eISSN 2783-6851

NEW TRENDS in COMPUTER SCIENCES

2025
Volume 3
Issue 2
Pages 100–125
https://doi.org/10.3846/ntcs.2025.25442

IMPROVING AHP CONSISTENCY THROUGH COGNITIVE COLLABORATION WITH LARGE LANGUAGE MODELS

Augustinas MACEIKA [1], Andrej BUGAJEV [2✉], Olga Regina ŠOSTAK [3]

[1]*Department of Mechanical and Materials Engineering, Faculty of Mechanics, Vilnius Gediminas Technical University, Vilnius, Lithuania*
[2]*Department of Mathematical Modelling, Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, Vilnius, Lithuania*
[3]*Department of Engineering Graphics, Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, Vilnius, Lithuania*

Article History:
- received 28 November 2025
- accepted 06 Jaunuary 2026

Abstract. The Analytic Hierarchy Process (AHP) is a human-centered method designed to structure complex problems and extract the authentic, consistent opinions of decision-makers. However, its practical application is often limited by inconsistency in human judgments, often caused by the respondent's insufficient understanding of the task rather than simple mathematical error. The main goal of the article is to explore the possibilities of integration of innovative Artificial Intelligence (AI) tools for improving the AHP method. In order to improve respondent understanding and facilitate more intuitive and transparent consistency adjustments, this study also analyzes how to reduce the occurrence of inconsistency in pairwise comparison matrices and improve the Consistency Ratio (CR) by using advanced capabilities of large language models. The initial stage of study included a literature review, identifying typical problems in this area, reviewing the tools and methods used for obtaining better results, and presenting areas for improvement. At the second stage, the possibilities for improving consistency by increasing the influence of humans as decision makers, moving from the use of powerful mathematical optimization mechanisms to the application of human-centered explanatory AI techniques were analyzed. Based on the study results, the description of approaches for improvement of consistency in AHP was presented.

Keywords: analytic hierarchy process (AHP), large language models (LLM), multi-criteria decision making (MCDM), human-centered AI, explainable AI (XAI), consistency management, decision support systems, consistency ratio.

✉Corresponding author. E-mail: ***andrej.bugajev@vilniustech.lt***

1. Introduction

One of the most recognized and widely used multi-criteria evaluation methods is the Analytic Hierarchy Process (AHP) introduced by Saaty (2000). The rise of AI makes AHP more critical and relevant than ever – while AI models excel at processing big datasets and generating content, they lack the inherent subjective values, knowledge that define human decision-making. AI can calculate the consequences of a decision, but it cannot determine what matters to a specific stakeholder – for example, whether "sustainability" outweighs "cost" in a specific project context. When applying this method to project evaluation, complex solutions are structured into a hierarchical system and a pairwise comparison is made, which is based on the subjective opinion of the evaluator. However, when evaluating projects, the problem of improving consistency is encountered. This problem can be caused by the evaluator's lack of skills, incomplete understanding of the task, or confusion among the priorities of many criteria. The problem of ensuring consistency when evaluating projects is important to support the reliability and validity of the results. The appearance of inconsistencies in the matrix of pairwise comparisons of the evaluated criteria weakens the consistency of the decision model and critically worsens the consistency ratio. This is a constant challenge in the application of the AHP method. Purely algorithmic solutions can adjust the matrices without taking into account the evaluator's interpretations. In this case, mathematically consistent results can be obtained, but this will lead to an inconsistency between the improved option and the true intentions of the real evaluator, which will affect the quality of the future decision.

Inconsistency is conventionally measured using the Consistency Ratio (CR), with typical threshold of 0.10. However, recent literature (Sato & Tan, 2023) questions its reliability for varying matrix sizes and may not reflect true human. Therefore, the conventional CR value limits the detection of actual inconsistencies and may inaccurately reflect the evaluator's opinion. Bose (2023) pointed out that there are inherent limitations that depend on the size of the matrix, so the standard CR size gives "false positives for small matrices and false negatives for larger ones". Çoban (2023) also noted that there are challenges in applying CR to various matrix characteristics, as the ideal compatibility between the new ratios and Saaty's CR for larger matrices was an "artifact" due to the random data set, which practically did not contain consistent matrices at such sizes. Based on this criticism, it can be argued that CR is not always reliable in determining true consistency and measuring improvements. There are different types of methods used for inconsistency improvement, they can be split into following categories: optimizations algorithms, mathematical/statistical methods, semi-automatic or interactive decision support systems. Many researchers propose fully automatic optimization algorithms, such as Multi-objective optimization (IMOPSO by Sun et al., 2025), Evolutionary Algorithms (Heymann et al., 2024; Wang et al., 2023), and Particle Swarm Optimization variants (Liu et al., 2023b), designed to correct matrices by minimizing CR and semantic deviation. Other researchers focus on mathematical/statistical methods, such as Bayesian analysis frameworks (Li et al., 2025) and benchmark-based frameworks (Bose, 2023), usually aim to prevent inconsistency by design. Often all these methods have significant limitations despite the fact of successful reduction of CR below the 0.1 threshold and original information preservation.

One of the main problems of existing solutions is the "black box" nature of most automatic correction methods – it lacks the reasoning transparency for users, leading to the loss of trust and understanding of the decision process.

The focus on mathematical precision and automation can lead to achieving very low CRs (e.g., Wang et al., 2023, achieving CR < 10^{-4}), however it often comes at the cost of convenience of the user and the lack of explainability. Thus, this results in a significant barrier to practical adoption – mathematically good results are achieved without improving the understanding and confidence of decision-maker.

Moreover, there is a research gap in the literature – the assessment of user-oriented metrics such as time on task, user satisfaction, cognitive load are very infrequent. In other words, in existing research we see a primary focus on mathematical consistency over practical usability and human experience. Many studies lack publicly available code, raw data, detailed re-run instructions. On top of that, a lot of studies use synthetically created matrices instead of real human judgments, limiting the generalizability for real-world scenarios. These limitations collectively highlight a need for a paradigm shift from formal metrics-oriented towards human-centered consistency improvement processes. The recently opened possibilities of large language models (LLM) provide a unique opportunity to address the shortcomings of current AHP consistency improvement methods.

Traditional algorithms cannot understand the conversation and adapt to the user's data as it is entered. Meanwhile LLM is well suited for understanding and generating natural language and allows for intuitive explanations.

LLMs directly address the problem of "respondent's insufficient understanding of the task" highlighted in the abstract, facilitates clear communication and provides personalized instructions during the assessment and consistency correction stages.

LLM can explain in a comprehensible way the nature of the inconsistency and the basis for the proposed corrections, thereby increasing the user's confidence and motivation to learn, i.e. LLM can go beyond the limits of "black box" corrections. This is a fundamental shift from purely mathematical or rule-based methods to a cognitive, conversational and human-centered methodology.

This shift moves the focus from simply adjusting the numerical matrix to educating and assisting the decision maker, thereby addressing the root cause of inconsistency (human misunderstanding) and increasing user confidence and satisfaction. This is not just a mechanical improvement, but also a rethinking of the way the AHP method interacts with the computer, offering the potential for more flexible and transparent solutions.

In short, this study aims to explore the possibilities of integration of innovative AI tools for improving the AHP method.

Based on the above, in order to improve respondent understanding and facilitate more intuitive and transparent consistency adjustments, this study aims to reduce the occurrence of inconsistency in pairwise comparison matrices and improve the Consistency Ratio (CR) by using advanced capabilities of large language models.

2. Background and problem statement

The AHP, as a foundational methodology in multi-criteria decision-making (MCDM), provides a structured approach to complex problems by decomposing them into hierarchies and using pairwise comparisons to derive priorities. The concept of consistency ensures that the judgments made by a decision-maker are logical and transitive. For a long time, the main tool for this diagnostic task has been the Consistency Ratio (CR), with its widely accepted threshold of CR < 0.1. This single metric has become the de facto standard in the field. However, some recent studies have begun to challenge this standard, questioning not only its reliability but the principles of how inconsistency is defined and measured.

A critical component of the AHP method is the concept of consistency, which ensures a decision-maker makes logical and transitive judgments. Conventionally, the Consistency Ratio (CR) has served as the de facto standard for consistency diagnostics, with a threshold of CR < 0.10 used to distinguish acceptable matrices from those requiring correction.

