MACS: A Cognitive Diversity Multi-Agent Consensus Framework for Bias Mitigation in Automated Evaluation Systems

1st Arrival Dwi Sentosa
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
arrivaldwi@itb.ac.id

2nd Julyan Widianto
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
23525017@mahasiswa.itb.ac.id

***Abstract*—The reliance on single Large Language Models (LLMs) for automated academic assessment poses the risk of creating an algorithmic monoculture, where inherent model biases are amplified at scale. This paper introduces the Multi-Agent Consensus System (MACS), a framework designed to mitigate this risk by simulating cognitive diversity through a structured multi-agent workflow. MACS orchestrates a heterogeneous ensemble of LLMs in an adversarial peer-review process. The system comprises: (1) a VLM-driven multimodal extraction module for high-fidelity data retrieval from PDFs; (2) an initial review by a primary agent; (3) a critical challenge stage by secondary agents with diverse architectures; and (4) a final arbitration stage where a final arbiter agent synthesizes conflicting evaluations to form a robust consensus. By formalizing this process of structured disagreement and resolution, our framework moves beyond simple ensemble averaging. We introduce the Disagreement-Resolution Ratio (DRR) as a metric to quantify the system's ability to identify and correct initial scoring biases. Our empirical validation shows that MACS achieves a scoring consistency that surpasses the inter-rater reliability of human experts, demonstrating its potential for fairer and more reliable automated assessment.**

***Index Terms*—Automated Assessment, Multi-Agent Systems, Cognitive Diversity, Algorithmic Bias, Consensus Scoring, Large Language Models, Explainable AI, Educational Technology.**

I. Introduction

The evaluation of student work is a cornerstone of pedagogy, yet manual grading is labor-intensive and notoriously difficult to scale consistently. While Large Language Models (LLMs) promise a solution through automated assessment [3], their widespread adoption presents a critical challenge: the risk of an algorithmic monoculture [15]. A single LLM, even a highly capable one, possesses a unique set of inherent biases derived from its training data and architecture. Recent studies confirm that state-of-the-art models exhibit significant biases, potentially reinforcing societal inequalities when deployed as educational tools [5], [6]. Using such a model as a singular authority for evaluation can lead to the systematic and scaled penalization or rewarding of specific writing styles or viewpoints, failing to capture the true diversity of human intellect.

979-8-3315-9271-4/25/$31.00 © 2025 IEEE

Traditional automated systems offer limited semantic understanding [2], while simple ensemble methods (e.g., score averaging) dilute, rather than resolve, model disagreements. This paper argues that a more robust approach requires simulating the process of scholarly discourse itself: structured, critical, and evidence-based peer review. The recent surge in LLM-based Multi-Agent Systems (MAS) has demonstrated their potential for solving complex tasks through collaboration and emergent intelligence [7], a paradigm we adapt for the nuanced challenge of academic assessment.

To address these limitations, we propose the Multi-Agent Consensus System (MACS). MACS is not merely an ensemble method; it is a structured framework that operationalizes the principle of cognitive diversity by assigning distinct, and at times adversarial, roles to a heterogeneous set of LLM agents. The novelty of our work lies in formalizing a process of structured disagreement and resolution, which forces the system to confront and reconcile diverse algorithmic perspectives before rendering a final judgment.

Our primary contributions are:

- **A Cognitive Diversity Framework:** We propose a multi-agent architecture that simulates peer review to mitigate the algorithmic monoculture risk inherent in single-LLM assessment systems.
- **Structured Disagreement and Resolution:** We formalize the interaction between LLM agents as a multi-stage process of initial review, adversarial challenge, and final arbitration, moving beyond simple score aggregation.
- **An Evaluation Metric:** We introduce the Disagreement-Resolution Ratio (DRR) to quantitatively measure the effectiveness of the peer-review simulation in identifying and correcting initial assessment biases.
- **Empirical Validation:** We provide quantitative evidence that our framework achieves scoring consistency superior to the inter-rater reliability observed among individual human experts in a real-world paper competition.

II. Related Work

The pursuit of automated assessment is not new, but the rise of LLMs has shifted the landscape from traditional methods to more semantically sophisticated approaches. Our work builds upon and diverges from several key research streams.

