# Extended Nomological Deductive Reasoning (eNDR) for Transparent AI Outputs*

Gedeon Hakizimana [1,*,†]

[1]Universidad Carlos III de Madrid
Department of Computer Science and Engineering
Av. Universidad, 30, 28911 Leganés (Madrid), Spain

**Abstract**

Extended Nomological Deductive Reasoning (eNDR) is a novel Explainable AI framework that integrates causal domain knowledge with deductive logic to deliver transparent, trustworthy predictions in dynamic, high-stakes environments. Building on the original NDR model, eNDR handles continuous data, uncertainty, and real-time inference by expressing domain laws as continuous functions, modeling conditions as constraints, and generating explanations through differentiable reasoning and probabilistic integration. Early results show that eNDR produces human-readable, domain-aligned explanations without sacrificing predictive performance, offering a pathway toward AI systems that are both accurate and interpretable across domains such as healthcare, finance, and criminal justice.

## 1. Context and Motivation for the Research

The integration of Explainable AI (XAI) into machine learning (ML) has become crucial due to growing demands for transparency, fairness, and trust, especially in high-stakes areas like healthcare, finance, autonomous systems, and criminal justice. As a matter of fact, many high-performing ML models, such as deep learning architectures, lack interpretability, raising concerns about their decision-making processes.

However, most current XAI methods struggle to produce explanations that are both human-understandable and aligned with domain knowledge. This research builds on the previously proposed Nomological Deductive Reasoning (NDR) framework, which combines deductive logic with causal domain knowledge to improve interpretability, accuracy, and trust. At its core, the Nomological Deductive Knowledge Representation (NDKR) allows AI systems to use structured knowledge bases, enabling predictions supported by clear, logically consistent explanations.

The initial NDR approach was limited to static data. This research aims to extend it (eNDR) to handle continuous data, uncertainty, and complex real-world conditions, leading to more generalizable, robust, transparent and accountable explainable AI model.

## 2. Key Related Work

The field of Explainable AI (XAI) has produced over 200 techniques aimed at improving model transparency. Popular model-agnostic methods like LIME and SHAP offer post hoc explanations by approximating black-box decision boundaries but often lack domain-specific reasoning, limiting their effectiveness in expert-driven fields.

Inherently interpretable models—such as decision trees and rule-based systems—offer more understandable outputs but typically sacrifice predictive performance, highlighting the persistent trade-off between accuracy and interpretability in XAI.

*Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, Turkey*

100476007@alumnos.uc3m.es (G. H. )

A promising alternative is the Nomological Deductive Reasoning (NDR) framework, which balances performance with explanatory depth by integrating deductive logic and causal domain knowledge through the Nomological Deductive Knowledge Representation (NDKR). Inspired by Hempel's covering law model, NDR aims to provide logically structured, cognitively satisfying explanations.

Originally proposed by Hakizimana and Ledezma Espino, NDR was limited to static, deterministic settings. To address real-world complexity, the extended version—eNDR—introduces probabilistic reasoning, continuous data handling, and real-time inference, making it more applicable to dynamic, high-stakes domains like healthcare, finance, and criminal justice.

# 3. Research Questions, Hypothesis, and Objectives

## 3.1. Research Questions

This research will explore the following key questions:

1. How can the Nomological Deductive Reasoning (NDR) framework be extended to handle continuous data and domain-specific uncertainty while preserving the interpretability and trustworthiness of the AI model's explanations?
2. How can the integration of structured knowledge and causal reasoning improve the transparency and explainability of AI systems in high-stakes applications?

## 3.2. Hypothesis

By modeling the Nomological Deductive Reasoning approach through the expression of laws as continuous functions; conditions as constraints on data instances; predictions as a function of data instances, laws, and conditions; and explanation as an integral summing over all possible laws and conditions capturing the cumulative effect of how they interact with the data, the NDR framework can handle continuous data and complex real-world scenarios. In addition, using probabilistic settings and considering the prediction task as an optimization process can enhance the interpretability and transparency of NDR-based predictions in domains with complex data and uncertainty.