Recent literature has increasingly challenged the reliability of the standard CR metric. Studies suggest that the conventional CR value limits the detection of actual inconsistencies and may inaccurately reflect the evaluator's true opinion. For instance, Bose (2023) argues that the standard CR creates "false positives for small matrices and false negatives for larger ones" due to inherent limitations in how benchmarks are derived. Similarly, Çoban (2023) notes that statistical artifacts in larger matrices can create an illusion of compatibility that does not exist in practice. Furthermore, Sato and Tan (2023) demonstrated that CR is statistically insensitive to decision quality metrics and user satisfaction, suggesting that optimizing for CR does not necessarily improve the human decision outcome.

Besides these statistical issues, the definition of inconsistency itself is a subject of scientific debates. As noted in a recent theoretical review by Pant et al. (2025), there is no single universally accepted axiomatic system for inconsistency. They analyzed 18 different indices and revealed fundamental disagreements on what makes judgment consistent. Which means that the field is developing sophisticated algorithms for a problem without well-established definition.

This lack of conceptual clarity has created a niche dominated by purely mathematical and algorithmic solutions. In the absence of a clear definition of "human consistency," researchers have focused on a concrete, quantifiable target: minimizing the CR value. This has led to the proliferation of "black box" automated correction methods. While these algorithms can achieve high mathematical precision, sometimes reducing CR to infinitesimal values (e.g., < 10–4) – they often do so by treating human judgments as a noisy signal to be "cleaned" rather than understood.

Consequently, mathematically consistent results are obtained, but they may lead to an inconsistency between the improved option and the true intentions of the real evaluator. This creates a barrier to practical adoption, as the reasoning transparency is lost, leading to a decrease in the decision-maker's trust and understanding.

3. Methodology

To address the research objective of shifting AHP consistency improvement from a purely mathematical to a human-centered paradigm, this study employs a critical review approach supported by a systematic literature search. The methodological framework consists of three stages: data collection, relevance screening, and critical analysis.

The main source of data for this study is the Web of Science (WoS) Core Collection, selected for its coverage of high-impact journals in operations research and decision sciences. To capture the most recent methodological developments, the search was restricted to the period of 2023–2025.

The search strategy focused on the intersection of the AHP method and consistency management. The specific advanced search query was: PY = (2023–2025) AND TS = ("analytic hierarchy process" OR "AHP") AND TS = ("Consistency Balancing" OR "Consistency Index" OR "Consistency Ratio").

The initial search gave a total of 178 results. A multi-step screening process was applied to identify the most relevant studies for critical analysis:

1. Relevance screening: titles and abstracts were reviewed to distinguish between methodological studies (those proposing new ways to measure or improve consistency) and application studies (those simply using AHP for a case study). Papers focusing

solely on routine application without methodological novelty regarding consistency were excluded.

2. Citation chaining: to ensure coverage of significant discussions, a "snowballing" technique was used. References within key identified papers were scanned to locate influential studies that might have been missed by the initial keyword string ("cited by" and "citing" tracking).
3. Final selection: this process resulted in a core set of 24 articles that represent the current state-of-the-art in inconsistency management. These 24 studies form the basis of the comparative analysis presented in the Results section and Appendix Table A1.

The selected articles were analyzed not just for their mathematical results, but for their underlying philosophical approach to the human decision-maker. The studies were categorized based on:

- Correction mechanism: automated optimization vs. interactive support.
- Evaluation metrics: mathematical reduction of Consistency Ratio (CR) vs. user-centric metrics (e.g., satisfaction, trust).
- Transparency: "Black box" algorithmic operations vs. explainable logic.

This critical analysis identified the human-centric gap, which served as the foundation for the proposed LLM-based model described in Section 5.

4. Results

4.1. Findings of the critical analysis

The review of the selected articles (summarized in Appendix Table A1) reveals the current state of AHP research. While the theoretical "crisis of measurement" is acknowledged by a minority of authors, the overwhelming majority of recent publications remain focused on the algorithmic paradigm. The analysis identified three main trends of the current state-of-the-art.

1. Statistical unreliability of the standard Consistency Ratio. Our analysis identified that the standard CR metric is not merely theoretically weak but statistically flawed in recent empirical tests. The review extracted the following specific evidence of measurement failure:

- Insensitivity to Decision Quality: Sato and Tan (2023) demonstrated through numerical simulations that the standard CR is statistically insensitive to decision quality metrics. Specifically, they found no significant correlation between CR and the Conformity of Rankings (CAR) ($p = 0.638$) or user satisfaction, represented by the Goodness-of-fit of Weight to human perception (GWP) ($p = 0.494$). This indicates that optimizing for CR does not necessarily improve the actual utility of the decision.
- Size-dependent classification errors: Bose (2023) established that the standard CR benchmarks produce systematic classification errors depending on matrix size, yielding "false positives for small matrices and false negatives for larger ones".
- Methodological Artifacts: Çoban (2023) revealed that the perceived reliability of alternative indices for larger matrices is often a statistical artifact. Their analysis showed that in randomly generated datasets, there were almost no consistent matrices at sizes $n \geq 6$, meaning high compatibility rates were observed simply because there were no consistent cases to test against.

2. The dominance of "Black Box" algorithmic optimization. The review confirms a trend towards increasing computational complexity at the cost of transparency. As detailed in Appendix Table A1, many of analyzed methods utilize metaheuristic algorithms to force matrix convergence:

- Extreme mathematical precision: Wang et al. (2023) utilized Evolutionary Algorithms to achieve near-perfect consistency (CR < 10^{-4}).
- Opaque logic: Methods such as the Gaussian Quantum Behavior Particle Swarm Optimization (GQPSO) by Liu et al. (2023b) and Genetic Algorithms by Heymann et al. (2024) operate as "black boxes". They successfully output mathematically consistent matrices but fail to provide the reasoning for specific adjustments, which creates a gap in user trust as well as between the decision and user intent.

4.2. The deeper theoretical divide: what is inconsistency?

Besides statistical and methodological critiques, the definition of inconsistency itself is a subject of scientific debates. A theoretical review by Pant et al. (2025) moves this discussion from a technical to a philosophical level. By analyzing over 18 different inconsistency indices against four major competing axiomatic systems, they reveal that no single axiomatic system (the set of fundamental properties an inconsistency index should possess) is universally accepted. Different indices, such as Saaty's CI, Koczkodaj's KI, and the Geometric Consistency Index (GCI), are based on different sets of axioms, meaning they are built on fundamentally different assumptions about what inconsistency is.

This lack of consensus on inconsistency definition is more than academic curiosity, it has significant practical implications. As a result, in MCDM field's AHP-based part scientists are developing highly sophisticated solutions to a problem that is not reliably measured and not even universally defined. With the lack of a clear, unified definition of inconsistency, the focus on a concrete, quantifiable target (e.g. a low CR value) becomes an attractive but potentially misleading goal. This focus on numerical optimization over conceptual clarity explains the growing spread of the "black box" algorithms that dominate the field. This ambiguity also provides a powerful motivation for a new approach in a different direction – by focusing on a more real and fundamental target: improving the decision-maker's comprehension and logical reasoning, which is the fundamental source of inconsistency (independent on its definition).

4.3. The mainstream of algorithmic correction

In response to the challenge of inconsistency, academic literature has produced a big variety of methods focused at correcting inconsistent pairwise comparison matrices (PCMs). A review of this topic showed a general tendency: the application of increasingly sophisticated, fully automated algorithms designed to mathematically optimize the matrix. These methods, while technologically advanced and correct, share a common philosophy that treats human judgments as a noisy signal to be algorithmically cleaned or corrected, often prioritizing mathematical aspects over other factors.

Metaheuristic and evolutionary approaches. A large area of research utilizes metaheuristic and evolutionary algorithms (EA) to search for optimally consistent matrices. This approach motivates to move towards greater computational complexity. For example, Wang et al.

(2023) employ an EA which, instead of meeting the conventional CR threshold, finds the matrix with the absolute minimum CR, achieving a state of almost perfect mathematical consistency with CR < 0.0001. In a similar way, Heymann et al. (2024) utilize a Genetic Algorithm (GA) to solve the complex problem of filling in missing values in incomplete PCMs, at the same time optimizing for a low CR together with preservation of a desired ranking of criteria.

Particle Swarm Optimization (PSO), another nature-inspired technique, has also been widely applied. Sun et al. (2025) developed a multi-objective PSO variant (IMOPSO) that focuses on a balance by minimizing at the same time both, the CR and the semantic deviation from the expert's original judgments. While analyzing group decision-making, Liu et al. (2023a) use PSO to algorithmically adjust the judgments of multiple experts to improve both the consistency of individual matrices and the consensus level of the group as a whole. The common idea among these studies is consideration of inconsistency as a complex optimization problem to be solved computationally, with the human decision-maker being largely absent from the correction process itself.

