FROM HUMAN INTUITION TO CAUSAL GRAPHS: A DOLCE-BASED HUMAN-ENSEMBLED NEURO-SYMBOLIC ARCHITECTURE

Manojshyaam C J
MS Student, Dept. of Applied Mechanics and Biomedical Engineering
Indian Institute of Technology Madras
am24s055@smail.iitm.ac.in

ABSTRACT

We find ourselves at a peculiar moment: systems that outperform humans on benchmarks yet falter at the genuinely unexpected. This brittleness is not only a data problem but also an architectural one—current AI lacks any principled mechanism to integrate the intuitive, pattern-based cognition that humans deploy when facing novelty. Kahneman's Dual-Process Theory distinguishes System 1 intuition from System 2 deliberation, yet AI architectures remain bifurcated, capturing neither the capacity to let intuition constrain reasoning nor the judgment of what matters to guide why it matters.

We propose a reconciliation through the DOLCE foundational ontology, treating human intuition as **endurant selection**—identifying objects, agents, and events that persist through time—and machine reasoning as **perdurant inference**—the temporal processes and causal dependencies that unfold upon them. Through Vector Symbolic Architectures, we bind these into a unified framework where causal graphs are constructed from human judgment rather than purely learned from correlation, grounding symbols in cognitive categories while constraining the combinatorial explosion of pure symbolic methods. We outline an evaluation protocol on ARC-AGI-2 and present a qualitative case study on task `898e7135`, a multi-step compositional reasoning problem that remains challenging for frontier models. This paper offers the theoretical architecture and proof-of-concept; empirical validation at scale is ongoing. We submit this as a Blue Sky contribution, a meditation on human-machine collaborative reasoning at the intersection of formal ontology, cognitive science, and neuro-symbolic AI.

1 THE ARCHITECTURAL VOID

The last decade has produced AI systems of remarkable, almost uncanny capability. They translate across languages, generate photorealistic imagery, and defeat grandmasters at games of perfect and imperfect information. Yet we have all witnessed the same brittleness: the language model that hallucinates citations with perfect confidence, the vision system that fails to recognize a stop sign spray-painted with graffiti, the recommendation engine that cannot comprehend why a pandemic might alter purchasing patterns. These are not isolated bugs in the training data. They are symptoms of an architectural absence—an AI that reasons without grounding, that correlates without causation, that pattern-matches without understanding *what matters* in a given situation.

The cognitive science literature offers a diagnostic lens. Kahneman's Dual-Process Theory posits two modes of human cognition: System 1, fast and intuitive, pattern-based and associative; and System 2, slow and deliberative, logical and sequential (Kahneman, 2011). Contemporary AI architectures map awkwardly onto this framework. Deep neural networks approximate System 1, capturing statistical regularities through gradient descent, yet they lack the symbolic scaffolding to represent the causal structure underlying those patterns. Symbolic AI, conversely, implements System 2 reasoning through explicit representations and logical inference, yet it remains disconnected from the perceptual grounding required to identify which entities merit reasoning about in the first place.

What neither approach captures—and what we argue is essential for robust AI—is the *collaboration* between intuition and deliberation. When a human encounters a novel situation, System 1 rapidly identifies salient entities and potential causal relationships; System 2 then evaluates, constrains, and constructs explicit causal models. The intuition guides the reasoning; the reasoning corrects the intuition. Current AI has no such feedback loop. It is as if we have built the fuselage of System 2 without the sensory-perceptual apparatus of System 1 to populate it, or the neural correlates of System 1 without the symbolic machinery to organize them.

2 A RECONCILIATION: DOLCE AS COGNITIVE ARCHITECTURE

We propose that a promising path forward lies not in more data or more parameters, but in a fundamental rethinking of how AI systems represent the world—specifically, through the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) (Gangemi et al., 2002). DOLCE is a top-level foundational ontology developed to capture the ontological categories underlying natural language and human cognition, with a clear orientation toward language and cognition rather than purely engineering convenience. Unlike domain-specific ontologies that catalog medical terms or geological classifications, DOLCE provides the *categories of being*—what exists in the world and how we categorize it.

