Configuring Neuro-Symbolic Systems with Guaranteed Properties with Application in Hybrid Intelligence

Johannes E. Bendler
Vrije Universiteit Amsterdam
j.e.bendler@vu.nl

Abstract

Neuro-symbolic systems are a promising paradigm for Hybrid Intelligence (HI), combining the complementary strengths of symbolic reasoning and machine learning in human-AI interactions. However, designing architectures that satisfy desired properties currently relies on informal reasoning, rather than a rigorous framework. Our research addresses this gap by developing a formal intermediate layer that links architecture design to system properties. Building on the visual pattern language of the Boxology for neuro-symbolic systems, we develop a mathematical formalization that is grounded in Category Theory, in order to enable proof-based guarantees of properties. The soundness of this formalism will be demonstrated through empirical case studies of existing Hybrid Intelligence systems. This work advances the theoretical foundations of neuro-symbolic AI by supporting principled design of HI systems.

1 Introduction

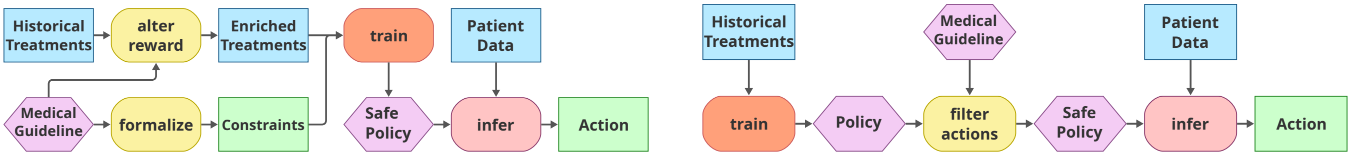
Hybrid Intelligence (HI) systems represent a synergy between human and machine intelligence, wherein their complementary strengths are combined to achieve shared goals. These can range from simple decision support over sizeable hybrid teams to human-in-the-loop systems. Ideally, HI systems should satisfy so-called CARE properties: ***C**ollaborative, **A**daptive, **R**esponsible, **E**xplainable* [Akata *et al.*, 2020]. Neuro-symbolic artificial intelligence serves as a promising paradigm to realize HI systems, since humans use symbolic language (e.g. natural language, graphs) to communicate and machine learning systems use sub-symbolic ”neural” components to operate. The Boxology [Van Bekkum *et al.*, 2021], a visual diagram notation, provides a taxonomically organized vocabulary for architectures, processes, and data structures used in these HI systems. Figure 1a shows such a Boxology representation for a neuro-symbolic HI model in the medical domain. It displays a system, described by [den Hengst *et al.*, 2024], in which medical guidelines are used to guide the training process of a machine learning model. An alternative architecture, depicted in Figure 1b, uses medical guidelines to filter actions proposed by a statistical policy model. Compared to the former, this second system can be described as *responsible*, as all unsafe actions can reliably be filtered out, and as *adaptive*, because the medical guidelines are easily exchangeable without retraining the model. This illustrates that the choice of architecture can influence the CARE properties of these systems.

However, it is not generally possible to deterministically derive which properties an architecture satisfies, nor to methodically design a system to fulfill a required set of properties. This is due to an abstraction gap between high-level architecture design, low level implementation, and vague property descriptions. In short: There exists no unifying formal framework to describe complex neuro-symbolic HI systems and their CARE attributes.

Our research aims to close this research gap by addressing the following questions:
RQ1: How can the Boxology framework and the CARE properties of HI systems be formalized into a shared mathematical-logical structure?
RQ2: How can properties be derived from a given neuro-symbolic architecture in a deterministic manner and how can such systems be configured to exhibit specific properties?
RQ3: To what extent does the developed framework enable the description and improvement of existing HI systems in practice?

2 Approach

Boxology: We argue that the aforementioned Boxology framework is the most suitable to approach these research questions. [Van Bekkum *et al.*, 2021] have shown in their work how their pattern language can model a vast amount of diverse neuro-symbolic architectures. Rather than through fixed behavioural classification, it approaches the description of these systems from an architectural perspective, which we see as crucial for the previously mentioned derivation of CARE properties. Furthermore, the Boxology already lays the foundation for modelling human interaction, as required for HI scenarios. What contrasts the Boxology to frameworks such as UML, is the existing semantic commitment to the domain of neuro-symbolic systems, such as the data/symbol, or statistical/semantic model dichotomy. Additionally, the authors of the Boxology themselves highlight the need for a



(a) A system that uses medical guidelines to discourage unsafe actions during the training process of the statistical model.

