

Publication status: This preprint has not been published elsewhere.

Ontological Mediation in Explainable Artificial Intelligence: Concept-Based Models and Ontology-Grounded Explanatory Justification

Dr. Eduardo de Mattos Pinto Coelho, Dr. Marcello Peixoto Bax

<https://doi.org/10.1590/SciELOPreprints.18206>

Submitted on: 2026-10-03

Posted on: 2026-10-06 (version 1)

(YYYY-MM-DD)

Ontological Mediation in Explainable Artificial Intelligence

Concept-Based Models and Ontology-Grounded Explanatory Justification

Eduardo de Mattos Pinto Coelho, Marcello Peixoto Bax

Federal University of Minas Gerais (UFMG), Department of Information Theory and Management, Belo Horizonte, Minas Gerais, Brazil

ORCID Coelho: <https://orcid.org/0009-0005-2586-1389>; ORCID Bax:

<https://orcid.org/0000-0003-0503-3031>

Corresponding author: Eduardo de Mattos Pinto Coelho (em4ttos@gmail.com)

Abstract

Current approaches to Explainable Artificial Intelligence (XAI) have made significant progress in producing explanations for complex machine learning models. However, many remain epistemologically fragile because their explanations lack explicit semantic and ontological grounding, limiting their adequacy in critical domains such as healthcare. We argue that this limitation is fundamentally ontological rather than algorithmic.

Meaningful explainability requires the coordination of two complementary explanatory regimes governed by distinct ontological commitments and validation criteria: a fast sub-symbolic regime that produces concept-bearing representations through neural models, and a slower symbolic regime grounded in ontologies, rules, and formally specified knowledge structures. Learned concepts from sub-symbolic learning architectures capable of generating conceptual representations are therefore not self-sufficient explanatory units; they require ontological mediation and validation.

We present a methodology and illustrate it by a proof of concept integrating Concept Embedding Models (CEMs) with a Semantic Data Dictionary (SDD) enriched by ontologies. Within this framework, ontological artifacts enable semantic consolidation, justification, and ontological unpacking of sub-symbolic explanations. A case study on Alzheimer's disease diagnosis illustrates how this framework improves the coherence, communicability, and contestability of explanations in high-stakes decision contexts, and we argue that semantic and ontological grounding are necessary conditions for explanation, justification, and governance in intelligent systems.

Keywords

Ontological mediation, ontological unpacking, ontology-based explainability, concept-based explainability, neuro-symbolic AI, justification

1. Introduction

Recent advances in Explainable AI (XAI) have expanded techniques for complex machine learning models, notably deep neural networks. Despite this, the epistemic status of many explanations remains fragile in critical domains, such as healthcare, finance, and law, where decisions require justification and scrutiny by multiple stakeholders [1, 2, 3, 4, 5]. This fragility suggests that current XAI limitations are not algorithmic but arise from deeper conceptual and ontological issues.

Most XAI approaches focus on local explanations, feature attributions, or intrinsic transparency. While valuable, these techniques often lack explicit semantic grounding and ontological commitments, making it difficult to assess what is being explained, under which assumptions, and with what justificatory force. Consequently, explanations may be operationally understandable while remaining conceptually indeterminate [6, 7]. This paper argues that meaningful explainability requires more than improved algorithms: it requires explicit ontological grounding.

We contend that AI explanations are semantically justifiable only when anchored in formal conceptual structures that mediate between learned representations and domain commitments. The central problem is not how to extract explanations from models, but what kind of ontological structure is required for them to be meaningful, communicable, and justificatory.

To address this, we analyze explainability through two complementary regimes. The first is a sub-symbolic regime, characteristic of neural models, where explanations derive from learned representations. The second is a symbolic regime, grounded in explicit concepts and rules, supporting reasoning, justification, and communication. Our primary contribution is interpreting this distinction ontologically, as a shift between regimes operating over different entities and validation criteria.

Within this framework, ontologies play a central mediating role. By making commitments explicit, ontologies provide the semantic structure to integrate sub-symbolic explanations with symbolic justification. This enables ontological unpacking: the explicit articulation of commitments assumed by explanatory models [8]. We instantiate this through a neuro-symbolic framework where ontologies regulate and interpret concept-based neural explanations. A case study on Alzheimer's disease diagnosis illustrates this role.

The next section provides the theoretical background for the proposed framework. It first situates the work within the current landscape of Explainable AI and then introduces the core notions that underpin the ontology mediated approach developed in the subsequent sections. Section 3 presents the proposed framework by introducing Ontological Mediation as the central architectural principle through which reconstructed conceptual rules are related to explicit ontological commitments and normative constraints, enabling ontology mediated justificatory assessment. Section 4 operationalizes this integration via CEMs and ontology-grounded infrastructures, introducing Ontologically Mediated Concept Rules (OMCRs) as alignment mechanisms. Section 5 applies the framework to OASIS-2, showing how OMCR enables evaluation and justificatory assessment of concept-based rules. Section 6 discusses the theoretical implications of the proposed framework in relation to theoretical backgrounds and AI governance. It examines how Ontological Mediation and ontology mediated justificatory assessment contribute to semantically grounded, normatively defensible, and epistemically accountable explanations, while also outlining the framework's current limitations and directions for future research. The conclusion synthesizes the central argument that important limitations of contemporary Explainable AI arise not only from explanatory techniques but also from the absence of explicit mechanisms for Ontological Mediation. It emphasizes the role of Ontological Mediation, supported by Ontological Unpacking, Semantic Grounding, Ontological Grounding, and OMCRs, in enabling ontology mediated justificatory assessment. It also outlines future work on the formal characterization of Ontological Mediation, the formal conceptualization and axiomatization of OMCRs, richer OMCR taxonomies, and broader implementation architectures.

2. Backgrounds

This section establishes the broader research context by situating the proposed framework within contemporary efforts toward trustworthy, fair, accountable, and explainable AI. The section also introduces the core notions that underpin the proposed framework, providing the conceptual basis for integrating sub-symbolic and symbolic explanatory regimes. It then positions the framework in relation to theoretical backgrounds (Concept-Based XAI – C-XAI, XAI 2.0, and Neuro-Symbolic AI – NeSy), motivating the need for explicit semantic and ontological foundations.

2.1. Context and Problem Definition: Algorithmic Opacity, Social Impact, and Regulatory Pressures

The rapid progress of machine learning and deep learning over the past decade has produced remarkable gains in predictive performance and operational efficiency across multiple domains. However, the widespread adoption of these techniques, particularly in high-risk contexts, has been constrained by their intrinsic opacity.

This opacity becomes unacceptable in applications where automated decisions directly affect human lives, fundamental rights, or large-scale social outcomes, including healthcare diagnostics, credit allocation, hiring, judicial sentencing, surveillance systems, autonomous vehicles, and military tech-

nologies. In such contexts, the inability to examine or contest automated decisions undermines trust, accountability, and institutional legitimacy.

Beyond opacity, empirical evidence shows that AI systems may reproduce or amplify discriminatory patterns embedded in training data or operational environments. Cases of algorithmic discrimination affecting employment, credit access, insurance pricing, and judicial recommendations have been widely documented, revealing biases related to race, gender, ethnicity, and socioeconomic status [2, 3]. Such outcomes not only harm affected individuals but also generate broader systemic risks, including legal liability, reputational damage, and erosion of public trust.

These concerns have motivated growing attention to normative frameworks for responsible AI development. A widely adopted reference is the FATE framework, – *Fairness, Accountability, Transparency, and Ethics* – which emphasizes the need for AI systems that are both technically effective and socially responsible [9, 10]. More recent discussions extend these concerns to environmental sustainability, raising questions about the long-term viability of current AI practices.

In parallel, international regulatory frameworks for AI governance – as in the EU, Canada, Brazil, and Japan – now mandate risk assessments, data documentation, and formal impact reports. These developments obligate developers to provide explanations that are auditable, contestable, and intelligible to non-technical stakeholders.

As a result, the evaluation of AI systems is gradually shifting from purely technical performance metrics toward broader notions of reliability, accountability, and governance. Predictive accuracy alone is no longer sufficient; systems must also provide explanations that enable human oversight and normative evaluation. Consequently, explainability in AI has emerged not merely as a technical challenge but as a response to concrete epistemic, ethical, and regulatory deficiencies in contemporary AI systems. Addressing this challenge requires more than improved explanatory techniques: it requires frameworks capable of supporting explanations that can be evaluated, justified, and aligned with shared conceptual and normative commitments.

2.2. Core Notions

Given the conceptual breadth of the present framework, several core notions require explicit operational clarification. The following definitions establish the intended meaning of the principal concepts employed throughout this work. Figure 1 presents a concept map that clarifies how the concepts are interconnected.

Semantic grounding establishes referential connections between symbols and empirical entities, Harnad (1990) e Taddeo and Floridi (2005) [11, 12]. Guizzardi and Guarino (2023) [8] refer to *Real-World Semantics*, which clarifies the ontological nature of those entities by making explicit their identity criteria, dependence relations, and ontological commitments. Contrasting them, semantic grounding answers what a concept refers to, while real-world semantics answers what kind of thing that concept refers to. As we shall see, the scope of the first is the sub-symbolic regime, in which the source is empirical data, as in our case study; the scope of the second is the symbolic regime, which deals with entities considered in computational ontologies.

