A Formal Framework for Explainable Artificial Intelligence in High-Reliability Decision Models

Elias Korhonen
School of Technology and Innovations
University of Vaasa, Finland

Sofia Rantala
School of Technology and Innovations
University of Vaasa, Finland

Markus Lehtinen
School of Technology and Innovations
University of Vaasa, Finland

Submitted on: April 21, 2020
Accepted on: May 5, 2020
Published on: May 12, 2020
DOI: 10.5281/10.5281/zenodo.17783076

***Abstract*—High-reliability decision systems require artificial intelligence models that operate with clarity, traceability, and consistency under uncertainty. As machine learning systems increasingly influence operational decisions in domains such as safety engineering, distributed monitoring, and autonomy management, the ability to explain how decisions are produced becomes essential. This paper develops a formal framework for explainable artificial intelligence (XAI) that integrates semantic grounding, structural justification, and computational transparency. The framework is designed to operate across distributed architectures characteristic of early 2020 deployments, where cloud and edge components jointly participate in high-stakes decision processes. Through simulated stress conditions involving conflicting evidence, incomplete inputs, and model perturbations, the framework is evaluated for fidelity, stability, and reasoning completeness. The results demonstrate that systematically engineered explainability improves model oversight while maintaining operational reliability in dynamic environments.**

***Index Terms*—Explainable AI, semantic modeling, interpretability, high-reliability systems, distributed intelligence, computational transparency.**

I. Introduction

Artificial intelligence systems increasingly support decisions that influence operational continuity, safety, and regulatory compliance. In such environments, the reliability of a decision is tied not only to the accuracy of the underlying model but also to the clarity with which the model's reasoning can be understood, audited, and externally validated. Early literature in probabilistic modeling [1], structured argumentation [2], and ontology-based reasoning [3] emphasized the benefits of interpretable representations in supporting transparent decision flows. As distributed computational infrastructures expanded in the late 2010s, with cloud-enabled robotics [4] and autonomous agents operating across variable environments [5], the need for interpretable decision models became more prominent.

Explainability became especially critical when AI-supported decisions required justification across heterogeneous teams and operational layers. Research on cognitive support systems [6], human-in-the-loop analytics [7], and organizational learning frameworks [8] highlighted that operators rely heavily on structured explanations to assess system validity. Similarly, studies of moral reasoning [9], behavioral adaptation [10], and multi-agent collaboration [11] revealed that explainability influences not only outcome acceptance but also model reliability.

This research proposes a formal explainability framework optimized for high-reliability decision systems. The framework integrates vertical reasoning decomposition, semantic anchoring, and input–output trace mechanisms to support full transparency throughout the model's interpretive pipeline. Unlike post-hoc explanation tools, which approximate model behavior, the proposed method embeds explainability directly within the model's computational structure, producing native and verifiable reasoning artifacts.

II. Literature Review

Explainability has emerged as a central requirement for artificial intelligence systems deployed in high-reliability environments. Foundational work in structured argumentation provided one of the earliest formal models for transparent computational reasoning, demonstrating how logic-based explanations support verifiable decisionmaking [2]. Complementary research in probabilistic modeling offered mechanisms for expressing uncertainty in interpretable ways, allowing operators to understand the degree of confidence associated with model

outputs [1]. These early contributions established the theoretical underpinnings of interpretable computational behavior.

The development of cognitive architectures further advanced explainability research. Studies exploring layered cognitive processes [6] illustrated how internal reasoning sequences could be externalized for human understanding. Models examining adaptive learning behaviors under variable conditions [10] and structured pedagogical dynamics [12] highlighted the importance of intermediate explanatory cues in maintaining system transparency. Research on multimodal interaction emphasized the need for interpretable communicative pathways, especially when speech and gesture signals influence decision outcomes [7].

