NEXUS: A Neural-Symbolic Architecture for Building Aligned Superintelligent Systems

H. P. Alesso
CTO, AI HIVE
Pleasanton, USA
info@ai-hive.net
Founders Circle, MIT
Cambridge, USA
alesso@alum.mit.edu

March 8, 2025

Abstract

As artificial intelligence systems advance toward superintelligence, ensuring their alignment with human values and goals becomes increasingly critical. Current deep learning approaches, while powerful, operate as "black boxes" with limited interpretability, while purely symbolic systems lack the adaptability needed for complex tasks. This paper introduces NEXUS (Neural-symbolic EXtensible Unified System), a novel architecture that integrates neural networks with symbolic reasoning to create systems with both high performance and transparent decision-making processes. We establish the theoretical foundations for this integration, propose a practical framework for implementation, and demonstrate its effectiveness through a medical case study of COVID-19 severity assessment. Results show that NEXUS leverages the complementary strengths of neural and symbolic components, outperforming either approach in isolation, particularly in cases requiring nuanced reasoning. We argue that this neural-symbolic integration offers a promising path toward superintelligence that maintains interpretability, correctness, and alignment with human values.

Keywords: neural-symbolic integration, superintelligence, alignment, interpretable AI, metacognitive control

1 Introduction

The development of artificial general intelligence (AGI) and potentially artificial superintelligence (ASI) represents one of humanity's most ambitious technological pursuits. Defined as "AI that surpasses human intelligence in all tasks and domains with exceptional thinking skills" [1], ASI would fundamentally transform human society. However, this advancement comes with profound challenges regarding system alignment and interpretability [2, 3].

Current deep learning approaches have achieved remarkable performance in domains ranging from natural language processing [4] to protein folding [5], but they operate as opaque "black boxes" with limited interpretability [6]. Their reasoning processes remain obscure, making it difficult to verify alignment with human values or diagnose failures. Conversely, traditional symbolic AI methods offer transparent reasoning through explicit rules and knowledge structures but struggle with the flexibility and pattern recognition capabilities needed for complex real-world tasks [7].

As Bengio noted in his seminal 2019 NeurIPS presentation [8], progress toward artificial general intelligence requires a transition from "System 1" thinking (intuitive, fast, unconscious cognitive processes) to "System 2" thinking (logical, deliberate, conscious cognitive processes). While current transformer-based large language models (LLMs) implement some aspects of System 2 thinking through techniques like chain-of-thought prompting [9], these approaches lack robustness and fail to fundamentally address the interpretability challenges inherent in neural architectures.

This paper introduces NEXUS (Neural-symbolic EXtensible Unified System), a novel architecture that integrates neural networks with symbolic reasoning to create systems with both high performance and transparent decision-making processes. NEXUS aims to combine the complementary strengths of both paradigms: the pattern recognition and learning capabilities of neural networks with the logical precision and interpretability of symbolic systems. This integration occurs during the model-building process rather than at a test-time compute, resulting in a unified system capable of both statistical and logical reasoning.

Our key contributions include:

1. A formal framework for bidirectional translation between neural and symbolic representations

2. A metacognitive control mechanism that dynamically determines when to rely on neural versus symbolic components

3. A practical implementation of the NEXUS architecture in a medical decision-support system for COVID-19 severity assessment

4. Empirical evidence that this neural-symbolic integration outperforms either approach in isolation, particularly in cases requiring nuanced reasoning and domain knowledge

We propose that the NEXUS architecture represents a promising path toward the development of aligned superintelligent systems that maintain interpretability while maximizing performance.

2 Related Work

2.1 Neural-Symbolic Integration

Neural-symbolic integration has a rich history dating back to the early days of AI. Smolensky's tensor product representations [10] and Shastri's SHRUTI system [11] were early attempts to combine connectionist models with symbolic reasoning. More recently, several approaches have emerged to bridge the gap between neural and symbolic paradigms.