Central to DOLCE is the distinction between **endurants** and **perdurants**. Endurants are particulars that persist through time while maintaining their identity: objects, agents, substances, qualities. Perdurants are particulars that unfold through time, existing only as processes: events, actions, states, accomplishments. The distinction is not merely philosophical; it is cognitively fundamental. When you see a ball rolling down a hill, you perceive simultaneously the endurant (the ball, the hill) and the perdurant (the rolling motion, the descent).

Our key insight is to map this ontological distinction onto the System 1/System 2 divide. We treat **human intuition as endurant selection**—the rapid, pattern-based identification of which objects, agents, and events in a scene are ontologically salient, and thus candidates for causal consideration. We treat **machine reasoning as perdurant inference**—the deliberate construction of causal graphs representing how these endurants interact, transform, and influence one another over time. The human provides the *what*; the machine provides the *how* and *why*.

3 HUMAN-ENSEMBLED CAUSAL ARCHITECTURE

Operationally, we implement this division through a human-in-the-loop architecture grounded in Vector Symbolic Architectures (VSA) (Kanerva, 2009). VSAs provide a computational framework for binding and bundling distributed representations, allowing compositional structure to emerge from high-dimensional vectors.

3.1 FORMALIZATION

Let $\mathcal{X}$ denote the space of perceptual inputs (e.g., images or grids), and let $\mathcal{E}$ denote a finite universe of candidate endurants (objects, agents, and events) detectable in those inputs. For a given input $x \in \mathcal{X}$, we define the *endurant selection* function

$$S : \mathcal{X} \to 2^{\mathcal{E}}, \qquad S(x) = \{e_1, \ldots, e_k\}, \tag{1}$$

where $2^{\mathcal{E}}$ is the power set of $\mathcal{E}$. Intuitively, $S(x)$ returns the subset of entities that the human judges to be ontologically and causally salient in the scene.

Each endurant $e \in \mathcal{E}$ is encoded as a high-dimensional vector $\mathbf{v}(e) \in \mathbb{R}^d$ in a VSA. We assume a binding operation $\otimes : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ (e.g., elementwise multiplication or circular convolution) and a bundling operation $\oplus$ (e.g., vector addition followed by normalization). A simple role-filler structure such as an agent pushing an object is encoded as

$$\mathbf{r}_{\text{push}} = \mathbf{v}(\text{AGENT}) \otimes \mathbf{v}(\text{PUSHES}) \otimes \mathbf{v}(\text{OBJECT}). \tag{2}$$

Given an input x and the selected endurants $S(x)$, the system constructs a set of bound relational descriptors

$$R(x) = \{\mathbf{r}_1, \ldots, \mathbf{r}_m\}, \tag{3}$$

where each $\mathbf{r}_j$ is a bound representation of a potential perdurant involving elements of $S(x)$.

The *perdurant inference* stage then maps these descriptors to a causal graph

$$G(x) = f(R(x)), \tag{4}$$

where $G(x) = (\mathcal{V}, \mathcal{A})$ is a directed graph with node set $\mathcal{V} \subseteq S(x)$ and edge set $\mathcal{A} \subseteq \mathcal{V} \times \mathcal{V}$. Each edge $(e_i, e_j) \in \mathcal{A}$ is labeled with a causal relation type (e.g., CAUSES_MOTION, BLOCKS, SUPPORTS) and is derived from similarity comparisons between bound vectors in $R(x)$ and learned causal templates, in the spirit of structural causal models (Pearl, 2009).

Interventions are modeled by modifying $G(x)$ and propagating effects through the graph. For an intervention $do(e_i =)$ that removes an endurant e_i, we compute a counterfactual prediction by evaluating the modified graph $G'(x)$ with node e_i and its incident edges removed. This gives a concrete semantics for "what would happen if some endurant were removed or altered," while keeping the representational focus on DOLCE-typed endurants and perdurants (Harnad, 1990).

