS.2025.113691.

Y. Zhao, Y. Shen, X. Liu, Y. Luo, W. Zang, and X. Liu. 2023. "Spiking neural P systems with long-term potentiation and depression." *Inf. Sci.*, 640, 119082. doi:10.1016/J.INS.2023.119082.

A. Zharmagambetov and M. Á. Carreira-Perpiñán. 2022. "Semi-Supervised Learning with Decision Trees: Graph Laplacian Tree Alternating Optimization." In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022,*

NeurIPS 2022. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. http://papers.nips.cc/paper_files/paper/2022/hash/104f7b25495a0e40e65fb7c7eee37ed9-Abstract-Conference.html.

J. Zheng, Z. Qin, S. Wang, and D. Li. 2022. "Attention-based explainable friend link prediction with heterogeneous context information." *Inf. Sci.*, 597, 211–229. doi:10.1016/J.INS.2022.03.010.

K. Zheng, S. Yu, and B. Chen. 2024. "CI-GNN: A Granger causality-inspired graph neural network for interpretable brain network-based psychiatric diagnosis." *Neural Networks*, 172, 106147. doi:10.1016/J.NEUNET.2024.106147.

Y. Zheng, J. Ding, F. Liu, and D. Wang. 2023. "Adaptive neural decision tree for EEG based emotion recognition." *Inf. Sci.*, 643, 119160. doi:10.1016/J.INS.2023.119160.

E. Zhou, C. Vong, Y. Nojima, and S. Wang. 2023. "A Fully Interpretable First-Order TSK Fuzzy System and Its Training With Negative Entropic and Rule-Stability-Based Regularization." *IEEE Trans. Fuzzy Syst.*, 31, 7, 2305–2319. doi:10.1109/TFUZZ.2022.3223700.

L. Zhou, Y. Farag, and A. Vlachos. 2024. "An LLM Feature-based Framework for Dialogue Constructiveness Assessment." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*. Ed. by Y. Al-Onaizan, M. Bansal, and Y. Chen. Association for Computational Linguistics, 5389–5409. doi:10.18653/V1/2024.EMNLP-MAIN.308.

H. Zhu, G. Wang, Z. Liao, and W. Zhang. 2025. "AITL-Net: An adaptive interpretable transfer learning network with robust generalization for liver cancer recognition." *Knowl. Based Syst.*, 318, 113473. doi:10.1016/J.KNOSYS.2025.113473.

X. Zhu, Y. Du, Y. Mao, L. Chen, Y. Hu, and Y. Gao. 2023. "Knowledge-refined Denoising Network for Robust Recommendation." In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023*. Ed. by H. Chen, W. (Duh, H. Huang, M. P. Kato, J. Mothe, and B. Poblete. ACM, 362–371. doi:10.1145/3539618.3591707.

S. Zhuo, D. Wu, X. Hu, and Y. Wang. 2024. "ARDST: An Adversarial-Resilient Deep Symbolic Tree for Adversarial Learning." *Int. J. Intell. Syst.*, 2024. doi:10.1155/2024/2767008.

J. Zou, L. Wu, K. Liu, Y. Xu, S. He, and X. Bo. 2025. "A feature pair-based neural network embedded decision tree for synergistic drug combination prediction." *Pattern Recognit.*, 164, 111608. doi:10.1016/J.PATCOG.2025.111608.

A. Zunic, P. Corcoran, and I. Spasic. 2022. "The Case of Aspect in Sentiment Analysis: Seeking Attention or Co-Dependency?" *Mach. Learn. Knowl. Extr.*, 4, 2, 474–487. doi:10.3390/MAKE4020021.

X. Zuo, H. Yuan, B. Yang, H. Wang, and Y. Wang. 2023. "Exploring graph capsual network and graphormer for graph classification." *Inf. Sci.*, 640, 119045. doi:10.1016/J.INS.2023.119045.

A Extended Methodology

This appendix details the protocol, computational pipeline, and validation procedures underlying our systematic literature review on symbolic XAI for deep learning (January 2017–June 2025). We follow PRISMA 2020 guidance (Page et al. 2021) and document every step for transparency and exact reproducibility.

A.1 Protocol Overview

Scope and objectives. We target peer-reviewed, primary studies on *symbolic* explainable AI (XAI) applied to *deep learning/neural networks*. We explicitly include all subfields where deep models are used (e.g., CV, NLP, RL, KG/Reasoning), and we exclude purely non-symbolic XAI.