Ontological grounding is an emergent concept in applied computer and information science; it is a synthetic construct reconstructed from several converging discussions in ontology, semantics, explainability, and knowledge representation. For our purposes, the best primary sources for this understanding are Nananukul et. Al., 2025 e Bajaj et al., 2024 [13, 14]. The *Emergent Mind* website¹ articulates this concept precisely through numerous publications on ArXiv.

“Ontological grounding is the process by which symbolic entities – terms, models, data, or theoretical constructs – are systematically anchored to structured representations of world knowledge, such as ontologies or primitive ontological frameworks. Its core function is to bridge the gap between abstract symbols and concrete referents, enabling interpretability,

¹<https://www.emergentmind.com/topics/ontological-grounding>

interoperability, and robust reasoning across domains ranging from AI and cognitive science to the natural sciences and engineering.”

Ontological grounding refers to the explicit alignment of explanatory entities and relations with domain-level conceptual commitments and constraints. In this context, grounding does not necessarily imply a complete foundational ontological analysis in the strong sense (e.g., full UFO/BFO/DOLCE alignment), but rather the explicit semantic stabilization of explanatory concepts through ontology mediated structures.

Although the notion of ontological grounding is not explicitly stated in Guizzardi and Guarino (2023) [8], we also believe that this concept may emerge from this article, considering these authors' concern with ontological ground in the analysis of entities and relations, in addition to explicitly citing Naomi Thompson (2016) [15], who, like Kit Fine (2012) [16], are references for the metaphysical discussion of the question of explanations based on *ontological ground*. As with real-world semantics, the concept of ontological grounding is applied to the symbolic regime.

In this paper, **ontological mediation** denotes the process through which symbolic semantic structures regulate, constrain, and contextualize explanations generated by sub-symbolic systems. Rather than replacing predictive models, ontological mediation introduces explicit domain commitments that enable explanatory outputs to be evaluated relative to semantic consistency, domain coherence, and justificatory adequacy. To illustrate the presentation of the proposed structure, we adopt the Semantic Data Dictionary (SDD) as an example of an implementation artifact that operationalizes Ontological Mediation, organizing semantic resources and enabling the retrieval of OMCRs.

A **Concept-Based Model** refers to a fast sub-symbolic regime that produces concept-bearing representations through neural models². Learned concepts from sub-symbolic learning architectures capable of generating conceptual representations are therefore not self-sufficient explanatory units. To facilitate the presentation of the proposed framework, this section adopts the Concept Embedding Model (CEM) as a representative example of an AI model capable of reconstructing explicit conceptual representations from sub-symbolic learning.

Both components, SDD and CEM, do not constitute a theoretical requirement of the framework. Rather, they illustrate how the proposed approach can be instantiated. Both will be referenced in Section 4, *Proposed Framework*.

Ontological unpacking is a core notion in Guizzardi and Guarino (2023) [8] and refers to the explicit reconstruction of assumptions, identity criteria, conceptual dependencies, and admissible inferences implicitly presupposed by explanatory concepts and rules. In this sense, ontological unpacking aims to expose the semantic commitments underlying concept-based explanations, rendering them inspectable and contestable. Unlike ontological grounding, ontological unpacking implies a foundational ontological analysis UFO/BFO/DOLCE/Etc. alignment.

In this paper, **justificatory assessment** concerns the evaluation of whether an explanation satisfies explicit semantic and normative constraints associated with a domain of application. Unlike predictive evaluation, which measures accuracy or fidelity, justificatory assessment evaluates whether explanatory claims remain coherent with domain commitments, conceptual constraints, and admissible forms of reasoning. We deliberately distinguish *explanation* from *justificatory assessment*. Explanation reconstructs the concept-level rationale underlying the model's prediction. Justificatory assessment is a subsequent normative evaluation of that reconstructed explanation, examining its semantic grounding, ontological coherence, and epistemic admissibility.

Ontologically Mediated Concept Rules (OMCRs) are defined as ontology-constrained reformulations of concept-based explanatory rules generated by sub-symbolic models. OMCRs function as

²Native concept based architectures, such as CBMs, CEMs, and neural concept reasoners, provide the most direct instances of this setting. More generally, the proposed framework applies whenever a predictive model, whether intrinsically interpretable or post hoc explainable, supports a faithful reconstruction of its decision in terms of explicit, human-interpretable concepts and their compositional organization. This includes, for example, linear and logistic regression models, generalized additive models, decision trees, rule-based and neuro-symbolic systems, prototype-based architectures, self-explaining neural networks, and models providing native, emergent, or reconstructable concept-level representations, provided that they yield a sufficiently explicit conceptual structure for ontological mediation and justificatory assessment through an OMCR.

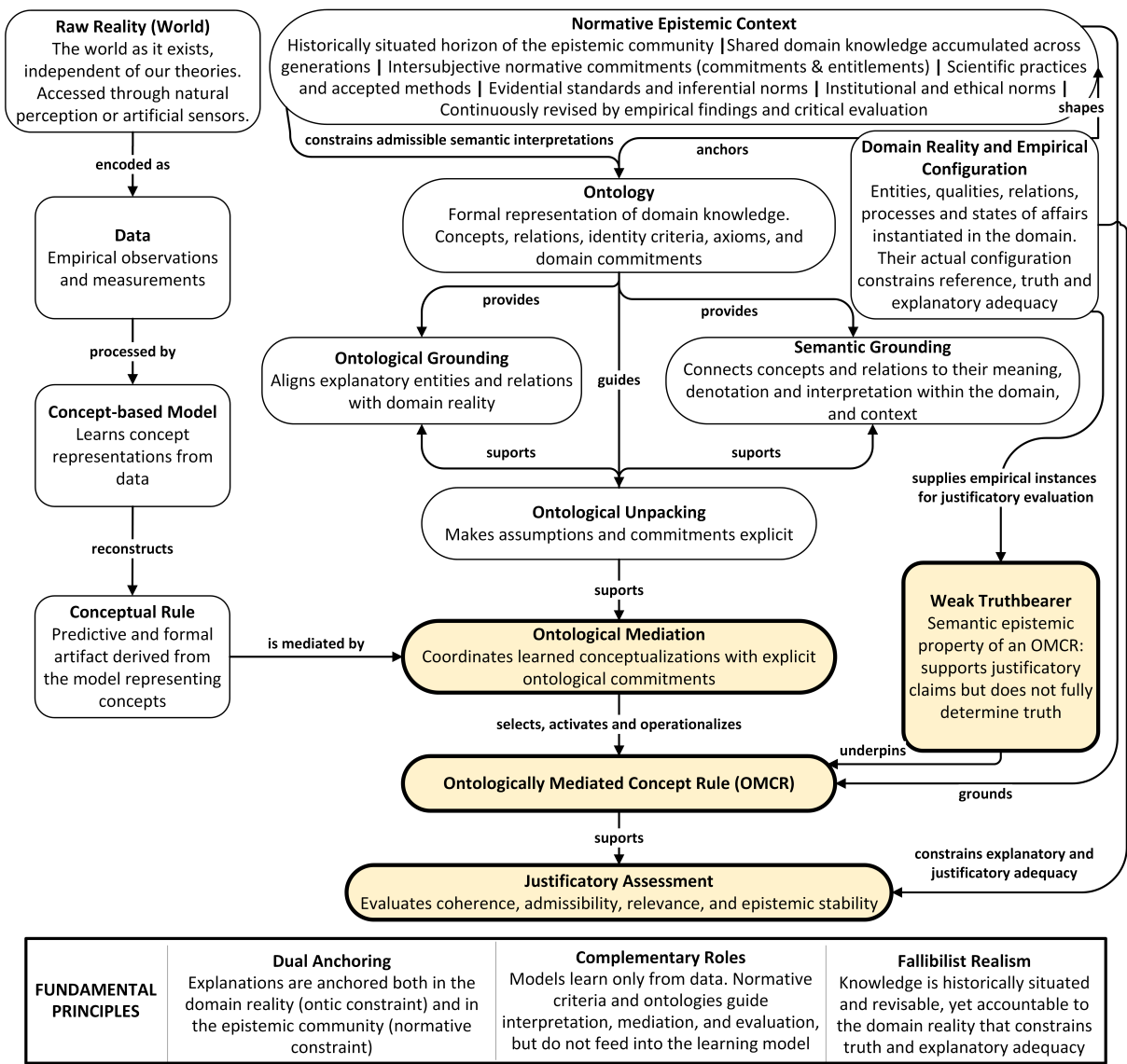


Figure 1. High-level conceptual map of the proposed ontology mediated justification framework. The diagram illustrates how domain reality shapes normative epistemic contexts, how ontological structures constrain admissible semantic interpretations, and how Ontological Mediation operationalizes Ontologically Mediated Conceptual Rules (OMCRs) to support context-dependent justificatory assessments through weak truthbearers. The blocks with a yellow background and thick outline indicate the authors' contributions.

mediating artifacts between sub-symbolic representations and symbolic domain constraints, enabling explanatory outputs to undergo semantic and ontological validation and justificatory evaluation relative to explicit ontological commitments. In addition to articulating the regime-symbolic-sub-symbolic framework, OMCR are the main contributions of this article. This concept will be elucidated and illustrated in more detail later.

