Article

Not peer-reviewed version

Heuristic Layering: Structuring AI Systems Beyond End-to-End Models

Rogério Figurelli *

Posted Date: 23 June 2025

doi: 10.20944/preprints202506.1782.v1

Keywords: heuristic layering; artificial intelligence; modular architecture; interpretability; explainable AI; cognitive systems

 doi:10.20944/preprints202506.1782.v1

Article

Heuristic Layering: Structuring AI Systems Beyond End-to-End Models

Rogério Figurelli

Independent Researcher, Brazil; figurelli@gmail.com

Abstract

Heuristic layering is introduced as a conceptual and architectural alternative to end-to-end modeling in artificial intelligence. Instead of treating intelligent systems as monolithic pipelines optimized solely through data-driven processes, this paradigm organizes AI into explicit, modular layers — each governed by transparent heuristics tailored to distinct cognitive or computational functions. The approach offers a pathway to increased interpretability, flexibility, robustness, and adaptability, addressing core limitations of end-to-end systems such as global brittleness and lack of targeted auditability. By mapping clear interfaces between layers and enabling local updates without retraining the entire system, heuristic layering supports both incremental innovation and systematic error isolation. This article frames the theoretical foundations of heuristic layering, draws parallels with modular design in software engineering and cognitive science, and discusses scenarios in which layered architectures may surpass traditional end-to-end models in transparency, maintainability, and domain transferability. The conclusion outlines future research avenues for the taxonomy, formal patterns, and practical deployment of heuristic-layered AI systems.

Keywords: heuristic layering; artificial intelligence; modular architecture; interpretability; explainable AI; cognitive systems

Subjects: artificial intelligence models; tools and applications; computer science; mathematics

Introduction

The rise of deep learning and transformer-based architectures has established end-to-end modeling as the prevailing paradigm in artificial intelligence. These unified pipelines, optimized by large-scale data-driven processes, have delivered unprecedented performance in domains ranging from language understanding to visual recognition and strategic decision-making. However, the very features that underpin their empirical success — massive parameterization, opaque intermediate representations, and tight global coupling — now reveal intrinsic limitations as AI systems are embedded in real-world, mission-critical, or high-stakes environments.

Among the most salient of these limitations are the chronic lack of interpretability, the brittleness of global performance under distributional shift or adversarial input, and the practical obstacles to incremental adaptation or targeted debugging. The opacity of internal mechanisms makes it difficult to understand, audit, or trust the behavior of deployed systems, while tightly coupled architectures amplify the systemic impact of localized errors, hindering safe evolution and robust maintenance.

Heuristic layering emerges in response to this predicament as a framework for decomposing intelligence into explicit, auditable layers — each governed by well-defined heuristics or modular procedures. By shifting from undifferentiated, end-to-end optimization to layered modularity, this paradigm offers a pathway toward explainable, resilient, and adaptive AI architectures. It opens new prospects for targeted innovation, systematic error isolation, and transparent self-reconfiguration under epistemic and operational stress. The sections that follow elaborate the conceptual foundations, architectural principles, and future scenarios in which heuristic layering may enable a generational leap in the design and governance of intelligent systems.