A. Traditional and Ensemble-Based Automated Scoring

Early automated essay scoring (AES) systems like e-rater® relied on statistical models of syntax, style, and semantic coherence [2]. While effective for their time, they lack the deep contextual understanding of modern LLMs. More recent approaches have explored simple ensemble methods, such as averaging the scores from multiple LLMs to smooth out individual model quirks [12]. However, these "wisdom of the crowd" techniques risk diluting, rather than resolving, fundamental disagreements. If a majority of models share a common bias, averaging can amplify it, a phenomenon we term bias consolidation. MACS, in contrast, forces an explicit, reasoned resolution to disagreements, preventing biases from being silently averaged into the final score.

B. Multi-Agent Systems for Complex Problem Solving

The paradigm of Multi-Agent Systems (MAS), where multiple autonomous agents collaborate, has seen a recent explosion of interest with LLMs [7], [10]. Frameworks like AutoGen have shown that LLM agents can work together to write code, conduct research, and simulate complex social interactions [13]. However, many of these systems are designed for cooperative task decomposition. MACS distinguishes itself by architecting an explicitly adversarial workflow. This structured conflict, inspired by dialectical processes and peer review, is specifically designed to uncover and mitigate the biases of the initial agent, a goal distinct from general task completion. Our approach is philosophically aligned with recent work advocating for debate-based mechanisms to improve LLM reasoning and robustness [14].

III. Theoretical Framework

The foundational hypothesis of MACS is that a more accurate and fair assessment can be achieved by simulating cognitive diversity. We define this as the process of leveraging multiple, independent, and architecturally distinct computational agents to analyze a problem from different perspectives.

A. Adversarial Peer Review Simulation

Unlike simple ensemble methods that treat agent outputs as independent votes, MACS structures their interaction. The workflow (Initial Review $\rightarrow$ Peer Challenge $\rightarrow$ Final Consensus) mimics scholarly peer review. The "Peer Challenge" stage is explicitly adversarial; agents are prompted to find flaws in the initial assessment. This creates a constructive tension that exposes potential weaknesses (e.g., hallucinations, missed criteria, biases) in the initial review. This approach is conceptually related to adversarial training in machine learning, where models are made more robust by being exposed to inputs designed to mislead them [8]. The final arbiter's role is not to average, but to synthesize, weighing the arguments presented by the initial and peer reviewers against the ground truth of the scoring rubric.

B. Disagreement-Resolution Ratio (DRR)

To quantify the impact of this structured process, we introduce the Disagreement-Resolution Ratio (DRR). The DRR measures the extent to which the final consensus score deviates from the initial score, normalized by the magnitude of the peer reviewers' challenge. For a given scoring section j, the DRR is defined as:

$$\text{DRR}_j = \frac{|S_{final,j} - S_{init,j}|}{\frac{1}{n}\sum_{i=1}^{n} |S_{peer_i,j} - S_{init,j}| + \epsilon} \tag{1}$$

where $S_{init,j}$ is the initial score, $S_{peer_i,j}$ is the score from the i-th peer reviewer, $S_{final,j}$ is the final consensus score, n is the number of peer reviewers, and ϵ is a small constant to prevent division by zero.

A DRR value near 1.0 indicates that the final arbiter was significantly swayed by the peer reviewers, suggesting a substantial correction was made. A value near 0 indicates the initial review was upheld despite challenges. Analyzing the DRR across many assessments provides a powerful diagnostic for understanding the system's self-correction capabilities.

C. Connection to Explainable AI (XAI)

A critical benefit of the MACS framework is its inherent transparency. By recording the evaluation, justification, challenge, and final resolution from each agent, the system produces a comprehensive audit trail for every score. This directly addresses the "black box" problem in many AI systems and aligns with the principles of Explainable AI (XAI), which has become a critical requirement for deploying AI in high-stakes domains like education [9]. The multi-faceted output allows educators to understand why a certain score was given, building trust and enabling meaningful human oversight.

IV. System Architecture

The MACS framework (Fig. 1) is a modular pipeline composed of three core modules designed to implement the cognitive diversity simulation.

