



Algebraic Structures for Explainable Artificial Intelligence and Decision-Making Systems

DR. SATHYAPRIYA S

DEPARTMENT OF MATHEMATICS AND COMPUTATIONAL SCIENCE
SRI KRISHNA KNOWLEDGE FOUNDATION
DEEMED TO BE UNIVERSITY

How to Cite this Article:

S, S. (2026). Algebraic Structures for Explainable Artificial Intelligence and Decision-Making Systems. International Journal of Creative and Open Research in Engineering and Management, 2(10), 1-9. <https://doi.org/10.55041/ijcope.v2i10.138>

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.



<https://doi.org/10.55041/ijcope.v2i10.138>

Abstract

Explainable Artificial Intelligence (XAI) has become an important research direction because intelligent systems are increasingly used to support decisions in complex and uncertain environments. However, the interpretability of a decision depends not only on the final prediction but also on the mathematical structure through which the prediction is obtained. This paper develops an algebraic framework for interpretable multi-criteria decision-making based on bounded ordered structures, weighted aggregation, and explicit criterion-level contributions. A finite set of alternatives is evaluated using normalized criteria, and a weighted score is assigned to each alternative. The framework represents criterion values in a bounded lattice and uses an order-preserving aggregation operator to obtain transparent rankings. Mathematical properties concerning monotonicity, boundedness, score decomposition, and stability under perturbations are established. An illustrative numerical example involving the selection of an artificial intelligence model demonstrates how the framework generates rankings and human-readable explanations. The approach provides a mathematically explicit foundation for decision support and can be extended to fuzzy, linguistic, and neutrosophic environments. The proposed framework and numerical example are illustrative; novelty and performance claims require literature verification and empirical validation.

Keywords: Abstract algebra; lattice theory; explainable artificial intelligence; multi-criteria decision-making; interpretable decision systems;

weighted aggregation.

1. Introduction

Artificial Intelligence (AI) has become an essential component of decision-support systems in engineering, healthcare, finance, education, and information technology. Machine learning models can identify complex patterns and produce predictions from large datasets. Nevertheless, the internal mechanisms of many models are difficult for human users to interpret. Explainable Artificial Intelligence attempts to make model behaviour and predictions understandable through explanations based on features, rules, examples, or mathematical representations.

Several approaches have been developed to improve model interpretability. Local Interpretable Model-Agnostic Explanations (LIME) provide local approximations of complex predictive models. Shapley Additive Explanations (SHAP) use concepts from cooperative game theory to attribute predictions to input features. These methods provide important foundations for explaining AI predictions, although their explanations depend on the model, data representation, and assumptions used in their construction.

A complementary approach is to design decision systems whose mathematical operations are transparent from the beginning. Algebraic structures offer a systematic way to represent ordered information, combine evaluations, and study the properties of decision rules. In particular, lattice theory provides an order-based representation of information, while weighted aggregation enables the combination of several criteria into an overall score.



The objective of this paper is to formulate a transparent multi-criteria decision framework using bounded lattices and weighted algebraic aggregation. The framework is designed to explain how each criterion contributes to the final decision. The objectives are to formulate an ordered algebraic representation for decision data; define an aggregation function; establish properties of the decision function; construct criterion-level explanations; and illustrate the method using an AI model selection problem.

The proposed framework is a transparent decision-support method rather than a replacement for all existing XAI techniques.

2. Related Work

2.1 Explainable artificial intelligence

Explainability is concerned with communicating why a model produces a particular output. LIME constructs interpretable local surrogate models around individual predictions. SHAP uses Shapley values to quantify feature contributions under a specified attribution framework. These approaches are valuable for analysing complex predictive models. A decision-support system may also benefit from a transparent aggregation rule that allows users to reconstruct the final decision directly from its inputs.

2.2 Algebraic structures and interpretability

Algebraic structures provide formal rules for combining elements and studying the behaviour of mathematical systems. Semigroups, lattices, ordered sets, and semirings have applications in computing, optimization, and information processing. A lattice is useful when information is represented through an ordering relation. Its meet and join operations describe the greatest lower bound and least upper bound of two elements, respectively.