 doi:10.20944/preprints202506.1782.v1

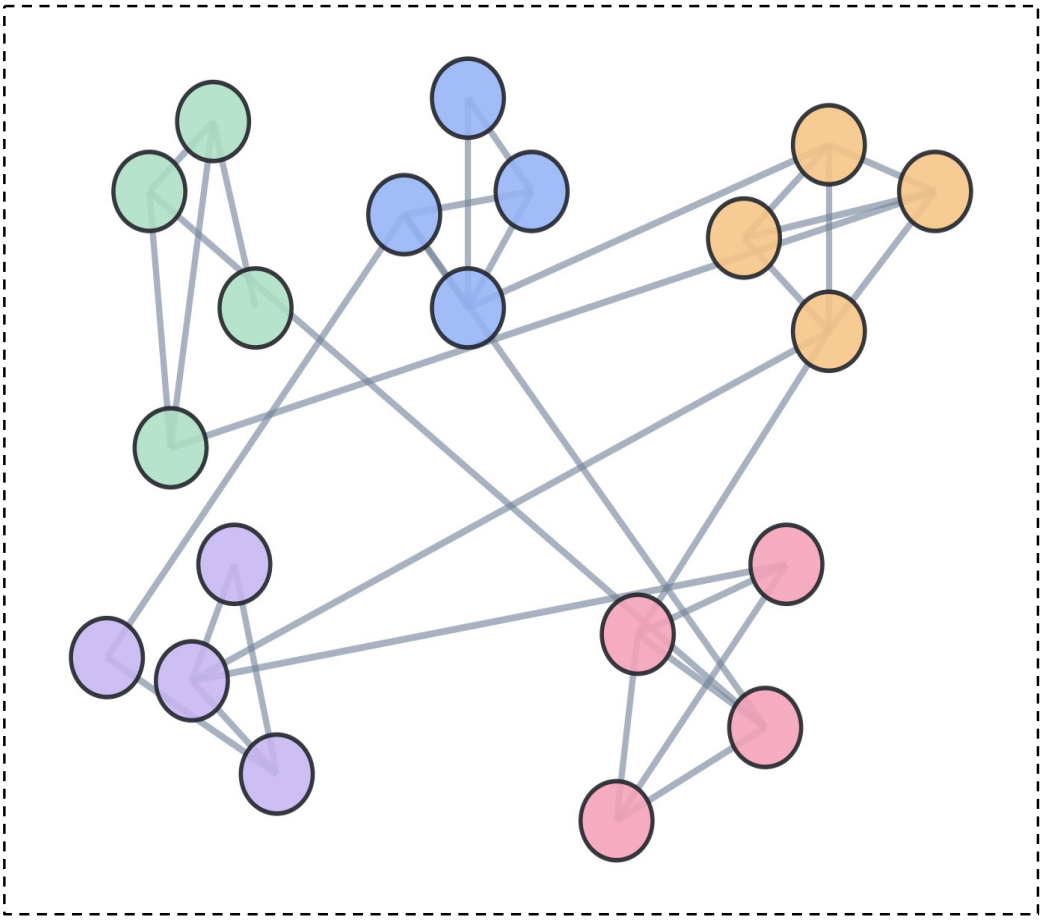
Conceptual Foundations

The intellectual landscape of artificial intelligence has been profoundly shaped by the emergence and dominance of end-to-end learning paradigms, particularly in the wake of breakthroughs in deep neural networks and transformer-based models [1]. These systems, celebrated for their capacity to learn representations directly from raw data, have catalyzed a powerful shift toward unified, data-centric architectures. Yet, the very seamlessness that makes them attractive introduces hidden epistemic costs. Their internal workings, characterized by opaque high-dimensional parameter spaces and nonlinear transformations, render them largely inscrutable even to their creators [2]. This opacity not only undermines interpretability but also magnifies systemic risk: errors propagate without clear boundaries, and targeted intervention becomes both laborious and imprecise [3].

Heuristic layering is proposed as a conceptual inversion of this trend. Rather than relying on undifferentiated flows of computation, heuristic layering constructs intelligent systems from modular strata — each composed of explicit, auditable heuristics or procedural rules.

This architectural philosophy finds resonance in the history of software engineering, where modularity and layered abstraction have long been pillars of robust system design [4]. In the cognitive sciences, layered and modular models have served as explanatory frameworks for understanding complex reasoning, perception, and action [5]. By importing these ideas into AI, heuristic layering promises architectures that are not only more interpretable but also capable of local adaptation, targeted repair, and systematic auditing. Each layer, governed by well-defined heuristics, can be individually developed, analyzed, and improved — breaking the monolithic opacity of end-to-end models and enabling a new form of epistemic accountability.

As visually abstracted in Figure 1, the heuristic layering paradigm can be represented as a network of interconnected modules, each cluster suggesting distinct heuristic domains whose boundaries remain permeable to adaptation and reconfiguration.



 doi:10.20944/preprints202506.1782.v1

Figure 1. Conceptual modular network representing heuristic layering. A purely visual diagram of interconnected clusters, each color-coded to suggest modular heuristic domains. The structure illustrates modularity, flexibility, and potential for adaptive recombination — emphasizing the epistemic contrast with monolithic, end-to-end models.