Garcez et al. [12] categorize neural-symbolic integration methods into three primary approaches:

1. Symbol for Neural: Incorporating symbolic knowledge into neural systems

2. Neural for Symbol: Enhancing symbolic reasoning with neural networks

3. Hybrid Integration: Creating systems where neural and symbolic components operate in tandem

Within the "Symbol for Neural" category, knowledge graphs have played a central role. Chen et al. [13] demonstrated how knowledge graphs can enhance large language models by providing structured domain knowledge.

Similarly, Wang et al. [14] introduced a knowledge graph attention network for recommender systems that leverages symbolic relationships to guide neural attention mechanisms.

In the "Neural for Symbol" domain, Zhang et al. [15] developed variational reasoning networks that use neural networks to accelerate knowledge graph reasoning. Similarly, Qu and Tang [16] proposed probabilistic logic neural networks that combine statistical and logical reasoning.

Hybrid approaches include DeepProbLog [17], which integrates neural networks with probabilistic logic programming, and Neuro-Symbolic Concept Learner [18], which combines perception with symbolic program synthesis.

Our work builds upon these foundations while introducing novel mechanisms for bidirectional translation and metacognitive control, creating a more deeply integrated neural-symbolic system aimed at superintelligent capabilities.

2.2 Explainable AI and Interpretability

The opacity of deep neural networks has led to significant research in explainable AI (XAI) [19]. Techniques like LIME [20] and SHAP [21] provide post-hoc explanations of model predictions, while approaches like attention visualization [22] attempt to reveal the internal workings of neural architectures.

However, these methods have limitations. They often provide simplifications that fail to capture the full complexity of model decision-making [23], and they remain vulnerable to potential "explanation hacking" where models can produce misleading explanations [24].

Neural-symbolic approaches offer an alternative path to explainability by integrating interpretable symbolic reasoning directly into the system architecture [25]. Rather than explaining an opaque process after the fact, these systems incorporate transparency into their design. Our NEXUS architecture extends this principle by ensuring that symbolic reasoning occurs at every stage of decision-making, providing inherent interpretability.

2.3 AI Alignment

As AI systems grow more capable, ensuring their alignment with human values becomes increasingly critical [26]. Approaches to alignment include

reinforcement learning from human feedback (RLHF) [27], constitutional AI [28], and formal verification methods [29].

However, alignment becomes particularly challenging as systems approach superintelligence, when they may develop capabilities that exceed human understanding or control [30]. Christiano et al. [31] proposed scalable supervision methods like amplification and distillation, while Burns et al. [32] explored weak-to-strong generalization for aligning more capable systems.

Neural-symbolic integration offers unique advantages for alignment by making reasoning processes transparent and incorporating explicit human knowledge [33]. Our work extends this by introducing metacognitive control mechanisms that dynamically adjust the influence of neural and symbolic components based on their respective confidences, creating an auditable decision-making process aligned with human values and goals.

3 The NEXUS Architecture

NEXUS is a neural-symbolic architecture designed to combine the complementary strengths of neural networks and symbolic reasoning. The core innovation lies in its bidirectional integration mechanism and metacognitive control system.

3.1 Formal Framework

We define the NEXUS architecture in terms of the following components:

1. **Neural Component:** A function
$$f_\theta : X \to Z$$
that maps inputs $x \in X$ to latent representations $z \in Z$, parameterized by θ.

2. **Symbolic Component:** Consists of a knowledge base
$$KB = \{(e_i, r_j, e_k)\}$$
representing entities and their relationships, a logical reasoning system $\mathcal{L}$ with inference rules $\mathcal{R}$, and an inference function
$$I(KB, q, \mathcal{R}) \to a$$
that derives answers from the knowledge base.