Registration and Materials. We pre-specified inclusion/exclusion criteria, search sources, and the computational pipeline before screening. We release (at submission) the code, configuration files, and machine-readable outputs (e.g., topic assignments, cluster statistics, selection metrics, etc.).

A.2 Information Sources and Search Strategy

Databases. We searched Embase, IEEE Xplore, Scopus, and Web of Science to prioritize curated, peer-reviewed venues. We excluded preprint aggregators (e.g., arXiv, Google Scholar).

Query. We designed a Boolean query spanning three concept domains (symbolic; neural/sub-symbolic; explainability), implemented with wildcard variants:

```
("symbol*" OR "logic*" OR "formal*" OR "knowledge*" OR "ontolog*" OR "decision tree*" OR
"decision list*" OR "decision set*" OR "constraint*" OR "rule" OR "rules" OR "reasoning")
  AND
```

```
("sub*symbolic" OR "deep*learning" OR "neural*" OR "network*" OR "layer*" OR "tensor*" OR
"differentiab*" OR "neuro*symbolic" OR "deep reinforcement learning")
  AND
("explainab*" OR "interpretab*" OR "accountab*" OR "explicab*" OR "transparen*" OR "translucen*"
OR "XAI" OR "IML")
```

Time window and language. We restricted results to January 2017–June 2025 (English only), capturing the modern XAI era that follows the DARPA XAI program and early interpretable deep learning work.

A.3 Eligibility Criteria

Inclusion. Peer-reviewed primary studies (observational or experimental), English language; deep models are central; symbolic elements appear in the explanation or in the pipeline (e.g., neurosymbolic reasoning)

Exclusion. Editorials, book chapters; non-English; studies that only consider non-symbolic XAI.

A.4 Deduplication, Quality Pre-filter, and PRISMA Flow

We exported all records, deduplicated with Mendeley Reference Manager,[11] and then applied a venue-quality filter to focus on high-impact, methodologically credible work:

- *Journals:* SCImago Artificial Intelligence area, Quartile 1 (Q1).[12]
- *Conferences:* CORE A* and A venues in AI-related categories;[13] the full list is in Table 20.

The PRISMA flow diagram is summarized in Figure 1. From 49,836 initial records we removed 19,719 duplicates, screened 30,117 records, quality-filtered to 3,749 screenable reports, sought retrieval for all, assessed eligibility, and included studies meeting our symbolic/deep-learning criteria with empirical content.

To improve topical completeness while preserving the reproducibility of our original database-driven pipeline, we followed PRISMA 2020 guidance to report an *additional* identification stream via "other methods" (distinct from database/register searching) (Page et al. 2021), illustrated on the right branch in Figure 1. Concretely, we incorporated *expert input during peer review* to address reviewer-identified gaps (e.g., missing seminal clusters within differentiable logic layers and concept bottleneck models). This expert-augmented stream is *additive*: it does not alter any steps, parameters, or outputs of the frozen database pipeline described above, and it is reported separately in the PRISMA flow diagram.

To keep the core methodology reproducible, we treat expert-augmented papers as an extension set that is *mapped onto* the existing topic/taxonomy structure rather than re-running the full discovery pipeline. Specifically, we assign each added paper to the closest existing topic representation using the same embedding model as in the main pipeline; we then update only descriptive statistics (e.g., per-topic reference lists and "key works" highlights), while leaving the original clustering and selection outputs intact.

A.5 Computational Screening and Analysis Pipeline

We implemented a three-stage pipeline to:

(1) induce semantically coherent topics,
(2) select representative papers per topic under explicit diversity/centrality trade-offs, and
(3) statistically validate that the selection is representative and unbiased.

A.5.1 Stage 1: Topic Modeling and Cluster Induction. We used BERTopic for topic discovery over titles, abstracts, and keywords.

[11] https://www.mendeley.com/
[12] https://www.scimagojr.com/
[13] https://portal.core.edu.au/conf-ranks/

Table 20. Selected conferences and ranks presented by categories.