The transition from end-to-end pipelines to heuristic layers thus constitutes more than an engineering choice; it is an epistemic repositioning. It reframes intelligence as a composition of interacting, semi-autonomous strata, each of which can be interrogated, modified, or replaced in isolation. The prospect is not only technical but philosophical: by rendering the scaffolding of intelligence visible and mutable, heuristic layering reclaims agency from the black box and reintroduces the possibility of reflective, principled design.

Architecture and Operation

A heuristic-layered artificial intelligence system is architected as a stack of functionally distinct modules, each responsible for a specific computational or cognitive task. In canonical implementations, this typically involves an Input Layer, which preprocesses and structures raw data; a Processing Layer, applying domain-specific heuristics such as clustering or semantic filtering; a Decision Layer, synthesizing outputs from previous stages using rule-based or logic-driven mechanisms; and a Meta Layer, which performs continuous monitoring, adaptive tuning, and self-evaluation [2]. The separation of layers enforces auditability and modularity, with each stratum exposing explicit interfaces for both input and output, thus ensuring that information flow remains both traceable and modifiable.

The operational core of this approach lies in the isolation of errors and targeted intervention. If a fault or misjudgment occurs in a specific module, its propagation can be confined, allowing for precise diagnostics and local repair without destabilizing the entire system [4]. This compartmentalization not only mitigates systemic risk but also facilitates incremental evolution: heuristics within any given layer may be updated, replaced, or even entirely restructured independently of the rest of the architecture [6].

In practical terms, communication between layers is governed by clearly defined protocols, and feedback can be introduced from higher to lower layers to reinforce adaptive behavior or impose corrective constraints. This architectural flexibility, grounded in modular abstraction and epistemic supervision, directly confronts the opacity and brittleness endemic to end-to-end pipelines, providing a path toward resilient, explainable, and sustainable artificial intelligence.

The organizational structure of heuristic layering extends beyond a simple sequence of discrete layers. In advanced systems, each functional module may belong to a cluster of related heuristics within a given layer, and dynamic connections can form not only between consecutive strata but also across different layers and domains. This flexible network topology is depicted in Figure 3, where modular clusters are distributed across vertical levels and interconnected by both sequential and cross-layer pathways. Such an arrangement enables non-linear information flow, lateral adaptation, and the recombination of specialized heuristics, supporting the emergence of robust, adaptive, and highly configurable AI architectures. Unlike monolithic end-to-end pipelines, this approach offers inherent pathways for innovation, error correction, and domain transferability.

 doi:10.20944/preprints202506.1782.v1

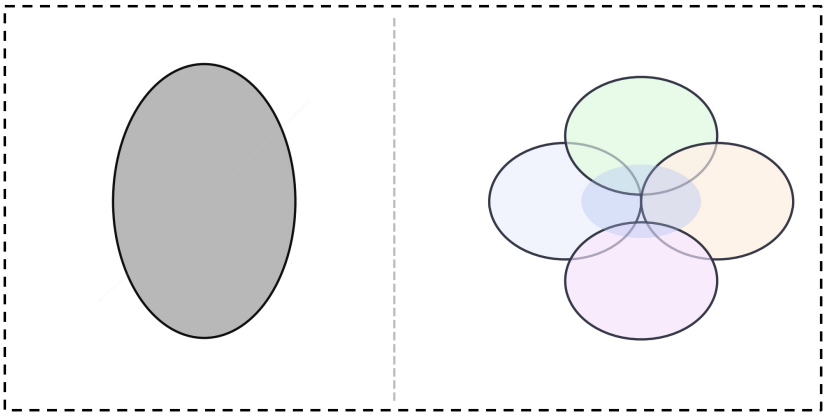


Figure 2. Conceptual Venn diagram contrasting end-to-end AI (left, isolated) with heuristic layering (right, overlapping domains). The visualization highlights how modular architectures facilitate the intersection of core system attributes, supporting incremental adaptation and local error containment.

