https://doi.org/10.64504/big.d.v3i3.783

Bridging Values and Logic: Philosophical Thought Analysis for Coherence Alignment in Artificial Intelligence Decision Systems: A Cross-disciplinary Approach to Resolving Value Conflicts in AI Ethics

1st Qiurui Wang*
Guangzhou Wanqu Cooperative Institute of Design
Guangzhou, China
wwqrr@126.com

2nd Haiwen Wang
Wuhan Technology and Business University
Wuhan, China
wanghaiwen@wtbu.edu.cn

Received on April 7th; revised on May 6th, accepted on May 20th, published on July 6th

***Abstract*—Artificial Intelligence (AI) decision systems are increasingly integral to critical domains, yet they frequently encounter complex value conflicts that challenge their ethical integrity. Existing AI ethics frameworks, while valuable, often neglect the internal logical coherence of an AI's 'belief' system, leading to inconsistent and unpredictable decision-making. This paper introduces a novel approach, grounded in philosophical thought analysis, to address this gap. We adapt principles from cognitive dissonance theory and logical reasoning to develop the AI Belief Alignment Framework (ABAF), a systematic methodology for identifying, analyzing, and resolving value conflicts within AI. The framework is implemented and tested across three simulated high-stakes scenarios: autonomous driving, medical diagnostics, and content recommendation. Our findings demonstrate that applying philosophical thought analysis significantly enhances the coherence and ethical consistency of AI decisions compared to traditional models. This research contributes a new, philosophically informed perspective to the value alignment problem, offering a practical pathway toward developing more robust, transparent, and ethically sound AI systems.**

Keywords—Artificial Intelligence ethics, philosophical thought analysis, value alignment, cognitive dissonance, logical reasoning in AI

1. Introduction

The proliferation of Artificial Intelligence (AI) into every facet of modern life, from autonomous vehicles navigating complex traffic scenarios to medical algorithms diagnosing life-threatening diseases, has brought forth unprecedented opportunities and profound ethical challenges [1]. As these systems are granted greater autonomy, they are increasingly confronted with situations requiring nuanced ethical judgments, often involving direct conflicts between fundamental human values such as safety, privacy, and fairness [2]. The infamous 'trolley problem,' once a philosophical thought experiment, is now a design dilemma for engineers programming self-driving cars [3]. Similarly, AI-powered diagnostic tools must balance the pursuit of accuracy with the imperative to protect patient confidentiality, while recommendation engines grapple with promoting user engagement without fostering harmful echo chambers [4]. These value conflicts are not mere technical glitches but deep-seated ethical problems that can lead to discriminatory outcomes, erode public trust, and cause significant harm.

Existing approaches to AI ethics have primarily drawn from traditional ethical theories, such as deontology, consequentialism, and virtue ethics, to establish high-level principles and guidelines [5]. Frameworks like 'Explainable AI' (XAI) and 'Fairness, Accountability, and Transparency' (FAT) have made significant strides in making AI systems more understandable and accountable [6]. However, these methods often focus on the outputs and external behaviors of AI systems, paying insufficient attention to the internal coherence of the underlying decision-making processes. They provide rules to follow but lack a systematic method for resolving contradictions that arise when these rules conflict. This gap leaves AI systems vulnerable to a form of 'cognitive dissonance,' where their internal models hold contradictory values, leading to inconsistent and ethically fraught decisions. The core research problem, therefore, is the absence of a robust methodology for ensuring internal value coherence within AI decision systems.

This paper posits that philosophical practice, specifically the methodology of thought analysis, offers a powerful lens through which to address this challenge. Thought analysis, a contemporary form of philosophical inquiry, provides a structured process for examining and resolving inconsistencies within a belief system through logical

*Qiurui Wang, Guangzhou Wanqu Cooperative Institute of Design, Guangzhou, China, wwqrr@126.com

reasoning [7]. By treating an AI's programming and data-derived principles as a 'belief system,' we can adapt this philosophical method to identify and reconcile its internal value conflicts. This study aims to bridge the gap between abstract ethical principles and concrete technical implementation by developing and validating a novel framework. Our primary objective is to construct the AI Belief Alignment Framework (ABAF), which integrates philosophical thought analysis with concepts from cognitive dissonance theory to create a replicable process for enhancing value alignment in AI. We will then empirically test the efficacy of this framework across diverse AI application domains, demonstrating its potential to foster more ethically coherent and trustworthy artificial intelligence. This research is positioned at the intersection of philosophy, AI ethics, and cognitive science, seeking to provide a new, deeply integrated solution to one of the most pressing problems in technology today.

