From a Neural-Symbolic Means to the Inductive-Instructional End

Loizos Michael
Open University of Cyprus &
CYENS Center of Excellence
Nicosia, Cyprus
loizos@ouc.ac.cy

ABSTRACT

Debates around neural-symbolic AI have productively explored how different representational substrates may be integrated within a single system. This paper argues that such debates risk mistaking a family of architectures for the underlying objective they serve. We propose a conceptual reframing: the central requirement for open-ended intelligence is not the integration of neural and symbolic representations per se, but the integration of inductive learning from experience with instructional learning from communicated information. Neural-symbolic approaches are best understood as a promising class of means toward this deeper epistemic end, which we characterize as inductive-instructional learnability. Clarifying this distinction helps unify disparate research threads and sharpens the criteria by which future AI systems should be evaluated.

1 INTRODUCTION

Neural-symbolic AI (NSAI) has emerged as a focal point for research aimed at overcoming the limitations of purely connectionist or purely symbolic systems. By combining subsymbolic learning with explicit symbolic reasoning, NSAI promises both robustness and interpretability. Unsurprisingly, much of the literature has focused on architectural questions: how representations should be structured, how information should flow between components, and how learning and reasoning modules should be composed.

This paper advances a complementary, underlying perspective. The integration that ultimately matters is not chiefly architectural but epistemic. What marks systems capable of open-ended intelligence is their ability to reconcile two modes of acquiring knowledge: *(i)* induction from experience, and *(ii)* instruction from external sources. Neural-symbolic architectures, we argue, are best viewed as tools for realizing this deeper integration, not as its definition.

By making this epistemic objective explicit, the paper aims to clarify what is at stake in current debates on open-ended artificial intelligence and to provide a more precise criterion for evaluating progress. We begin by characterizing the dominant architectural framing of NSAI and the way in which it treats representational integration as the primary locus of progress. We then make explicit the epistemic capability such architectures are implicitly intended to support: the integration of inductive learning from experience with instructional learning from communicated information.

Building on this distinction, we situate the inductive-instructional perspective within a broader intellectual context, showing both that instruction has long been recognized as a distinct mode of learning in AI, and that the integration of instruction and induction is a defining feature of human intelligence. We further argue that architectural choices alone do not determine whether this integration is achieved, and that many socially salient properties of AI systems, including explainability and contestability, are best understood as consequences of inductive-instructional integration rather than as architectural add-ons. We conclude by discussing the implications of this perspective for how progress in AI should be evaluated.

2 THE ARCHITECTURAL MEANS OF NSAI

Neural-symbolic AI has come to serve as a dominant framework for thinking about the integration of learning and reasoning in AI systems. Motivated by the complementary strengths and weaknesses of connectionist-only and symbolic-only approaches, NSAI seeks to combine subsymbolic learning mechanisms with explicit symbolic representations and inference. Within this framing, neural components are typically associated with robustness, generalization, and learning from data, while symbolic components are associated with abstraction, compositionality, and structured reasoning.

Much of the NSAI literature is organized around architectural questions. Core concerns include where symbolic representations should reside within a system, how neural and symbolic components should interact, how information should flow between them, and which parts of the system should be trainable or fixed. Progress is commonly assessed in terms of architectural expressiveness, scalability, interpretability, or task performance, and comparisons between systems are frequently cast in terms of alternative designs.

This architectural focus is made especially explicit in influential attempts to systematize the neural-symbolic design space. A prominent example is provided by Kautz [1], who offers a taxonomy of neural-symbolic systems based on how neural and symbolic components are arranged and interact within a system. Such taxonomies distinguish, among others, systems in which symbolic knowledge is translated into neural representations, systems in which neural outputs are lifted into symbolic form, hybrid architectures with parallel components that can invoke each other, and systems in which symbolic structure is compiled into subsymbolic models.

These distinctions classify systems according to representational location, direction of information flow, and degree of coupling between components, thereby providing a map of the space of possible implementations for combining neural and symbolic resources.