## 3.3. Objectives

1. To extend the NDR framework to accommodate continuous data and uncertainty through advanced mathematical methods like calculus and optimization techniques.
2. To test the effectiveness of the extended NDR framework in generating human-comprehensible explanations that are aligned with domain-specific knowledge.
3. To evaluate the performance of AI models using the NDR framework in real-world applications, such as healthcare, finance, or criminal justice.

# 4. Research Approach, Methods, and Rationale for Testing the Hypothesis

The research approach is structured around the extension of the Nomological Deductive Reasoning (NDR) framework.The inclusion of probabilistic models and uncertainty quantification within NDR allows the system to account for variations in the input data and make robust, well-grounded predictions even in the face of noise or uncertainty, enhancing both the model's reliability and its interpretability when making decisions on complex data.

### 4.1. Expressing Laws (L) as Continuous Functions

In our initial development of NDR framework, we have assumed the following:

**Laws (L)**
Let $L$ represent the set of laws or rules governing a certain real-world domain (e.g., healthcare diagnosis, bank credit score, traffic code for mobility applications, criminal justice, etc.). These laws are formalized as logical statements or principles that provide the foundation for reasoning in the system. Each law $L_i \in L$ corresponds to a specific rule or law within the system.
*Example (in medical settings):*

- $L_1$: "If a patient has high blood pressure and is over 60 years old, then they are at a high risk of cardiovascular disease."
- $L_2$: "If a treatment is an ACE inhibitor, it lowers blood pressure."

**Conditions (C)**
Let $C$ denote the set of antecedents or conditions that must hold true in order for a law to be applicable to a particular data instance. Each condition $C_j \in C$ is a prerequisite that must be satisfied for the corresponding law to be activated or relevant.
*Example (in medical settings):*

- $C_1$: "Patient has high blood pressure."
- $C_2$: "Patient is over 60 years old."

**Data Instances (D)**
Let $D = \{d_1, d_2, \ldots, d_k\}$ represent the set of input data fed into the AI system. Each $d_i \in D$ represents a specific data sample.
*Example (in medical settings):*

- $d_1$: A data sample where the patient has high blood pressure and is 65 years old.
- $d_2$: A data sample where the patient has normal blood pressure and is 45 years old.

**Hypothesis or Prediction (H)**
Let $H = \{h_1, h_2, \ldots, h_p\}$ represent the set of predictions or outcomes generated by the AI model. Each $h_i \in H$ corresponds to a specific prediction for the instance $d_i$.
*Example (in medical settings):*

- $h_1$: "The patient is at high risk for cardiovascular disease."
- $h_2$: "The patient is not at high risk for cardiovascular disease."

**Formalized Deductive Inference**
The key goal of the NDR framework is to use deductive reasoning to formalize how the AI model generates a prediction $h_i$ based on the combination of conditions $C$ and laws $L$ applied to the input $d_i$.

$$\forall d_i \in D,\ \exists h_i \text{ such that } (C_1 \land C_2 \land \cdots \land C_n \land L_1 \land L_2 \land \cdots \land L_m) \vdash h_i$$

Where:

- $d_i \in D$ is an input data instance.
- $C_1, C_2, \ldots, C_n$ are the conditions (e.g., patient characteristics like age, blood pressure).
- $L_1, L_2, \ldots, L_m$ are the domain laws (e.g., risk relationships).
- $\vdash$ denotes deductive reasoning leading to prediction $h_i$.

**Formalized Explanation Generation**

Once we have the laws, conditions, and input data, the explanation $E_i$ for the prediction $h_i$ can be expressed as:

$$E_i = f(L, C, d_i) \Rightarrow h_i$$

Where:

- $E_i$ is the explanation for prediction $h_i$.
- $f$ is the function describing how $L$, $C$, and $d_i$ combine to produce $h_i$.
- $\Rightarrow$ indicates the logical flow from input to prediction.