4 Case Study: ARC-AGI-2 Task `898e7135`

ARC-AGI-2 is a recent benchmark designed to be "relatively easy for humans, yet hard, or impossible, for current AI systems" (Chollet et al., 2025; Arc Prize Foundation, 2025; ARC Prize, 2024). Each task consists of a small number of colored grid input–output examples and one or more held-out test inputs that must be transformed according to the same underlying rule (Chollet, 2019). The ARC-AGI-2 technical report highlights task id `898e7135` as a representative example of multi-step compositional reasoning, where several large structured shapes and scattered smaller markers must be transformed into a new configuration following a latent multi-stage program (Arc Prize Foundation, 2025).

Figures 1 and 2 show the official training and test grids for this task as provided in the ARC-AGI-2 repository (Chollet et al., 2025). Frontier models perform poorly on ARC-AGI-2 as a whole (Chollet et al., 2025), and task `898e7135` in particular requires a level of structure discovery and reuse that purely pattern-matching systems struggle to capture. Human solvers, by contrast, reliably solve the task in one or two attempts, often by articulating an informal program over the large shapes and the background.

We sketch how the proposed human-ensembled architecture can be instantiated on this task. The goal here is not to present a full quantitative evaluation, but to show how human intuitive traces can be converted into a DOLCE-typed causal program that generalizes to the test grid in Figure 2.

Algorithm

Algorithm: Human-Ensembled Causal Reasoning on ARC-AGI-2 Task `898e7135`

Input: ARC-AGI-2 JSON specification for task `898e7135`, providing training pairs $\{(X_{\text{in}}^{(t)}, X_{\text{out}}^{(t)})\}_{t=1}^{T}$ as in Figure 1 and test inputs $\{X_{\text{test}}^{(k)}\}$ as in Figure 2; human solution logs for the training pairs (intermediate grids over time). *Output:* Predicted test outputs $\{\hat{X}_{\text{test}}^{(k)}\}$ and a stored causal program G^*.

(i) **Human solving phase (offline).** Human participants solve each training input of task `898e7135` using the ARC-AGI-2 interface. Their interactions produce a sequence of intermediate output grids, from a blank canvas to the final correct output. For each training pair, we align intermediate grids with the original input in Figure 1 and record which regions of the grid are edited at each step, yielding a trajectory of discrete operations (for example, "fill background with color c," "copy shape A into canonical position," "erase residual markers").

(ii) **Endurant discovery and DOLCE typing.** A low-level parser segments each input grid $X_{\text{in}}^{(t)}$ into primitive connected components, producing a set of candidate endurants $\tilde{\mathcal{E}}^{(t)}$. Using the human edit trajectories, we mark as endurants those components that are ever edited or referenced; all others are treated as ontologically inert background. Each selected component is assigned a DOLCE-inspired type label such as OBJECT, MARKER, or BACKGROUND-STRUCTURE, reflecting its role as a persistent