At the same time, the criteria by which neural-symbolic systems are categorized remain largely silent on the epistemic processes that govern knowledge acquisition. The taxonomies do not directly address how inductively acquired generalizations should interact with externally provided constraints, how instruction should override or reshape experience-driven beliefs, or how conflicts and tensions between these sources of knowledge should be resolved. As a result, systems realized using very different architectural arrangements may share the same epistemic limitations, while systems built on similar architectures may support divergent epistemic capabilities.

The point here is not to challenge the value of the NSAI paradigm or of any architectural taxonomy. Rather, it is to make explicit the

perspective they adopt. Neural-symbolic AI provides a vocabulary for describing *how* different computational resources are combined within a system. What it leaves largely implicit is *why* such combinations are required in the first place, and which underlying learning capabilities they are intended to support. Making this distinction explicit sets the stage for a shift in perspective, from architectural means to the underlying epistemic ends they are meant to serve.

3 THE INDUCTIVE-INSTRUCTIONAL END

To articulate the epistemic end toward which architectural means are directed, we distinguish two complementary modes of knowledge acquisition: inductive learning and instructional learning. The modes diverge in the source of justification for acquired knowledge.

Inductive learning refers to the acquisition of general constraints or regularities from experience. It encompasses statistical learning, pattern recognition, and generalization from data. Inductively acquired knowledge is justified by empirical support and is typically revised through further interaction with the environment.

Instructional learning, by contrast, refers to the uptake of knowledge communicated by an external source. Such knowledge may take the form of rules, explanations, objections, corrections, demonstrations, norms, or constraints, and is justified not by the learner's own experience, but by deference to another agent or authority.

Open-ended intelligence requires the integration of both modes of learning. A system that relies exclusively on induction is *epistemically isolated*: it may generalize effectively from its own experience, but lacks principled means for incorporating communicated constraints or exceptions, and is bound to rediscover much of what others already know. Conversely, a system that relies exclusively on instruction is *epistemically parasitic*: it may acquire content through communication, but lacks independent mechanisms for grounding, extending, or appropriately revising and qualifying that content.

We propose *inductive-instructional learnability* as a characterization of the underlying capability at stake. An inductive-instructional learner *(i)* extracts structure from data through inductive processes, *(ii)* incorporates externally provided information that may not be recoverable from data alone, and *(iii)* maintains coherence when inductive and instructional constraints interact or conflict.

This framing subsumes a wide range of existing paradigms, including supervised learning, learning from demonstration, interactive programming, and explanation-based learning, while preserving a principled distinction between experiential evidence and communicated knowledge. The inductive-instructional end is not merely to support these paradigms in parallel, but to reconcile them in a robust and revisable manner. Instruction may override or qualify inductive generalizations, while experience may contextualize, refine, or challenge communicated information. Neural-symbolic architectures constitute one important family of strategies for realizing this capability, but they do not exhaust the design space.

4 INSTRUCTION AS A LEARNING ABILITY

The idea that intelligent systems should be able to learn from communicated information is not new. In his seminal proposal of the Advice Taker, McCarthy [2] argued for AI systems capable of accepting declarative advice and modifying their behavior accordingly, without requiring major reprogramming or, as he memorably described it, "brain surgery". Advice, in this sense, is not experiential data, but communicated knowledge intended to be interpreted, integrated, and reasoned with. McCarthy's proposal thus distinguished instruction from inductive learning, treating advice as a first-class epistemic input rather than as an auxiliary training signal.

While the Advice Taker articulated the conceptual importance of instruction, it left open key questions regarding its formal development. In particular, it did not specify how advice is incorporated over time, how conflicting pieces of advice are handled, or what criteria determine the success of the knowledge acquisition process.

Recent work on the Advice Taker 2.0 and Machine Coaching [3, 4, 6] can be understood as a systematic response to these gaps. This line of work endows instruction with explicit learning semantics, specifying how advice is incorporated, maintained, or qualified, while respecting McCarthy's requirement to avoid "brain surgery".