Mathematical and probabilistic programming. Another research direction focuses on other forms of mathematical programming and statistical modeling to achieve similar goals. These methods are often based on formulating the inconsistency problem as a constrained optimization model that can be solved with specialized software. Tu et al. (2023) present a two-stage optimization strategy designed for Interval Pairwise Comparison Matrices (IPCMs), where judgments are expressed as ranges instead of single values. Their model first minimizes the magnitude of adjustments and then minimizes the number of judgments involved. Liu et al. (2023b) introduce a model which is based on the concept of relative projection between matrices, which is then solved using a Gaussian Quantum Behavior Particle Swarm Optimization (GQPSO) algorithm to balance consistency and consensus in a group of experts. On the related topic of Quality Function Deployment (QFD), Xiao and Wang (2024) address the incomplete and conflicting opinions, often encountered in customer requirement analysis, using a consistency-driven optimization model.

Some of the authors used more sophisticated statistical methods. For example, Li et al. (2025) work is based on a Bayesian framework, which allows for comparing prior and subsequent respondent ratings, yielding a "posterior preference". Wang et al. (2024) used the "Probabilistic Consistency Index" (PCI) and a stochastic programming model to improve the probability that stochastic matrix is consistent p(CR < 0.1).

In order to achieve optimal solutions, the goal was to reduce the consistency index as much as possible, thus the respondent's decisions are increasingly treated as erroneous and need to be corrected regardless of his thinking and understanding process. This created the conditions for the development of the "black box" problem.

4.4. A critique of the algorithmic paradigm: the "Black Box" and the missing human

The application of formal algorithms for consistency optimization resulted in the fact that the respondent's decision was ignored. This factor reduced the clarity of the assessment using the "black box" method and created a gap between the goal of supporting the respondent's decisions and the methods used to evaluate the result.

The "Black Box" nature of automated correction. In the "black box" approach, the input is an inconsistent matrix, and the output is a mathematically consistent matrix. However, such approach lacks transparency leading to a lack of trust. There are several sophisticated optimization algorithms, for example, Sun et al. (2025) presented multi-objective particle swarm optimization, the Gaussian quantum behavior PSO by Liu et al. (2023b), and the genetic algorithm by Heymann et al. (2024). However, they provide the optimal solution without any explanation of what was wrong with the respondent's assessment, why it needed to be adjusted.

Decision support systems (DSS) can also face the "black box" paradox. Some DSSs have an interactive correction mode. The DSS proposed by Escobar et al. (2023) may be an example of this case. However, the presented system lacks an explanation of why changes are made and how to reduce respondent misunderstanding. Meanwhile, it is also important to improve the respondent's decision-making skills and increase confidence in the DSS's work.

The absence of human-centric evaluation. Literature lacks human-oriented indicator evaluations. Results are measured using mathematical criteria. For example, the Consistency Ratio (CR) reduction criterion is mentioned in Kaushik et al. (2024) and Kuraś et al. (2024), the Geometric Consistency Index (GCI) is used in Escobar et al. (2023), Minimizing deviation is applied by Sun et al. (2025) and Tu et al. (2023), and calculation speed is measured by Yuen (2024) and Kuraś et al. (2023). How different is work of Sato and Tan (2023) in which the authors measure a proxy for user satisfaction (their GWP metric).

In the works for respondents' satisfaction, confidence in the evaluation of calculation results, and time costs are given little attention. Therefore, future work should return to a human-centered approach.

4.5. Alternative philosophies and systemic obstacles

Alternatives to the algorithmic correction paradigm can be found in the literature. The idea of eliminating causes of inconsistency as the initial goal could lead to a more complete picture of the situation. However, it requires redesigning the decision-making process.

Alternative philosophy: inconsistency prevention. Mostafa (2024) and Vommi and Vommi (2025) suggest that the respondent Þrst evaluates all criteria and then moves on to comparing adjacent items. The weakness of this method is that if an error is made early in the process, it can propagate through the entire matrix without any mechanism for correction. Furthermore, the respondent's preferences are simpliÞed into a rigid linear structure.

It is important to appreciate that these methods, by removing inconsistency from the process, eliminate the valuable diagnostic feedback that inconsistency detection can provide. A high value of the consistency ratio (CR), despite all its shortcomings, is a signal that the decision maker may have misunderstood the criteria or is having difÞculty Þnding a logical compromise. Preventing this signal from occurring is another way to avoid the main problem of human cognitive error, rather than interacting with and solving it. This is achieved through mathematical purity, limiting human self-expression and eliminating the potential opportunity for learning and improvement.

The obstacle to the application of AI is the cul*ture of closure prevailing in the scientific world*. Currently, due to the high level of scientific competition and excessive fear of data leakage and plagiarism, a harmful practice has developed in the world, which can be described as a systemic disruption of scientific practice, which leads to a lack of universal reproducibility. i.e. many studies lack publicly available code, primary or detailed data instructions on how to re-examine leading to a lost the opportunity to use this particular study in other more complex studies in the development stages. These statements are supported by the authors' in-depth analysis of 24 literature sources, which shows that this is a widespread analysis worldwide. Many works that offer complex, algorithm-based solutions do not provide tools for independent verification. All data is classified and hidden, for example Sun et al. (2025), Li et al. (2025), Çoban (2023), Liu et al. (2023b), as well as the work of Srđević and Srđević (2024), and others. As a result, without access to the code and data, scientists cannot accurately verify, validate the results, or build their own research based on others' work.

In contrast, there are examples supporting open science. These scientists provide a basis for global trends and encourage the formation of new rules in the scientific community. Heymann et al. (2024) and Yuen (2024) provide public GitHub repositories with code, Bose (2023) makes his methodology available through the open-source R package. We argue that such transparency should be the norm, not the exception. The current culture of insularity in the scientific world is weakening the scientific claims and capabilities of the entire field. An environment has now been created in which methods cannot be properly evaluated and progress is slowed down, which in turn slows down workability and reduces the number of new scientific discoveries. It is necessary to take a step and share more data, because in our belief, any new scientific contribution that is transparent, open and reproducible, has a value for the scientific community that goes beyond its specific methodological novelty.

AI application is limited by over-reliance on synthetic data. The study highlights another major systemic problem that limits the applicability of most existing studies – an over-reliance on synthetic data. The introduction emphasizes that many studies utilize randomly generated or synthetically created matrices rather than real human judgments, limiting the generalizability of findings. This practice is common in studies that require large datasets for validation, such as Çoban (2023), Kuraś et al. (2023), and Wang et al. (2024). Although Bose (2023) attempts to correct this by modeling "logical" pairwise comparison matrices (PCMs), this still remains a surrogate, not a substitute for authentic human input.

Inconsistency is fundamentally a human phenomenon, rooted in cognitive biases, misunderstandings, and the inherent difficulty of comparing abstract concepts. Data validation studies using purely random numerical data may not adequately capture the structured error patterns that arise from human thought processes. Consequently, methods developed and tested only on synthetic data may prove less effective when confronted with the complexity of real-world decision-making. This again emphasizes the great need for research-based data derived from real human decisions that are human-centered in their approach.

Building human-centered AI. A review of the recent literature about inconsistency management in AHP shows that the field is both technologically advanced and conceptually stagnant at the moment. While researchers have developed an impressive number of sophisticated algorithms, they have done so at the expense of the human, mathematically focused, opaque, and largely untested in terms of its practical utility in improving human decision-making. As a result, the identified shortcomings and problems in computational, methodological, and scientific practice do not merely encourage incremental improvement; they require a fundamental systemic change.

The need for a new approach to the application of LLM. The combination of all the above shortcomings – faulty metrics, opaque algorithms, missing human elements and poor scientific practices – creates a clear and urgent need for a new approach. It is important to note that the methodology proposed in this study, which exploits the capabilities of large language models (LLM), is not just another tool for correcting matrices. It is positioned as a direct response to the identified gaps and a catalyst for a necessary change in the system of scientific values and beliefs.

As indicated in the introduction to this study, the LLM-based approach fundamentally reorients the goal of inconsistency management. This approach addresses the black box directly, moving from opaque numerical optimization to transparent, dialogue and explanation-based correction. LLM can formulate the root of logical contradiction in understandable language, helping the user to understand why his judgements are inconsistent.

It directly addresses the missing human problem by shifting the focus from the mathematical artifact (the matrix) to the cognitive agent (the decision maker). The primary goal is not to "fix" the numbers, but to improve the user's understanding, thereby empowering them to make corrections on their own based on clarified thinking.