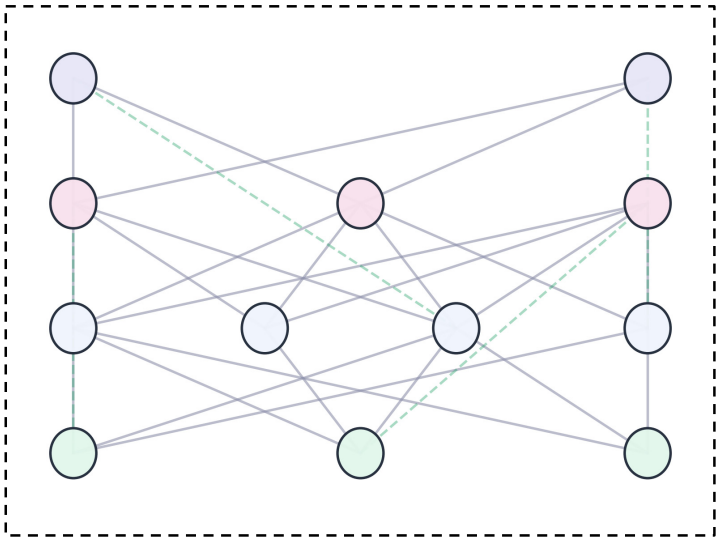


Figure 3. Hierarchical modular network of heuristic layering. A multi-layer network of clusters with both sequential and cross-layer connections, illustrating the flexible, recombinatory structure of advanced AI systems.

This networked modularity is foundational to the epistemic integrity and operational resilience advocated throughout the heuristic layering paradigm. By allowing information, influence, and adaptation to traverse not just top-down or bottom-up channels but also lateral and diagonal ones, the architecture supports emergent behaviors that are both explainable and evolvable. Specialized modules can collaborate, compete, or substitute for each other as system requirements change, while feedback from meta-level clusters can propagate improvements or corrections dynamically across the network. The resulting system is capable of local learning, targeted fault isolation, and even creative reconfiguration — features that are unachievable within rigid, globally-coupled models.

At a conceptual level, the hierarchical modular map exemplifies how symbolic agency and architectural transparency can coexist in artificial intelligence.

The clusters are not static containers but living zones of heuristic activity: they continually recompose as new knowledge is acquired or as domains evolve. Cross-layer links embody the principle that intelligence is not confined to any single representational stratum, but emerges from the interplay of diverse cognitive functions distributed across the system.

This configuration operationalizes the central thesis of the article: that robust, interpretable, and future-proof AI is not merely a product of component performance, but of the structural relations and adaptive dynamics among modular heuristics.

 doi:10.20944/preprints202506.1782.v1

Advantages over End-to-End

Heuristic layering offers a series of advantages that directly address foundational weaknesses of end-to-end AI architectures. Foremost among these is interpretability: the explicit delineation of modular layers governed by transparent heuristics makes it possible to audit, explain, and understand the inner workings of the system at every level [2,4]. Unlike end-to-end models — where decisions often emerge from opaque, high-dimensional interactions — layered architectures enable developers and auditors to trace causal chains, isolate sources of error, and justify outputs in operational terms [15].

Flexibility is another core strength. Because each layer operates as an autonomous module, updates or innovations can be introduced locally, without the need to retrain or destabilize the entire system. This supports incremental improvement, domain adaptation, and targeted debugging — capabilities that are costly or unattainable in monolithic pipelines [4]. Robustness is also enhanced: faults or failures that emerge within a single module can be contained and corrected before propagating, reducing systemic risk and simplifying maintenance routines [3].

As visually depicted in Figure 2, heuristic layering enables the convergence and overlap of properties such as flexibility, robustness, interpretability, and reusability — whereas end-to-end systems, by contrast, remain isolated and unable to naturally integrate these advantages within a single architecture.

A further benefit is reusability. Heuristics or modules that demonstrate effectiveness within one domain or context can be ported to new applications, encouraging composability and accelerating the development of novel solutions [2]. This modular inheritance stands in contrast to the often brittle transferability of end-to-end models, whose learned representations may resist decomposition or adaptation.

Taken together, these advantages frame heuristic layering not merely as an engineering convenience, but as an epistemic shift toward transparent, resilient, and maintainable artificial intelligence — an imperative as AI systems become increasingly embedded in safety-critical, regulated, or mission-driven environments.