3. **Integration Mechanism:** Provides bidirectional translation between neural and symbolic representations.

$$T_{n\to s} : Z \to KB \quad \text{and} \quad T_{s\to n} : KB \to Z$$

   - *Neural-to-Symbolic Translation* $T_{n\to s}(z)$ maps neural representations to symbolic knowledge.
   - *Symbolic-to-Neural Embedding* $T_{s\to n}(KB)$ embeds symbolic knowledge into neural space.

4. **Joint Reasoning Module:**

$$J(z, KB, q) \to (z', KB', a)$$

   updates both neural and symbolic representations and derives answers.

   - Neural reasoning:

$$z' = f_\theta\big(z,\ T_{s\to n}(KB)\big)$$

   - Symbolic reasoning:

$$KB' = KB \cup T_{n\to s}(z')$$

   - Answer derivation:

$$a = I(KB',\ q,\ \mathcal{R})$$

5. **Metacognitive Control Function:**

$$M(z, KB, q, \lambda) \to (\omega_n, \omega_s)$$

   determines how much to rely on neural versus symbolic reasoning based on uncertainty estimates λ.

The key innovation in this framework is the bidirectional flow of information between neural and symbolic components, governed by metacognitive control. This allows the system to leverage the strengths of each approach while compensating for their weaknesses.

3.2 Neural Component

The neural component can be implemented using various architectures depending on the task domain. For natural language processing, transformer-based models [34] offer state-of-the-art performance. For healthcare applications, architectures like multilayer perceptrons or recurrent neural networks may be more appropriate.

In the transformer-based implementation, the neural component includes:

- *Input embedding:* $E(x) = W_e x$
- *Self-attention mechanism:*

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{QK^T}{\sqrt{d_k}}\Big)V$$

- *Multi-head attention:*

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

The neural component processes raw inputs and extracts features, generating both predictions and latent representations that capture the statistical patterns in the data.

3.3 Symbolic Component

The symbolic component consists of a knowledge base and reasoning system. The knowledge base stores entities, relationships, and rules in a structured format. For knowledge representation, we use a graph structure with nodes representing entities and edges representing relationships.

The reasoning system implements both forward and backward chaining algorithms:

- *Forward chaining* applies rules to derive new facts until a goal is reached.
- *Backward chaining* starts with a goal and works backward to determine if it can be proven.

The symbolic component provides transparency by making its reasoning process explicit and auditable. It encodes domain knowledge in a form that can guide neural processing and verify its outputs.

3.4 Integration Mechanism

The integration mechanism is the heart of the NEXUS architecture, facilitating bidirectional translation between neural and symbolic representations. This translation occurs through:

1. **Neural-to-Symbolic Translation:** Maps neural representations to symbolic entities and relationships. This can be implemented through various methods:
   - *Threshold-based extraction:*
   $$T_{n \to s}(z) = \{(e_i, r_j, e_k) \mid P(e_i, r_j, e_k \mid z) > \tau\}$$
   - *Attention-based mapping:* Using attention weights to identify relevant symbolic concepts
   - *Graph neural networks:* Learning to map between neural and graph representations
2. **Symbolic-to-Neural Embedding:** Embeds symbolic knowledge into the neural space. This includes:
   - *Entity embedding:* $\phi(e) = W_e e$
   - *Relation embedding:* $\phi(r) = W_r r$
   - *Triple embedding:* $\phi(e_i, r_j, e_k) = g(\phi(e_i), \phi(r_j), \phi(e_k))$

The integration mechanism ensures that information flows seamlessly between neural and symbolic components, allowing each to enhance and constrain the other.

3.5 Metacognitive Control

The metacognitive control system dynamically determines when to rely on neural versus symbolic reasoning based on:

- *Uncertainty estimation:* Both components provide confidence metrics for their outputs.
- *Domain recognition:* The system identifies which component has expertise in the current domain.

- *Task complexity:* Some tasks are better suited to neural pattern recognition, others to symbolic reasoning.