2.3 Research motivation

This work considers a finite decision problem in which each alternative is evaluated against several criteria. Its methodological emphasis is the combination of bounded ordered criterion values, transparent weighted aggregation, exact decomposition of the aggregate score, and mathematical analysis of monotonicity and stability. The approach is a specific algebraic decision model; its originality relative to existing methods must be established through a comprehensive literature review and comparative experiments.

3. Mathematical Preliminaries

3.1 Partially ordered sets and lattices

A partially ordered set is a pair $(L, \leq)$ in which $\leq$ is a reflexive, antisymmetric, and transitive relation on L .

A lattice is a partially ordered set $(L, \leq)$ in which every pair x, y has a greatest lower bound $x \wedge y$ and a least upper bound $x \vee y$.

A bounded lattice is a lattice containing a least element 0_L and a greatest element 1_L , such that $0_L \leq x \leq 1_L$ for every x in L .

In this paper, the principal numerical domain is $L = [0, 1]$. It is a bounded lattice under the usual order, with $x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$. The lattice operations describe ordering and comparison. The aggregate score is defined separately by weighted addition; a weighted arithmetic mean is not the lattice join or meet.

3.2 Normalized criterion vectors

Let m be the number of criteria. A normalized criterion vector is $x = (x_1, x_2, \dots, x_m) \in [0, 1]^m$, where x_j denotes the normalized performance of an alternative under criterion j .

A criterion-weight vector $w = (w_1, w_2, \dots, w_m)$ satisfies $w_j \geq 0$ and $\sum_j w_j = 1$. A benefit criterion can be normalized by $x_j = (v_j - a_j)/(b_j - a_j)$, while a cost criterion can be normalized by $x_j = (b_j - c_j)/(b_j - a_j)$, assuming appropriate bounds $a_j < b_j$.



4. Proposed Algebraic Decision Framework

4.1 Decision information space

Let $A = \{A_1, A_2, \dots, A_n\}$ be a finite set of alternatives and $C = \{C_1, C_2, \dots, C_m\}$ a set of evaluation criteria. The decision information matrix is $X = [x_{ij}]_{n \times m}$, where $x_{ij} \in [0, 1]$ denotes the normalized performance of alternative A_i under criterion C_j . The complete decision representation is $D = (A, C, X, w)$.

4.2 Algebraic decision score

For alternative A_i , define its algebraic decision score by:

$$S(A_i) = \sum_{j=1}^m w_j x_{ij}.$$

The function assigns a score to each alternative. A higher score indicates better performance under the specified normalized criteria and weights.

4.3 Criterion contribution representation

For alternative A_i , define the contribution of criterion C_j by $e_{ij} = w_j x_{ij}$. The contribution vector $e_i = (e_{i1}, e_{i2}, \dots, e_{im})$ lists the individual contributions to the final score. The total score is recovered by $S(A_i) = \sum_j e_{ij}$.

This explanation is exact for the specified additive decision model. It should not be confused with a causal explanation of a real-world outcome.

4.4 Decision rule

$$A^* = \arg \max \{S(A_i) : A_i \in A\}.$$

If multiple alternatives have the same maximum score, the decision rule returns all maximizers or applies a separately specified tie-breaking rule.

5. Mathematical Properties

5.1 Boundedness

Theorem 1. For every alternative A_i , $0 \leq S(A_i) \leq 1$.

Proof. Since $x_{ij} \in [0, 1]$ and $w_j \geq 0$, we have $0 \leq w_j x_{ij} \leq w_j$. Summing over all criteria gives $0 \leq \sum_j w_j x_{ij} \leq \sum_j w_j = 1$. Therefore, $S(A_i) \in [0, 1]$.

5.2 Monotonicity

Theorem 2. Suppose two normalized criterion vectors x and y satisfy $x_j \leq y_j$ for every j . Then $S(x) \leq S(y)$.

Proof. For every j , non-negativity of w_j implies $w_j x_j \leq w_j y_j$. Summing these inequalities gives $\sum_j w_j x_j \leq \sum_j w_j y_j$. Hence the decision function is order-preserving.

If $w_j = 0$, changing criterion j alone does not change the aggregate score. If $w_j > 0$, increasing that criterion while keeping all others fixed cannot decrease the score.

5.3 Exact explanation decomposition

Theorem 3. For every alternative A_i , $S(A_i) = \sum_j e_{ij}$. Proof. By definition, $e_{ij} = w_j x_{ij}$, so $\sum_j e_{ij} = \sum_j w_j x_{ij} = S(A_i)$. Thus, the criterion contributions reconstruct the final score exactly.