(b) A system that uses medical guidelines to filter actions produced by a statistical model.

Figure 1: Two alternative architectures for personalized treatment planning systems, as described in [den Hengst *et al.*, 2024], with each a learning and a reasoning component. Both are exhibiting different CARE properties.

more formalized grammar and calculus for composition and verification, which we aim to provide.

Formalization: To develop a theory for configuring neuro-symbolic systems that meet specific attributes, it is essential to formalize the Boxology components, their compositions, and the CARE properties beyond their current informal version. We see Category Theory (CT) as the most promising approach to this. Historically, CT has been used to study the differences and similarities between many different fields of mathematics, such as algebra (the formalism of machine learning systems, in the form of vector algebra), and logics (the inspiration for many symbolic components) (**RQ1**). Crucially, a central concept in CT is the notion of compositionality. This can be beneficial to study different (sub-)symbolic Boxology components in isolation and in relation to others. It offers notions of structure-preserving (here: property-preserving) maps between objects, as well as a constructive proof-theoretic framework, specifying the constraints elements must fulfill (**RQ2**). To formalize the Boxology in CT, we can build on the work of [Mossakowski, 2022], who proposed a category-theoretic definition of the Boxology components, and their subsequent refinement, under the notion of colimits.

Evaluation: To assess the soundness of our approach, we evaluate along two dimensions: First, we show formal correctness by proving that composing architectural components has predictable effects on system properties and that property derivation can be done deterministically. Secondly, we demonstrate practical utility through case studies from the Hybrid Intelligence Centre [Akata *et al.*, 2020], which spans medical, cultural, and technical domains. In the respective case studies, we apply the framework to existing architectures, derive their properties, and explore whether alternative configurations could better satisfy the required attributes. Results are then discussed with the original system designers to validate their correctness and relevance (**RQ3**).

3 Ongoing and Future Work

In a first step, we addressed the semantic grounding of the Boxology framework by introducing the *Tool4Boxology* [Bendler *et al.*, 2026]. Our work provides a graphical editor for modelling and validating architectural diagrams, as well as a mapping from these diagrams to ontology-compliant knowledge graphs, which enable the querying of patterns and checking semantic correctness. Next, we aim to extend this toolbox with hierarchical components, dynamic human and team interactions, and control structures, as well as a more comprehensive library of common patterns. This is then followed by the formalization process.

Novelty and Impact: Our work will introduce a formal intermediate layer bridging the gap between architectural descriptions of neuro-symbolic systems and their behavioural properties. On the one hand, this will enable the post-hoc analysis of existing systems by modelling them in the Boxology, while on the other it informs design decisions by guaranteeing properties. This work presents a step towards a unified theory of neuro-symbolic systems, providing a new research line at the intersection of system architecture and formal methods, such as Category Theory. We expect the results of our research, such as proof strategies for guaranteed properties, to be applicable to a broader range of AI systems, as well as covering additional properties.

References

[Akata *et al.*, 2020] Zeynep Akata, Dan Balliet, et al. A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28, 2020.

[Bendler *et al.*, 2026] Johannes E. Bendler, Yashrajsinh Chudasama, et al. Tool4boxology: A semantic toolbox for constructing and analysing neuro-symbolic architectures. In *Proc. of the 23rd Intl. Conference on Semantic Web (ESWC)*, 2026. To appear.

[den Hengst *et al.*, 2024] Floris den Hengst, Martijn Otten, et al. Guideline-informed reinforcement learning for mechanical ventilation in critical care. *Artificial intelligence in medicine*, 147:102742, 2024.

[Mossakowski, 2022] Till Mossakowski. Modular design patterns for neural-symbolic integration: refinement and combination, 2022.

[Van Bekkum *et al.*, 2021] Michael Van Bekkum, Maaike De Boer, Frank Van Harmelen, André Meyer-Vitali, and Annette Ten Teije. Modular design patterns for hybrid learning and reasoning systems: A taxonomy, patterns and use cases. 51(9):6528–6546, 2021.