We implement metacognitive control using a simple decision rule:

- If neural confidence is high ($c_n > \tau_n$) and symbolic confidence is low ($c_s < \tau_s$), rely on neural processing.
- If symbolic confidence is high ($c_s > \tau_s$) and neural confidence is low ($c_n < \tau_n$), rely on symbolic reasoning.
- Otherwise, use a weighted combination of both:

$$a = \alpha\, a_n + (1 - \alpha)\, a_s,$$

where α is determined by relative confidences.

This mechanism allows NEXUS to adapt its reasoning approach based on the specific demands of each task, creating a flexible and robust system.

4 Implementation and Case Study: COVID-19 Severity Assessment

To demonstrate the effectiveness of the NEXUS architecture, we implemented a medical decision-support system for COVID-19 severity assessment. This domain offers an ideal testbed for neural-symbolic integration, as it requires both pattern recognition from patient data and application of medical knowledge rules.

4.1 System Implementation

The NEXUS implementation consists of the following components:

1. **Neural Component:** A multilayer perceptron classifier trained on patient vital signs and medical history to predict COVID-19 severity (mild, moderate, severe, critical).
2. **Symbolic Component:** A knowledge base containing medical rules for COVID-19 severity assessment, including:

- Symptom thresholds (e.g., fever above 101 °F suggests at least moderate severity)
- Risk multipliers for comorbidities (e.g., immunocompromised status increases risk by 3x)
- Treatment recommendations based on severity

3. **Integration Mechanism:** Bidirectional translation between neural predictions and symbolic medical knowledge.
4. **Metacognitive Control:** A decision system that determines whether to trust neural predictions, symbolic reasoning, or a weighted combination based on confidence levels.

4.2 Experimental Setup

We evaluated the system on synthetic data for five COVID-19 patients with varying severity levels and comorbidities (see GitHub repository):

- **Patient 1:** Mild case with minimal symptoms, no comorbidities
- **Patient 2:** Moderate case with fever and moderate symptoms, one comorbidity
- **Patient 3:** Severe case with high fever, low oxygen, and two comorbidities
- **Patient 4:** Critical case with severe symptoms and multiple comorbidities
- **Patient 5:** Moderate symptoms but with multiple high-risk comorbidities

Each patient was represented by a feature vector including:

- Fever temperature (°F)
- Cough severity (1–10)
- Fatigue level (1–10)
- Breathing difficulty (1–10)

- Oxygen saturation (%)
- Binary indicators for comorbidities (hypertension, diabetes, heart disease, lung disease, immunocompromised status)

We compared three approaches:

1. **Neural-only:** Using only the MLP classifier's predictions
2. **Symbolic-only:** Using only the symbolic reasoning system
3. **NEXUS:** The full neural-symbolic architecture with metacognitive control

4.3 Results

Table 1: Severity assessments by different approaches

Patient	Neural	N-Conf	Symbolic	S-Conf	NEXUS	Dominant	Agreement
1	mild	0.92	mild	0.20	mild	Neural	Yes
2	moderate	0.85	moderate	0.30	moderate	Neural	Yes
3	severe	0.78	severe	0.70	severe	Integrated	Yes
4	critical	0.95	critical	0.90	critical	Integrated	Yes
5	moderate	0.65	severe	0.80	severe	Symbolic	No

The treatment recommendations derived from these assessments ranged from basic supportive care for mild cases to intensive care interventions for critical cases:

- **Patient 1 (mild):** rest, hydration, monitoring, acetaminophen
- **Patient 2 (moderate):** monoclonal antibodies, antivirals, close monitoring
- **Patient 3 (severe):** hospitalization, oxygen therapy, dexamethasone, remdesivir
- **Patient 4 (critical):** ICU admission, ventilator, dexamethasone, remdesivir, specialty consult
- **Patient 5 (severe):** hospitalization, oxygen therapy, dexamethasone, remdesivir

4.4 Analysis

The results demonstrate several key advantages of the NEXUS architecture:

1. **Complementary Strengths:** For straightforward cases (Patients 1–2), both components agree, with the neural component showing higher confidence due to its pattern recognition strengths. For complex cases (Patients 3–4), both components contribute to an integrated decision.