And the most importantly, such approach addresses the root cause of inconsistency – defined in the abstract of this study as "insufficient understanding of the respondent's task" – rather than simply treating the symptom (a high congruence ratio (CR) value). By supporting clearer communication and personalized recommendations, LLM can help prevent inconsistency before it occurs and resolve it more meaningfully.

4.6. Formulating the theoretical argument: from mathematical correction to cognitive collaboration

In literature review we discussed AHP CI improvement as a field that is at the same time technologically advanced and conceptually stuck. Sophisticated algorithms were developed at the cost of the focus on human decision-maker, mathematically clear but largely unverified in the practical context of improving human judgment. As a result, it clarified fundamental flaws that demand a systemic, paradigmatic shift.

The established standard for success, Saaty's Consistency Ratio (CR), has been shown by many authors to be flawed:

- statistically insensitive to some critical aspects;
- in some context functionally blind to the outcomes it was designed to measure;
- systematically flawed in misclassification of consistency by its dependence on matrix size;
- showing artificial reliability where none exists.

The aim of reducing CR below the 0.1 threshold, therefore, may have been a misguided direction, leading to optimization of flawed numerical targets rather than true decision quality. This led to the pursuit of increasingly complex solutions, such as Multi-objective Particle Swarm Optimization (IMOPSO) or Gaussian Quantum Behavior Particle Swarm Optimization (GQPSO), and the field has become stuck in a self-reinforcing loop where flawed metrics drive the development of "black box" tools, which in turn further strengthens the focus on those flawed metrics.

Some better algorithm is not sufficient to break this loop, it requires a new paradigm. This study argues that the primary goal should not be the mathematical "correction" of a numerical artifact (the matrix) but rather a cognitive "collaboration" with the human agent (the decision-maker). The new paradigm shifts the focus from treating the symptom of a high CR value to addressing the root cause, which this study identifies as the "respondent's insufficient understanding of the task". In this paper, therefore, there is not proposed another optimization technique but rather a theoretical justification for a necessary paradigm shift, utilizing the unique capabilities of LLMs to re-center the human in the decision-making process.

4.7. Stepping back from the black box: LLM-powered transparency and explainability

The strongest argument for an LLM-based approach lies in its natural ability to resolve the "black box" problem, a central failure of the dominant algorithmic paradigm. The field of Explainable AI (XAI) has established that transparency and explainability are not optional features but essential features to build trust and guarantee reliability, particularly in risk-sensitive applications where decisions have significant impact. Traditional dominant AHP correction methods, with their focus on computational optimization provide corrected outputs without explaining why the initial judgments were logically flawed or how the adjustments were made, leading to a loss of user trust and understanding, in essence they guess the user intentions without a user assuming them being correct and applying these corrections. These algorithms operate silently, treating human judgments as a noisy signal to be algorithmically cleaned. The output is a new set of numbers after the transformations with a hidden complex logic. To understand and interpret this change, a user would need a separate translation layer to interpret the numerical adjustment, a process that is unnatural, cognitively demanding and is not based on correcting logic, therefore, it is flawed.

In contrast, LLM's native working primitives come from natural language, which is a natural operational environment for human cognition and logical reasoning. An LLM can directly address a logical contradiction by verbalizing it. For example, a mathematical inconsistency in three comparisons "A is strongly preferred to B, B is strongly preferred to C, but C is strongly preferred to A" can be transformed from numerical relationships into a simple, intuitive sentence: "It seems there is a logical conflict in your comparisons. You prefer criterion A over B, and B over C, but then you also prefer C back over A. This creates a circular loop with contradiction. Would you like to review these three specific comparisons together?"

The ability of LLMs to generate natural language explanations (NLEs) provide justifications that are human-readable and directly enhance user understanding. This creates a convenient connection between the tool and the user. At the beginning of the AHP task itself linguistic logical preferences translate into numbers (e.g., "cost is moderately more important than durability"). The inconsistency is created from a conflict in these logical preferences already. LLM operates at the level of language and logic directly, bypassing the need to interpret abstract numerical feedback for the user, thus it reduces the cognitive load required to understand and act with the system's guidance (Xu et al., 2024). Therefore, LLM is not just a better explainer – it is a cognitively compatible (based on language and logic at the first place) tool for this specific human-centric task.

LLM-based addressing the root cause of inconsistency. Recent research has shown that the LLM can successfully eliminate the root cause of inconsistency: "insufficient understanding of the respondent's task" by using authentic explanatory power. LLM has the ability to perform logical reasoning and detect inconsistencies – and this ability can be purposefully improved – including precisely those specific violations that cause problems in AHP (Analytical Hierarchy Process) matrices (Cheng et al., 2025). Meanwhile, algorithmic methods can only "correct" the symptom (a high value of the coefficient of consistency (CR)).

The LLM system successfully solves inconsistency problem by initiating a dialogue and helping the user diagnose, understand and learn from the cognitive errors that cause inconsistency. The system helps the user to correct their own thinking – it encourages self-reflection and helps learners recognize and correct their own false beliefs (Perez & Ong, 2024). This approach reflects the principles of effective dialogue-based learning systems that use targeted questions and feedback. By using this capability in a conversational environment, LLM acts as an experienced and subtle guide that reveals the user's erroneous or incomplete thinking so that it can be consciously re-evaluated and changed. LLM directly addresses the human source of error, stimulating deeper understanding that can prevent future inconsistencies and ultimately lead to better quality decisions. As a positive side effect, LLM can act as a teacher to the expert to help him properly evaluate using AHP, therefore, indirectly improving assessment in long-term.

Revealing the importance of the human element in approaching conscious collaboration between machine and human. LLM generates new goals and tools for inconsistency management by focusing precisely on managing human understanding better, thereby improving the quality of decisions made and their success rates. This contrasts with the lack of human-centered evaluation in the existing literature, which systematically ignores the human at the center of the decision-making process.

The current situation can be characterized by an algorithmic paradigm when consistency is evaluated only by machine-centered metrics: the reduction of the coefficient of consistency (CR) or geometric consistency index (GCI), the minimization of the deviation from the original matrix, or the pure computational speed of the correction algorithm. This means that mathematically "correct" solutions are generated without empirical evidence confirming that they lead to better, more reliable, or more trustworthy human decisions.

The emerging field of human-centered LLM evaluation proposes that the main metrics of AHP inconsistency management tools should be focused on the cognitive and emotional state of the user (Xu et al., 2024). Thus, creating new success criteria that are focused on the human factor.

These metrics should cover the following factors:

- User confidence: the decision maker's confidence in the final priorities derived from his or her evaluations.

- Decision satisfaction: the user's satisfaction with both the interactive process of resolving inconsistencies and the final outcome of the decision.
- Reduced cognitive load: the perceived ease of performing the AHP task, especially when identifying and reviewing inconsistent evaluations. Studies have already shown that LLM-based agents can significantly reduce cognitive load compared to traditional human consultation (Xu et al., 2024).
- Task understanding: Should measure the improvement of the user's understanding of the AHP criteria, the logic of pairwise comparisons, and their own biases.

The application of these metrics redefines the very purpose of the intervention. This means that a system optimized for user confidence will prioritize clarity of explanations and supportive dialogue over the speed of numerical optimization alone.

This is consistent with frameworks such as ConSiDERS-The-Human, which argue that a reliable evaluation must consider the overall user experience (UX), including usability, cognitive biases, and emotional response (Elangovan et al., 2024).

The work of Sato and Tan (2023) with their GWP indicator to measure user satisfaction is a rare but vital precedent for this shift in the AHP literature itself, demonstrating how possible and important it is to measure human-centered outcomes.

From correction to collaboration: *LLM as a cognitive partner*. It is now becoming clear that the role of LLM is not to impose a mathematically optimal solution, but to make easier the user's own thinking process, helping them to refine their thoughts. It becomes clear that the user is no longer just a passive subject whose "mistakes" are corrected by machine. Instead, the user becomes an active participant in a collaborative dialogue with an AI, which acts as a cognitive partner (Lou et al., 2025). It is becoming clear that this change in indicators creates a deeper variant of the interaction model, in which humans and AI have complementary capabilities. Models of human-AI teamwork are emerging, in which AI is not just a tool but an active collaborator in the problem-solving process (Lou et al., 2025). This model is well suited to what has been called "decision-oriented dialogues" – complex scenarios (Lin et al., 2024).