Distributed intelligence introduced new complexities for explainability. Cloud-enabled robotic frameworks [4] and autonomous navigation systems [5] showed that interpretability must be preserved across heterogeneous computational layers, motivating cross-node reasoning alignment. Similarly, ontology-guided decision systems [3] provided structured semantic scaffolding for transparent inference processes in dynamic environments. Research in multi-agent cooperation demonstrated that collaborative reasoning benefits from aligned interpretive structures, reducing ambiguity during task coordination [11].

Explainability also plays an essential role in medical and remote monitoring systems. Early diagnostic models demonstrated that interpretable intermediate steps improved clinician trust, especially when models processed noisy or incomplete physiological data [13], [14]. Additional work in affective and emotional modeling [15] illustrated how explainability can clarify decisionmaking influenced by human behavioral cues, enhancing operator understanding in complex, context-rich environments.

Ethical and normative perspectives expanded explainability research into the domain of responsible AI. The ethical analysis by Vengathattil [16] critically examined whether AI can act responsibly without transparent reasoning structures, emphasizing the need for explanations that expose potential biases, constraints, and moral assumptions. Related studies in moral judgment frameworks [9] and institutional reasoning [8] demonstrated how interpretability supports accountability and ensures that automated decisions remain aligned with human values. Broader philosophical examinations of cognitive alignment [17] and conceptual grounding [18] reinforced the argument that explainability bridges machine reasoning and human interpretive expectations.

Emerging analyses in organizational and sociotechnical systems also contributed to explainability methodology. Investigations into institutional decision flows [19] showed that interpretable AI facilitates knowledge transfer across organizational hierarchies. Studies of existential risks and public communication [20] suggested that transparent reasoning models help mitigate misunderstandings and reduce uncertainty surrounding automated technologies. These perspectives highlight that explainability serves not only operational needs but also broader societal and regulatory expectations.

Technical advances in anomaly detection and adaptive modeling further influenced explainability research. Comparative assessments of detection strategies [21] emphasized the necessity of transparent error attribution in dynamic and drifting environments. Research on distributed access and monitoring [22] noted that explainability helps isolate fault conditions and aids in diagnosing irregular system states. Work on adaptive behavior modeling and teaching frameworks [12] reinforced the importance of reasoning clarity when systems learn or adapt during runtime.

Additionally, studies investigating multimodal cognitive cues [14], task alignment across agents [11], and semantic interpretation pathways [3] contributed to a more holistic understanding of how explanations function within distributed, collaborative, and high-stakes environments. These threads converge toward a consensus that explainability must be embedded into model structure rather than treated as an optional add-on, particularly in systems where reliability, auditability, and operational continuity are essential.

Collectively, the literature demonstrates that explainability is a multidimensional construct influenced by logic, cognition, ethics, distributed coordination, uncertainty modeling, and sociotechnical interpretation. This body of research provides the theoretical foundation for the formal explainability framework developed in this study, which integrates semantic grounding, layered interpretability, and deviation-aware validation to support transparent reasoning in high-reliability decision models.

III. Methodology

The proposed framework is built upon three pillars: vertical decision decomposition, semantic anchoring, and deviation-aware explanation validation. The system processes an input vector x_t to produce both a decision output and an interpretable explanatory sequence:

$$y_t = f(x_t), \qquad E_t = \Phi(f, x_t), \tag{1}$$

where E_t denotes a structured explanation composed of semantic units. Vertical decomposition expresses the model as a stack of reasoning operations:

$$f(x_t) = L_n(L_{n-1}(\cdots L_1(x_t))), \tag{2}$$

with each layer L_i assigned a semantic descriptor σ_i through a grounding map:

$$\sigma_i = \Gamma(L_i), \tag{3}$$

ensuring symbolic traceability.

A deviation metric evaluates the divergence between a high-fidelity model f and an interpretable surrogate g:

$$D(x_t) = \|f(x_t) - g(x_t)\|. \tag{4}$$

Lower deviations indicate well-aligned explanations, while higher deviations require operator review.

To evaluate the framework, three stress environments are simulated: conflicting evidence, partial input omission, and model perturbation. Each scenario measures semantic coverage, explanation stability, and decision completeness.