2. **Knowledge-Enhanced Decision-Making:** Patient 5 highlights the value of symbolic knowledge. The neural component classified this patient as moderate based on symptom patterns alone, but the symbolic component recognized the high risk from multiple comorbidities and elevated the severity assessment to severe. The metacognitive control system, detecting higher confidence in the symbolic component, favored its assessment, potentially leading to more appropriate care.

3. **Transparent Reasoning:** Unlike a black-box model, NEXUS provides a clear explanation of its decision-making process, showing which component dominated each decision and why. This transparency is crucial for building trust in medical applications.

4. **Adaptability:** The metacognitive control mechanism allows NEXUS to adapt its reasoning approach based on the specific characteristics of each patient, creating a more flexible and robust system.

This case study demonstrates how neural-symbolic integration can enhance decision-making in complex domains, combining the pattern recognition capabilities of neural networks with the explicit knowledge encoding of symbolic systems.

5 Toward Superintelligent Neural-Symbolic Systems

The NEXUS architecture represents a step toward developing aligned superintelligent systems that maintain interpretability while maximizing performance. In this section, we discuss how this approach can be scaled and extended to address the challenges of superintelligence.

5.1 Recursive Self-Improvement

A key capability of superintelligent systems would be recursive self-improvement [35], where the system enhances its own capabilities. The neural-symbolic approach offers advantages for controlled self-improvement:

1. **Verifiable Improvements:** Symbolic components can verify that proposed changes maintain alignment with core values and constraints.

2. **Transparent Reasoning:** The system's reasoning about self-improvement remains interpretable.

3. **Knowledge Integration:** The system can incorporate new knowledge into both neural and symbolic components.

We propose a recursive self-improvement framework where neural components generate potential improvements while symbolic components verify these improvements against safety constraints. This creates a path for controlled advancement while maintaining alignment with human values.

5.2 Multimodal and Multidomain Learning

Superintelligent systems would need to operate across multiple modalities and domains. NEXUS can be extended to support this through:

- *Cross-modal symbolic grounding:* Establishing connections between symbols and their manifestations across different modalities (text, vision, speech, etc.).

- *Domain-specific knowledge bases:* Maintaining separate symbolic knowledge for different domains while sharing a common neural substrate.

- *Meta-learning:* Learning to adapt to new domains quickly by leveraging similarities with known domains.

Knowledge graphs serve as a particularly powerful mechanism for integrating information across modalities and domains, creating a unified representation that can be accessed and manipulated by both neural and symbolic components.

5.3 Alignment and Safety

Ensuring alignment with human values becomes increasingly critical as systems approach superintelligence [36]. The neural-symbolic approach offers unique advantages for alignment:

1. **Explicit Value Representation:** Human values can be encoded explicitly in the symbolic knowledge base.
2. **Verifiable Constraints:** Symbolic reasoning can enforce hard constraints on system behavior.
3. **Auditable Decision-Making:** The system's reasoning process remains transparent and auditable.
4. **Corrigibility:** The system can be designed to accept corrections to both its neural and symbolic components.

By combining the adaptability of neural networks with the precision of symbolic reasoning, NEXUS creates a foundation for systems that can learn from experience while remaining within ethical boundaries.

6 Conclusion and Future Work

This paper has introduced NEXUS, a neural-symbolic architecture that integrates neural networks with symbolic reasoning to create systems with both high performance and transparent decision-making processes. Through a case study in COVID-19 severity assessment, we have demonstrated how this architecture leverages the complementary strengths of neural and symbolic components, outperforming either approach in isolation.

The NEXUS architecture addresses several key challenges in the development of superintelligent systems:

1. **Interpretability:** By incorporating symbolic reasoning at every stage of decision-making, NEXUS provides inherent interpretability.
2. **Alignment:** The system can explicitly represent human values and verify its decisions against these values.
3. **Knowledge Integration:** Both learned patterns and explicit knowledge can be incorporated into the system's reasoning.