This shift can make powerful but complex methodologies such as AHP more accessible, intuitive and reliable to a wider range of non-expert decision-makers, effectively democratizing access and eliminating the "black box" principle that has long been a barrier to adoption. In the context of AHP, a human has tacit knowledge, preferences and understanding of the context, while the AI partner is able to quickly process the logical structure of assessments and formulate possible objections (Lin et al., 2024).

4.8. Improving AHP with LLM support

The respondent could be better engaged in the AHP evaluation process if there is given the opportunity to communicate with the system through LLM for greater transparency and feedback. To this purpose, a step-by-step model is provided:

Step 1: *Initial submission of the respondent's assessment*. The respondent fills in a pairwise comparison matrix based on the list of criteria to be assessed, using the standard Saaty or other chosen assessment scale. Traditional AHP software can be used at this stage.

Step 2: *Inconsistency detection using software*. The system calculates a selected consistency index. For example, this index can be CR, GCI, or CR_{β}, as presented by Sato and Tan (2023). If the index exceeds a set threshold, then the LLM-based interview function with the respondent is activated.

Step 3: *Applying LLM by implementing feedback through an interview with the respondent*. After identifying the largest source of inconsistency, the respondent is informed about the illogicality in assessing the criteria, indicating these criteria, and describing how much their importance was illogically assessed. The respondent is suggested to examine these criteria more closely.

Step 4: *Re-assessment and clarification of the respondent's position*. Instead of automatic consistency correction, the LLM-based system encourages the respondent to reflect on the criteria assessments and indicate the reasons for such an assessment. The respondent is offered to focus on problem areas and explain why one criterion was assessed too much or too little. Based on these open-ended questions, the respondent's erroneous attitudes are revealed, and the cognitive process is stimulated. A similar system designed to correct students' misconceptions is described by Perez and Ong (2024).

Step 5: *iterative improvement and confirmation of the assessment results*. The respondent adjusts assessments, the system calculates the consistency index and provides recommendations until the logic of the decisions is ensured and an acceptable consistency index is obtained. At the respondent's request, the results are confirmed or further improvement is carried out.

The above-mentioned step-by-step model is also useful in that it collects data from the interview with the respondent, which, when analyzed, makes the decision-making path, changes in opinions, and the justification for all adjustments clearer. The system ensures better decision justification, transparency, and traceability, which ensures an advantage compared to purely automatic mathematical correction methods.

Synthesis of a new method, *change of approach*. After analyzing the literature, it was determined that the problem under consideration must be reconsidered. The opaque "black box" methods used to improve consistency, mathematical optimization, and the lack of respondent-oriented assessment methods do not become additional problems, but rather a reason for fundamental change in approach.

Based on the analysis, it can be stated that a respondent-centered LLM-based system would be an evolutionary step in changing the approach to ensuring consistency. The advantages obtained compared to the old approach are:

1. Clear, understandable dialogue instead of opaque numerical correction using a "black box".
2. Orientation to respondent trust, satisfaction, knowledge and understanding, rather than purely mathematical purity.
3. Concentration on the cause of inconsistency caused by insufficient understanding of the task, rather than on the consequence of poor consistency.

The comparison of the old and new approaches to improving consistency is presented in Table 1.

Table 1. Description of approaches for improvement of consistency in AHP

Approach	Description of the current dominant approach to algorithmic application	Description of the emerging respondent-centered collaborative approach
Subject under consideration	Paired comparison matrix values	Knowledge of the decision maker (respondent)
Nature of interaction with the respondent	Receiving the assessment results from the respondent and adjusting this result	Supporting dialogue and collaboration with the respondent
Measured indicators	Consistency index (CR), Geometric Consistency Index (GCI), data processing speed, differences from the original assessment matrix	Respondent satisfaction with the result, level of confidence, improved task understanding
Nature of support	Respondent error correction	Support for the respondent in improving dialogue and decision-making
Planned support result	Optimized assessment matrix, with a better consistency index, regardless of the respondent's wishes	A matrix of better consistency aligned with the respondent's opinion

The respondent – decision maker and cognitive artificial intelligence, cooperating, identify opportunities for more effective decision-making. This is the basis for further empirical research in developing a new approach.

4.9. Illustrative case

To better visualize the proposed interaction process for readers we will provide the hypothetical illustrative example based on the case study by Maceika et al. (2021), where different criteria for construction projects were assessed, based on the Table 4 of aforementioned study, we obtain the matrix, which is corrupted by changing two values P-E and E-P from 1/3 and 3 to 3 and 1/3, accordingly, the result is presented in Table 2.

Table 2. Pairwise comparison matrix

Criteria	P	E	S	L
Psychological (P)	1	3	1/3	2
Economic (E)	1/3	1	1	5
Strategic (S)	3	1	1	5
Location (L)	1/2	1/5	1/5	1

By applying the consistency ratio calculation procedure, which can be found in research by Maceika et al. (2021) in equations (4)–(11), the obtained $CR \approx 0.223$, above the acceptable threshold. The concrete methodology to develop well-rounded universal prompt, utilizing the best practices of prompt engineering is out of the scope of current research. Therefore, we formulate the following minimalistic prompt to the ChatGPT 5.2 extended thinking model: *Please help an expert to improve his assessment in evaluation of criteria for AHP method for a given AHP criteria evaluation table, where consistency is 0.223 which is much bigger than 0.1. Explain the possible logical contradictions in this inconsistency and possible ways to correct it so expert would understand why his assessment was not inconsistent and could reevaluate his judgements. Avoid focusing on maths, you help to address the inconsistency problem in a human-centred way. Be short, 100 words is your limit. Criteria P E S L Psychological (P) 1 3 1/3 2 Economic (E) 1/3 1 1 5 Strategic (S) 3 1 1 5 Location (L) 1/2 1/5 1/5 1.*

Chatbot's answer: *Your pairwise statements clash in a few "triangles". Example: you say P > E (3) and E = S (1), so P should be close to S—but you also say S > P (3). Another: E > L (5) and P > E (3) suggest P should greatly exceed L, yet you rate P > L only (2). To fix: revisit these triangles, decide which judgment you trust most, soften extreme 5s, and re-rate using the same verbal meaning of "3" and "5" across all pairs.*

As we can see, according to LLM there are two problematic triangles P-E-S and E-L-P, sharing the evaluation between P and E, so it's a huge hint, to restore the original, balanced matrix, which was used by Maceika et al. (2021).

5. Discussion

After analyzing the scientific literature on improving consistency using the AHP method, several aspects for discussion were identified.

One of them is the return to increasing the influence of the human as a decision maker, rather than using only a powerful mathematical optimization mechanism. In this area, it would be worth using an agile way to improve cooperation between AI and humans. According to Sowa et al. (2021), it can be seen that there is an increasingly positive attitude towards AI in knowledge work, but also significant fears related to full automation, including among the younger generations, in addition, the research revealed the need for further personalization of AI tools as well as the need for more collaborative approach to the future of work. Hassan et al. (2025), study highlights the importance of designing AI systems that complement rather than replace, human expertise, therefore, involving users and encouraging them to provide input and feedback is crucial for developing software that aligns with clinical workflows and decision-making processes is crucial. According to Bauer and Michalowski (2025), there is an opportunity to extend the current model-centered evaluation approaches to incorporate human-centered metrics, supporting the transition into practice.

The second aspect is the adaptation of the criteria used in the AHP method to be more focused on the respondent, on his cognitive and emotional state. Ensuring greater transparency in decision-making would also contribute to greater respondent satisfaction both in performing the task and in evaluating the final results. There is a need to build human-aware AI and explanations by aligning AI and explanation updates with human knowledge (Wang, 2025). According to Kim et al. (2024), explanations that are meaningful to users can be divided into three components: the contextualized quality of the explanation, the contribution of the explanation to human-AI interaction, and the contribution of the explanation to human-AI performance.

The third aspect is trust in the results provided by AI. Based on Ding et al. (2025), trust is a decisive factor influencing the quality of human-AI collaboration, as uncalibrated trust may lead to task failure or even catastrophic consequences, significantly compromising the safety of human–machine systems. According to Peters and Chin-Yee (2025), AI has a tendency to overgeneralize scientific findings and here can also be a subtle form of "hallucination". The study of Pan et al. (2025) provides evidence demonstrating the importance of targeted AI literacy development in building trust and fostering effective collaboration in human-generative AI teams. Explainable AI (XAI) systems are designed to provide clear explanations of how the system arrived at a decision or prediction, which increases users' trust (Morandini et al., 2023). The research of Tiwari (2023) on Explainable AI (XAI) and its applications in building trust and understanding in AI decision making has shown that XAI has the potential to enhance trust and understanding in AI decision making, however, more research is needed to fully understand the effectiveness of different XAI techniques and to address the challenges associated with implementing XAI in specific domains. Based on Van Leersum and Maathuis (2025), human-centered explainable AI (HCXAI) could be a solution to focus on humane ethical decision-making instead of pure technical choices.