Category	Acronym	Title	Rank
Artificial Intelligence	AAAI	AAAI Conference on Artificial Intelligence	A*
Artificial Intelligence	IJCAI	International Joint Conference on Artificial Intelligence	A*
Artificial Intelligence	AIED	International Conference on Artificial Intelligence in Education	A
Artificial Intelligence	AISTATS	International Conference on Artificial Intelligence and Statistics	A
Artificial Intelligence	ECAI	European Conference on Artificial Intelligence	A
Artificial Intelligence	UAI	Conference on Uncertainty in Artificial Intelligence	A
Machine Learning	COLT	Conference on Learning Theory	A*
Machine Learning	ICML	International Conference on Machine Learning	A*
Machine Learning	ICLR	International Conference on Learning Representations	A*
Machine Learning	NeurIPS	Advances in Neural Information Processing Systems	A*
Logic	CAV	Computer Aided Verification	A*
Logic	LICS	ACM/IEEE Symposium on Logic in Computer Science	A*
Logic	ICLP	International Conference on Logic Programming	A
Logic	JELIA	Logics in Artificial Intelligence, European Conference	A
Knowledge	KDD	ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	A*
Knowledge	KR	International Conference on Principles of Knowledge Representation and Reasoning	A*
Knowledge	CIKM	ACM International Conference on Information and Knowledge Management	A
Knowledge	ECML PKDD	European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases	A
Knowledge	LAK	International Conference on Learning Analytics and Knowledge	A
Knowledge	TARK	Theoretical Aspects of Rationality and Knowledge	A
Computer Vision	CVPR	IEEE/CVF Conference on Computer Vision and Pattern Recognition	A*
Computer Vision	ECCV	European Conference on Computer Vision	A*
Computer Vision	ICCV	IEEE International Conference on Computer Vision	A*
Natural Language Processing	ACL	Association for Computational Linguistics	A*
Natural Language Processing	EMNLP	Empirical Methods in Natural Language Processing	A*
Natural Language Processing	EACL	European Association for Computational Linguistics	A
Natural Language Processing	NAACL	North American Association for Computational Linguistics	A
Data	ICDE	International Conference on Data Engineering	A*
Data	ICDM	IEEE International Conference on Data Mining	A*
Data	SIGIR	International ACM SIGIR Conference on Research and Development in Information Retrieval	A*
Data	SIGMOD	ACM SIGMOD International Conference on Management of Data	A*
Data	VLDB	International Conference on Very Large Databases	A*
Data	WWW	ACM World Wide Web Conference	A*
Other	AAMAS	International Joint Conference on Autonomous Agents and Multiagent Systems	A*
Other	ICAPS	International Conference on Automated Planning and Scheduling	A*
Other	ICRA	IEEE International Conference on Robotics and Automation	A*
Other	FOCS	IEEE Symposium on Foundations of Computer Science	A*

Embedding model. Sentence-BERT `all-MiniLM-L6-v2` for 384-dimensional embeddings.

Dimensionality reduction. UMAP with `n_components=256`, `n_neighbors=15`, `min_dist=0.0`, cosine metric.