In this research, the first step is to model the domain laws as continuous functions. These laws describe relationships between input variables and outcomes or predictions, and can be formalized as continuous functions that capture gradual changes in dynamic systems or continuous data. For example, in the medical domain:

$$L_i : \mathbb{R}^n \to \mathbb{R}, \quad L_i(x) = \text{some relationship between conditions}$$

*Example:*

$$L_1(x) = \frac{\text{BloodPressure} \cdot \text{Age}}{1 + \text{Age}}, \quad \text{where } x = [\text{BloodPressure}, \text{Age}]$$

This equation models how the interaction between a patient's blood pressure and age influences cardiovascular risk.

## 4.2. Conditions (C) as Constraints on Data Instances

Conditions are modeled as constraints that must hold true for the corresponding laws to be applicable. These are represented as indicator functions:

$$C_j(x) = \begin{cases} 1 & \text{if condition } j \text{ holds true for instance } x \\ 0 & \text{otherwise} \end{cases}$$

*Example (in medical settings):*

- $C_1(x) = \begin{cases} 1 & \text{if blood pressure is high} \\ 0 & \text{otherwise} \end{cases}$

This ensures that the system only applies relevant laws when the conditions are satisfied.

## 4.3. Data Instances (D) and Their Continuous Representation

Data instances $x$ are treated as vectors in an $n$-dimensional space, representing different entities in the real world. For example, in healthcare, the vector

$$x = (\text{BloodPressure}, \text{Age}, \ldots)$$

could represent a particular patient's characteristics. The data set $D$ is formalized as:

$$x \in D \subset \mathbb{R}^n$$

Each $x$ represents an instance with $n$ features that are used in the model.

### 4.4. Hypothesis or Prediction (H) as a Function of Data

The AI model generates a prediction $h_i$ as a function of the data instances, laws, and conditions. This can be represented as:

$$h_i(x) = f(L_1, L_2, \ldots, L_m, C_1, C_2, \ldots, C_n, x)$$

Where $f$ is a function that maps input data to a prediction. This function can take various forms, such as linear, non-linear, or other suitable formulations depending on the model's complexity.

### 4.5. Formalized Deductive Inference Using Differentiable Functions

To formalize the inference process, we use calculus to express how the prediction $h_i(x)$ changes with respect to the input data $x$. The derivative of $h_i(x)$ with respect to $x$ is computed as:

$$\frac{dh_i(x)}{dx} = \sum_{j=1}^{m} \frac{\partial L_j(x)}{\partial x} \cdot C_j(x)$$

This derivative represents how sensitive the prediction is to changes in the input features. The term $\frac{\partial L_j(x)}{\partial x}$ represents how the laws $L_j$ change with respect to the input, and $C_j(x)$ ensures that the law applies only when the condition holds.

### 4.6. Formalized Explanation Generation As an Integration Task

The explanation $E_i$ for a prediction $h_i$ can be generated by integrating over the laws and conditions that contributed to the prediction. This can be expressed as:

$$E_i = \int_C \int_L f(L, C, x)\, dL\, dC$$

This integral sums over all possible laws and conditions, capturing the cumulative effect of how they interact with the data. In a probabilistic setting, we can also use a Bayesian approach:

$$E_i = \int_L P(L \mid x)\, dL$$

Where $P(L \mid x)$ represents the posterior probability of law $L$ given the data instance $x$.

### 4.7. Formalized Deductive Inference as Optimization

In AI systems, particularly those utilizing machine learning, the inference process is often optimized through a loss function. Hence, the optimization problem can be formalized as:

$$\hat{h}_i = \arg\min_{\theta} \sum_{i=1}^{N} \mathcal{L}(\hat{h}_i(x; \theta), y_i)$$

Where $\mathcal{L}$ is the loss function (e.g., mean squared error), $\hat{h}_i$ is the predicted outcome based on model parameters $\theta$, and $y_i$ is the true label. The parameters $\theta$ represent the weights associated with the laws, conditions, and other model components. We propose the architecture of the extended NDR framework as per the illustration in Figure 1.