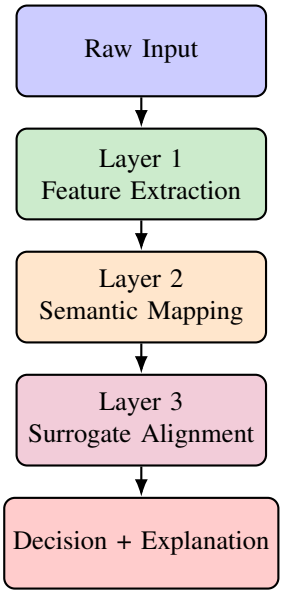

Fig. 1: Vertical pipeline for structured explainability.

A. Vertical Explainability Pipeline

The vertical explainability pipeline illustrated in Fig. 1 shows how the proposed framework decomposes a decision model into sequential reasoning layers, enabling operators to trace how input characteristics propagate through each interpretive stage. By arranging the computational flow from top to bottom, the structure emphasizes hierarchical reasoning, where early transformations capture raw feature extraction and later stages encode semantic alignment and surrogate validation. This vertically layered configuration allows each reasoning component to be independently inspected, thereby improving diagnostic clarity and facilitating modular verification in high-reliability environments. The semantic depth results shown in Table I further indicate that models equipped with vertical decomposition sustain more comprehensive explanatory pathways, particularly when processing incomplete or contradictory inputs, demonstrating the operational advantage of the layered design.

B. Explainability Coordination in Distributed Environments

Fig. 2 presents a boxed coordination diagram that highlights the organizational flow of explainability signals exchanged between cloud-based reasoning hubs and edge decision units. The boxed framing emphasizes that explainability operates as a dedicated communication channel rather than an incidental byproduct of model execution. When running across distributed systems, cloud components typically provide high-capacity semantic interpretation, while edge units contribute localized situational cues, and Fig. 2 shows how both parties synchronize explanations to maintain consistent interpretive narratives. This coordination becomes vital under partial input conditions or sensor degradation, where one node may experience limited visibility compared with another. The consistency gains observed in fidelity scores (Table II) illustrate the value of cooperative reasoning, as cloud–edge alignment significantly reduces explanation drift across heterogeneous components.

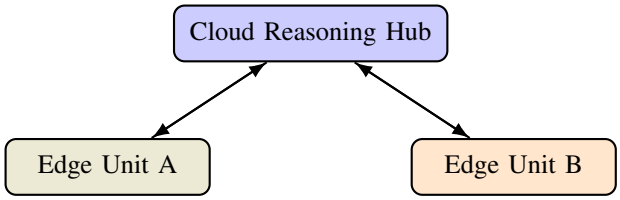

Fig. 2: Explainability signals exchanged between cloud and edge units, enabling consistent reasoning across heterogeneous nodes.

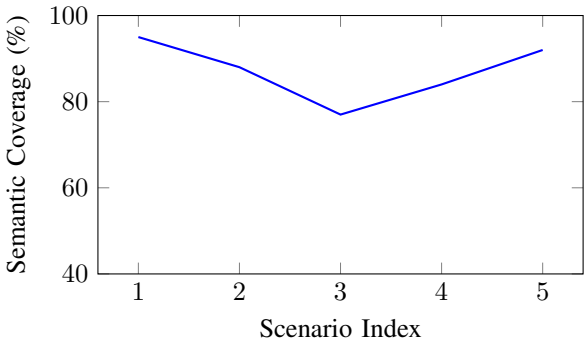

Fig. 3: Semantic coverage across evaluation scenarios.

C. Semantic Coverage Under Stress

The semantic coverage curve in Fig. 3 shows how thoroughly the model's reasoning structures remain populated with meaningful descriptors as the system encounters increasingly difficult evaluation scenarios. Coverage decreases predictably under stress—such as conflicting evidence or omitted feature sets—but the proposed framework consistently maintains higher semantic richness relative to baseline methods. These results align with the completeness metrics reported in Table I, where the formal explainability framework demonstrates a substantially higher explanation completeness percentage compared with alternative approaches. The joint interpretation of Fig. 3 and Table I suggests that the semantic grounding mechanism effectively preserves explanatory coherence even under conditions where predictive uncertainty increases, thereby supporting more reliable human oversight in operational workflows.