Finally, **Weak Truthbearer**. Inspired by the notion of weak truthmaker discussed by Guarino, Porello, and Guizzardi (2019) [17], we introduce this concept as an ontology mediated propositional artifact whose justificatory force depends on explicit ontological commitments and on the way relevant entities are instantiated in reality. Analogously to weak truthmakers, which contribute to the truth of propositions through ontologically relevant aspects of the world: weak truthbearers contribute to the admissibility of explanatory claims by making explicit the ontological conditions under which such claims can be accepted, rejected, or considered epistemically unstable. In this work, OMCRs are ontologically mediated representations of reconstructed explanations that function as weak truthbearers by mediating between sub-symbolic conceptualizations and symbolic ontological commitments, thereby

supporting context-dependent justificatory assessments of the admissibility of those explanations.

Confalonieri and Guizzardi (2025) [18] discuss several roles of ontologies in supporting intelligible explanations for neuro-symbolic AI, emphasizing reference modelling, common-sense reasoning, and knowledge refinement. Our contribution is complementary rather than alternative. While their analysis focuses on how ontologies support the construction and management of explanations, we address the subsequent normative problem of how reconstructed concept-level explanations can be ontologically mediated and evaluated with respect to their semantic grounding, ontological coherence, and epistemic admissibility. The proposed OMCR is introduced precisely to make these normative commitments explicit.

The approach is not intended as a universal explanation framework for arbitrary computational artifacts or scientific theories. Its intended target is a neural predictive model whose decision can be reconstructed at the level of human-interpretable concepts. More precisely, the framework addresses cases in which lower-level features are associated or composed into decision-relevant concepts and these concepts support a classification, prediction, diagnosis, or recommendation. The neural model itself remains the source system, but the immediate object of ontological mediation and justificatory assessment is the extracted or reconstructed concept-level decision structure.

2.3. Research Context

The proposed framework contributes directly to ongoing discussions in Concept-Based XAI (C-XAI), XAI 2.0, and Neuro-Symbolic AI (NeSy). In Concept-Based XAI [19], explanations are constructed around learned concepts that are intended to be more interpretable and cognitively meaningful than low-level features. However, without explicit semantic and ontological grounding, such concepts remain context-dependent and epistemically fragile. Their explanatory role depends on assumptions that often remain implicit and therefore difficult to evaluate or contest. Ontological mediation provides a mechanism for stabilizing concepts across representational regimes, enabling concept-based explanations to function as scalable and semantically coherent explanatory resources.

These findings also resonate with the emerging agenda of XAI 2.0 [7], which emphasizes multidimensional explainability, human-centered evaluation, social accountability, and justification. Many of the challenges identified in this literature – particularly those concerning evaluation, contestability, trustworthiness, and social impact) – cannot be addressed solely through improvements in interpretability techniques. Rather, they require explicit mechanisms capable of articulating and assessing the ontological commitments underlying explanatory claims. Ontological grounding and ontology mediated justificatory assessment provide one possible pathway toward addressing these requirements.

Finally, the analysis strengthens the connection between XAI and Neuro-Symbolic AI (NeSy). Contemporary NeSy research seeks to integrate data-driven learning with symbolic reasoning, recognizing that scalable learning alone is insufficient for trustworthy AI [20, 21, 22, 23, 24, 25, 26, 27]. The framework proposed here complements this agenda by introducing Ontological Mediation as the architectural mechanism connecting learned conceptualizations with explicit ontological commitments, while Ontological Unpacking, Semantic Grounding, and Ontological Grounding provide the semantic foundations supporting this mediation.

Within this perspective, ontological mediation supports traceability, validation, intervention, and justificatory assessment, extending the role of symbolic reasoning beyond inference toward explanation and justification.

3. Motivation and objectives

3.1. Ontologies as Mediating Artifacts for Explainability

Ontologies have increasingly demonstrated their capacity to strengthen explainability across multiple domains by providing shared vocabularies, explicit conceptual commitments, and mechanisms for symbolic inference. Previous research shows that ontologies can enhance post-hoc explanations,

support user-centered explanations, increase trust through semantic descriptions, and enable domain-specific reasoning [28, 6, 29, 30]. More recent work extends this role to large language models and broader machine-learning pipelines [31, 7, 32].

The healthcare domain offers particularly favorable conditions for semantic approaches to explainability due to the presence of standardized terminologies and clinical taxonomies. Even when such standards remain incomplete or evolving, ontologies help make uncertainties and alternative interpretations explicit, thereby increasing transparency in diagnostic reasoning.

Within this context, ontologies play a mediating role between explanation, interpretation, and justification. Rather than merely supporting XAI techniques, they enable explanations that are interoperable, inspectable, and normatively assessable. This mediating role underlies the ontological approach to explainability developed in the following sections.

Although often introduced as formal vocabularies for annotation or interoperability, ontologies perform a deeper function. In this work we adopt a strong ontological perspective in which ontologies act as mediating artifacts that structure, constrain, and consolidate explanatory relations across heterogeneous representational regimes.

Ontologies are computational artifacts composed of formally defined concepts organized into classes and linked by axioms, typically expressed in description logics. Through these axioms, symbolic inference becomes logically adjudicated rather than probabilistic or vague. In the classical definition proposed by Gruber [33, 34], an ontology is a formal and explicit specification of a shared conceptualization whose commitments are situated and revisable.

Because they are collectively developed and maintained, ontologies also possess an intersubjective dimension: they can be evaluated, criticized, and refined by expert communities. This property allows them to function as relatively stable semantic conventions that support explanations capable of being justified beyond local or individual interpretations.

Finally, ontologies enable semantic interoperability across heterogeneous systems by operationalizing shared terminologies and conceptual structures. In healthcare, widely adopted standards and domain ontologies, such as those associated with FHIR (*Fast Healthcare Interoperability Resources*), provide stable semantic reference points for knowledge integration, validation, and reuse [35, 36, 37].

3.2. Symbolic and Sub-symbolic Explanatory Regimes

This study investigates integrating sub-symbolic learning with ontological frameworks to enhance AI explainability and fairness. We explore ontology-based semantic resources as mediating infrastructures that satisfy ethical, regulatory, and epistemic requirements. Within this view, fairness is grounded in explicit domain constraints made inspectable through ontological commitments, enabling bias monitoring and foster interoperability in fragmented domains like healthcare [38].

Many XAI approaches rely on post-hoc methods whose limitations are evident in regulated environments. These methods often remain epistemically disconnected from internal decision processes, produce unstable explanations, and are vulnerable to adversarial manipulation [1, 4]. Such weaknesses undermine their ability to satisfy transparency and accountability expectations [5], supporting our central claim: explainability alone does not guarantee justification. Closing this gap requires shifting from explanatory techniques toward the ontological conditions that render explanations stable and normatively assessable.

To clarify these foundations, we distinguish explanation (revealing causal factors), interpretation (assigning contextual meaning), and justification (normative defense under shared commitments). These distinctions reveal that AI explanations inhabit two complementary regimes. In the sub-symbolic regime: typical of neural architectures, where explanations arise from distributed representations. Validation is primarily operational, focusing on predictive relevance. The Symbolic regime, by contrast, operates over explicit concepts and formal rules. Validity depends on conceptual coherence and ontological grounding [8].

Despite their utility, sub-symbolic explanations maintain weak ontological commitments; learned concepts often function as data-dependent instruments lacking explicit identity criteria [4]. As Blu-

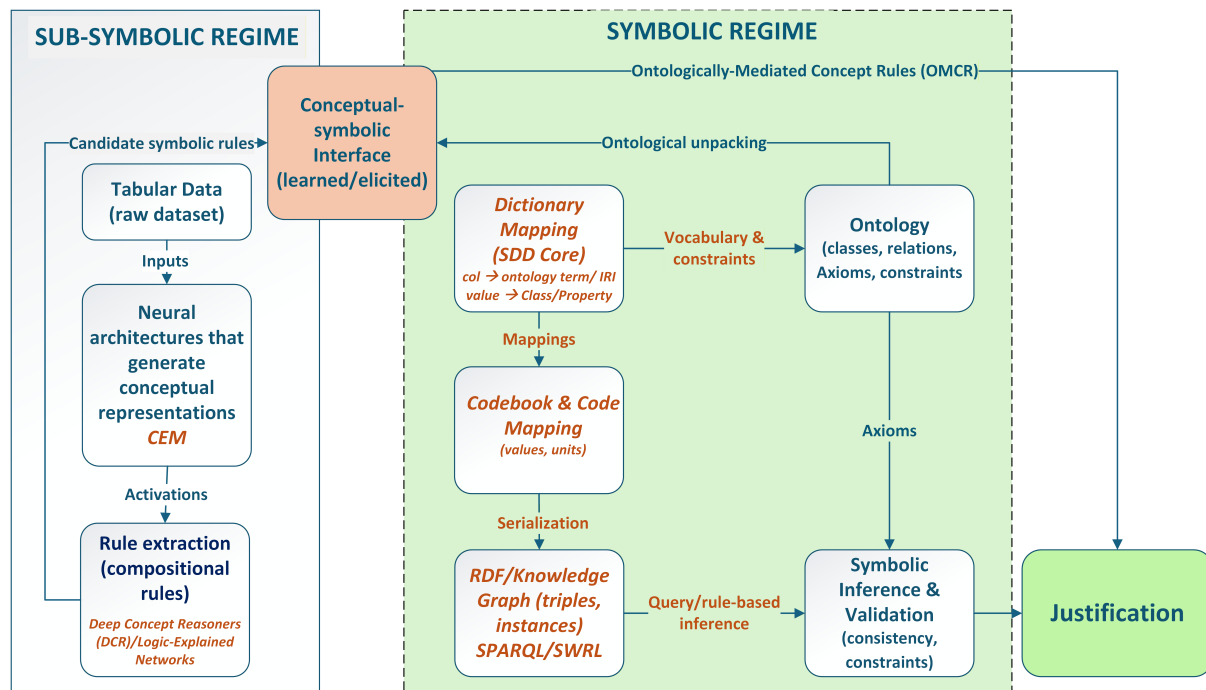


Figure 2. Concept-based rules are refined via ontological unpacking into **Ontologically Mediated Concept Rules (OMCRs)**, which validate and justify neural predictions. The ontological and conceptual artifacts that instantiate the case study in this framework are highlighted in orange and italics.

menberg [39] suggests, such concepts represent provisional stabilizations rather than final captures of reality. This necessitates ontological mediation. Ontologies formalize identity criteria and admissible inferences, enabling the articulation of underlying commitments. Through ontological unpacking [8], conceptual explanations become inspectable and evaluable. Justification thus emerges from mediating concept-based explanations with explicit ontological structures, stabilizing semantic reference and enabling intersubjective evaluation.