5.4 Lipschitz stability

Theorem 4. Let $x, y \in [0, 1]^m$ and suppose $\max_j |x_j - y_j| \leq \epsilon$. Then $|S(x) - S(y)| \leq \epsilon$.

Proof. By the triangle inequality, $|S(x) - S(y)| = |\sum_j w_j (x_j - y_j)| \leq \sum_j w_j |x_j - y_j| \leq \sum_j w_j \epsilon = \epsilon$, since the weights sum to one.



5.5 Ranking stability

Corollary. Suppose $S(A_p) - S(A_q) > 2\epsilon$. If every normalized criterion value of each alternative changes by at most ϵ , their ordering cannot reverse. Proof. Each perturbed score differs from its original score by at most ϵ , so $S'(A_p) - S'(A_q) \geq S(A_p) - S(A_q) - 2\epsilon > 0$.

The stability result concerns perturbations in normalized criterion values with fixed weights. It does not establish robustness against changes in weights, normalization bounds, or criteria.

6. Explainable Decision-Making Algorithm

The procedure consists of identifying alternatives and criteria, normalizing criterion measurements to $[0, 1]$, assigning non-negative weights summing to one, calculating aggregate scores and criterion contributions, and ranking alternatives while reporting explanations.

Input: decision matrix X and weights w .

Verify that every x_{ij} lies in $[0, 1]$, every w_j is non-negative, and the weights sum to one.

For each alternative A_i , calculate $e_{ij} = w_j x_{ij}$ and $S(A_i) = \sum_j e_{ij}$.

Sort alternatives by decreasing aggregate score.

Output scores, ranking, and explanation vectors.

For a dense decision matrix, score computation requires $O(nm)$ arithmetic operations. Sorting the alternatives requires $O(n \log n)$ comparisons. The overall time complexity is $O(nm + n \log n)$.

7. Illustrative Numerical Example

7.1 Problem description

Consider a hypothetical organization selecting an AI model for a decision-support application. Three alternatives A_1, A_2, A_3 are evaluated using predictive performance C_1 , interpretability C_2 , computational efficiency C_3 , and reliability C_4 . Assume all measurements are normalized to $[0, 1]$ and the weights are $w = (0.40, 0.25, 0.20, 0.15)$. These weights sum to one and represent hypothetical preferences.

7.2 Decision matrix

Alternative	C_1	C_2	C_3	C_4
A_1	0.90	0.70	0.80	0.85
A_2	0.82	0.90	0.75	0.88
A_3	0.86	0.78	0.92	0.80

The values are assumed for demonstration and are not measurements obtained from a real experiment.

7.3 Score calculation

For A_1 : $S(A_1) = 0.40(0.90) + 0.25(0.70) + 0.20(0.80) + 0.15(0.85) = 0.360 + 0.175 + 0.160 + 0.1275 = 0.8225$.

For A_2 : $S(A_2) = 0.40(0.82) + 0.25(0.90) + 0.20(0.75) + 0.15(0.88) = 0.328 + 0.225 + 0.150 + 0.132 = 0.8350$.

For A_3 : $S(A_3) = 0.40(0.86) + 0.25(0.78) + 0.20(0.92) + 0.15(0.80) = 0.344 + 0.195 + 0.184 + 0.120 = 0.8430$.



7.4 Ranking of alternatives

Alternative	Aggregate score	Rank
A ₁	0.8225	3
A ₂	0.8350	2
A ₃	0.8430	1

According to the assumed data and weights, alternative A₃ receives the highest score.

7.5 Criterion-level explanations

Alternative	C ₁ contribution	C ₂ contribution	C ₃ contribution	C ₄ contribution
A ₁	0.3600	0.1750	0.1600	0.1275
A ₂	0.3280	0.2250	0.1500	0.1320
A ₃	0.3440	0.1950	0.1840	0.1200

The explanation vector for A₃ is (0.344, 0.195, 0.184, 0.120). Its components sum to $0.344 + 0.195 + 0.184 + 0.120 = 0.843$. Thus, the model provides an exact arithmetic reconstruction of the aggregate score.