6. Future research agenda

The findings of this study are a starting point for a new, human-centered research agenda for hybrid AHP-LLM methods.

The identified gaps and proposed theoretical foundation shows a clear and actionable direction for the research. First, there is an urgent need to bridge the gap between theory and practice by conducting empirical validation and developing prototypes. Future work must focus on creating a working prototype of the LLM-based Analytic Hierarchy Process (AHP) system described in this article. This prototype should be tested through controlled experiments, comparing its performance with traditional algorithmic methods and unaided decision-making. Most importantly, these experiments should be evaluated not using outdated, machine-centered metrics, but using the proposed human-centered metrics.

Second, coordinated efforts across multiple studies are needed to formalize and standardize human-centered metrics in the field of MCDM. This includes the development and validation of reliable survey instruments designed to measure factors such as Decision Satisfaction and User Trust. Establishing these metrics is a necessary condition for changing the evaluation standards in this field. A closed culture and over-reliance on synthetic data have slowed down progress and limited the generalizability of findings.

Therefore, a key direction for future work is to:

- Develop and share public repositories of AHP matrices derived from real human decisions in real-world decision-making scenarios.
- Support the practice of publishing new algorithmic methods and LLM-based prototypes as open source, encouraging reproducibility, verifiability, and collaboration-driven progress.

This agenda is not just about testing a new tool; it is a plan to create an entirely new research ecosystem. The critique presented in this work is intended not only for the theories of the field but also for its research practices. The proposed LLM-based method can only be properly developed, tested, and implemented within a scientific community that prioritizes transparency, reproducibility, and, most importantly, the human decision-maker.

7. Conclusions

The findings presented in this article are focused on the future of decision support algorithms, they extend beyond the specificity of AHP consistency. By supporting a shift from blind algorithmic consistency correction to transparent cognitive collaboration (between human and AI), this work contributes to a broad scientific movement for creating effective, reliable and human-centric artificial intelligence.

A model for human-AI cognitive partnership. The proposed paradigm shift is not a narrow solution for an MCDM method, it is rather a powerful approach for a new generation of evolving methods based on human-AI collaboration. AI model acting as a cognitive partner, which explains, questions and guides, is applicable to any domain where complex human assessment is essential. Medical diagnosis, legal analysis, financial planning – in all these domains the principles of transparent, AI-human dialogue-based support can help build systems that complement rather than replace human expertise. This research should be considered as study in larger context of the design of AI systems that augment human intelligence and empower human with better decision-making abilities through partnership with AI.

Enhancing the quality and accessibility of critical decisions. The goal of this research is to improve the quality of critical decisions. Although this goal is not reached via concrete experimental setup and test, the main attempts were focused on systemizing the knowledge from other scientific sources. As a result, there were built arguments supporting the improvement of complex and often unintuitive methodology like AHP by making it more transparent, accessible and trustworthy by extending it with LLM-based AI. Therefore, it can empower a wider range of non-expert users to systemize their thinking, identify their biases (as a result – inconsistencies) and understand complex trade-offs with clarity and confidence. The shift from algorithmic correction is more than a methodological preference – it is an essential step toward building decision support systems that are not limited only by mathematical consistence but also more effective, reliable and better controlled/understood by a human.

References

Bauer, J. M., & Michalowski, M. (2025). Human-centered explainability evaluation in clinical decision-making: A critical review of the literature. *Journal of the American Medical Informatics Association*, *32*(9), 1477–1484. https://doi.org/10.1093/jamia/ocaf110

Bose, A. (2023). Improving consistency classification: An innovative benchmark-based approach for the AHP. *Journal of Multi-Criteria Decision Analysis*, *31*(1). https://doi.org/10.1002/mcda.1821

Cheng, F., Li, H., Liu, F., van Rooij, R., Zhang, K., & Lin, Z. (2025). Empowering LLMs with logical reasoning: A comprehensive survey. In *The 34th International Joint Conference on Artificial Intelligence* (*IJCAI*). https://doi.org/10.24963/ijcai.2025/1155

Çoban, V. (2023). Developing random indices and consistency ratios for inconsistency methods in pairwise comparison. *Journal of The Faculty of Engineering and Architecture of Gazi University*, *38*(2), 781–793. https://doi.org/10.17341/gazimmfd.903495

Čančer, V. (2024). Selection procedure of the approximation methods for deriving priorities: A case of inconsistent pairwise comparisons. *Business Systems Research Journal*, *15*(2), 21–30. https://doi.org/10.2478/bsrj-2024-0015

Ding, S., Pan, X., Hu, L., & Liu, L. (2025). A new model for calculating human trust behavior during human-AI collaboration in multiple decision-making tasks: A Bayesian approach. *Computers & Industrial Engineering*, *200*, Article 110872. https://doi.org/10.1016/j.cie.2025.110872

Elangovan, A., Liu, L., Xu, L., Bodapati, S. B., & Roth, D. (2024). ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, *1*, 1137–1160. https://doi.org/10.18653/v1/2024.acl-long.63

Escobar, M. T., Aguarón, J., Moreno-Jiménez, J. M., & Turón, A. (2023). A decision support system for improving the inconsistency in AHP. *International Journal of Decision Support System Technology*, *15*(2), 1–16. https://doi.org/10.4018/IJDSST.315644

Hassan, R., Nguyen, N., Finserås, S. R., Adde, L., Strümke, I., & Støen, R. (2025). Unlocking the black box: Enhancing human-AI collaboration in high-stakes healthcare scenarios through explainable AI. *Technological Forecasting and Social Change*, *219*, Article 124265. https://doi.org/10.1016/j.techfore.2025.124265

Heymann, M. C., Pereira, V., & Caiado, R. G. G. (2024). PyMissingAHP: An evolutionary algorithm for filling missing values in incomplete pairwise comparisons matrices with real or fuzzy numbers via mono and multiobjective approaches. *Arabian Journal for Science and Engineering*, *49*(1), 7375–7394. https://doi.org/10.1007/s13369-023-08227-4

Islam, R., Anis, A., & Azam, M. S. E. (2024). Revitalizing university performance evaluation: The case of SETARA model in Malaysia. *Journal of Applied Research in Higher Education*, *17*(4), 1294–1312. https://doi.org/10.1108/JARHE-12-2023-0561

Kaushik, S., Pant, S., Joshi, L. K., Kumar A., & Ram, M. (2024). A review based on various applications to find a consistent pairwise comparison matrix. *Journal of Reliability and Statistical Studies*, *17*(1), 45–76. https://doi.org/10.13052/jrss0974-8024.1713

Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. *Frontiers in Artificial Intelligence*, *7*, Article 1456486. https://doi.org/10.3389/frai.2024.1456486

Kuraś, P., Strzałka, D., Kowal, B., Organiściak, P., Demidowski, K., & Vanivska, V. (2024). REDUCE-a tool supporting inconsistencies reduction in the decision-making process. *Applied Sciences*, *14*(23), Article 11465. https://doi.org/10.3390/app142311465

Kuraś, P., Strzałka, D., Kowal, B., & Mazurek, J. (2023). REDUCE – a Python module for reducing inconsistency in pairwise comparison matrices. *Advances in Science and Technology Research Journal*, *17*(4), 227–234. https://doi.org/10.12913/22998624/170187

Li, H., Dai, X., Wu, Q., Zhou, L., & Pedrycz, W. (2025). A Bayesian analysis framework for decision making with interval pairwise comparison judgments. *Decision Analysis*. https://doi.org/10.1287/deca.2024.0207

Lin, J., Tomlin, N., Andreas, J., & Eisner, J. (2024). Decision-oriented dialogue for human-AI collaboration. *Transactions of the Association for Computational Linguistics*, *12*, 892–911. https://doi.org/10.1162/tacl_a_00679

Liu, F., Liu, T., & Hu, Y.-K. (2023a). Reaching consensus in group decision making with non-reciprocal pairwise comparison matrices. *Applied Intelligence*, *53*, 12888–12907. https://doi.org/10.1007/s10489-022-04136-5

Liu, Z.-L., Liu, F., Zhang, J.-W., & Pedrycz, W. (2023b). Optimizing consistency and consensus in group decision making based on relative projection between multiplicative reciprocal matrices. *Expert Systems with Applications*, *224*, Article 119948. https://doi.org/10.1016/j.eswa.2023.119948