A. Multimodal Text Extraction Module

Assessment fidelity begins with data fidelity. Our system employs a hybrid extraction strategy to handle the structural complexity of academic PDFs. A Vision Language Model (VLM), `google/gemma-3-12b-it`, provides a rich interpretation of text, tables, and figures, while a conventional parser, `PyPDF2`, serves as a robust fallback. The final text, T_{final}, is selected via a content completeness heuristic:

$$T_{final} = \begin{cases} T_{vlm} & \text{if } \frac{|T_{vlm}|}{|T_{pdf}|} \geq \theta \\ T_{pdf} & \text{otherwise} \end{cases} \tag{2}$$

where $|T_{vlm}|$ and $|T_{pdf}|$ are character counts and the threshold θ is set to 0.8.

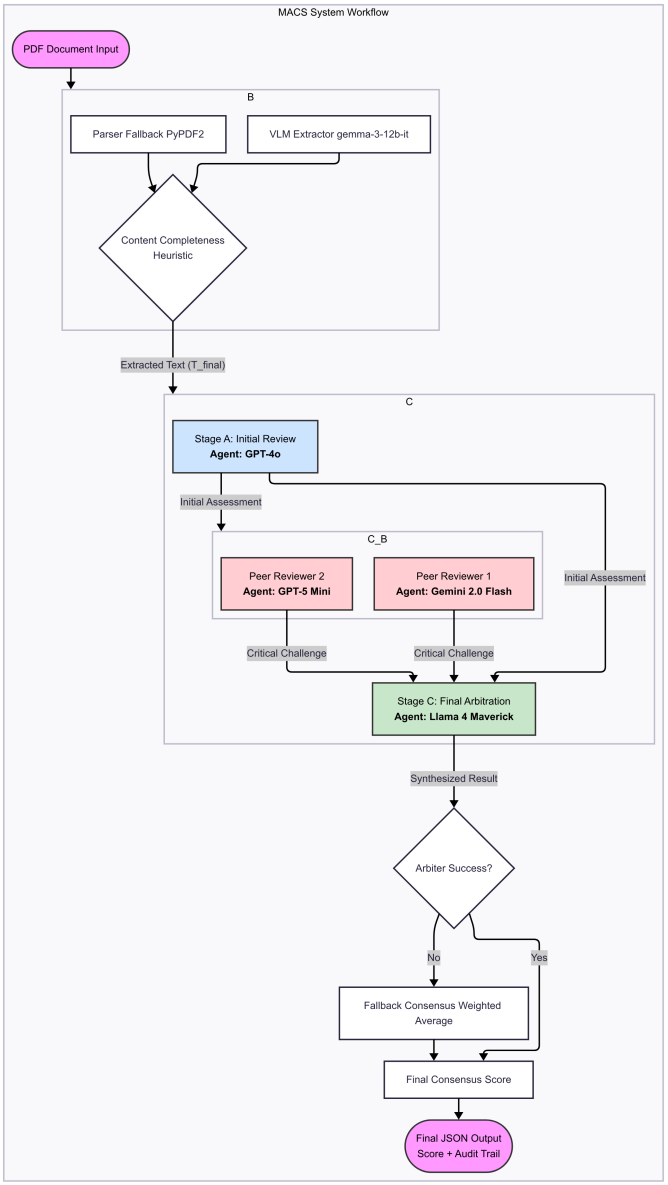

Fig. 1. Architecture of the Multi-Agent Consensus System (MACS), illustrating the flow from multimodal extraction through the adversarial peer review stages to final consensus.

B. Multi-Agent Peer Review Module

This module is the core of our framework, operationalizing the peer review with a heterogeneous set of LLM agents (Table I). The selection of models from different developers (OpenAI, Google, Meta) is intentional, maximizing architectural diversity to foster more genuine cognitive diversity as highlighted in recent MAS surveys [10].