Although A₃ has the highest computational-efficiency score, it has a lower reliability contribution than A₂. The final ranking reflects the combined effect of all four criteria and their assigned weights.

7.6 Pairwise explanation

The score difference between A₃ and A₂ is $0.8430 - 0.8350 = 0.0080$. The criterion-wise contribution differences are C₁: $0.344 - 0.328 = 0.016$; C₂: $0.195 - 0.225 = -0.030$; C₃: $0.184 - 0.150 = 0.034$; and C₄: $0.120 - 0.132 = -0.012$. Their sum is $0.016 - 0.030 + 0.034 - 0.012 = 0.008$.

This comparison explains why A₃ ranks above A₂: its advantages in predictive performance and computational efficiency outweigh its disadvantages in interpretability and reliability under the assumed weights. The score difference is small, so the ranking should not be interpreted as evidence that A₃ is universally or statistically superior to A₂.

8. Discussion

The proposed framework demonstrates how algebraic representations can support transparent decision-making. Its principal strength is that the score and criterion-level explanations are derived from the same explicit mathematical expression. Boundedness guarantees that aggregate scores remain in $[0, 1]$. Monotonicity ensures that improving a criterion cannot reduce the score when all other values and weights remain fixed. Exact decomposition guarantees score reconstruction, and the stability theorem bounds the effect of small perturbations in criterion values.

Potential applications include selecting AI models, ranking software systems, evaluating suppliers, prioritizing engineering projects, and assessing technology alternatives. Nevertheless, the weighted sum is compensatory: a high value in one criterion can offset a low value in another. This may be inappropriate when certain criteria represent mandatory safety or reliability requirements. Such constraints should be handled separately or incorporated through a justified non-compensatory decision rule.

The framework assumes that criteria can be normalized and that weights meaningfully represent decision preferences. Poor normalization or subjective weights can affect the ranking. The model does not automatically explain a black-box machine-learning prediction; it explains a decision produced by the specified aggregation function. An algebraic explanation is not necessarily a causal explanation.



9. Future Research Directions

Fuzzy algebraic decision models: represent imprecise criterion evaluations with fuzzy membership information.

Neutrosophic decision systems: represent truth, indeterminacy, and falsity when the application requires explicit treatment of uncertainty and incomplete knowledge.

Lattice-valued aggregation: investigate decision rules based directly on lattice operations and establish conditions under which resulting rankings are meaningful.

Integration with machine learning: use algebraic decision functions to explain or audit predictive-model outputs, distinguishing surrogate explanations from faithful explanations of the original model.

Empirical validation: evaluate ranking stability, explanation fidelity, computation time, and agreement with domain experts using real datasets.

Comparative analysis: compare the framework with established MCDM approaches, including TOPSIS, ELECTRE, and suitable interpretable machine-learning methods.

10. Conclusion

This paper formulated an algebraically structured framework for explainable multi-criteria decision-making using a bounded lattice of normalized criterion values and a weighted aggregation function. The framework produces rankings and criterion-level explanations within a common mathematical representation. Boundedness, monotonicity, exact score decomposition, and Lipschitz stability were established mathematically. An illustrative AI model selection problem demonstrated the calculation of aggregate scores and the reconstruction of pairwise ranking differences from individual criterion contributions.

The results establish mathematical properties of the specified decision rule, rather than empirical superiority over existing AI or MCDM techniques. Future work should investigate non-compensatory algebraic operators, uncertain criterion information, and validation on real-world datasets.

References

- [1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You? Explaining the Predictions of Any Classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- [2] Lundberg, S. M., & Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions." Advances in Neural Information Processing Systems, 30.
- [3] Nakanishi, T. (2026). "AIME²: Toward a Unified Algebraic Theory of Explainability via Approximate Inverse Operators." Communications AI & Computing, 1, Article 3. <https://doi.org/10.1038/s44488-026-00004-0>. Verify final bibliographic details against the publisher record.
- [4] Shi, C., & Yao, Y. (2025). "Explainable multi-criteria decision-making: A three-way decision perspective." International Journal of Approximate Reasoning, 187, Article 109528. <https://doi.org/10.1016/j.ijar.2025.109528>
- [5] A related 2023 article on algebraic views of tree ensemble classifiers was mentioned in the draft source list, but its full bibliographic record has not been verified. Identify and verify the exact publication before citing it in a submission.