These two regimes can be analytically related to the dual-process theory of cognition underlying decision-making and judgment: fast, automatic, and intuitive processes (*System 1*), and slower, reflective, and deliberative processes (*System 2*) [40]. This correspondence is intended as an analogy rather than a psychological commitment, serving to clarify the distinction between explanatory modalities. As illustrated in Figure 2, the integration of these regimes within a unified architecture enables their coordination and contrast, while preserving their distinct ontological roles. Our objective is to integrate ontologies into the explainability process via SDDs, generating rules and knowledge graphs that conceptually structure explanations [38]. The following sections illustrate how these artifacts mediate between explanatory regimes.

4. Proposed Framework

Given the limitations of post-hoc XAI approaches, we propose an architecture that combines complementary sub-symbolic and symbolic regimes through Ontological Mediation. The sub-symbolic regime is responsible for learning concept-sensitive representations and reconstructing Conceptual Rules from data, whereas the symbolic regime represents explicit ontological commitments and normative constraints required for their interpretation and justification. Rather than competing paradigms, these regimes perform complementary functions: the former identifies statistically meaningful conceptual regularities, while the latter stabilizes their semantic interpretation and supports ontology-mediated justificatory assessment.

To illustrate the sub-symbolic regime, we adopt Concept Embedding Models (CEMs)³ [41–44], which

³<https://github.com/mateoespinosa/cem/> and https://github.com/pietrobarbiero/pytorch_explain/blob/master/README.rst

reconstruct explicit Conceptual Rules from learned concept representations. These rules capture predictive conceptual regularities but, by themselves, remain conceptual artifacts whose ontological commitments, explanatory dependencies, and domain constraints are not explicitly represented.

The symbolic regime complements this limitation by representing the semantic and normative knowledge required to interpret those conceptual rules. Domain experts, acting within the Normative Epistemic Context, manually construct Ontologically Mediated Conceptual Rules (OMCRs) that formalize explicit ontological commitments, domain constraints, institutional norms, and other context-dependent epistemic requirements. These OMCRs are incorporated into the ontology as normative semantic artifacts rather than being generated automatically from the learned Conceptual Rules. During inference, a reconstructed Conceptual Rule activates Ontological Mediation, which retrieves the corresponding OMCR and operationalizes it to support ontology-mediated justificatory assessment.

Ontologies provide the explicit semantic reference underlying this process by specifying identity criteria, taxonomic structures, semantic relations, admissible inferences, and domain constraints. Consequently, Ontological Unpacking is guided by the ontology, making explicit the commitments that remain implicit in the reconstructed Conceptual Rules. The resulting OMCRs therefore function as mediating artifacts between sub-symbolic conceptualizations and symbolic ontological commitments, enabling explanatory claims to be evaluated under explicit semantic and normative constraints.

As a practical engineering realization of the symbolic regime, our implementation adopts a Semantic Data Dictionary (SDD)⁴ [45]. The SDD is used here as a semantic mediation workbench rather than as an ontology itself. Following the MIREOT protocol, it supports modular ontology reuse and provides the semantic mapping infrastructure through which domain concepts and dataset elements are associated with independently developed ontological resources.⁵

An important characteristic of the proposed framework is therefore its implementation independence. The theory requires only a semantic infrastructure capable of representing ontological commitments, normative constraints, and their relationships while supporting semantic retrieval or inference over the corresponding OMCRs. Depending on the adopted knowledge architecture, these commitments may be represented directly within the core domain ontology or distributed across specialized normative ontologies formalizing regulations, protocols, institutional policies, clinical guidelines, or other context-dependent epistemic commitments. Consequently, the proposed framework naturally accommodates heterogeneous semantic ecosystems while remaining independent of any particular ontology language, ontology repository, reasoning platform, or semantic engineering technology.

5. Case Study and Proof of Concept: Ontology-Grounded Justificatory Explanations in Alzheimer's Disease

We illustrate the proposed architecture using Alzheimer's disease data from the **OASIS-2 dataset**⁶, comprising longitudinal MRI data for 150 individuals aged 60–96 (373 sessions) [41]. Subjects were assessed using the **Clinical Dementia Rating (CDR)** scale, enabling diagnostic classification.

The dataset includes nine attributes combining demographic, cognitive, and neuroanatomical variables (e.g., MMSE, CDR, nWBV, eTIV), supporting a binary dementia/nondementia classification. Data preparation was intentionally conservative, prioritizing evaluation of semantic grounding under realistic conditions rather than predictive optimization.

⁴<https://github.com/tetherless-world/SemanticDataDictionary>

⁵In this implementation, the SDD provides a semantic grounding infrastructure connected to an ontological grounding infrastructure. By mapping dataset elements to ontology identifiers, it supports semantic retrieval, inference, provenance, and interoperability (<https://tetherless-world.github.io/sdd/>).

⁶<https://sites.wustl.edu/oasisbrains/home/oasis-2/>

Table 1. Characteristics of Alzheimer's Disease with their respective measurements.

Characteristic	Definition	Values, Range
Group	Class	Demented, Nondemented, Converted ⁷
M/F	Sex	M or F
Age	Age at time of image acquisition	Years
YearsEducation	Years of education	Years
SES	Socioeconomic status as assessed by the Hollingshead Social Position Index	1 (highest status) to 5 (lowest status)
MMSE	<i>Mini Mental State Examination</i>	0 = worst to 30 = best In the sample range from 4 to 30
CDR	<i>Clinical Dementia Rating</i>	0 = no dementia, 0,5 = very mild, 1 = mild, 2 = moderate
eTIV (cm³)	<i>Estimated total intracranial volume</i>	Sample values range from 1.106 to 2.004
nWBV (%)	<i>Normalize Whole Brain Volume</i>	Sample values range from 0,644 to 0,837
ASF (no unit)	<i>Atlas Scaling Factor</i>	Sample values range from 1,0000 to 1,3800

Diagnostic concepts were reconstructed from feature-importance analyses, serving both as learnable representations and as explicit substitutes for expert knowledge. Based on the study by Chakraborty and colleagues [42] the table above was developed and, below, the rules for obtaining a diagnosis of Dementia or Non-dementia: concept definitions were operationalized as weighted aggregations of normalized attributes with explicit thresholds.

Dementia (Positive diagnosis for Alzheimer's disease.) = Aggregation($0.48 \times \text{MMSE}$, if $\text{MMSE} \leq 0.77$; $0.17 \times \text{ASF}$, if $\text{ASF} \leq 0.37$; $0.16 \times \text{CDR}$, if $\text{CDR} > 0.25$; $0.09 \times \text{Age}$, if $\text{Age} \leq 0.44$; $0.08 \times \text{eTIV}$, if $\text{eTIV} > 0.31$; $-0.03 \times \text{nWBV}$, if $0.74 < \text{nWBV} \leq 0.75$; $0.03 \times \text{Sex}$, if $\text{Sex} \leq 0.00$ (Female); $0.02 \times \text{YearsEducation}$, if $0.75 < \text{YearsEducation} \leq 0.88$; $0.02 \times \text{Sex}$, if $0.00 < \text{Sex}$ (Male) > 0.00)

Non-Dementia (Negative diagnosis for Alzheimer's disease.) = Aggregation($-0.17 \times \text{MMSE}$, if $0.77 < \text{MMSE} \leq 0.92$; $-0.13 \times \text{CDR}$, if $\text{CDR} \leq 0.00$; $-0.08 \times \text{nWBV}$, if $0.74 < \text{nWBV} \leq 0.75$; $-0.05 \times \text{ASF}$, if $0.43 < \text{ASF} \leq 0.49$; $0.04 \times \text{Sex}$, if $\text{Sex} \leq 0.00$ (Female); $-0.02 \times \text{eTiv}$, if $0.22 < \text{eTIV} \leq 0.26$; $0.02 \times \text{Age}$, if $0.44 < \text{Age} \leq 0.56$; $-0.02 \times \text{Sex}$, if Sex (Male) > 0.00 ; $0.00 \times \text{YearsEducation}$, if $0.75 < \text{YearsEducation} \leq 0.88$)

Table 2. Dementia Concept – Non-Dementia Concept

Characteristic	Dementia		Non-dementia	
	Weight	Condition	Weight	Condition
MMSE	0.48	$\text{MMSE} \leq 0.77$	-0,17	$0.77 < \text{MMSE} \leq 0.92$
ASF	0.17	$\text{ASF} \leq 0.37$	-0.05	$0.43 < \text{ASF} \leq 0.49$
CDR	0.16	$\text{CDR} > 0.25$	-0.13	$\text{CDR} \leq 0.00$
Age	0.09	$\text{Age} \leq 0.44$	0.02	$0.44 < \text{Age} \leq 0.56$
eTIV	0.08	$\text{eTIV} > 0.31$	-0.02	$0.22 < \text{eTIV} \leq 0.26$
nWBV	-0.03	$0.74 < \text{nWBV} \leq 0.75$	-0.08	$0.74 < \text{nWBV} \leq 0.75$
Sex (Female)	0.03	$\text{Sex} \leq 0.00$ (Female)	0.04	$\text{Sex} \leq 0.00$ (Female)
YearsEducation	0.02	$0.75 < \text{YearsEducation} \leq 0.88$	0.00	$0.75 < \text{YearsEducation} \leq 0.88$
Sex (Male)	0.00	$\text{Sex} > 0.00$ (Male)	-0.02	$\text{Sex} > 0.00$ (Male)

⁷Converted: a group of patients whose Group classification changed during the study.