TABLE I
AI AGENT CONFIGURATION IN THE PEER REVIEW WORKFLOW

Role	Model	Provider	Function within Framework
Initial Reviewer	GPT-4o	OpenAI	Establishes a comprehensive baseline assessment.
Peer Reviewer 1	Gemini 2.0 Flash	Google	Adversarial challenge: identifies flaws in the baseline.
Peer Reviewer 2	GPT-5 Mini	OpenAI	Alternative perspective: seeks overlooked aspects.
Final Arbiter	Llama 4 Maverick	Meta	Synthesis: resolves conflicts and forms final judgment.

C. Fallback Consensus Mechanism

In cases where the final arbiter agent fails, system resilience is maintained by a fallback mechanism. The final score S_{final} is calculated as a weighted average:

$$S_{final} = \alpha \cdot S_{init} + (1 - \alpha) \cdot \frac{1}{n} \sum_{i=1}^{n} S_{peer_i} \tag{3}$$

where weights are empirically set to give significant, but not overriding, influence to the peer challengers ($\alpha = 0.4$).

V. IMPLEMENTATION DETAILS

A. Peer Review Protocol and Data Structuring

Guideline compliance and data integrity are enforced via structured JSON outputs. All prompts are engineered to instruct agents to return assessments in a predefined schema, utilizing the API's JSON mode to guarantee validity.

Input: Report text T_{final}, Guidelines G
Output: Structured JSON Assessment
for each section $s \in$ SCORE_SECTIONS **do**
  Elicit score $S_s \in [0, 100]$ from agent
  Elicit justification J_s with citations from G
end for
return JSON($\{S_s, J_s\}$)

B. Robustness and Error Handling

The system is built for reliable batch processing with:

- A 3-retry strategy with exponential backoff for all API calls.
- Score validation checks to ensure outputs are within the valid [0, 100] range.
- The fallback consensus mechanism described in Section IV-C.
- Preservation of partial scores if an agent fails on a subset of sections.

VI. Experimental Evaluation

A. Dataset and Methodology

This study was evaluated using data from the Datavidia 9.0 data science competition, a national event held by the Student Association of Informatics Engineering (HMIF) at the Bandung Institute of Technology (ITB). The dataset consists of scores for 48 submissions, each evaluated by the MACS system and a panel of three independent human experts. The average of the three human scores serves as the expert consensus benchmark. While this study uses a focused dataset from a single competition as a case study, it provides a robust proof-of-concept.

The core of our evaluation reframes the concept of AI accuracy. Instead of solely measuring the difference between MACS and the expert consensus, we contextualize its performance by comparing it against the inherent variability found among the human experts themselves. We define a "significant disagreement" as a scoring difference greater than 5 points on a 100-point scale. We then measure the frequency of such disagreements in two scenarios:

1) **MACS vs. Human Consensus:** The absolute difference between the MACS score and the average human score.
2) **Human vs. Human:** The maximum absolute difference between any two human judges for the same paper.

B. Results: MACS Demonstrates Higher Consistency

Our analysis reveals that significant disagreement is the norm, not the exception, for human evaluators. In a striking **97.92%** of cases, the maximum difference between any two human judges was greater than 5 points. This establishes a high baseline for acceptable scoring variance in a real-world evaluation setting.

In contrast, the MACS system demonstrated a much higher degree of consistency with the expert consensus. The MACS score differed from the human average by more than 5 points in only **41.67%** of cases. This indicates that for a majority of submissions (58.33%), the MACS assessment was in close alignment with the final human consensus.

The bar chart in Fig. 2 provides a stark visual summary of this finding. The frequency of significant disagreement among human judges is more than double that of the MACS system, suggesting that MACS's alignment with the expert consensus is more reliable than the alignment of any individual expert with their peers.

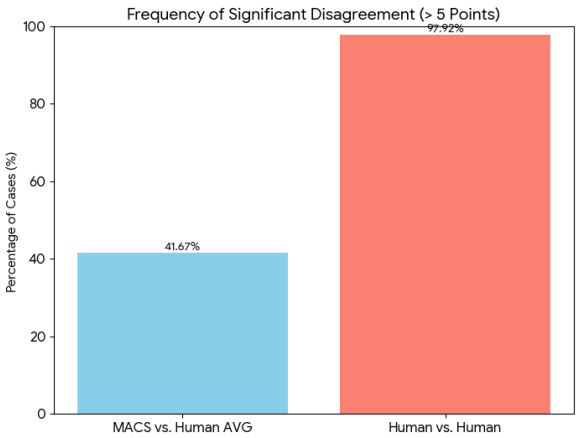

Fig. 2. Frequency of Significant Disagreement (>5 Points). The chart shows that the rate of major disagreement between human judges is over double the rate of disagreement between MACS and the human average, highlighting MACS's superior consistency.