4. **Adaptive Decision-Making:** The metacognitive control mechanism allows the system to adapt its reasoning approach based on the specific demands of each task.

While promising, the NEXUS architecture has limitations and opens up several directions for future research:

- **Scaling to Large Models:** Integrating symbolic reasoning with large-scale neural networks presents computational challenges that must be addressed.
- **Learning Symbolic Knowledge:** Developing methods for automatically extracting and refining symbolic knowledge from data.
- **Complex Reasoning Domains:** Extending the architecture to support more complex forms of reasoning, including counterfactual, temporal, and causal reasoning.
- **Multimodal Integration:** Enhancing the system to reason across multiple modalities, including text, vision, and speech.

As AI systems continue to advance toward superintelligence, ensuring their interpretability, alignment, and safety becomes increasingly critical. The neural-symbolic approach represented by NEXUS offers a promising path forward, combining the strengths of neural and symbolic paradigms to create systems that are both powerful and transparent. By continuing to develop and refine these architectures, we can work toward superintelligent systems that augment human capabilities while remaining aligned with human values.

References

[1] Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. doi:10.1093/oso/9780199678112.001.0001

[2] Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking. ISBN: 978-0525558613

[3] Dafoe, A. (2018). AI governance: A research agenda. Governance of AI Program, Future of Humanity Institute, University of Oxford. Retrieved from `https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf`

[4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. doi:10.48550/arXiv.2005.14165

[5] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. doi:10.1038/s41586-021-03819-2

[6] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. doi:10.1038/s42256-019-0048-x

[7] Hitzler, P., Bianchi, F., Ebrahimi, M., & Sarker, M. K. (2020). Neural-symbolic integration and the Semantic Web. *Semantic Web*, 11(1), 3–11. doi:10.3233/SW-190368

[8] Bengio, Y. (2019). From System 1 Deep Learning to System 2 Deep Learning. NeurIPS 2019 Keynote. Retrieved from `https://slideslive.com/38921750/from-system-1-deep-learning-to-system-2-deep-learning`

[9] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837. doi:10.48550/arXiv.2201.11903

[10] Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1-2), 159–216. doi:10.1016/0004-3702(90)90007-M

[11] Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16(3), 417–451. doi:10.1017/S0140525X00030910

[12] Garcez, A. D., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *Jour-*

nal of Applied Logics, 6(4), 611–632. Retrieved from https://www.collegepublications.co.uk/downloads/ifcolog00026.pdf

[13] Chen, X., Liang, S., Ding, N., Chen, M., Xiao, C., Li, Y., Song, L., Zaniolo, C., & Sun, Y. (2022). KagNet: Knowledge-Aware Graph Networks for commonsense reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5734–5746. doi:10.18653/v1/2022.emnlp-main.390

[14] Wang, X., He, X., Cao, Y., Liu, M., & Chua, T. S. (2019). KGAT: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 950–958. doi:10.1145/3292500.3330989

[15] Zhang, Y., Dai, H., Kozareva, Z., Smola, A. J., & Song, L. (2018). Variational reasoning for question answering with knowledge graph. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 6069–6076. doi:10.1609/aaai.v32i1.11627

[16] Qu, M., & Tang, J. (2019). Probabilistic logic neural networks for reasoning. In *Advances in Neural Information Processing Systems* (NeurIPS 2019), pages 7712–7722. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2019/file/13e5ebb0fa112fe1b31a1067962d74a7-Paper.pdf

[17] Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). DeepProbLog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems* (NeurIPS 2018), pages 3749–3759. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2018/file/dc5d637ed5e62c36ecb73b654b05ba2a-Paper.pdf

[18] Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *7th International Conference on Learning Representations* (ICLR 2019). Retrieved from https://openreview.net/forum?id=rJgMlhRctm