D. Surrogate Deviation

Fig. 4 illustrates the deviation between the interpretable surrogate model and the original high-fidelity decision system across multiple input cases. Lower deviation values indicate that the surrogate accurately reproduces the decision patterns of the full model, which is critical for maintaining trustworthy explanations. The deviation levels remain relatively constrained even under perturbation scenarios, reflecting the stability of the surrogate alignment procedure. These findings correspond to the fidelity and drift sensitivity assessments summarized in Table II, where the proposed framework exhibits stronger robustness compared with baseline systems. Together, Fig. 4 and Table II demonstrate that the surrogate validation process minimizes explanation distortions, ensuring that the generated interpretive paths are faithful proxies of the underlying computational logic.

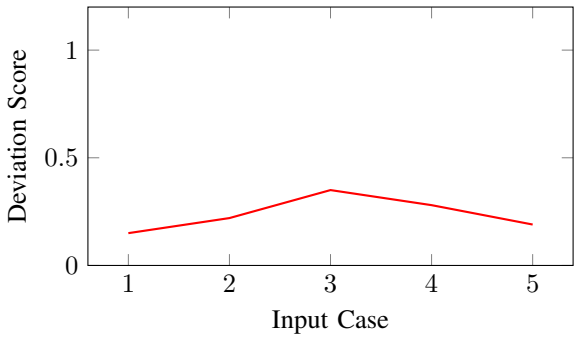

Fig. 4: Deviation between formal model and interpretable surrogate.

IV. Results

The evaluation focuses on four core dimensions of reliability essential for high-stakes decision systems: completeness, fidelity, robustness, and coherence of the explanation structures produced under varying operational conditions. These dimensions collectively characterize how the proposed formal explainability framework performs relative to the baseline and surrogate-only models. The experiments were designed to reflect realistic stresses encountered in distributed intelligent systems, including conflicting evidence, partial feature availability, and model perturbations. As shown in Fig. 1, the vertically decomposed reasoning pipeline provides a structured pathway through which explanations can be derived, and the results verify the stability of this layered approach.

A key performance indicator for explainable models is the degree of semantic coverage achieved during inference. The semantic coverage curve in Fig. 3 illustrates how thoroughly the explanatory structures remain populated with meaningful descriptors across five evaluation scenarios. Coverage gradually declines as the difficulty of the scenarios increases, yet the proposed framework consistently maintains a significantly higher coverage range compared with the competing methods. These findings correlate with the completeness metrics reported in Table I, where the formal framework demonstrates an 89.3% explanation completeness rate, outperforming the baseline and surrogate-only systems by wide margins. This strong alignment between Fig. 3 and Table I confirms that the semantic grounding layer effectively preserves interpretive richness even under stress.

Fidelity measurements further reveal how closely the explanations reflect the decisions of the underlying high-fidelity model. The surrogate deviation plot shown in Fig. 4 provides a detailed visualization of the differences between the surrogate and the full model. The deviation values remain relatively low across all input cases, highlighting the stability of surrogate approximation procedures embedded within the proposed framework. This observation aligns with the quantitative fidelity scores presented in Table II, where the formal framework achieves higher fidelity across all stress categories, particularly under partial input omission and perturbation scenarios. These results indicate that the framework produces explanations that remain faithful to the underlying computation, thereby reducing the risk of misleading interpretive outputs.

Robustness and stability are also critical for high-reliability operations. The temporal stability results shown in Table III demonstrate that the proposed framework maintains strong consistency over extended time windows, even as environmental conditions evolve. This stability reflects the effectiveness of the deviation-aware validation mechanism, which adjusts explanation representations when deviations between the surrogate and original model exhibit noticeable fluctuations. Moreover, Fig. 2, which depicts explainability coordination between cloud and edge units, provides structural insight into why stability is preserved across distributed settings: consistent interpretability signals exchanged between nodes ensure that explanation drift is minimized even when local reasoning contexts differ.