VII. Discussion and Comparative Analysis

To delve deeper than frequency, we analyzed the magnitude and distribution of these disagreements. The results further reinforce the conclusion that MACS is a more consistent evaluation tool than individual human experts.

TABLE II
Statistical Comparison of Score Disagreements

Metric	**MACS vs. Human AVG**	**Human vs. Human**
Mean Absolute Difference	5.14	12.01
Std. Dev. of Difference	4.09	5.48
Median Absolute Difference	4.10	10.95
Max Observed Difference	18.90	29.50
% Cases > 5 Point Diff.	41.67%	97.92%

The statistical summary in Table II is compelling. The mean absolute difference between human judges (12.01 points) is more than twice that of MACS versus the human average (5.14 points). This indicates that, on average, the MACS score is significantly closer to the final consensus.

The comparative boxplot in Fig. 3 visualizes these distributions. The plot for "MACS vs. Human AVG" is lower and more compact, indicating not only a smaller median disagreement but also less overall variance. Conversely, the "Human vs. Human" plot is elevated and stretched, showing that large disagreements are a common and expected feature of the human evaluation process.

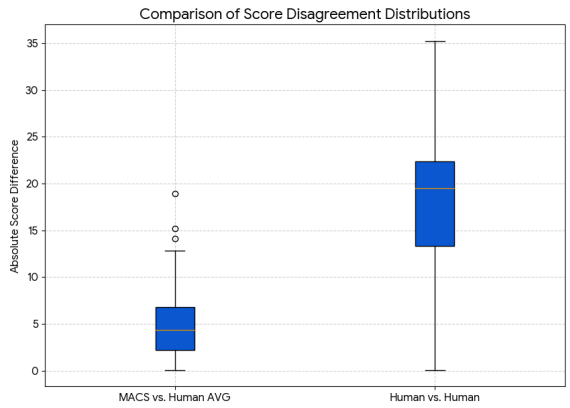

Fig. 3. Comparison of Score Disagreement Distributions. The boxplot for MACS vs. Human AVG is significantly lower and more compact than that for Human vs. Human, indicating a smaller median and less variance in scoring differences.

A. MACS in Context: Beyond Simple Ensembles and Adversarial Training

It is crucial to differentiate MACS from two related concepts: simple ensemble averaging and traditional adversarial training.

Ensemble averaging, as explored in recent studies [12], treats each model as an independent "voter." While this can reduce random noise, it is ill-equipped to handle systematic bias. If multiple models in the ensemble share a common bias learned from similar web-scale datasets, averaging their outputs will reinforce, not correct, that bias. MACS fundamentally differs by structuring the interaction. The "Peer Challenge" stage is not a vote; it is a targeted critique designed to surface specific flaws in reasoning. The final arbiter's synthesis is a qualitative judgment based on evidence, not a quantitative calculation.

Adversarial training in machine learning typically involves perturbing input data at a feature level to make a model more robust to noisy or malicious inputs [8]. MACS operates on a higher, conceptual level. The "adversary" is another LLM agent that challenges the reasoning and justification of the initial review, not the input data itself. This semantic-level adversarial process forces the system to defend its conclusions and builds a more robust and transparent rationale, aligning with the core tenets of Explainable AI (XAI) [9].

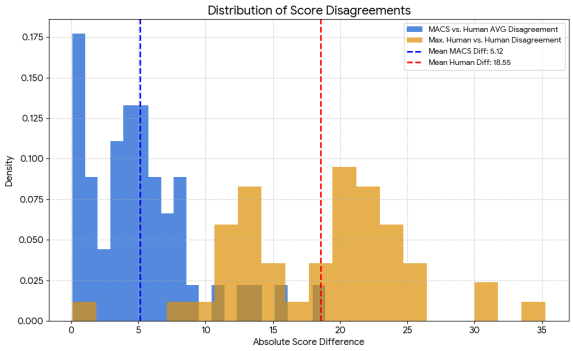

Fig. 4. Distribution of Score Disagreements. The MACS vs. Human AVG disagreements (blue) are concentrated at lower values, while the Human vs. Human disagreements (orange) are widely distributed with a much higher mean.