The significance of Machine Coaching lies not in its particular representational choices, but in its treatment of instruction as epistemically meaningful: something that can be accepted, contested, refined, or overridden according to principled criteria. By making the learning semantics of instruction explicit, Machine Coaching sets the stage for its integration with other modes of learning.

The importance of instruction as a mode of learning is further underscored in a broader cognitive context. Valiant [9] argues that the capacity to learn from communicated information, alongside learning from experience and reasoning with acquired knowledge, is central to what distinguishes human cognition from that of other animals. From this perspective, instruction is not a peripheral convenience but a core component of open-ended intelligence, and the significance of formalizing it extends beyond historical continuity with early AI proposals: it reflects an attempt to capture a fundamental epistemic capability characteristic of human learners.

5 THE MEANS DO NOT PREEMPT THE END

Having posited that inductive-instructional competence is an epistemic requirement that cannot be inferred from architectural form alone, we now provide a concrete demonstration of that claim. We show, in particular, that even within a fixed neural-symbolic architecture, learning behavior can vary substantially depending on how communicated and experiential information are treated.

This point is illustrated through the analysis of autodidactic and coachable neural architectures [5]. The systems studied in that work share an architectural organization corresponding to a single class in Kautz's taxonomy: a symbolic policy consuming the outputs of a neural component. Holding this organization constant, qualitatively different modes of learning can nevertheless be obtained by assigning different epistemic roles to the symbolic policy, which may function in some cases as a descriptive model of the environment and in others as an encoding of user intent or constraints.

The epistemic role assigned to the policy determines which inference capabilities are relevant, the process by which the symbolic policy can be acquired, and how the policy, whether during or after its acquisition, interacts with the neural module and generates learning signals. Crucially, differences in learning behavior, including whether learning proceeds inductively or instructionally, arise

during training rather than deployment and depend on the role played by the policy rather than on architectural structure itself.

This analysis highlights a limitation of architecture-centered taxonomies of neural-symbolic AI. Systems that are indistinguishable at the level of representational organization may nonetheless implement very different learning semantics, while systems with different architectures may realize similar inductive-instructional competence. Architectural choices enable particular forms of interaction, but they do not determine the epistemic character of learning in advance. The relationship between means and ends is therefore one of underdetermination: architectural form does not preempt the inductive-instructional end it is intended to realize.

6 EXPLAINABILITY AND CONTESTABILITY

If inductive-instructional competence is the appropriate target for open-ended intelligence, prevailing criteria for evaluating AI systems require revision. Architectural integration, benchmark performance, and even post-hoc interpretability offer, at best, indirect evidence of whether a system reconciles learning from experience with learning from instruction. From the inductive-instructional perspective, the decisive issue is not representational content, but how epistemic authority is allocated, revised, and contested.

This shift has immediate implications for explainability. Explanations are often treated as artifacts derived from symbolic structure or transparent internal states. On the present view, however, explainability is better understood as a consequence of instructional competence. A system capable of learning from instruction must already possess mechanisms for interpreting communicated content, integrating it with prior commitments, and justifying its acceptance or rejection. These same mechanisms support the production of explanations that are responsive to queries, sensitive to context, and grounded in the system's epistemic commitments, rather than merely exposing internal computations without normative force.

A similar shift applies to contestability. Contestability is often framed as a matter of interface design, governance, or external oversight. From the inductive-instructional perspective, however, contestability depends on a system's capacity to treat communicated objections as epistemically meaningful inputs. A system that supports instructional learning must be able to register objections as reasons, integrate them with existing commitments, and revise its behavior accordingly. These same capacities enable contestability, allowing challenges to function as grounds for modification rather than as mere data or undifferentiated learning signals.