Applications and Future Directions

Heuristic layering is especially promising for artificial intelligence systems operating in regulated or safety-critical environments, such as healthcare, finance, transportation, or automated governance. In these domains, the need for transparency, targeted validation, and incremental adaptation is paramount.

Layered architectures enable both regulatory compliance and robust human-in-the-loop oversight, as each module can be individually audited, tested, and, if necessary, reconfigured without destabilizing the larger system [4,15]. This capability is particularly relevant in sectors where explainability and traceability are formal requirements, and where domain experts or regulators must be able to intervene at the level of specific cognitive functions [16].

The migration from legacy end-to-end systems to heuristic-layered frameworks may proceed incrementally: existing pipelines can be decomposed into modules, with new heuristic layers interposed to monitor, adapt, or override behaviors as necessary. This approach supports gradual modernization and mitigates the risks of wholesale system replacement [3]. Beyond compliance and maintainability, heuristic layering facilitates the introduction of advanced meta-level capabilities, such as dynamic self-assessment, performance monitoring, and adaptive governance protocols.

Looking ahead, research into systematic taxonomies, design patterns, and domain-specific metrics for layered architectures is essential. Formal methodologies for evaluating modular effectiveness, transparency, and resilience must be developed and standardized. As artificial intelligence continues to expand its reach and societal impact, the heuristic layering paradigm offers a foundation for building systems that are not only more explainable and adaptable, but also

 doi:10.20944/preprints202506.1782.v1

inherently future-proof — capable of integrating new knowledge, accommodating evolving standards, and supporting sustained innovation.

Conclusions

Heuristic layering represents a substantive epistemic and architectural advance for artificial intelligence, transcending the constraints of end-to-end pipelines by reintroducing modular agency and symbolic accountability into the design of cognitive systems. By enabling each layer to be explicable, auditable, and independently adapted, this paradigm establishes a foundation for robust, resilient, and transparent AI — properties that are essential as such systems move ever deeper into safety-critical, regulated, and human-centered domains [4]. The shift from monolithic optimization to explicit modularity marks not only a technical reconfiguration, but a philosophical repositioning of what it means to construct and govern artificial intelligence.

The arguments and scenarios presented throughout this article point toward a new generation of AI architectures, where transparency, maintainability, and incremental innovation are structurally embedded. Future research into systematic taxonomies, design patterns, and epistemic metrics will be vital to realizing the full promise of heuristic layering. Ultimately, this approach invites the broader scientific community to reclaim agency over intelligent systems, fostering a culture of design where resilience and interpretability are not afterthoughts, but core operational imperatives.

Limitations and Expanded Discussion

Despite its conceptual and operational advantages, heuristic layering is not a universal solution; its adoption entails trade-offs that must be acknowledged and addressed. One primary limitation lies in the challenge of optimal modular decomposition: poorly chosen boundaries or interfaces can introduce new forms of brittleness or inefficiency, undermining the intended benefits of flexibility and auditability [2]. Achieving the right level of granularity in each layer — neither too coarse to obscure local adaptation nor too fine to fragment systemic coherence — requires both domain expertise and iterative refinement [4].

Implementation complexity is another practical concern. Designing, maintaining, and coordinating multiple heuristic modules may increase initial development overhead, especially in large-scale or legacy environments. Additionally, inter-module communication protocols and feedback mechanisms must be carefully engineered to prevent emergent failure modes or unanticipated side effects.

Epistemically, there remains the risk that modular transparency may mask, rather than resolve, deeper ambiguities or ontological tensions — particularly when high-level heuristics inherit or aggregate the blind spots of their constituent layers [14]. The discipline of systematically auditing, validating, and governing layered systems is itself an open research area, demanding new methodologies, formal metrics, and institutional practices.

Finally, heuristic layering, like any architectural innovation, operates within specific historical, technical, and organizational contexts. Its success depends on ongoing dialogue between theoretical ambition and empirical rigor, as well as sustained engagement with the broader social, ethical, and regulatory dimensions of artificial intelligence.

Where Was the Gap?