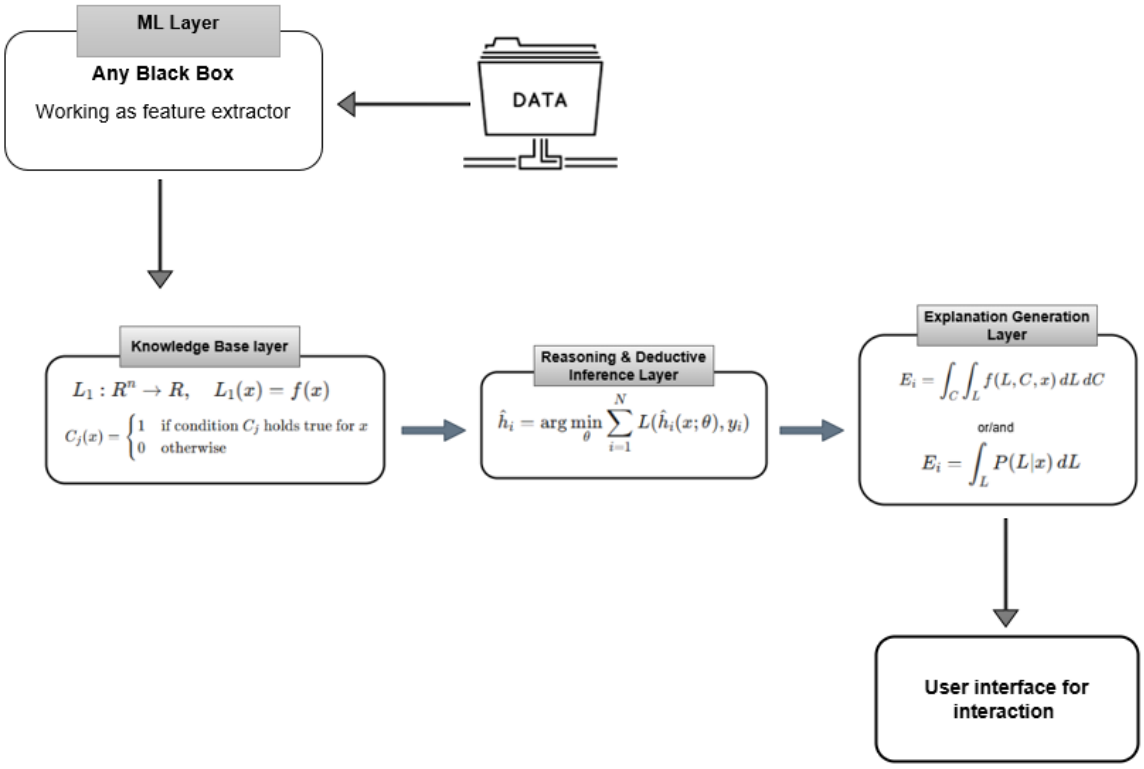


**Figure 1:** Block diagram of the extended Nomological Deductive Reasoning (eNDR) framework.

## 4.8. eNDR Application Scenario

Let's consider the task of predicting the likelihood of a heart attack or stroke based on various risk factors. In this scenario, eNDR framework can be used to model causal relationships from cardiovascular theories, linking the disease to various risks. eNDR then applies the deductive reasoning on complex factors, handles uncertainty, and explains predictions based on the laws governing the medical domain (e.g., age, cholesterol, smoking habits). Any black-box learning model can be used to extract features from the dataset, and by eNDR the output is human-readable knowledge-based explanations.as captured on figure 2.

```
Sample 1:
Model prediction: High risk of CVD (probability: 0.898)
True label: Has CVD
Reasoning based on medical domain laws:
  1. Gender indicates male (activation: 0.871) which is risky according to medical laws
  2. Age (value: 56.0 years) is over the threshold 50 years (activation: 0.843) which is risky according to medical
laws
  3. Max heart rate (value: 98.0 bpm) is under the threshold 120 bpm (activation: 0.743) which is risky according
to medical laws
  4. Cholesterol (value: 203.0 mg/dl) is below the threshold 240 mg/dl (activation: 0.031) and is not risky
according to medical laws
```

**Figure 2:** Sample output of eNDR human-readable knowledge-based explanations.

## 4.9. Metrics for Model Validation

To validate the extended NDR framework with a focus on complex data and uncertainty, a complex health dataset (e.g., the Framingham Heart Study dataset) will be used. This dataset involves multiple features with both causal relationships and uncertainty (due to missing data, variable progression, and

noise), which makes it an excellent candidate for demonstrating the effectiveness of eNDR in providing transparent and interpretable explanations grounded in causal knowledge.