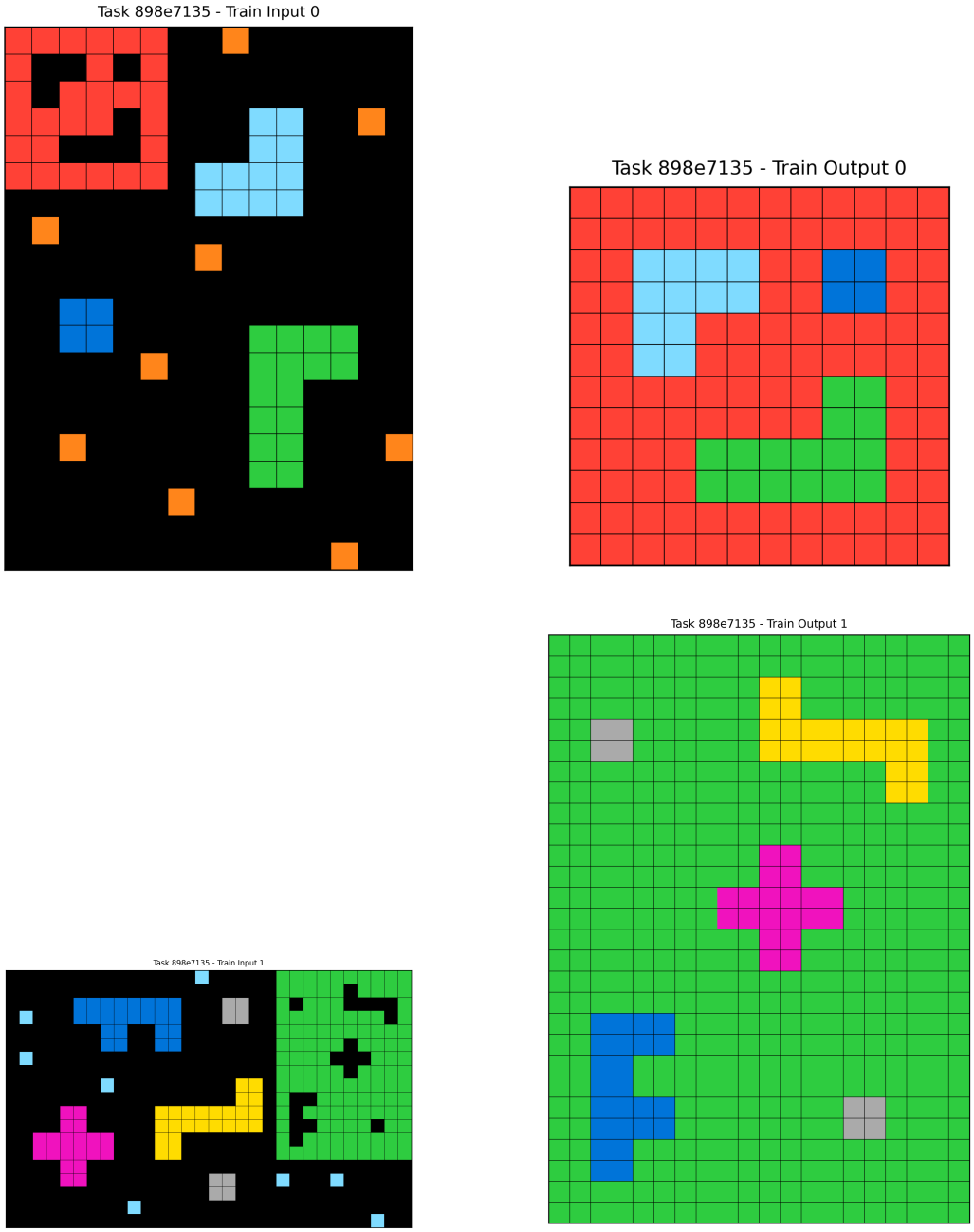

Figure 1: ARC-AGI-2 task `898e7135`: two training input–output pairs. Large structured shapes and scattered smaller blocks are transformed into dense fields with specific colored shapes on a uniform background. Grids are reproduced directly from the official ARC-AGI-2 repository.

entity (Gangemi et al., 2002). This yields a set $S^{(t)} \subseteq \tilde{\mathcal{E}}^{(t)}$ of causally relevant endurants for each training example.

(iii) **Perdurant extraction from human edits.** For each pair of consecutive intermediate grids in a human trajectory, we compute their difference and express it as a transformation applied to the selected endurants. A typical step in task `898e7135` may be summarized as "extend shape e_i along its local axis until it reaches the frame defined by endurant e_j" or "paint the background around all large shapes with a uniform color." Each such transformation is treated as a perdurant candidate and