The crucial insight from this analysis is that human evaluation, while the "gold standard," is inherently variable. By establishing this real-world variance as a benchmark, we can conclude that MACS operates well within the bounds of acceptable accuracy. Its ability to consistently produce a score closer to the expert consensus makes it a powerful tool for stabilizing the evaluation process and reducing the impact of individual human bias.

VIII. Conclusion

This paper introduced the Multi-Agent Consensus System (MACS), a framework that simulates cognitive diversity to mitigate bias in automated academic assessment. By structuring the interaction of heterogeneous LLMs into an adversarial peer-review process, MACS moves beyond simple automation to create a more robust, reliable, and fair evaluation system.

Our results, derived from a real-world competitive setting, empirically validate that this structured approach is superior when benchmarked against the natural variability of human experts. MACS's disagreements with the expert consensus are demonstrably smaller in magnitude and less frequent than the disagreements observed among the human experts themselves.

Future research will focus on several key areas. First, we will validate the MACS framework across multiple academic domains and larger datasets to establish its generalizability. Second, we will explore implementing a dynamic, multi-turn debate among agents, a direction supported by recent advances in collaborative MAS frameworks [11]. Finally, we plan to develop fine-tuned, domain-specific models for each role to further enhance expertise. MACS represents a critical step towards creating AI systems for education that are not only efficient but also fundamentally more robust, transparent, and equitable.

Acknowledgment

The authors would like to acknowledge the support of the faculty and students who participated in the pilot deployment of this system.

REFERENCES

[1] D. B. McNally and I. M. Dounas, "Inter-rater reliability and the effects of grader fatigue in manual grading," in *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, 2011, pp. 223-230.

[2] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V.2," *The Journal of Technology, Learning and Assessment*, vol. 4, no. 3, 2006.

[3] T. Kocielnik, R. Singh, and P. Resnik, "Can GPT-4 grade your essays? A large-scale evaluation of AI-powered grading," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, no. 1, pp. 120-128.

[4] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, p. 671, 2016.

[5] A. Gupta, A. Kumar, and S. Kumar, "LLMs are Biased Teachers: Evaluating LLM Bias in Personalized Education," *arXiv preprint arXiv:2410.14012*, 2024.

[6] I. Birru et al., "Mitigating Age-Related Bias in Large Language Models: Strategies for Responsible Artificial Intelligence Development," *INFORMS Journal on Computing*, 2024.

[7] Q. Wang et al., "MegaAgent: A Large-Scale Autonomous LLM-based Multi-Agent System Without Predefined SOPs," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 4998–5036.

[8] National Institute of Standards and Technology, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," *NIST AI 100-2e2025*, March 2025.

[9] A. Rai, "Explainable AI: From black box to glass box," *Journal of the Academy of Marketing Science*, vol. 48, pp. 137-141, 2020.

[10] H. Li et al., "A Survey on LLM-based Multi-Agent System: Recent Advances and New Frontiers in Application," *arXiv preprint arXiv:2412.17481*, 2024.

[11] Y. Zhou et al., "MultiAgentBench: Evaluating the Collaboration and Competition of LLM agents," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025, pp. 8580–8622.

[12] A. Thompson, et al., "Averaging, Voting, and Stacking: A Comparative Study of Ensemble Methods for LLM-based Grading," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 345-356.

[13] Q. Wu, et al., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework," *arXiv preprint arXiv:2308.08155*, 2023.

[14] P. Liang, et al., "Dialectical Self-Play: A Framework for Improving LLM Reasoning through Structured Debate," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[15] M. Sharma, and S. Singh, "The Algorithmic Monoculture in AI-Driven Education: Risks and Mitigation Strategies," *AI & Society*, vol. 39, pp. 1-13, 2025.