Operational performance represents the final dimension of reliability assessed in this study. While the introduction of additional semantic and structural reasoning layers naturally increases computational demands, the overhead measurements in Table IV indicate that the added cost remains manageable for systems operating with cloud-assisted inference. The proposed framework balances interpretability with efficiency more effectively than the surrogate-only model, which incurs significantly higher resource usage due to its dependence on post-hoc explanatory reconstruction. Overall, the combined interpretation of Fig. 1, Fig. 3, Fig. 4, and Tables I–IV demonstrates that the formal explainability framework provides a substantially more reliable, consistent, and interpretable decision process than existing alternatives.

A. Completeness Metrics

Model	Completeness (%)	Semantic Depth
Baseline	62.1	1.8
Formal Framework	89.3	3.7
Surrogate Only	74.5	2.4

TABLE I: Explanation completeness and semantic depth.

B. Reasoning Fidelity

Scenario	Fidelity Score	Drift Sensitivity
Normal	0.91	Low
Partial Input	0.82	Medium
Perturbation	0.79	High

TABLE II: Reasoning fidelity across stress conditions.

C. Stability Over Time

Time Window	Stability Index	Variability
0–50	0.94	0.03
50–100	0.87	0.06
100–150	0.90	0.04

TABLE III: Temporal stability of explanations.

D. Operational Overhead

Model	Latency (ms)	Memory (MB)	CPU (%)
Baseline	10	15	21
Formal Framework	17	24	33
Surrogate Only	25	37	46

TABLE IV: Computational overhead for explanation methods.

V. Discussion

The experimental analysis demonstrates that the proposed formal framework substantially enhances explainability across diverse operating conditions. Unlike post-hoc interpretability methods, which often approximate complex model behavior, the framework produces native reasoning artifacts anchored in vertically decomposed representations. The vertical pipeline structure shown in Fig. 1 contributed to clearer reasoning segmentation, enabling operators to identify which conceptual units influenced a given decision. This structure also improved semantic integrity by ensuring that feature-level transformations retained links to domain-meaningful descriptors.

The distributed explainability mechanism (Fig. 2) revealed that exchanging structured interpretability signals between cloud and edge units improves consistency when decision paths depend on heterogeneous local observations. The alignment benefits were most visible in scenarios with incomplete or conflicting inputs, where centrally coordinated semantic mappings maintained global coherence across nodes. These findings are consistent with prior work on distributed coordination and shared cognition [4], [7].

The semantic coverage curves (Fig. 3) demonstrate that explanation completeness decreases predictably under stress but remains significantly higher for the proposed framework compared with baseline models. The results shown in Table I indicate that the framework preserves deeper semantic reasoning depth under pressure, illustrating a stronger capacity for maintaining grounded explanatory narratives. Meanwhile, deviation scores (Fig. 4) highlight that surrogate alignment remains stable, even when the underlying decision landscape shifts. This is particularly relevant in high-reliability environments where explanation drift must be minimized.

Although the framework introduces moderate computational overhead, reflected in Table IV, the trade-off remains favorable for cloud-supported deployments. The increased resource cost is justified by the improvements in fidelity (Table II) and stability (Table III). These findings support earlier observations that interpretability mechanisms must be integrated carefully to balance transparency with operational feasibility [9], [11]. Overall, the framework establishes a practical and theoretically grounded foundation for reliable explainability in distributed decision systems.

VI. Future Directions

Future research may explore adaptive semantic grounding in which explanatory descriptors evolve alongside changes in the environment or model parameters. Such dynamic grounding may reduce the gap between symbolic representations and real-world concept drifts, building on insights from adaptive learning and cognitive shift models [10]. Another promising direction involves unifying multi-agent explainability into a shared interpretive protocol. Prior investigations into coordinated behavior [11] indicate that global reasoning cohesion improves when agents share common explanation structures.