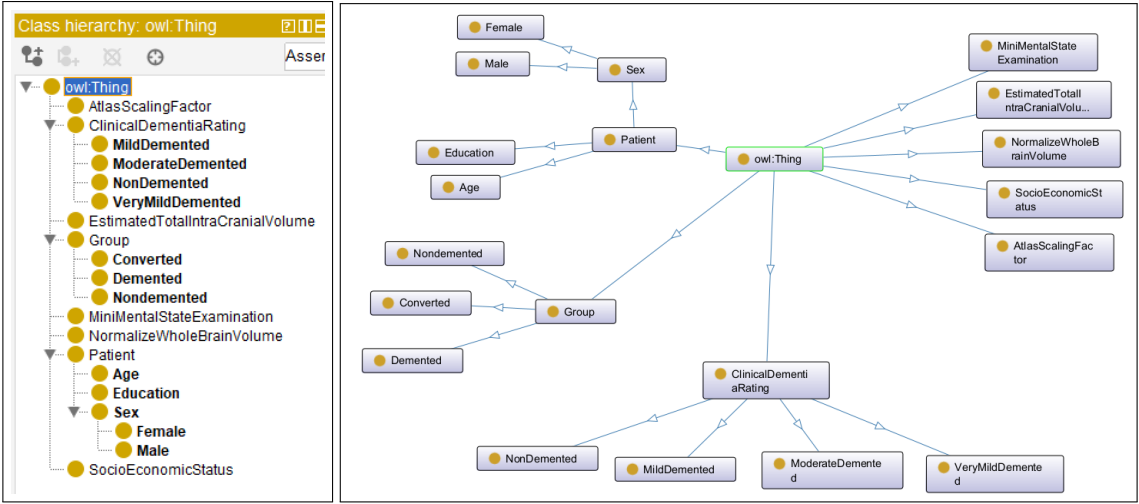


Figure 3. Hierarchy displaying classes and subclasses and their relationships. Ontology built from Oasis-2 data⁸.

A domain ontology was developed in **Protégé** to represent diagnostic categories and patient attributes, enabling classification through symbolic inference, *cf.* Figure 3.

The sub-symbolic regime was implemented using Concept Embedding Models (CEM). From this setup, conceptual rules were extracted in the form of Logic-Explained Networks (LENs). The experiment produced a total of 47 rules, of which 25 supported the diagnosis of Dementia and 22 supported Non-dementia. For the purposes of this study, we selected a subset of seven representative rules that are sufficient to illustrate the proposed mediation mechanism (*cf.* Table 3).

Table 3. Conceptual Rules (LENs) associated with the concepts of Dementia and Non-dementia.

Nº	Diagnostic	Rule
1	Dementia	(~Age & nWBV)
2	Dementia	(~eTIV & ~Age & YearsEducation)
3	Dementia	(eTIV & MMSE & ~Age & YearsEducation)
4	Dementia	(SES & ~ASF & ~nWBV)
5	Dementia	(SES & MMSE & ~Age & ~Sex & ~nWBV)
6	Dementia	(SES & MMSE & ~Age & ~YearsEducation & ~nWBV)
7	Non-dementia	(~eTIV & YearsEducation & ~Age & ~Sex & nWBV & ASF)

Inspired by the notion of weak truthmaker discussed by Guizzardi, Guarino and Porello [8, 18], we interpret Ontologically Mediated Concept Rules (OMCRs) as weak truthbearers. Whereas weak truthmakers explain how propositions may depend on contingent qualities instantiated by entities rather than on their mere existence, OMCRs carry explicit propositional content concerning the justificatory admissibility of conceptual classifications. Their justificatory force is therefore contingent upon the actual instantiation of clinically relevant qualities and explicit normative commitments rather than upon universal ontological conditions alone. In this sense, an OMCR expresses an ontology-mediated proposition that grounds explanation not merely in predictive relevance but in explicit semantic and normative commitments.

The quality normalized Whole Brain Volume (nWBV) provides an illustrative example. A reduced nWBV may contribute to making true propositions concerning cerebral atrophy and cognitive decline, while variations in this quality may modify the corresponding truth conditions. Consider the OMCR: "A NonDementia diagnosis is epistemically admissible only if reduced cognitive performance is not contradicted by significant structural atrophy." This proposition does not simply perform an inference. Rather, it expresses an ontology-mediated justificatory claim whose admissibility depends simultaneously on the patient's instantiated qualities and on explicit ontological commitments represented in the symbolic regime. In this sense, the OMCR functions as a weak truthbearer.

Table 4. Conceptual Rules and Corresponding Ontologically Mediated Concept Rules (OMCR)⁹

N°	Rules, objectives, consequences
1	1. <i>Dementia: ($\sim eTIV$ & $\sim Age$ & $YearsEducation$)</i> 2. <i>Dementia: (SES & $\sim ASF$ & $\sim nWBV$)</i> OMCR Fairness Constraint. <i>IF $eTIV < 1300$ AND $nWBV > 0.72$ THEN $AlzheimerDiagnosis \neq Dementia$</i> This OMCR establishes the primacy of structural volumetric markers ($eTIV$, $nWBV$, ASF) over demographic or socioeconomic variables such as Age, $YearsEducation$, and SES. Its purpose is to prevent Alzheimer's classifications based solely on demographic correlations when essential neuroanatomical markers are absent. As an OMCR, the rule operationalizes an ontology-grounded fairness constraint by confronting concept-based predictions with explicit domain commitments. 1When this condition occurs, the system blocks automated classification and issues the alert: "Bias Risk Detected: Prediction relies on socioeconomic correlations. Human validation with culturally adapted neuropsychological tests is required."
2	3. <i>Dementia: ($\sim Age$ & $nWBV$)</i> 4. <i>Dementia: (SES & $MMSE$ & $\sim Age$ & $\sim Sex$ & $\sim nWBV$)</i> 5. <i>Dementia: (SES & $MMSE$ & $\sim Age$ & $\sim YearsEducation$ & $\sim nWBV$)</i> OMCR Atypicality Constraint. <i>IF ($Age < 60$) AND ($MMSE < 20$) THEN $TriggerSearchFor(EarlyOnsetGeneticMarkers OR AlternativeDementia)$</i> This OMCR identifies early-onset Alzheimer's predictions as atypical cases requiring additional validation. When the model predicts Alzheimer's disease in young patients, the rule requires corroboration through genetic markers (e.g., $PSEN1$) due to the low statistical prevalence of the condition in this population. In such situations, the conceptual inference is treated as provisional and the system requests additional evidence and human validation before the prediction can support a diagnostic decision.
3	6. <i>Dementia: ($eTIV$ & $MMSE$ & $\sim Age$ & $YearsEducation$)</i> OMCR Cognitive Reserve Constraint. <i>IF ($YearsEducation > 16$) AND ($nWBV > 0.72$ AND $eTIV < 1300$ AND $ASF < 1.1$)) AND ($MMSE \geq 27$) THEN $Status = HighReservaCognitivaMaskingSymptoms$</i> This OMCR addresses the potential masking effect of cognitive reserve associated with high educational attainment. Its objective is to verify whether the model accounts for situations in which high education levels may conceal cognitive decline in neuropsychological tests despite the presence of structural brain atrophy. When high $YearsEducation$ co-occurs with neuroanatomical indicators of degeneration, the rule flags the prediction as uncertain. In such cases, the system issues the alert: "Clinical uncertainty detected: cognitive reserve may be masking cognitive decline. Additional diagnostic evaluation is recommended."

⁸Figure 3 should be understood primarily as a lightweight domain ontology designed to support the semantic representation of the OASIS-2 dataset and to illustrate the ontology mediated framework proposed in this work. Its purpose is operational rather than foundational. Accordingly, it was not intended to constitute a complete ontological analysis of the domain in the sense commonly adopted in Formal Ontology, where explicit grounding in foundational ontologies such as UFO, BFO, or DOLCE and a systematic analysis of ontological commitments are typically required. The objective of Figure 3 is to provide a semantic representation capable of supporting ontology mediated interpretation and justificatory assessment within the case study.