Recent work on explanatory compliance [7] provides a concrete illustration of this connection between instructional learning and socially relevant system properties. By specifying how communicated information is incorporated, qualified, or rejected, explanatory compliance makes clear that both explainability and contestability depend on the same underlying learning semantics. A system can explain its behavior or accommodate challenges only insofar as it has principled procedures for managing communicated inputs: tracking their epistemic status, integrating them with existing commitments, and revising those commitments when warranted. On this view, explanation and contestation are not auxiliary capabilities layered on top of an otherwise complete learning system, but natural consequences of the system's capacity to be instructable.

7 IMPLICATIONS AND OPEN PROBLEMS

With inductive-instructional learnability taken to be the relevant epistemic objective, progress toward open-ended intelligence should be evaluated in terms of a system's capacity to manage epistemic interactions across the system's lifetime: how it integrates experiential and instructional inputs, how it resolves conflicts between them, and how it responds to challenges from external agents.

Making the inductive-instructional end explicit brings a set of open problems into sharper focus. One such challenge concerns the calibration of instructional force: when communicated information should override or guide experience; when it should be contextualized by experience; and how the normative authority and weight of instructional sources should be learned, adjusted, and maintained.

A closely related but distinct issue concerns the principled management of tensions among accumulated commitments, including decisions about revision, exception handling, or suspension. These questions connect naturally to belief revision, which develops criteria for incorporating, prioritizing, or discarding information [8].

These challenges are compounded by scale and plurality. Instructional learning must remain coherent in the presence of multiple instructors, heterogeneous sources of authority, and long temporal horizons. How instructional commitments are tracked, compared, and revised under such conditions remains largely unexplored.

The perspective advanced here does not reject neural-symbolic AI, nor does it prescribe a single architectural solution. Instead, it reframes the design challenge for AI systems operating in social and normative environments. The central question shifts from integrating neural and symbolic components to constructing systems whose learning dynamics respect distinctions between direct experience and external advice, evidential support and deference to authority, and hypothesis revision and normative compliance.

Addressing this challenge will likely require closer connections between machine learning, knowledge representation, belief revision, human-AI interaction, and insights from cognitive psychology and learning theory. Making the inductive-instructional end explicit provides a unifying target for this work and a clearer basis for evaluating progress toward genuinely educable AI systems.

REFERENCES

[1] Henry A. Kautz. 2022. The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine* 43, 1 (2022), 105–125.

[2] John McCarthy. 1959. Programs with Common Sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes* (London, U.K.).

[3] Loizos Michael. 2017. The Advice Taker 2.0. In *Proceedings of the 13th International Symposium on Commonsense Reasoning* (London, U.K.).

[4] Loizos Michael. 2019. Machine Coaching. In *Proceedings of the Workshop on Explainable Artificial Intelligence @ IJCAI* (Macao, China).

[5] Loizos Michael. 2023. Autodidactic and Coachable Neural Architectures. In *Compendium of Neurosymbolic Artificial Intelligence*, Pascal Hitzler, Md Kamruzzaman Sarker, and Aaron Eberhart (Eds.). Frontiers in Artificial Intelligence and Applications, Vol. 369. IOS Press, 235–248.

[6] Loizos Michael. 2023. Explainability and the Fourth AI Revolution. In *Handbook of Research on Artificial Intelligence, Innovation and Entrepreneurship*, Elias Carayannis and Evangelos Grigoroudis (Eds.). Edward Elgar Publishing, 102–120.

[7] Loizos Michael. 2024. Explanatory Compliance. In *Proceedings of the 5th Workshop on Explainable Logic-Based Knowledge Representation @ KR* (Hanoi, Vietnam).

[8] Pavlos Peppas. 2008. Belief Revision. In *Handbook of Knowledge Representation*, Frank van Harmelen, Vladimir Lifschitz, and Bruce Porter (Eds.). Foundations of Artificial Intelligence, Vol. 3. Elsevier, 317–359.

[9] Leslie Valiant. 2024. *The Importance of Being Educable: A New Theory of Human Uniqueness.* Princeton University Press.