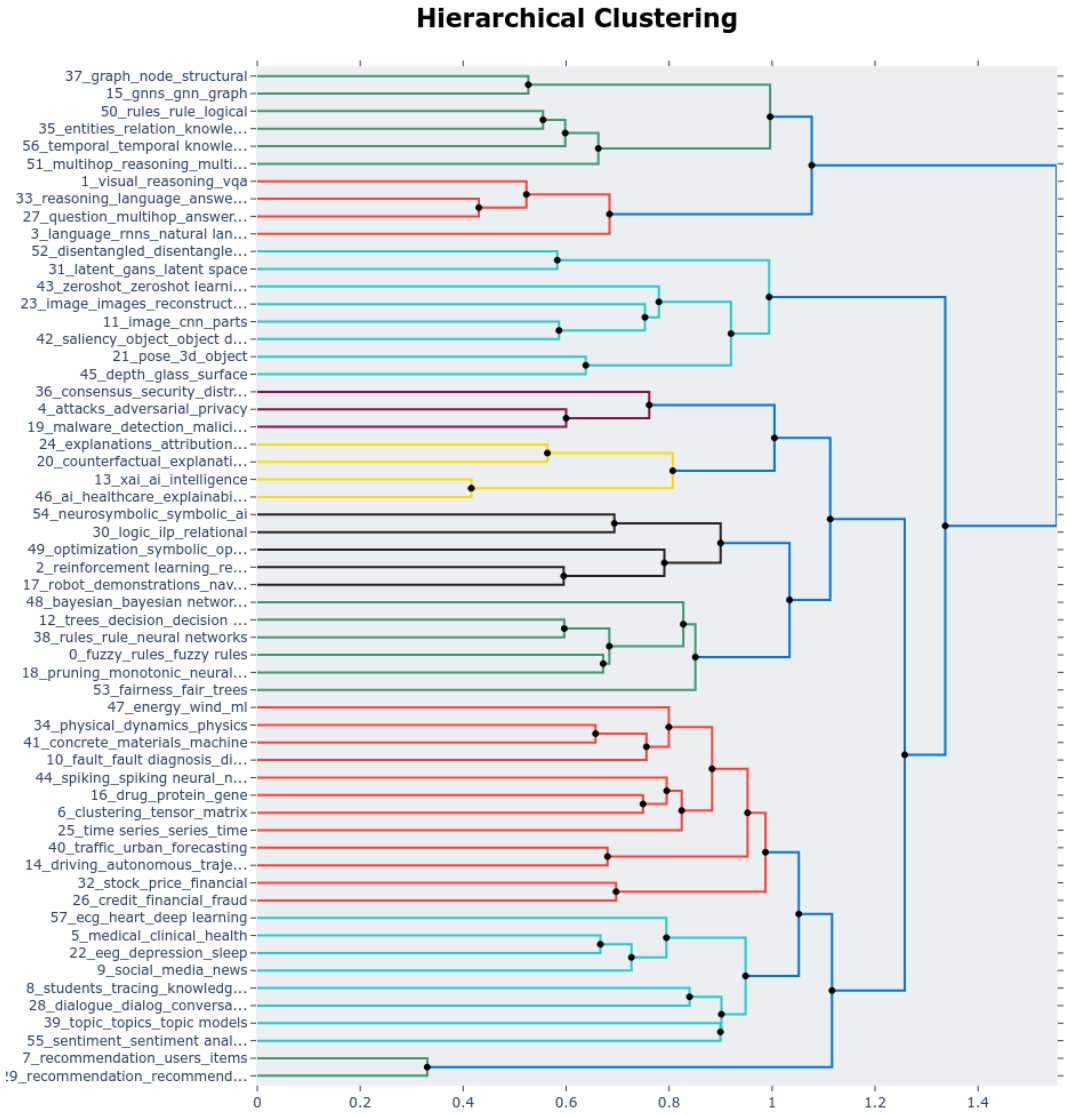

Fig. 4. HDBSCAN hierarchical topic clustering.

Clustering. HDBSCAN with `min_cluster_size=16`, `max_cluster_size=128`, Euclidean metric (BERTopic default workflow). Figure 4 illustrates the hierarchical clustering of all these topics.

Outlier handling. We formed c-TF-IDF centroids and reassigned outliers via a three-step procedure:

(1) centroid mapping (threshold 0.1),
(2) embedding-similarity assignment (threshold 0.7), and
(3) distributional plausibility checks.

The final model yielded 58 topics with a 0% outlier rate.

Topic Labeling. We proposed labels from top c-TF-IDF terms and bigrams, then verified the largest topics manually (human-in-the-loop quality assurance).

A.5.2 Stage 2: Diversity-Aware Representative Selection. Let cluster T contain papers $i \in T$ with embedding vectors $\mathbf{e}_i$ and centroid $\mathbf{c}_T$.

Centrality (topic-core representativeness). We define:

$$\text{centrality}_i \;=\; \cos\big(\mathbf{e}_i, \mathbf{c}_T\big).$$

Diversity (within-topic uniqueness). We define:

$$\text{diversity}_i \;=\; 1 - \max_{j \in T, j \neq i} \cos(\mathbf{e}_i, \mathbf{e}_j),$$

and min-max normalize both metrics to $[0, 1]$ per topic.

Representativeness score. We combine centrality and diversity with a coverage-first weight:

$$\text{rep\_score}_i \;=\; 0.55 \cdot \text{diversity}_i \;+\; 0.45 \cdot \text{centrality}_i.$$

To select representatives, the weight $\alpha = 0.55$ was determined through internal validation on a hold-out set of 10 manually annotated "ground truth" key papers identified by the authors; the papers are: (Barbiero, Ciravegna, et al. 2022; C. Chang et al. 2023; Fischer and Vreeken 2021; Y. Jiang and Bansal 2019; S. Liu et al. 2024; Mao et al. 2023; Nauta et al. 2021; Peng et al. 2022; J. Xu et al. 2018; Q. Zhang et al. 2019). We tested $\alpha \in \{0.4, 0.45, 0.5, 0.55, 0.6\}$ and selected the value maximizing F1-score against expert judgments. The 0.55 weight prioritizes diversity slightly while preserving centrality, acknowledging that representative papers should avoid redundancy while capturing the semantic core of a topic.