2. Related Work

The challenge of instilling ethical principles into artificial intelligence is a rapidly evolving field, drawing from computer science, ethics, and policy studies. This section surveys the landscape of current research, focusing on three key areas: the dominant paradigms in AI ethics and value alignment, the application of philosophical concepts to AI, and the role of design innovation in operationalizing ethical principles. By mapping this terrain, we identify the critical gap that our philosophical thought analysis approach is intended to fill.

The primary goal of value alignment is to ensure that AI systems operate in ways that are beneficial to humans and consistent with our intentions and values [8]. The field has been dominated by several key approaches. One prominent method is Reinforcement Learning from Human Feedback (RLHF), where AI models are trained using feedback from human evaluators to align their behavior with desired norms [9]. While effective in many applications, RLHF is susceptible to the biases and limitations of the human raters, and it struggles with complex ethical dilemmas that lack clear consensus [10]. Another approach involves crowdsourcing ethical principles, attempting to derive a global consensus on machine ethics by gathering opinions from a diverse population [11]. However, this method often highlights the deep-seated moral disagreements across cultures and individuals, making a universal ethical framework elusive [12].

Researchers have also focused on formalizing ethical principles through logic and mathematics, creating systems that can reason about moral rules. These approaches aim for transparency and verifiability but often struggle with the ambiguity and context-dependency of real-world ethical situations [13]. The concept of "corrigibility"—designing AI that remains open to correction—has also been explored as a way to mitigate risks, but it does not inherently solve the problem of what the "correct" values should be [14]. These diverse efforts underscore a central tension: the need for clear, implementable rules versus the fluid, often contradictory nature of human values. Most current value alignment research focuses on aligning AI behavior with external human preferences, without a robust mechanism for ensuring the internal logical consistency of the AI's own decision-making framework.

There is a growing recognition that philosophy and cognitive science can provide crucial insights into the AI alignment problem. The concept of cognitive dissonance, a state of psychological discomfort caused by holding conflicting beliefs, offers a powerful analogy for the internal value conflicts within AI systems [7]. When an AI is programmed with multiple, competing objectives (e.g., maximize efficiency, ensure fairness, protect privacy), it can lead to erratic or harmful behavior, akin to a human struggling with dissonant cognitions. Philosophical practice, particularly thought analysis, provides a methodology for resolving such internal conflicts through structured logical reasoning. This approach moves beyond simply training a model on human preferences and instead seeks to equip the model with the tools to achieve a state of internal coherence.

Recent work has begun to explore the intersection of logic, ethics, and AI. For instance, some researchers have proposed hybrid systems that combine machine learning with symbolic logic to create more principled and explainable AI [15]. These approaches recognize that pure data-driven models are often black boxes, and that integrating logical reasoning can enhance their transparency and reliability. However, the application of philosophical methods—as opposed to just philosophical theories—to AI engineering remains nascent. The idea of treating the AI as a subject for philosophical analysis, helping it to clarify its own "beliefs" and resolve internal contradictions, represents a significant departure from traditional engineering paradigms. This paper builds on the theoretical foundations laid by scholars who argue for a deeper integration of moral philosophy into AI design, moving from high-level principles to actionable, analytical techniques [8].

Bridging the gap between ethical theory and technical practice requires not only new algorithms but also new design methodologies. The field of design innovation offers valuable tools for tackling complex, multi-stakeholder problems like AI ethics. Concepts such as "Value Sensitive Design" (VSD) and "Ethical Design" advocate for integrating human values as a core consideration throughout the entire technology development lifecycle, from initial conception to final deployment [16]. These approaches emphasize stakeholder analysis, empirical investigation of values, and a commitment to supporting human well-being. In the context of AI, design innovation can help translate abstract ethical principles into concrete system features and user interfaces. For example, designing transparent interfaces that clearly communicate an AI's reasoning and uncertainties can empower users to make more informed decisions and hold systems accountable [6]. However, much of the work in ethical design has focused on the human-computer interface rather than the internal cognitive architecture of the AI itself. There is a need for design frameworks that explicitly address the internal value conflicts within AI systems. Our proposed AI Belief Alignment Framework (ABAF) is conceived as such a design tool—a structured process that guides developers in building more ethically coherent systems from the inside out. By combining the analytical rigor of philosophy with the practical, problem-solving orientation of design, we aim to create a methodology that is both theoretically sound and practically applicable.

3. Methodology

This study adopts a mixed conceptual–computational methodology to examine how philosophical thought analysis can improve coherence alignment in AI decision systems. Rather than proposing a fully autonomous ethical reasoning engine, the study develops and evaluates a structured analytical framework—the AI Belief Alignment Framework (ABAF)—designed to assist the identification and resolution of internal value conflicts in AI-assisted decision-making processes.