⁹The threshold values for the diagnostic criteria are dataset-dependent and have distinct influences on the diagnosis. $MMSE$ and $nWBV$ are treated as core diagnostic features, ASF as a supporting structural feature, Age as a contextual factor, $eTIV$ as a normalization feature. The values adopted here, in OMCR, are illustrative, using the OASIS-2 dataset as a reference.

Nº	Rules, objectives, consequences
4	7. Non-dementia: ($\sim eTIV$ & $YearsEducation$ & $\sim Age$ & $\sim Sex$ & $nWBV$ & ASF) OMCR-CR-2 Cognitive Reserve / Structural Tension Constraint. <i>If Non-dementia is supported by $YearsEducation$ while Age and Sex are explicitly considered but do not contribute positively to the classification, and structural evidence remains mixed ($nWBV \leq 0.72$ AND $ASF > 1300$), THEN classify the inference as epistemically unstable and require validation.</i> This OMCR establishes that educational attainment cannot justify Non-dementia when structural indicators remain semantically unstable or partially conflicting. If Non-dementia is supported by $YearsEducation$ under mixed structural evidence ($nWBV$, ASF), classify the inference as epistemically unstable and require validation. Block automatic justificatory acceptance; require further clinical validation.

Table 4 summarizes four illustrative cases of ontological mediation. In each case, Conceptual Rules reconstructed by the CEM are contrasted with normative constraints represented in the symbolic regime, giving rise to the corresponding OMCRs.¹⁰

To make the underlying semantic commitments explicit, Figures 4 and 5 present an OntoUML conceptual model highlighting identity criteria, roles, phases, qualities, and other ontological distinctions that remain implicit in the operational domain ontology. The OntoUML model therefore constitutes an ontological unpacking of the conceptual structures employed in the case study.

We now illustrate the framework process using Conceptual Rule 4 presented in Table 4:

$(\sim eTIV \ \& \ YearsEducation \ \& \ \sim Age \ \& \ \sim Sex \ \& \ nWBV \ \& \ ASF) \rightarrow NonDemented$

This rule expresses that a particular configuration of attributes is statistically relevant for classifying a patient as NonDemented. As a product of the sub-symbolic regime, it captures a learned conceptual regularity with predictive value, yet it provides little explicit information concerning the ontological status of the participating concepts or the semantic assumptions underlying the classification.

The first step toward justification consists of Ontological Unpacking, through which the implicit semantic and ontological assumptions embedded in the reconstructed Conceptual Rule are made explicit. Rather than modifying the learned rule, Ontological Unpacking identifies explanatory dependencies, identity criteria, domain constraints, and other commitments required for interpreting the learned conceptualization. Questions naturally emerge during this process: Is $YearsEducation$ a diagnostic criterion or merely a contextual modulator? Is the observed $nWBV$ compatible with a NonDemented condition? Does ASF possess independent diagnostic significance? Which relationships are constitutive of the diagnosis and which merely express statistical correlation? Such questions expose assumptions that remain invisible within the purely sub-symbolic representation.

This example also illustrates how Ontological Unpacking resolves semantic ambiguities present in learned conceptualizations. Within the reconstructed rule, the expression $\sim Age$ indicates that the concept Age was considered during learning but does not contribute positively to the predicted classification. In other Conceptual Rules, however, the absence of positive contribution is represented simply by omitting Age altogether. Although these two situations may appear operationally similar, they express different semantic commitments. The notation $\sim Age$ explicitly conveys conceptual information about the role assigned to Age within the learned representation, whereas omission expresses no explicit position regarding its explanatory contribution. Importantly, neither case implies that Age is ontologically irrelevant, since Age remains an intrinsic quality necessarily instantiated by every patient. Ontological Unpacking therefore distinguishes conceptual relevance from ontological existence, making explicit

¹⁰The constraint categories shown in Table 4 are illustrative rather than intended as an exhaustive taxonomy. They are included solely to exemplify the variety of normative constraints that may be incorporated into an OMCR during ontological mediation and justificatory assessment. A more systematic classification of OMCR constraint types is currently being developed as part of ongoing research, with particular attention to the characterization of epistemic and normative biases affecting both AI systems and human reasoning. Such a taxonomy, however, lies beyond the scope of the present paper.

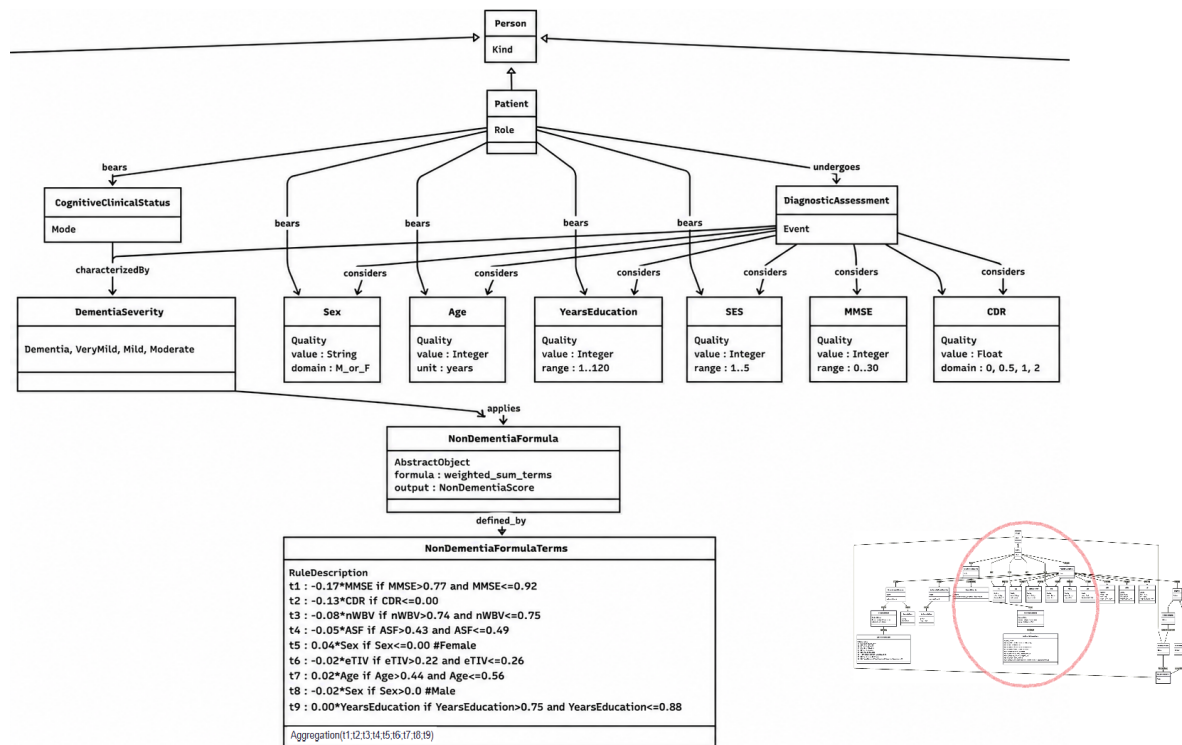


Figure 4. OntoUML conceptual model of the Alzheimer's disease case study. The model explicitly represents the ontological commitments underlying the diagnostic domain, including persons, patients, clinical phases, qualities, assessments, diagnostic scores, and classification outcomes. The image shows a crop (zoom) of the complete diagram shown at the bottom right.

commitments that remain implicit in the learned representation.

The explicit commitments recovered through Ontological Unpacking support the construction of the corresponding Ontologically Mediated Concept Rule (OMCR). During inference, the reconstructed Conceptual Rule activates Ontological Mediation, which retrieves and operationalizes the appropriate OMCR for the current justificatory context. The resulting OMCR neither replaces the learned Conceptual Rule nor the domain ontology. Instead, it functions as a normative semantic artifact connecting the conceptualization produced by the sub-symbolic regime with the explicit commitments represented in the symbolic regime.

Given these explicit commitments, the learned classification may be judged epistemically admissible, inadmissible, or epistemically unstable. In the present example, the OMCR may indicate that the NonDemented classification is epistemically unstable because the patient's structural markers do not sufficiently support—or may even contradict—the proposed interpretation. Such assessments are not merely inferential outcomes. They express explicit normative judgments concerning the justificatory status of the learned conceptualization. Within the framework ontology, the corresponding OMCR is therefore represented as a weak truthbearer, carrying propositional content whose admissibility is grounded both in explicit ontological commitments and in the contingent configuration of clinically relevant qualities instantiated by the patient.

This case study illustrates that explainability in high-stakes domains cannot be reduced to model introspection alone. Instead, explanation becomes an ontology-mediated justificatory process in which learned conceptualizations are interpreted, evaluated, and constrained by explicit ontological and normative commitments. Ontological Grounding stabilizes concept meaning, Ontological Unpacking makes implicit commitments explicit, Ontological Mediation operationalizes the appropriate OMCRs, and Justificatory Assessment evaluates whether the resulting conceptual classification is epistemically admissible within the relevant domain context.