Selection procedure. Within each topic we rank by rep_score and select under a duplicate-avoidance rule (cosine similarity threshold > 0.95). Quotas scale with topic size using a 15% target selection ratio. From 3,749 screened reports, we selected 535 representatives across 58 topics (overall ratio 14.27%).

Centroid vs. medoid selection. The high Spearman correlation (0.8707) indicates that both methods are highly consistent in how they rank papers by "centrality." Papers near the theoretical centroid are almost always those also identified as medoids or near-medoids. The moderate overlap (59.7%) in the top-5 selection suggests that while the "global" ranking is stable, the specific "local" choice of the top-1 or top-5 representatives can vary depending on whether one prioritizes the mathematical mean or an existing exemplar. This comparison validates the use of centroid-based selection in our pipeline. Since papers highly similar to the centroid are also those identified by the medoid approach, using centroids as a reference point successfully extracts papers that are representative in our terms.

Research-axis alignment (diagnostic only). To audit topical balance we compute prototype-similarity scores for three axes—XAI, symbolic, and sub-symbolic—by averaging the embeddings of curated term lists and taking cosine similarity per paper. We min–max normalize the scores to $[0, 1]$ corpus-wide. We do not use these scores for selection to avoid bias.

A.5.3 Stage 3: Statistical Validation and Bias Assessment. We compare distributions of selected vs. non-selected papers for six variables: similarity-to-centroid, diversity score, representativeness score, and the three alignment scores:

- *Hypothesis tests:* Two-sided Mann–Whitney U tests with Benjamini–Hochberg correction over six tests ($\alpha = 0.05$).
- *Effect sizes.* Cohen's d with 95% bootstrap CIs (10,000 resamples).
- *Temporal bias.* Kolmogorov–Smirnov (KS) test on publication-year distributions.

Results. The analysis confirms the selection method's effectiveness:

Variable	p-value	Significative?	Cohen's d	Magnitude	95% CI
Diversity Score	< 0.001	**Yes**	0.573	Medium	[0.480, 0.665]
Representativeness	< 0.001	**Yes**	0.636	Medium	[0.543, 0.729]
Similarity to Centroid	0.749	No	0.033	Negligible	[−0.059, 0.124]
XAI Alignment	0.694	No	−0.014	Negligible	[−0.105, 0.078]
Symbolic Alignment	0.966	No	0.012	Negligible	[−0.080, 0.103]
Sub-symbolic Alignment	0.440	No	−0.040	Negligible	[−0.132, 0.051]

Temporal bias. The KS test gives $p = 0.206$, indicating no significant temporal bias. Thus, the selected papers are temporally representative across 2017—2025.

Interpretation. Medium effect sizes for diversity and representativeness indicate practically meaningful selection differences. Non-significant alignment differences show balanced coverage across XAI approaches. Figures 5 and 6 report the metric distribution for all papers vs. selected representatives and statistical validation, respectively.

A.6 Data Extraction and Coding

We used a pre-piloted standardized form and extracted:

(1) citation metadata;
(2) study objectives and research questions;
(3) experimental design (datasets, architectures, training/evaluation);
(4) analysis methods and statistical approaches;
(5) explanation types—input modalities and output formats;
(6) scope (local vs. global);
(7) integration (ante-hoc vs. post-hoc; model-specific vs. model-agnostic);
(8) key findings;
(9) methodological quality (internal/external validity, stated limitations); and
(10) future research directions.

Two authors piloted the schema on 30 papers and refined definitions by consensus; disagreements during full extraction were resolved with the third author.