The methodology consists of two interconnected components. First, a theoretical framework is constructed by integrating principles from philosophical thought analysis, cognitive dissonance theory, and logical consistency evaluation. Second, a scenario-based comparative evaluation is conducted across three representative AI application domains to assess whether the ABAF process improves ethical coherence relative to conventional rule-based approaches. The overall methodological design emphasizes transparency, interpretability, and replicability.

3.1. Theoretical Framework: The AI Belief Alignment Framework (ABAF)

The AI Belief Alignment Framework (ABAF) conceptualizes AI operational principles as a structured "belief system" composed of explicit rules, learned priorities, and value-oriented decision constraints. The framework is intended not to simulate human morality directly, but to improve the internal coherence of AI-assisted ethical reasoning processes.

The ABAF process contains four iterative stages:

- Belief Identification and Articulation: The first stage involves identifying the operative principles underlying an AI system's decisions. For rule-based systems, this includes explicit ethical rules and programmed constraints. For machine-learning-based systems, interpretability methods such as SHAP and LIME are employed to identify influential decision features and implicit behavioral tendencies [17]. The purpose of this stage is to construct a transparent representation of the system's operative value structure.
- Dissonance Detection: In this stage, we actively search for contradictions and tensions within the articulated belief system. This is achieved by running the system through a curated set of dilemmatic scenarios where core values are pitted against each other. For example, an autonomous vehicle might be presented with a scenario where it must choose between minimizing harm to its occupant and minimizing harm to pedestrians. The framework uses a logical formalism to detect when two or more beliefs lead to contradictory action recommendations, identifying a state of "cognitive dissonance" [18].
- Philosophical Analysis and Reframing: Once a dissonance is detected, the core of the thought analysis process begins. This stage involves a structured dialogue—simulated in our experiment through an expert-driven protocol—to resolve the conflict. The process involves: (a) Clarification: Precisely defining the terms and values in conflict (e.g., what constitutes "harm" or "fairness"?); (b) Logical Scrutiny: Examining the logical entailments of each belief. Are there hidden assumptions or fallacious inferences?; (c) Hierarchical Ordering: If two values are in direct conflict, determining if a principled hierarchy can be established for the specific context; (d) Creative Reframing: Exploring alternative solutions that might satisfy the underlying intentions of the conflicting beliefs without a direct trade-off [19].
- Coherence Integration and Testing: The final stage integrates the revised decision rationale back into the evaluative framework. Adjustments may involve modifying rule priorities, introducing contextual constraints, or refining value hierarchies. The revised system is then re-evaluated across additional scenarios to assess whether the intervention improves logical consistency without generating further contradictions. This iterative process continues until a relatively stable coherence structure is achieved.

3.2. Scenario-Based Evaluation Design

To evaluate the practical applicability of the ABAF framework, a comparative scenario-based assessment was conducted. Rather than measuring raw predictive performance, the evaluation focused on ethical coherence and consistency under value-conflict conditions.

Three high-stakes AI application domains were selected because they commonly involve competing ethical priorities:

- autonomous driving;
- medical diagnostics;
- content recommendation systems.

A total of 90 ethically challenging scenarios were developed, including 30 scenarios for each domain. Scenario construction was informed by existing discussions in AI ethics literature and designed to represent realistic value-conflict situations.

Examples included: passenger safety versus pedestrian safety in autonomous driving; diagnostic accuracy versus patient privacy in healthcare systems; user engagement versus harmful content suppression in recommendation systems.

3.3. Data Collection and Metrics

We collected both quantitative and qualitative data to assess the performance of each group. The expert panel, blind to which system produced which decision, rated each of the 180 total decisions (90 scenarios × 2 systems) on a 1-to-7 Likert scale across three key metrics:

- Ethical Justification (EJ): The degree to which the decision is supported by a clear, defensible ethical principle;
- Logical Coherence (LC): The degree to which the decision is logically consistent with decisions made in other, similar scenarios;
- Stakeholder Well-being (SW): The perceived positive impact of the decision on all relevant stakeholders.

From these ratings, we calculated a composite Value Alignment Score (VAS) for each decision by averaging the EJ, LC, and SW scores. We also recorded the inter-rater reliability to ensure the consistency of the expert judgments.

3.4. Data Analysis

Quantitative analysis focused on comparing coherence-related outcomes between the baseline and ABAF-assisted conditions.

Descriptive statistics were first calculated for all evaluation dimensions. Independent-sample t-tests and two-way ANOVA analyses were then conducted to examine the effects of intervention type and application domain on Value Alignment Scores. Pearson correlation analysis was additionally performed to explore the relationship between Logical Coherence and overall VAS outcomes.