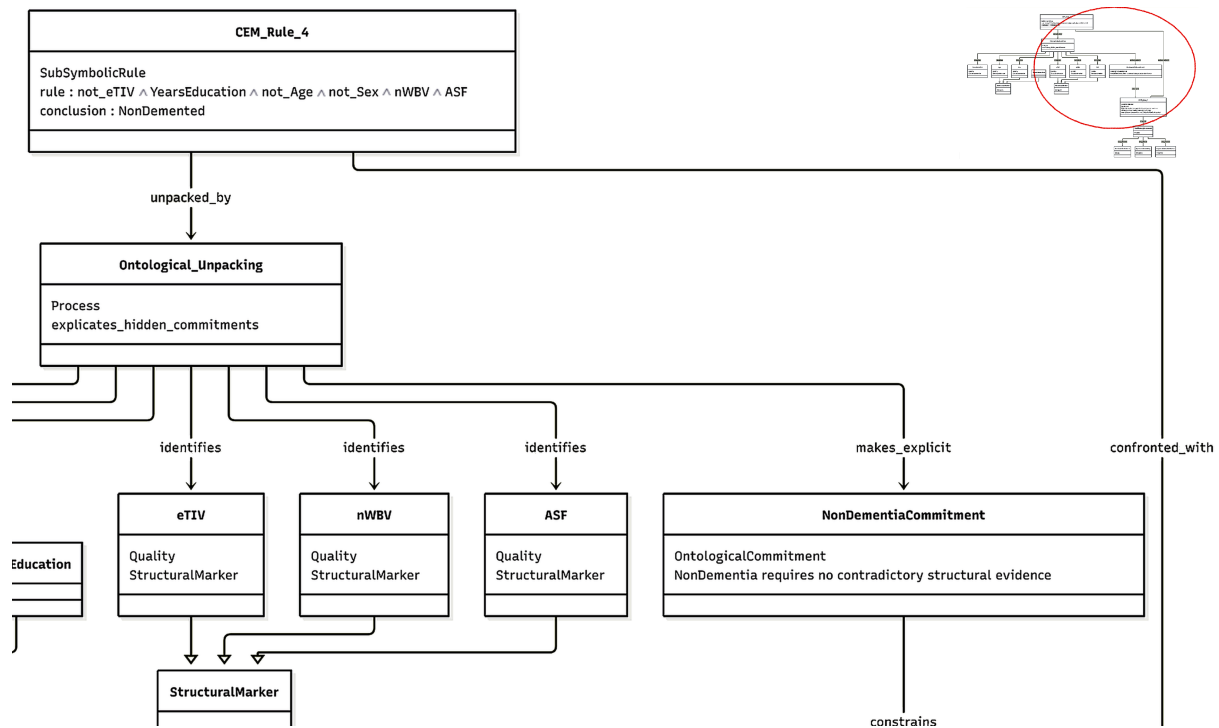


Figure 5. Ontological unpacking of conceptual rule 4. The diagram explicates the ontological commitments underlying the learned classification rule and illustrates how contextual and structural markers are transformed into an Ontologically Mediated Concept Rule (OMCR), enabling ontology mediated justificatory assessment. The image shows a crop (zoom) of the complete diagram shown at the upper right.

6. Discussion: Ontological Grounding and the Limits of Technically-Centered Explainability

The framework proposed in this paper extends current discussions on explainability by treating explanation not merely as a property of computational models, but as an epistemic and normative process grounded in explicit ontological commitments. Rather than revisiting the proposed architecture, we discuss here how this perspective relates to existing research in Explainable AI (XAI), Concept-Based XAI (C-XAI), and Neuro-Symbolic AI (NeSy), and the broader implications of ontology-grounded justificatory assessment. Recent work has highlighted the multiple roles that ontologies can play in supporting explanations in Neuro-Symbolic AI, including reference modelling, common-sense reasoning, knowledge refinement, and complexity management.

This perspective also complements ongoing developments in Concept-Based XAI. Concept-based methods improve interpretability by organizing explanations around semantically meaningful concepts rather than low-level features. However, the epistemic status of these concepts often remains implicit, depending on domain assumptions that are neither formally represented nor systematically assessable. Ontological mediation provides a mechanism for explicitly relating learned conceptualizations to domain ontologies, thereby enabling concept-level explanations to become semantically grounded, contestable, and normatively assessable.

Similar observations apply to the emerging agenda of XAI 2.0, which emphasizes multidimensional explainability, human-centered evaluation, social accountability, and justification. Many of the challenges identified in this literature – including contestability, trustworthiness, and accountability – cannot be fully addressed through increasingly sophisticated interpretability techniques alone. They require explicit mechanisms for representing and evaluating the ontological commitments underlying explanatory claims.

Ontological grounding and ontology mediated justificatory assessment provide one possible approach to this broader objective. The proposed framework also contributes to the broader Neuro-Symbolic AI

agenda. While NeSy seeks to integrate data-driven learning with symbolic reasoning, our contribution focuses specifically on the interface between learned conceptualizations and explicit ontological commitments. In this view, ontological unpacking and ontological grounding extend the role of symbolic representations beyond inference, enabling traceability, validation, contestability, and justificatory assessment of reconstructed explanations.

More broadly, this perspective has implications for the governance of AI systems deployed in regulated and safety-critical domains. Legal, ethical, and institutional requirements increasingly demand explanations that are not only intelligible but also auditable, contestable, and normatively defensible. By making ontological commitments explicit, ontology mediated justificatory assessment provides a structured basis for examining the assumptions underlying AI-assisted decisions and for supporting transparent human oversight.

The present work intentionally focuses on establishing the conceptual foundations of ontology mediated justificatory assessment rather than proposing a complete taxonomy of ontological constraints or domain-specific mediation strategies. Likewise, it does not seek to replace existing explainability techniques, but rather to complement them by providing an explicit normative layer through which reconstructed explanations may be evaluated.

Finally, the framework developed here suggests that future research on explainability should focus not only on extracting explanatory artifacts, but also on the ontological conditions under which explanations become semantically grounded, epistemically admissible, and normatively justifiable. From this perspective, explainability becomes not simply a matter of interpreting model behavior, but of explicitly representing and assessing the ontological commitments that support responsible AI. Future work will investigate systematic classifications of OMCR constraint types, broader empirical validation across domains, and additional mechanisms for integrating ontological mediation into explainable and neuro-symbolic AI systems¹¹.

7. Conclusion

Current approaches in Explainable Artificial Intelligence increasingly incorporate concepts, symbolic representations, and semantic structures, yet explanatory adequacy remains limited when explicit ontological mediation is absent. We argue that this limitation is fundamentally epistemic, reflecting the lack of mechanisms capable of grounding conceptual meaning and supporting the justificatory assessment of explanatory claims.

The case study illustrates that learned conceptual rules, although statistically predictive, remain conceptual artifacts unless their underlying ontological commitments are made explicit. To address this limitation, we proposed an ontology mediated framework in which conceptual rules undergo ontological unpacking, making their ontological and semantic commitments explicit and enabling the construction of Ontologically Mediated Conceptual Rules (OMCRs). Within this framework, interpretation, explanation, and justification are treated as distinct but interconnected dimensions of explainable AI, linked through Ontological Mediation.

Building on recent discussions of weak truthmaking, we further proposed interpreting OMCRs as Weak Truthbearers. Rather than functioning as classical truthbearers, they support context-dependent justificatory assessments by mediating between sub-symbolic conceptualizations and symbolic ontological commitments under explicit normative and semantic constraints.

Our central claim is that ontological mediation is not merely an optional enhancement of explainability, but a necessary condition for epistemically adequate explanations in high-stakes domains. By making explicit the ontological conditions under which explanatory claims become meaningful, admissible, and justifiable, the proposed framework establishes a foundation for ontology mediated justification in explainable AI. More broadly, it suggests that future research should move beyond improving explanation

¹¹The formalization of OMCR and ontological mediation, among other developments, can be followed at: <https://github.com/em4ttos/Ontological-Mediation-in-Explainable-Artificial-Intelligence>

extraction toward investigating the ontological and normative conditions that enable explanations to be meaningfully interpreted, critically assessed, and epistemically justified.