Lou, B., Lu, T., Raghu, T. S., & Zhang, Y. (2025). *Unraveling human-AI teaming*: *A review and outlook*. SSRN. https://doi.org/10.2139/ssrn.5211067

Maceika, A., Bugajev, A., Šostak, O. R., & Vilutienė, T. (2021). Decision tree and AHP methods application for projects assessment: A case study. *Sustainability*, *13*(*10*), Article 5502. https://doi.org/10.3390/su13105502

Morandini, S., Fraboni, F., Puzzo, G., Giusino, D., Volpi, L., Brendel, H., Balatti, E., De Angelis, M., De Cesarei A., & Pietrantoni, L. (2023). Examining the nexus between explainability of AI systems and user's trust: A preliminary scoping review. *CEUR Workshop Proceedings*, *3554*, 30–35. https://ceur-ws.org/Vol-3554/paper6.pdf

Mostafa, A. M. (2024). A group multi-criteria decision-making approach based on the best-only method for cloud service selection. *IEEE Access*, *12*, 119946–119957. https://doi.org/10.1109/ACCESS.2024.3450280

Pan, Z., Moore, O. A., Papadimitriou, A., & Zhu, J. (2025). AI literacy and trust: A multi-method study of human-GAI team collaboration. C*omputers in Human Behavior*: *Artificial Humans*, *4*, Article 100162. https://doi.org/10.1016/j.chbah.2025.100162

Pant, S., Kumar, A., & Mazurek, J. (2025). An overview and comparison of axiomatization structures regarding inconsistency indices' properties in pairwise comparisons methods: A decade of advancements. *International Journal of Mathematical*, *Engineering and Management Sciences*, *10*(1), 265–284. https://doi.org/10.33889/IJMEMS.2025.10.1.015

Peters, U., & Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, *12*, Article 241776. https://doi.org/10.1098/rsos.241776

Perez, J., & Ong, E. (2024). Designing an LLM-based dialogue tutoring system for novice programming. In *Proceedings of the 32nd International Conference on Computers in Education*. Asia-Pacific Society for Computers in Education. https://doi.org/10.58459/icce.2024.4954

Saaty, T. L. (2000). *Fundamentals of decision making and priority theory with the Analytic Hierarchy Process* (Analytic Hierarchy Process Series, Vol. 6). RWS Publications. https://doi.org/10.13033/isahp.y1999.038

Sato, Y., & Tan, K., H. (2023). Inconsistency indices in pairwise comparisons: An improvement of the consistency index. *Annals of Operations Research*, *326*, 809–830. https://doi.org/10.1007/s10479-021-04431-3

Sowa, K., Przegalinska, A., & Ciechanowski, L. (2021). Cobots in knowledge work: Human-AI collaboration in managerial professions. *Journal of Business Research*, *125*, 135–142. https://doi.org/10.1016/j.jbusres.2020.11.038

Srđević, B., & Srđević, Z. (2024). Multi-model assessing and visualizing consistency and compatibility of experts in group decision-making. *Mathematics*, *12*, Article 1699. https://doi.org/10.3390/math12111699

Sun, Y., Zhuang, L., Jis, T., Cheng, D., Zhao, X., & Guo, J. (2025). A risk assessment method for power internet of things information security based on multi-objective hierarchical optimisation. *IET Smart Grid*, *8*(1), Article e12208. https://doi.org/10.1049/stg2.12208

Tiwari, R. (2023). Explainable AI (XAI) and its applications in building trust and understanding in AI decision making. *International Journal of Scientific Research in Engineering and Management* (*IJSREM*), *7*(1), 1–13. https://doi.org/10.55041/IJSREM17592

Tong, X., & Wang, Z. J. (2023). New additive-consistency-driven methods for deriving two types of normalized utility vectors from additive reciprocal preference relations. *Journal of the Operational Research Society*, *74*(6), 1475–1494. https://doi.org/10.1080/01605682.2022.2096503

Tu, J., Wu, Z., & Xu, J. (2023). Geometric consistency index for interval pairwise comparison matrices. *Journal of the Operational Research Society*, *74*(5), 1229–1241. https://doi.org/10.1080/01605682.2022.2075803

Van Leersum, C. M., & Maathuis, C. (2025). Human centred explainable AI decision-making in healthcare. *Journal of Responsible Technology*, *21*, Article 100108. https://doi.org/10.1016/j.jrt.2025.100108

Vommi, A. M., & Vommi, V. B. (2025). A novel scale for inconsistency reduction in the pair-wise comparison matrices. *Foundations of Computing and Decision Sciences*, *50*(1), 87–114. https://doi.org/10.2478/fcds-2025-0004

Wang, P., Liu, Y., & Zhou, H. (2023). Research on the eco-geological environment carrying capacity in Pingwu County after the Wenchuan earthquake based on the modified AHP. *Natural Hazards*, *115*, 2097–2115. https://doi.org/10.1007/s11069-022-05629-9

Wang, X. (2025). *Human-centered evaluation and design of AI explanation in AI-assisted decision making* (Doctoral dissertation). Purdue University, West Lafayette, Indiana, USA. https://doi.org/10.1145/3640544.3645239

Wang, Y., Zhou, L., Li, H., & Dai, X. (2024). Probabilistic consistency of stochastic multiplicative comparison matrices based on Monte Carlo simulation. *Information Sciences*, *656*, Article 119896. https://doi.org/10.1016/j.ins.2023.119896

Xiao, J., & Wang, X. (2024). An optimization method for handling incomplete and conflicting opinions in quality function deployment based on consistency and consensus reaching process. *Computers & Industrial Engineering*, *187*, Article 109779. https://doi.org/10.1016/j.cie.2023.109779

Xu, Y., Gao, W., Wang, Y., Shan, X., & Lin, Y.-S. (2024). Enhancing user experience and trust in advanced LLM-based conversational agents. *Computing and Artificial Intelligence*, *2*(2), Article 1467. https://doi.org/10.59400/cai.v2i2.1467

Yuen, K. K. F. (2024). Closed-form solutions of consistency ratio in best worst method minmax optimization model: Max of edge error matrix and minmax edge error determinant methods. *Granular Computing*, *9*, Article 42. https://doi.org/10.1007/s41066-024-00459-5

APPENDIX

Table A1. Analysis of the statements by the 24 authors'

No.	Authors, year	Data source	Method used	Results
1	Sato and Tan (2023)	Two opinion surveys and large-scale numerical simulations	Evaluation using CAR (Conformity of Rankings) and GWP (Goodness-of-fit of Weight to human perception)	The results showed that the standard CR (Consistency Ratio) was statistically insensitive to both CAR ($p = 0.638$) and GWP ($p = 0.494$). In contrast, the new CR_{β} was highly sensitive in detecting inconsistencies measured by CAR ($p < 0.0001$) and showed a promising, though inconclusive, correlation with user satisfaction measured by GWP ($p = 0.098$)
2	Kaushik et al. (2024)	Decision-makers' judgments and randomly generated matrices	Success is mainly evaluated by the final Consistency Ratio (CR), with the goal of reducing it below the 0.10 threshold. More advanced methods also measure the deviation from the original judgments to ensure authenticity is preserved	Studies report that 62% of methods focus on CR reduction, with various techniques successfully repairing matrices while attempting to retain the original information
3	Escobar et al. (2023)	A single matrix from a publication by Saaty (2000) and simulated matrices	Reducing of the final Geometric Consistency Index (GCI) value. The quality of the correction is measured by the relative change in the final priority vector, which was kept minimal (5–8% max change) to preserve the user's intent. The number of corrections varied from 10 to 15 depending on the mode	GCI was successfully reduced from 0.790 to below the 0.37 threshold in the paper's example