A.7 Quality Appraisal and Risk of Bias

We employed the ROBIS tool (Whiting and al. 2016). Phase 2 domains—eligibility criteria; identification/selection; data collection/appraisal; synthesis/findings—map well onto our workflow. We provide completed signaling-question answers, domain judgments, and overall risk ratings in the supplementary repository.[2]

Mitigations. Because ROBIS targets intervention/diagnostic reviews, we complemented it with structured internal/external validity checks recorded in the extraction sheet. Our SCImago/CORE filter increases methodological

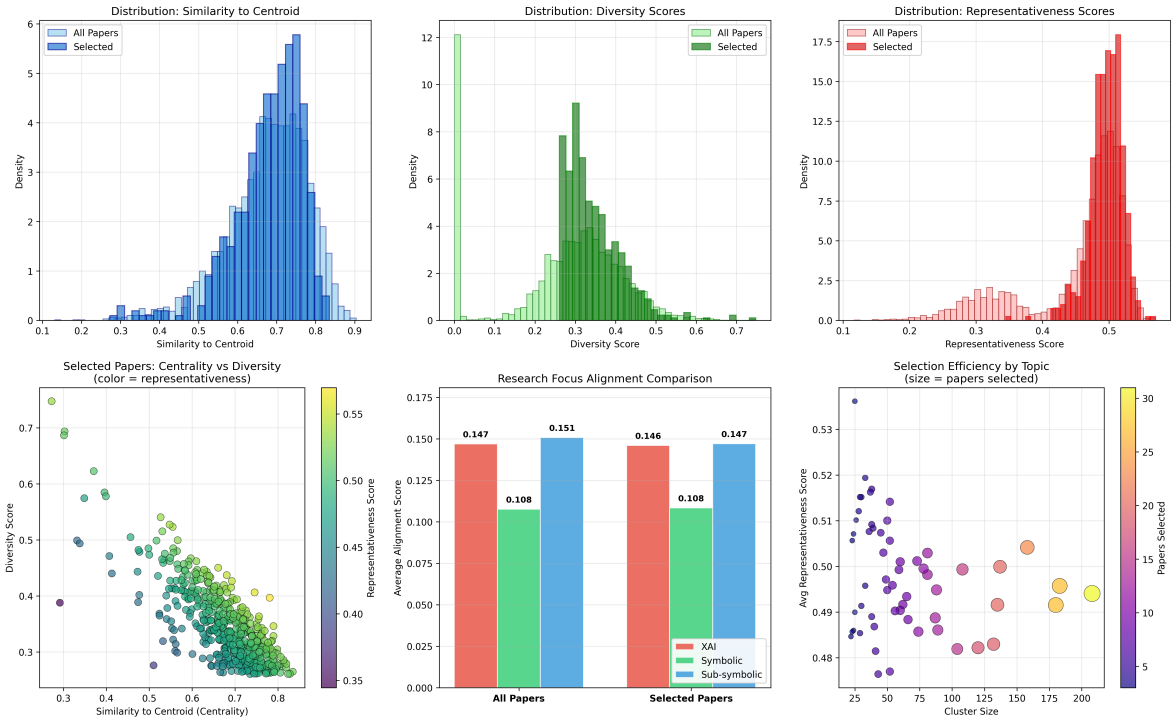

Fig. 5. Metric distributions for all papers vs. selected representatives. Top: histograms for similarity to centroid, diversity, and representativeness. Bottom left: centrality–diversity scatter for selected papers (color = representativeness). Bottom center: research focus—XAI, symbolic, and sub-symbolic—alignment comparison. Bottom right: topic-wise selection efficiency.

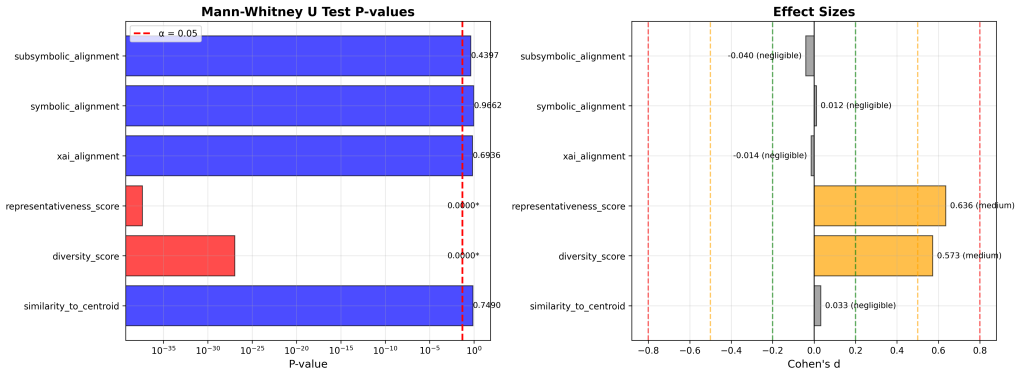

Fig. 6. Statistical validation. Left: Mann–Whitney U p-values (log scale, dashed line at $\alpha = 0.05$). Right: Cohen's d effect sizes with conventional thresholds (0.2/0.5/0.8).

reliability; to evaluate potential exclusion bias, we ran a sensitivity analysis without the venue filter and found no change in qualitative conclusions.