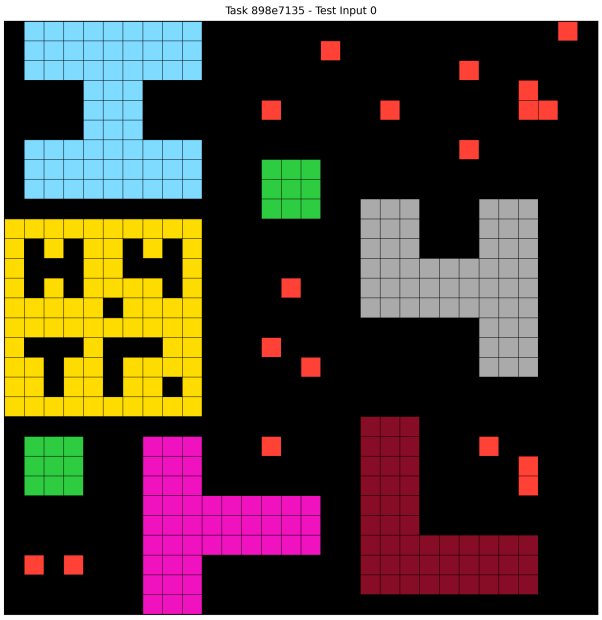

Figure 2: ARC-AGI-2 task `898e7135`: held-out test input, containing a richer composition of large shapes and scattered markers. Our architecture instantiates a human-derived causal program on this grid to generate the corresponding output.

encoded as a VSA binding

$$\mathbf{r} = \mathbf{v}(\text{SOURCE}) \otimes \mathbf{v}(e_i) \otimes \mathbf{v}(\rho) \otimes \mathbf{v}(\text{CONTEXT}(e_j)), \tag{5}$$

where ρ denotes a relation type (e.g., EXTEND, FILL, COPY) and $\text{CONTEXT}(e_j)$ encodes contextual endurants that condition its application.

(iv) **Inducing a causal program G^*.** Across all training pairs and human trajectories, we cluster the perdurant descriptors $\{\mathbf{r}\}$ in VSA space and identify a small set of recurrent transformation motifs that suffice to reproduce the human solutions for `898e7135`. We represent these motifs as a directed causal graph $G^* = (\mathcal{V}, \mathcal{A})$, where nodes $\mathcal{V}$ are DOLCE-typed schemas for endurants (for example, "large colored shape," "small marker") and perdurants (for example, "extend shape along axis," "flood-fill background"), and edges $\mathcal{A}$ encode how the application of one perdurant enables or conditions the next. In effect, G^* is a compact causal program distilled from human behavior on this task.

(v) **Test-time instantiation and execution.** For each test input $X_{\text{test}}^{(k)}$ in Figure 2, we re-run the low-level parser to obtain candidate endurants $\tilde{\mathcal{E}}^{\text{test}}$ and apply an automatic endurant selector trained to imitate the human selections observed in the offline phase, yielding $S^{\text{test}} \subseteq \tilde{\mathcal{E}}^{\text{test}}$. We then match the abstract endurant schemas in G^* to concrete endurants in S^{test} via their DOLCE types and simple visual features (color, shape, size). Once bound, we execute G^* as a sequence of perdurants, applying the learned transformations to the bound endurants and rendering the result as a predicted output grid $\hat{X}_{\text{test}}^{(k)}$.

(vi) **Evaluation and counterfactual probing.** We compare $\hat{X}_{\text{test}}^{(k)}$ to the ground-truth test outputs under the ARC-AGI-2 success criterion (Chollet et al., 2025). Because G^* is explicit and DOLCE-typed, we can also ask counterfactual questions (for example, "What if a particular large shape were absent?" or "What if markers were removed before extension?") by intervening on nodes or edges in G^* and re-executing the program, a capability that is absent from most current ARC-AGI-2 solvers that treat the task as a direct pattern-completion problem.

5 Related Work

Our proposal connects several strands of work that argue for more human-like, causally structured AI. Cognitive scientists have articulated criteria for machines that "learn and think like people," emphasizing intuitive physics, intuitive psychology, causal models, compositionality, and learning-to-learn (Lake et al., 2017). Judea Pearl's theory of causal models and interventions provides the formal backbone for many of these aspirations (Pearl, 2009), while developmental work highlights the early emergence of object-based representations and core knowledge that resemble DOLCE-style ontological categories.