There is a recurring paradox in scientific and technological innovation: the most transformative advances often arise not from inventing what is radically new, but from naming, formalizing, or articulating what was already implicit — what many practitioners may have sensed but left undescribed.

In this work, the framework of heuristic layering is presented as just such a case. At first glance, separating intelligent systems into modular, auditable strata may appear almost self-evident, a mere extension of common engineering sense. Yet, the widespread dominance of end-to-end pipelines, the

 doi:10.20944/preprints202506.1782.v1

inertia of habitual design, and the absence of a shared vocabulary for this paradigm left a crucial epistemic gap — one that only becomes visible once described and formalized. The process of surfacing the "gap" is both technical and philosophical. It requires not only the capacity to abstract recurring patterns from practice, but also the willingness to risk stating what may seem obvious. History shows that, in many fields, the act of naming and formalizing the apparent is what enables new forms of rigor, critique, and innovation. The gap, then, was not a lack of cleverness or technical ingenuity, but a missing articulation — a structure waiting to be made visible, discussed, and improved upon.

Reflecting on this dynamic, it becomes clear that scientific progress is often less about inventing new content and more about transforming the invisible into the visible, the tacit into the explicit. In offering heuristic layering as both a framework and an invitation, this article aims not to claim novelty for its own sake, but to catalyze a broader conversation about the epistemic work required to close such gaps — wherever they may remain, and however obvious they may seem in retrospect.

Conflicts of Interest: The author declares no conflicts of interest. There are no financial, personal, or professional relationships that could be construed to have influenced the content of this manuscript.

Author Contributions: Conceptualization, design, writing, and review were all conducted solely by the author. No co-authors or external contributors were involved.

Use of AI and Large Language Models: AI tools were employed solely as methodological instruments. No system or model contributed as an author. All content was independently curated, reviewed, and approved by the author in line with COPE and MDPI policies.

Ethics Statement: This work contains no experiments involving humans, animals, or sensitive personal data. No ethical approval was required.

Data Availability Statement: No external datasets were used or generated. The content is entirely conceptual and architectural.

References

1. Lecun, Y., Bengio, Y., & Hinton, G. "Deep learning." *Nature*, vol. 521, pp. 436–444, 2015.
2. Doshi-Velez, F., & Kim, B. "Towards a rigorous science of interpretable machine learning." *arXiv preprint* arXiv:1702.08608, 2017.
3. Amodei, D., Olah, C., Steinhardt, J., et al. "Concrete problems in AI safety." *arXiv preprint* arXiv:1606.06565, 2016.
4. Parnas, D. L. "On the criteria to be used in decomposing systems into modules." *Communications of the ACM*, vol. 15, no. 12, pp. 1053–1058, 1972.
5. Anderson, J. R., & Lebiere, C. "The Atomic Components of Thought." Erlbaum, 1998.
6. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. "Building machines that learn and think like people." *Behavioral and Brain Sciences*, vol. 40, e253, 2017.
7. Figurelli, R. "Heuristic Physics: Foundations for a Semantic and Computational Architecture of Physics." Trajecta White Paper, 2025.
8. Schmidhuber, J. "Deep Learning in Neural Networks: An Overview." *Neural Networks*, vol. 61, pp. 85–117, 2015.
9. Pearl, J. "Causality: Models, Reasoning, and Inference." Cambridge University Press, 2009.
10. Simon, H. A. "The Sciences of the Artificial." MIT Press, 3rd ed., 1996.
11. Holland, J. H. "Adaptation in Natural and Artificial Systems." University of Michigan Press, 1975.
12. Chaitin, G. J. "Algorithmic Information Theory." Cambridge University Press, 1987.
13. Wolfram, S. "A New Kind of Science." Wolfram Media, 2002.
14. Marcus, G. "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence." *arXiv preprint* arXiv:2002.06177, 2020.
15. Brachman, R. J., & Levesque, H. J. "Knowledge Representation and Reasoning." Morgan Kaufmann, 2004.

 doi:10.20944/preprints202506.1782.v1

16. Marr, D. "Vision: A Computational Investigation into the Human Representation and Processing of Visual Information." Freeman, 1982.