A.8 Reproducibility, Configuration, and Environment

Configuration as code. All parameters live in YAML files tracked by version control. A typical configuration snippet:

```
stage_1:
  embedding_model: {name: all-MiniLM-L6-v2, batch_size: 64}
  umap_params: {n_components: 256, n_neighbors: 15, min_dist: 0.0, metric: cosine}
  hdbscan_params: {min_cluster_size: 16, max_cluster_size: 128, metric: euclidean}
  outlier_reduction: {enabled: true, strategies: [c-tf-idf, embeddings, distributions]}

stage_2:
  metrics: {diversity_weight: 0.55}
  selection_strategy: {method: diverse_representative, target_selection_ratio: 0.15}

stage_3:
  statistical_analysis:
    tests:
      mann_whitney_u: {alpha: 0.05, alternative: two-sided}
      cohens_d: {confidence_level: 0.95}
      temporal_analysis: {test: kolmogorov_smirnov}
    multiple_testing_correction: {method: fdr_bh, alpha: 0.05}
pipeline: {reproducibility: {random_seed: 42}}
```

Determinism and audit trail. We fix random seeds (`seed=42`); log timestamps, package versions, and hyperparameters; and persist intermediate artifacts (embeddings, UMAP projections, cluster assignments, per-paper metrics, selection sets, and statistical outputs). All quantitative metrics are normalized to $[0, 1]$.

Software stack. Python $\geq$ 3.11; `sentence-transformers`, `umap-learn`, `hdbscan`, `bertopic`, `numpy`, `scipy`, `scikit-learn`, `statsmodels`, `pandas`, `matplotlib`/`seaborn`, `PyYAML`. We provide a `pyproject.toml` for exact replication in the supplementary repository.[2]

A.9 Human Oversight and Ethics

No human subjects were involved. We used bibliographic exports and publicly available PDFs. Two authors reviewed topic labels and edge cases (e.g., borderline symbolic content); disagreements were resolved by discussion with the third author.

A.10 Limitations (Methodological)

Our database selection may under-cover specialized venues; the English-only filter may under-represent non-English communities; and the 2017–2025 window omits earlier foundations by design. The venue-quality filter may exclude innovative early-stage work; our sensitivity check suggests this does not affect qualitative conclusions. Finally, while we minimized algorithmic bias via non-parametric tests and temporal checks, residual corpus bias (e.g., field-specific publication practices) is possible.

To address concrete omissions surfaced during peer review without altering the reproducible database pipeline, we additionally report an expert-augmented identification stream using PRISMA 2020's "Identification of studies via other methods" branch (Page et al. 2021). This expert-augmented stream improves topical completeness but should not be conflated with systematic citation-based retrieval: expert-augmented identification is compatible with, but not equivalent to, systematic backward/forward snowballing. Following established guidelines, future

work can extend the "other methods" branch into *structured* citation chasing (backward and forward snowballing) from an explicit seed set (e.g., topic medoids and key surveys) (Wohlin 2014). In this revision, we restrict the "other methods" stream to expert input (peer review) to address reviewer-identified gaps while maintaining a crisp separation between the frozen database pipeline and the additive augmentation step.

A.11 Outputs Provided with the Paper

We release:

(1) YAML configuration files;
(2) two BIB files containing the 3,749 records sought for retrieval and the 11 records included in the expert-identified corpus (during peer review);
(3) (reproducible) topic assignments and per-topic statistics;
(4) per-paper metrics (centrality, diversity, representativeness, alignment) for all 3749 papers;
(5) full statistical reports (test statistics, FDR-corrected p-values, Cohen's d with 95% CIs, KS results); and
(6) the complete pipeline and plotting code.

The supplementary repository[2] contains all these artifacts.

Summary. Our pipeline balances semantic coverage (topic induction), diversity (dispersion within topics), and typicality (centrality to topic cores). Statistical validation shows that selected representatives are significantly more diverse and representative while remaining unbiased with respect to alignment axes and publication years.

Received 28 February 2026; accepted 04 June 2026