In neuro-symbolic AI, there is a growing body of work that combines neural perception with symbolic reasoning (Garcez and Lamb, 2020; Stammer et al., 2024). The Neuro-Symbolic Concept Learner (NS-CL), for example, builds object-centric scene graphs and executes symbolic programs over them to answer visual questions (Mao et al., 2019). Our architecture is closest in spirit to such systems, but differs in two respects: we import the endurant/perdurant distinction from DOLCE as an explicit ontological backbone, and we treat human intuitive selection as a first-class mechanism for grounding and pruning the causal graph, rather than relying solely on learned perceptual modules.

Cognitive architectures such as ACT-R and Soar model the interaction of declarative and procedural knowledge in human cognition (Anderson et al., 2004; Laird, 2012). More recent work in reinforcement learning from human feedback shows how human preferences can shape complex policies (Christiano et al., 2017), and systems like AlphaGo demonstrate how search and value networks can be combined to achieve superhuman performance in structured domains (Silver et al., 2016). Our contribution is orthogonal: we focus on how human intuitive judgment about *what exists* and *what matters* can be integrated with causal reasoning in open-ended problems like ARC-AGI-2, complementing these architectures and learning paradigms rather than competing with them.

6 Discussion and Future Directions

This case study illustrates how human solution traces on a single challenging ARC-AGI-2 task can be converted into a DOLCE-typed causal program and then reused to solve held-out inputs. We emphasize that this is a qualitative demonstration rather than a full benchmark evaluation: we have not yet measured performance across the entire ARC-AGI-2 suite, nor optimized the learning and execution of G^* under realistic annotation budgets.

Several open questions remain. First, how reliably can automatic endurant selectors imitate human judgment across tasks, and what is the sample complexity of learning such selectors? Second, to what extent can causal programs like G^* transfer across superficially different tasks that share deeper ontological structure (for example, tasks that all involve "extending shapes to a frame")? Third, how should we combine DOLCE-typed graphs with large pretrained models, using the latter as perceptual front-ends or proposal mechanisms while maintaining explicit causal structure for reasoning (Lake et al., 2017; Garcez and Lamb, 2020; Stammer et al., 2024)?

7 Conclusion

We have proposed a cognitive architecture that bridges the gap between intuitive and deliberative reasoning in AI. By treating human judgment as endurant selection and machine inference as perdurant reasoning, grounded in the DOLCE ontology and implemented through Vector Symbolic Architectures, we offer an architectural ingredient that may help AI systems better handle genuinely unexpected situations through collaborative human–machine reasoning.

Our qualitative application to ARC-AGI-2 task `898e7135` sketches one way in which human intuitive traces can be distilled into an explicit, reusable causal program. While much remains to be done, we believe that taking seriously the ontological categories of human cognition—endurants and perdurants, objects and processes—is a promising step toward neuro-symbolic systems that do not merely pattern-match, but ask, guided by human intuition, what matters and why.

REFERENCES

Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Laure Vieu. Sweetening ontologies with DOLCE. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management*, 2002.

Pentti Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2):139–159, 2009.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1–3):335–346, 1990.

Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic AI: The 3rd wave. *arXiv preprint arXiv:2012.05876*, 2020.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2019.

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

François Chollet et al. ARC-AGI-2: A new challenge for frontier AI reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

Arc Prize Foundation. ARC-AGI-2 technical report. Technical report, 2025. Available at `https://arcprize.org/blog/arc-agi-2-technical-report`.

ARC Prize. ARC Prize guide. Online documentation, 2024. Available at `https://arcprize.org/guide`.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision. In *International Conference on Learning Representations (ICLR)*, 2019.

John R. Anderson, Dan Bothell, Christian J. Lebiere, and Michael Matessa. An integrated theory of the mind. *Psychological Review*, 111(4):1036–1060, 2004.

John E. Laird. *The Soar Cognitive Architecture*. MIT Press, 2012.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Gary Marcus and Ernest Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon, 2019.

Wolfgang Stammer, Thomas L. Griffiths, and Artur d'Avila Garcez. Neuro-Symbolic AI: Explainability, Challenges, and Future Directions. *arXiv preprint arXiv:2411.04383*, 2024.