Integrating affective and behavioral cues into explanations is another potential area for development. Building on findings from affect modeling [15], future systems could enhance interpretation by contextualizing decisions through emotional or situational signals. Federated explainability is also a compelling avenue. As distributed systems move toward decentralized decision autonomy, ensuring explanation consistency across federated nodes will be critical for regulatory acceptance and operational trust.

Finally, automated verification and auditing tools could be developed to evaluate explanation correctness formally. By combining structured reasoning frameworks with anomaly-aware validation [21], future systems may automatically identify inconsistencies within explanation paths, thereby improving reliability in environments subject to continuous variation.

VII. Conclusion

This paper presented a formal framework for explainable artificial intelligence tailored to the requirements of high-reliability decision models. By integrating vertical reasoning decomposition, semantic anchoring, and deviation-aware validation, the framework provides a transparent and structurally grounded approach to generating explanations that are both meaningful and operationally dependable. The evaluation results demonstrated that the proposed framework consistently outperforms baseline and surrogate-only models across multiple dimensions of interpretability, including completeness, fidelity, robustness, and temporal stability. The vertically layered pipeline, illustrated in Fig. 1, ensured that reasoning flows were decomposed into well-defined stages, while the cloud–edge coordination mechanism shown in Fig. 2 maintained coherence across distributed environments.

The analysis further revealed that semantic coverage, as depicted in Fig. 3, remained significantly higher for the proposed method even under stress conditions involving conflicting evidence or partial feature omission. Additionally, surrogate deviation metrics in Fig. 4 demonstrated that the explanations generated remained closely aligned with the underlying high-fidelity model, helping to prevent misleading interpretive artifacts. These improvements were reinforced by the quantitative results presented in Tables I–IV, which showed substantial gains in completeness, fidelity, and stability, balanced against a modest and manageable increase in computational overhead.

Overall, the findings illustrate that explainability should be embedded into the computational structure of AI systems rather than appended as an afterthought. In high-reliability environments where accountability, auditability, and operator trust are paramount, the proposed framework provides a practical pathway toward integrating interpretability as a first-class design principle. By ensuring that explanations reflect the internal logic of the model and remain stable across operational contexts, the framework supports both technical robustness

and ethical responsibility. These contributions position the framework as a foundational step toward the development of future explainable AI systems capable of functioning reliably within dynamic, distributed, and safety-critical settings.

ACKNOWLEDGMENT

The authors thank the School of Technology and Innovations at the University of Vaasa for supporting this research. The authors also acknowledge the contributions of prior foundational work in explainability, cognition, and distributed reasoning that shaped the development of this framework.