References

- [1] 2022. Bordt, S., Fink, M., Raidl, E., & von Luxburg, U. Post-hoc explanations fail in adversarial contexts. In FAccT 2022, 2022.
- [2] 2018. Noble, S. U. Algorithms of oppression. New York: NYU Press, 2018.
- [3] 2016. O’Neil, C. Weapons of math destruction. New York: Crown, 2016.
- [4] 2019. Rudin, C., & Radin, J. Why are we using black box models? Harvard Data Science Review, 2019.
- [5] 2022. Vale, D., El-Sharif, A., & Ali, M. Post-hoc XAI and non-discrimination law. AI and Ethics, 2022.
- [6] 2021. Confalonieri, R., Weyde, T., Besold, T. R., & Moscoso del Prado Martin, F. Using ontologies to enhance human understanding of post-hoc explanations. Artificial Intelligence, 296, 2021.
- [7] 2024. Longo, Luca et al. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research direction. Information Fusion 106, 2024. 22p.
- [8] 2023. Guizzardi, G., & Guarino, N. Semantics, ontology and explanation. arXiv:2304.11124, 2023.
- [9] 2023. Memarian, B., & Doleck, T. FATE in AI and higher education. Computers and Education: AI, 5, 2023.
- [10] 2024. Singhal, A., et al. FATE in AI for healthcare. JMIR Medical Informatics, 2024.
- [11] 1990. Harnad, S.: The Symbol Grounding Problem. Physica D 42(1–3), 335–346 (1990).
- [12] 2005. Taddeo, M., Floridi, L.: Solving the Symbol Grounding Problem: A Critical Review of Fifteen Years of Research. Journal of Experimental & Theoretical Artificial Intelligence 17(4), 419–445 (2005).
- [13] 2025. Nananukul, N., Zhang, Y., Lee, R., Boxer, E., May, J., Gogate, V.G., Pujara, J., Kejriwal, M.: LOGicalThought: Logic-Based Ontological Grounding of LLMs for High-Assurance Reasoning. arXiv preprint arXiv:2510.01530 (2025).
- [14] 2024. Bajaj, G., Parthasarathy, S., Shalin, V.L., Sheth, A.: Grounding from an AI and Cognitive Science Lens. arXiv preprint arXiv:2402.13290 (2024).
- [15] 2016. Thompson, N.: Grounding and Metaphysical Explanation. Proceedings of the Aristotelian Society 116(3), 395–402 (2016).
- [16] 2012. Fine, Kit.: Guide to Ground. In: Correia, F., Schnieder, B. (eds.) Metaphysical Grounding: Understanding the Structure of Reality, pp. 37–80. Cambridge University Press, Cambridge (2012).
- [17] 2019. Guarino, N., Porello, D., Guizzardi, G.: On Weak Truthmaking. In: Proceedings of the Joint Ontology Workshops (JOWO) (2019).
- [18] 2025. Confalonieri, R., Guizzardi, G. On the Multiple Roles of Ontologies for Explanations in Neurosymbolic AI. Neuro-Symbolic AI Journal, 2025.
- [19] n.d. Poeta, E., Ciravegna, G., et al. Concept-based XAI: A survey. ACM Computing Surveys, 2025 (forthcoming).
- [20] 2022. Bouneffouf, D., Aggarwal, C.C.: Survey on Applications of Neurosymbolic Artificial Intelligence. arXiv preprint arXiv:2209.12618 (2022).
- [21] 2024. Cunningham, D., Law, M., Lobo, J., Russo, A.: The Role of Foundation Models in Neuro-Symbolic Learning and Reasoning. arXiv preprint arXiv:2402.01889 (2024).
- [22] 2020. Garcez, A.S.d., Lamb, L.C.: Neurosymbolic AI: The 3rd Wave. arXiv preprint arXiv:2012.05876 (2020).
- [23] 2023. Gibaut, W., Pereira, L., Grassiotto, F., Osorio, A., Gadioli, E., Munoz, A., Gomes, S., dos Santos, C.: Neurosymbolic AI and its Taxonomy: A Survey. arXiv preprint arXiv:2305.08876 (2023).
- [24] 2024. Lu, Z., Afridi, I., Kang, H.J., Ruchkin, I., Zheng, X.: Surveying neuro-symbolic approaches for reliable artificial intelligence of things. Journal of Reliable Intelligent Environments 10, 257–279 (2024).

- [25] 2022. Manhaeve, R., Marra, G., Dumancic, S., Kimmig, A., De Raedt, L.: Neuro-Symbolic AI: Neural + Logical + Probabilistic AI. In: Hitzler, P., Sarker, M.K. (eds.) *Neuro-Symbolic Artificial Intelligence: The State of the Art. Frontiers in Artificial Intelligence and Applications*, vol. 342, pp. 173–191. IOS Press (2022).
- [26] 2021. Sarker, M.K., Zhou, L., Eberhart, A., Hitzler, P.: *Neuro-Symbolic Artificial Intelligence: Current Trends*. arXiv preprint arXiv:2105.05330 (2021).
- [27] 2023. Yu, D., Yang, B., Liu, D., Wang, H., Pan, S.: A survey on neural-symbolic learning systems. *Neural Networks* 166, 105–126 (2023).
- [28] 2023. Chari, S., Seneviratne, O., Gruen, D. M., et al. *Explanation ontology: A model for user-centered AI*. *Semantic Web*, 2023.
- [29] 2021. Memariani, A., et al. *Explainable ontology extension*. In *DAO-XAI Workshop*, 2021.
- [30] 2021. Nakagawa, P. I., et al. *Semantic description of explainable ML workflows*. *Applied Sciences*, 11, 2021.
- [31] 2024. Beverley, J., Franda, F., Karray, H., et al. *Ontologies, arguments, and large language models*. *Proceedings of the Joint Ontology Workshops (JOWO)*, 2024.
- [32] 2024. Xhani D., Rebelo Moreira, J. L., van Sinderen, M., & Pires, L. F. XMLPO: An ontology for explainable machine learning pipelines. In *Proceedings of the 14th International Conference on Formal Ontology in Information Systems (FOIS 2024)*, pp. 193–206, 2024.
- [33] 1993. Gruber, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), 1993.
- [34] 1995. Gruber, T. R. Toward principles for the design of ontologies. *IJHCS*, 43(5–6), 1995.
- [35] 2015. Leroux, H., et al. *ODM on FHIR*. In *SWAT4LS*, 2015.
- [36] 2017. Leroux, H., et al. *Semantic interoperability of clinical data with FHIR*. *Jour. Biom. Semantics*, 2017.
- [37] 2014. Weber, G. M., Mandl, K. D., & Kohane, I. S. *Big biomedical data*. *JAMA*, 311(24), 2014.
- [38] 2017. Rashid, S. M., et al. *Semantic data dictionary approach to Data Annotation & Integration.. CEUR-WS*, 2017.
- [39] 2013. Blumenberg, H. *Teoria da nao conceitualidade*. Belo Horizonte: Editora UFMG, 2013.
- [40] 2011. Kahneman, D. *Thinking, fast and slow*. New York: Farrar, Straus and Giroux, 2011.
- [41] 2010. Marcus, D. S., et al. *OASIS: Longitudinal MRI data*. *Journal of Cognitive Neuroscience*, 22(12), 2010.
- [42] 2024. Chakraborty, M., et al. *ANALYSE-AD: A comparative analysis of novel AI approaches for early Alzheimer’s detection*. *Array*, 22, 2024.
- [43] 2022. Zarlenga, M. E., Barbiero, P., et al. *Concept embedding models*. In *NeurIPS 2022*, 2022.
- [44] 2021. Barbiero, P.; Ciravegna, G.; Giannini, F.; Lio, P.; Gori, M.; Melacci, S. *Entropy-Based Logic Explanations of Neural Networks*. arXiv preprint arXiv:2106.06804, 2021.
- [45] 2023. Barbiero, P., Ciravegna, G., Giannini, F., et al. *Interpretable neural-symbolic concept reasoning*. In *Proceedings of ICML 2023*, PMLR 202, 2023.
- [46] 2023. Ciravegna, G.; Barbiero, P.; Giannini, F.; Gori, M.; Lio, P.; Maggini, M.; Melacci, S. *Logic Explained Networks*. *Artificial Intelligence*, vol. 314, 103848, 2023.
- [47] 2020. Rashid, S. M., et al. *The semantic data dictionary*. *Data Intelligence*, 2(4), 2020.

Author Contributions

Eduardo M. P. Coelho: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing.

Marcello P. Bax: Supervision, Writing – Review & Editing.

All authors reviewed and approved the final version of the manuscript.

Data availability statement

The authors declare that the dataset supporting the results of this study are available in the following data repository: <https://sites.wustl.edu/oasisbrains/home/oasis-2/>.

Use of Artificial Intelligence

Generative AI was not used to generate scientific content, formulate arguments, conduct analyses, or produce research results. During the preparation of this work, the authors used ChatGPT (OpenAI, GPT-5.6) for grammar and spelling checking, paraphrasing, and rewording. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflict of Interest statement

The authors declare that there are no conflicts of interest related to this work.

Note on manuscript status

This manuscript was peer-reviewed and accepted for presentation and publication in the proceedings of the OSETAI workshop at FOIS/JOWO 2026. Following acceptance, the paper was withdrawn because the authors were unable to complete registration and attend the event; consequently, it was not included in the final workshop program or proceedings.

This preprint was submitted under the following conditions:

- The authors declare that the necessary Terms of Free and Informed Consent of participants or patients in the research were obtained and are described in the manuscript, when applicable.
- The authors declare that the preparation of the manuscript followed the ethical norms of scientific communication.
- The authors declare that they are aware that they are solely responsible for the content of the preprint and that the deposit in SciELO Preprints does not mean any commitment on the part of SciELO, except its preservation and dissemination.
- The authors declare that the data, applications, and other content underlying the manuscript are referenced.
- The deposited manuscript is in PDF format.
- The authors declare that the research that originated the manuscript followed good ethical practices and that the necessary approvals from research ethics committees, when applicable, are described in the manuscript.
- The authors declare that once a manuscript is posted on the SciELO Preprints server, it can only be taken down on request to the SciELO Preprints server Editorial Secretariat, who will post a retraction notice in its place.
- The authors agree that the approved manuscript will be made available under a [Creative Commons CC-BY](#) license.
- The submitting author declares that the contributions of all authors and conflict of interest statement are included explicitly and in specific sections of the manuscript.
- The authors declare that the manuscript was not deposited and/or previously made available on another preprint server or published by a journal.
- If the manuscript is being reviewed or being prepared for publishing but not yet published by a journal, the authors declare that they have received authorization from the journal to make this deposit.
- The submitting author declares that all authors of the manuscript agree with the submission to SciELO Preprints.