The key metrics will include prediction accuracy to measure the algorithmic performance, uncertainty handling and rule coverage to measure its loyalty to Knowledge Base, as well as reasoning transparency and user trust measurement to evaluate eNDR trustworthiness.

## 5. Preliminary Results and Contributions to Date

Preliminary results show that integrating domain-specific laws into machine learning models enhances interpretability without compromising performance. Early tests in finance reveal that the NDR framework produces explanations aligned with human reasoning, offering clear insights into decision-making. Unlike methods such as Causal Inference, Neuro-Symbolic Reasoning, Knowledge Graphs, LIME, and SHAP, which often cater to technical users, NDR focuses on intuitive, domain-grounded explanations, which is key to trust in AI-powered solutions.

## 6. Expected Next Research Steps and Final Contribution to Knowledge

The next steps involve refining the NDR framework to handle larger, more complex datasets and incorporate probabilistic reasoning to account for uncertainty in real-world data. We will also evaluate the framework in additional domains, such as finance and criminal justice, to assess its generalizability.

The final contribution of this research will be a novel framework for Explainable AI that combines deductive reasoning with domain-specific knowledge, offering a pathway for building more transparent and trustworthy AI systems. This work will also provide insights into how structured knowledge and causal reasoning can be embedded into machine learning models without compromising performance.

## Declaration on Generative AI

The author has used ChatGPT for grammar and content spelling check, after what he reviewed and edited the content as needed. The author takes full responsibility for the publication's content

## References


[1] Caruana, R., Gehrke, J., Koch, P., & Stiksma, M. (2000). Rules for Interpretable Classification Models. In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 278-287). ACM.

[2] Cai, J., Li, H., & Wei, Z. (2020). Rule-based Machine Learning Models for Knowledge Representation and Inference. *Computational Intelligence and Neuroscience*, 2020, Article 1583720.

[3] Gottfried, S., & O'Reilly, T. (2017). Knowledge-Based Systems: A Practical Introduction. Springer.

[4] Guidotti, R., Monreale, A., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black-box Models. *ACM Computing Surveys (CSUR)*, 51(5), 93.

[5] Hakizimana, G.; Ledezma Espino, A. Nomological Deductive Reasoning for Trustworthy, Human-Readable, and Actionable AI Outputs. *Algorithms* 2025, **18**, 306. https://doi.org/10.3390/a18060306

[6] Hendrix, E., & Peeters, S. (2019). Interpretable Machine Learning Models: A Survey of the State of the Art. In Proceedings of the International Conference on Machine Learning (ICML) (pp. 232-241).

[7] Liao, Q. V., & Christodoulou, E. (2019). Model-agnostic Interpretability for Rule-based Systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-10). ACM.

[8] Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267, 1-38.

[9] Preece, S., & Gunning, D. (2019). Explainable AI: A Survey of the State of the Art. *IEEE Transactions on Neural Networks and Learning Systems*, 30(10), 2911-2924.

[10] Vilone, G.; Longo, L. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Machine Learning and Knowledge Extraction* 2021, **3**(3), 615–661. https://doi.org/10.3390/make3030032

[11] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135-1144). ACM.

[12] Vasilenko, A., & Gamanenko, O. (2016). Knowledge-Based Systems and Knowledge Engineering for Explainable AI Applications. *Artificial Intelligence Review*, 45(3), 329-356.

[13] Zhou, S., & Xie, L. (2020). Explainable Artificial Intelligence (XAI) Methods for Healthcare: A Review. *Journal of Healthcare Engineering*, 2020, Article 8706534.

[14] Hempel, C.G.; Oppenheim, P. Studies in the Logic of Explanation. *Philosophy of Science* 1948, 15(2), 135-175.

[15] Feng, S., & Zhang, W. (2020). A Survey on Rule-based Machine Learning Methods for Explainable Artificial Intelligence. *Artificial Intelligence Review*, 53(3), 1621-1644.

[16] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4765-4774).