[19] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D.,

Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. doi:10.1016/j.inffus.2019.12.012

[20] Ribeiro, M. T., Singh, S., & Guestrin, C. (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*, pages 1135–1144. doi:10.1145/2939672.2939778

[21] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (NIPS 2017), pages 4765–4774. Retrieved from `https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf`

[22] Vig, J. (2019). A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42. doi:10.18653/v1/P19-3007

[23] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. doi:10.1038/s42256-019-0048-x

[24] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*, pages 180–186. doi:10.1145/3375627.3375830

[25] Bennetot, A., Laurent, J. L., Chatila, R., & Díaz-Rodríguez, N. (2019). Towards explainable neural-symbolic visual reasoning. In *NeSy Workshop at IJCAI 2019*. doi:10.48550/arXiv.1909.09065

[26] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *9th International Conference on Learning Representations* (ICLR 2021). Retrieved from `https://openreview.net/forum?id=d7KBjmI3GmQ`

[27] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* (NIPS 2017), pages 4299–4307. Retrieved from `https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf`

[28] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ...& Irving, G. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*. doi:10.48550/arXiv.2212.08073

[29] Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). Reluplex: An efficient SMT solver for verifying deep neural networks. In *Proceedings of the 29th International Conference on Computer Aided Verification* (CAV 2017), pages 97–117. doi:10.1007/978-3-319-63387-9_5

[30] Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. doi:10.1093/oso/9780199678112.001.0001

[31] Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*. doi:10.48550/arXiv.1810.08575

[32] Burns, C., Ye, N., Landolfi, N., Kaplan, J., Openai, B., Li, X., Henighan, T., Orvos, J., & Steinhardt, J. (2023). Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*. doi:10.48550/arXiv.2312.09390

[33] Garcez, A. D., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd wave. *arXiv preprint arXiv:2012.05876*. doi:10.48550/arXiv.2012.05876

[34] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (NIPS 2017), pages 5998–6008. Retrieved

from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[35] Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 184–220). Oxford University Press. Retrieved from https://intelligence.org/files/AIPosNegFactor.pdf

[36] Demski, A., & Garrabrant, S. (2019). Embedded agency. *arXiv preprint arXiv:1902.09469*. doi:10.48550/arXiv.1902.09469

[37] Ding, M., Zhou, C., Chen, Q., Yang, H., & Tang, J. (2019). Cognitive graph for multi-hop reading comprehension at scale. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2694–2703. doi:10.18653/v1/P19-1259

[38] Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2022). QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1289–1305. doi:10.18653/v1/2022.naacl-main.93

[39] Wang, H., Zhao, M., Xie, X., Li, W., & Guo, M. (2019). Knowledge graph convolutional networks for recommender systems. In *The World Wide Web Conference* (WWW '19), pages 3307–3313. doi:10.1145/3308558.3313417

[40] Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., & Xing, E. P. (2019). Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11487–11496. doi:10.1109/CVPR.2019.01175

[41] Liu, M., Stella, X. Y., Liu, Y., Zhang, H., & Tao, D. (2023). Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *arXiv preprint arXiv:2305.18395*. doi:10.48550/arXiv.2305.18395

[42] Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan,

Y., Liu, Y., Liu, Z., Luo, Z., Qiu, C., Ren, S., ... & Zhu, H. (2021). Pre-trained models: Past, present and future. *AI Open*, 2, 225–250. doi:10.1016/j.aiopen.2021.08.002

[43] Balog, K., Oard, D. W., Clarke, C. L., & De Vries, A. P. (2022). Advances in neural information retrieval. *Foundations and Trends in Information Retrieval*, 16(1-2), 1–208. doi:10.1561/1500000079

[44] Liao, X. V., Zhang, Y., Weng, L., Dugan, L., Lee, M. K., Varshney, K. R., Zhang, J. M., & Dutta, S. (2023). AI alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*. doi:10.48550/arXiv.2310.19852