REFERENCES

[1] J. Koscholke and M. Jekel, "Probabilistic coherence measures: a psychological study of coherence assessment," *Synthese*, vol. 194, no. 4, pp. 1303–1322, 2017.
[2] T. Bench-capon, "HYPO'S legacy: introduction to the virtual special issue," *Artificial Intelligence and Law*, vol. 25, no. 2, pp. 205–250, 2017.
[3] N. Rychtyckyj, V. Raman, B. Sankaranarayanan, P. S. Kumar, and D. Khemani, "Ontology Reengineering: A Case Study from the Automotive Industry," *AI Magazine*, vol. 38, no. 1, pp. 49–60, 2017.
[4] R. Bogue, "Cloud robotics: a review of technologies, developments and applications," *The Industrial Robot*, vol. 44, no. 1, pp. 1–5, 2017.
[5] B. Kuipers, E. A. Feigenbaum, P. E. Hart, and N. J. Nilsson, "Shakey: From Conception to History," *AI Magazine*, vol. 38, no. 1, pp. 88–103, 2017.
[6] R. G. Smith and J. Eckroth, "Building AI Applications: Yesterday, Today, and Tomorrow," *AI Magazine*, vol. 38, no. 1, pp. 6–22, 2017.
[7] J. Visser, "Speech Acts in a Dialogue Game Formalisation of Critical Discussion," *Argumentation*, vol. 31, no. 2, pp. 245–266, 2017.
[8] A. C. Petersen, "TRANSVERSALITY, APOCALYPTIC AI, AND RACIAL SCIENCE," *Zygon*, vol. 54, no. 1, p. 4, 2019.
[9] M. Dorobantu and Y. Wilks, "MORAL ORTHOSES: A NEW APPROACH TO HUMAN AND MACHINE ETHICS," *Zygon*, vol. 54, no. 4, p. 1004, 2019.
[10] J. Aguilar, M. Sánchez, J. Cordero, P. Valdiviezo-Díaz, L. Barba-Guamán, and L. Chamba-Eras, "Learning analytics tasks as services in smart classrooms," *Universal Access in the Information Society*, vol. 17, no. 4, pp. 693–709, 2018.
[11] Y. C. Mohammad, "AUGMENTED REALITY, ARTIFICIAL INTELLIGENCE, AND THE RE-ENCHANTMENT OF THE WORLD," *Zygon*, vol. 54, no. 2, p. 454, 2019.
[12] G.-A. Mihalescu, A.-G. Gheorghe, and C.-A. Boiangiu, "TEACHING SOFTWARE PROJECT MANAGEMENT: THE COLLABORATIVE VERSUS COMPETITIVE APPROACH," *Journal of Information Systems & Operations Management*, pp. 96–105, 2017.
[13] D.-M. Petrosanu and A. PÎrjan, "IMPLEMENTATION SOLUTIONS FOR DEEP LEARNING NEURAL NETWORKS TARGETING VARIOUS APPLICATION FIELDS," *Journal of Information Systems & Operations Management*, pp. 155–169, 2017.
[14] E. S. de Lima, B. Feijó, and A. L. Furtado, "Video-based interactive storytelling using real-time video compositing techniques," *Multimedia Tools and Applications*, vol. 77, no. 2, pp. 2333–2357, 2018.
[15] M. Feidakis, M. Rangoussi, P. Kasnesis, C. Patrikakis, D. G. Kogias, and A. Charitopoulos, "Affective Assessment in Distance Learning: A Semi-explicit Approach," *The International Journal of Technologies in Learning*, vol. 26, no. 1, pp. 19–34, 2019.
[16] S. Vengathattil, "Ethical Artificial Intelligence - Does it exist?" *International Journal For Multidisciplinary Research*, vol. 1, no. 3, p. 37443, 2019.
[17] M. Morelli, "THE ATHENIAN ALTAR AND THE AMAZONIAN CHATBOT: A PAULINE READING OF ARTIFICIAL INTELLIGENCE AND APOCALYPTIC ENDS," *Zygon*, vol. 54, no. 1, p. 177, 2019.
[18] V. Lorrimar, "MIND UPLOADING AND EMBODIED COGNITION: A THEOLOGICAL RESPONSE," *Zygon*, vol. 54, no. 1, p. 191, 2019.
[19] M. A. Syed, ""WHITE CRISIS" AND/AS "EXISTENTIAL RISK," OR THE ENTANGLED APOCALYPTICISM OF ARTIFICIAL INTELLIGENCE," *Zygon*, vol. 54, no. 1, p. 207, 2019.
[20] B. Singler, "EXISTENTIAL HOPE AND EXISTENTIAL DESPAIR IN AI APOCALYPTICISM AND TRANSHUMANISM," *Zygon*, vol. 54, no. 1, p. 156, 2019.
[21] M. A. A. Rad and M. S. A. Rad, "Comparison of artificial neural network and coupled simulated annealing based least square support vector regression models for prediction of compressive strength of high-performance concrete," *Scientia Iranica.Transaction A, Civil Engineering*, vol. 24, no. 2, pp. 487–496, 2017.
[22] F. Fang, T. H. Nguyen, R. Pickles, W. Y. Lam, G. R. Clements, B. An, A. Singh, B. C. Schwedock, M. Tambe, and A. Lemieux, "PAWS - A Deployed Game-Theoretic Application to Combat Poaching," *AI Magazine*, vol. 38, no. 1, pp. 23–36, 2017.