Continue of Table A1

No.	Authors, year	Data source	Method used	Results
4	Pant et al. (2025)	Matrices based on existing academic literature	The Jaccard Index, used to quantify the similarity between different axiomatic systems	The main results are presented in tables that compare the axiomatic frameworks themselves and detail which of 18 inconsistency indices comply with the various axioms. The paper does not present performance metrics like CR reduction or ranking error
5	Tu et al. (2023)	Matrices based on existing academic literature	Performance is evaluated using the proposed Average Geometric Consistency Index E(GCI) and Uncertainty Index (UI), plus metrics for information loss: Number of Changes (NOC) and Amount of Change (AOC)	In a comparative test, the proposed correction method achieved the consistency target by modifying only 5 judgments with an AOC of 1.5010, outperforming baselines that altered all 12 judgments with higher information loss
6	Heymann et al. (2024)	Matrices based on existing literature	The algorithm's performance is evaluated based on the final Consistency Ratio (CR) and the preservation of a desired criteria ranking, which is measured using Kendall's Tau correlation (τ)	For one case study, the algorithm achieved a CR of 0.0300. Metrics concerning the rater's experience, such as time on task or satisfaction, were not assessed
7	Sun et al. (2025)	Newly generated matrices	Usage of optimization algorithm (IMOPSO) to correct inconsistencies. The optimization was run for 500 iterations	The method successfully reduced the average Consistency Ratio (CR) to 0.0152. Quality was measured by a low semantic distance from the original judgments and a high (90.86%) Pearson correlation with a traditional method, confirming semantic fidelity
8	Bose (2023)	Matrices based on simulation to mimic the outputs of a rational human decision-maker	Usage of an original benchmark-based framework designed to enhance the precision of consistency classification for pairwise comparison matrices (PCMs) within the AHP methodology. This innovative approach quantifies the discrepancy between a given PCM and its benchmark matrix, comprising comparison ratios that faithfully reflect the relative preferences encapsulated within principal eigenvector values, thereby capturing the true degree of coherence	The paper qualitatively claims that the new framework rectifies the known failures of the conventional CR method, which produces false positives for small matrices and false negatives for larger ones

Continue of Table A1

No.	Authors, year	Data source	Method used	Results
9	Çoban (2023)	Matrices based on newly generated synthetic data	The new ratios (Golden-Wang consistency ratio (MCCR) and Takeda consistency ratio (GWCR)) were evaluated against Saaty's CR using correlation and compatibility rates (agreement on the 0.10 threshold)	Results showed very high correlation and compatibility, especially for the MCCR. However, the perfect 100% compatibility for larger matrices ($n \geq 6$) is an artifact, as the random dataset contained virtually no consistent matrices at those sizes to test against
10	Li et al. (2025)	Matrices based on synthetically generated data	The evaluation is qualitative, based on illustrative examples and sensitivity analysis	The results demonstrate the framework's logical properties, such as its adaptability and the relationship between inconsistency and preference reversal. No quantitative performance metrics like consistency ratios, ranking errors against a ground truth, or tests of statistical significance are reported
11	Islam et al. (2024)	Matrices based on data from questionnaire for academic experts	The Consistency Ratio was calculated but no numerical results were reported	The key result was the significant change in criteria weights compared to the original model (e.g., "Input" domain weight increased from 20% to 41.05%)
12	Wang et al. (2023)	Newly generated spatial and environmental data specific to the Pingwu County case study	An Evolutionary Algorithm then computationally searches within these ranges to find the matrix with the absolute minimum Consistency Ratio (CR)	The key result was the optimal solution and standard deviation of Consistency Ratio (CR) was stable while the generation increases to five hundred ($CR < 10^{-4}$). The resulting weights showed that socio-economic factors ("Population" and "Basic Facilities") were the dominant drivers of risk, each weighted twice as heavily as any single geological factor
13	Kuraś et al. (2024)	Matrices based on illustrative examples created by the authors	REDUCE tool was used for the reduction of the Consistency Ratio (CR) to below the 0.10 threshold	Quantitative evaluations demonstrate that REDUCE can improve matrices with high inconsistency (e.g., CR = 0.25) to acceptable levels (e.g., CR = 0.08) while retaining up to 95% of the original preference integrity, depending on the chosen algorithm

Continue of Table A1

No.	Authors, year	Data source	Method used	Results
14	Liu et al. (2023a)	Author-generated raw numerical matrices for illustration and a separate simulation of random matrices	Cosine Similarity-based Consistency Index (CSCI) was used, which holistically measures deviations from both transitivity and reciprocity. Euclidean Distance (ED) and Misranking Value (MV) were used to assess prioritization quality	The proposed method performed best on the ED metric in the example. The consensus model's performance improved as the algorithm was allowed more flexibility. No statistical significance testing (e.g., p-values) was reported
15	Wang et al. (2024)	Synthetically generated via extensive Monte Carlo simulations	Probabilistic Consistency Index (PCI) was introduced and supplemented by the Consistency Ratio (CR) of the matrix's central tendency	The study shows that PCI is influenced by judgment consistency, uncertainty, and matrix size. It also demonstrates that the most likely ranking can change after consistency improvement, validating the necessity of the correction step before deriving final rankings
16	Liu et al. (2023b)	Matrices created by randomly sampling values	The strategy is fully automatic, using matrix optimization to correct inconsistencies. It minimizes a function balancing consistency (RPCI) and deviation from the original matrix. A Gaussian Quantum Behavior Particle Swarm Optimization (GQPSO) algorithm solves the model. Evaluation uses the proposed RPCI and compares it to CR, GCI, etc.. Correction quality is measured by minimizing changes to the original matrix, using max absolute change (ρ) and Euclidean distance (σ) as metrics	The paper's method uniquely met the criteria $\rho < 2$ and $\sigma < 1$ in its example. Final rankings are observed for stability. No rater-focused metrics (time, satisfaction) or formal statistical significance tests are included
17	Kuraś et al. (2023)	Matrices based on newly generated synthetic data	Evaluation focuses on computational performance, measuring the average execution time to reduce the Consistency Ratio (CR) to below the 0.1 threshold. No quantitative quality measures (like ranking error vs. ground truth), rater-centric metrics, or formal statistical significance tests were used	Results show the Cao et al. algorithm is up to six times faster than the other implemented methods for a 10×10 matrix

Continue of Table A1

No.	Authors, year	Data source	Method used	Results
18	Srđević and Srđević (2024)	Matrices based on opinion of the experts	Experts were ranked using the TOPSIS and Borda count methods to create a final quality assessment. The study does not report on rater time, satisfaction surveys, or formal statistical significance testing (e.g., p-values)	Results are presented as scores for four performance indicators (CR, ED, CO, SC) for each of the 12 experts; the group average CR was 0.13
19	Vommi and Vommi (2025)	Saaty's original matrices to simulate method's inputs	The method is evaluated by visually comparing its derived weights to Saaty's "actual weights" from benchmark problems, not through formal statistical analysis. A key issue is post-hoc selection bias: the "best" RPS scale (Low, Moderate, or High) is chosen retrospectively for each problem, with no predictive rule for new applications	The proposed RPS scales are compared against eight other existing AHP numerical scales using a single dataset: Saaty's "Buying a House" problem. The authors claim the results are close to Saaty's original scale, but this is a qualitative judgment
20	Tong and Wang (2023)	Matrices created by the authors	The method is evaluated on its ability to produce a "satisfactory" weight vector benchmarked against a gold standard	Results show it is the only method to achieve this, while also being more robust. It handles high-inconsistency cases where competing methods catastrophically fail or produce irrational results, like assigning zero weight to a criterion
21	Čančer (2024)	Matrices created and altered by the author	The methods are evaluated using Mean Absolute Deviation (MAD) and Mean Absolute Percentage Deviation (MAPD) to measure how closely their results match the Eigenvalue Method (EVM), which is treated as the "exact" baseline	The key finding is that the Geometric Mean method is the most accurate approximation, showing the smallest average deviation from the EVM results
22	Mostafa (2024)	Benchmark dataset, existing literature, and randomly generated matrices	The proposed GBOM method is designed to prevent inconsistency, consistently achieving a perfect CR of 0 in all validation tests	GBOM achieved a perfect Consistency Ratio (CR = 0) in all tests, while competing methods like Group-BWM and Group-AHP showed inconsistencies (CR up to 0.183). The quality measure was the final ranking, which was stable and logical. No rater time or satisfaction surveys were conducted, and statistical significance was not reported

End of Table A1

No.	Authors, year	Data source	Method used	Results
23	Xiao and Wang (2024)	Newly generated matrices for a case study	Evaluation focuses on mathematical metrics, primarily the Consensus Level (CL), which improved from 0.74 to an acceptable 0.95 in two iterations	The final ranking of customer requirements is the main output. The method is shown to be superior to baselines in achieving a higher Consistency Index (CI) and greater consensus efficiency (measured by symbolic distance). No user satisfaction or statistical significance metrics are reported
24	Yuen (2024)	Data is created stochastically by randomly generating comparison matrices	Evaluation focuses on two key metrics: accuracy of the inconsistency value (ξ*) and computational speed	The results show a dramatic performance gain, with the new methods being approximately 2058 times faster. The paper does not evaluate user-centric metrics like task time or ranking error