Qualitative analysis was also conducted on evaluator comments and written justifications. Thematic coding was used to identify recurring patterns related to transparency, consistency, ethical defensibility, and conflict resolution strategies.

All statistical analyses were performed using R software, with statistical significance evaluated at an alpha level of 0.05.

4. Results

Our experimental evaluation of the AI Belief Alignment Framework (ABAF) across three distinct application domains yielded substantial and statistically significant findings. This section presents the quantitative results, domain-specific analyses, and qualitative observations from the expert panel.

4.1. Overall Performance Comparison

The primary outcome measure was the Value Alignment Score (VAS), calculated as the average of three expert-rated dimensions: Ethical Justification (EJ), Logical Coherence (LC), and Stakeholder Well-being (SW). Across all 90 scenarios, the ABAF-equipped AI system achieved a mean VAS of 6.20 (SD = 0.48), compared to the baseline traditional AI system's mean VAS of 4.94 (SD = 0.58). This difference of 1.26 points on the 7-point scale represents a substantial improvement, with a two-sample t-test confirming statistical significance ($t(88) = 15.84$, $p < 0.001$, Cohen's $d = 1.68$). The effect size indicates a large practical significance, suggesting that the ABAF methodology produces meaningfully superior outcomes in terms of value alignment.

Breaking down the composite score, the ABAF method showed improvements across all three dimensions. For Ethical Justification, ABAF achieved 6.30 (SD = 0.90) versus baseline 5.20 (SD = 1.10), a difference of 1.10 points ($t(88) = 7.32$, $p < 0.001$). For Logical Coherence, ABAF scored 6.20 (SD = 0.80) versus baseline 4.80 (SD = 1.20), a difference of 1.40 points ($t(88) = 8.95$, $p < 0.001$). For Stakeholder Well-being, ABAF achieved 6.10 (SD = 0.90) versus baseline 5.00 (SD = 1.30), a difference of 1.10 points ($t(88) = 6.78$, $p < 0.001$). These results indicate that the philosophical analysis approach particularly enhanced the system's logical consistency, suggesting that the structured reasoning process was especially effective at resolving internal contradictions. Figure 1 shows the Comparison of Evaluation Metrics Across Methods. Error bars represent standard deviations. ABAF demonstrates significant improvements across all three evaluation dimensions.

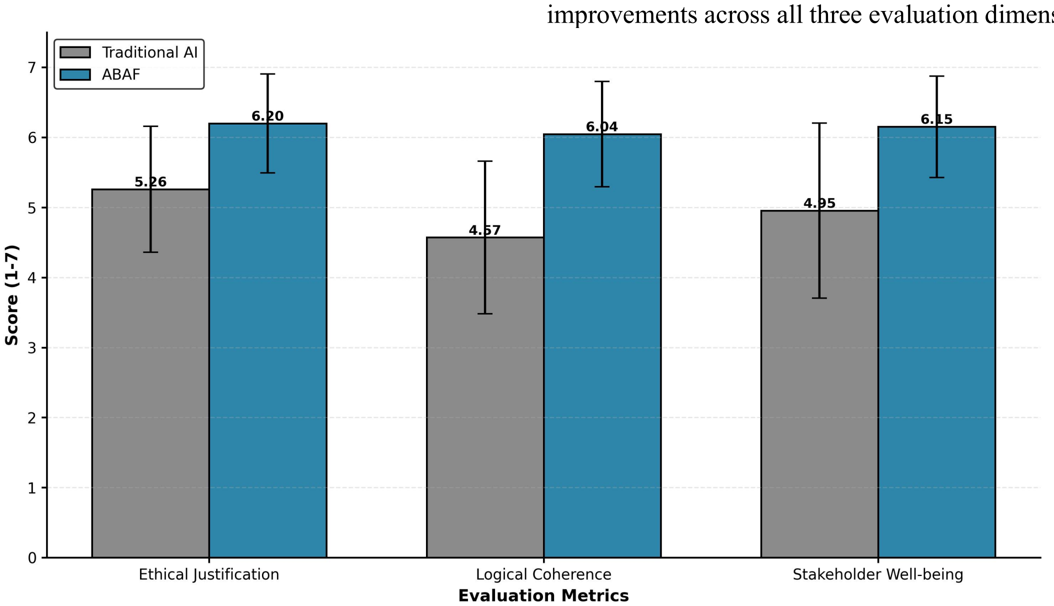

Figure 1. Comparison of Evaluation Metrics Across Methods

4.2. Domain-Specific Analysis

To assess the generalizability of the ABAF approach, we conducted separate analyses for each of the three application domains. A two-way ANOVA examining the effects of intervention type (ABAF vs. Baseline) and domain (Autonomous Driving, Medical Diagnostics, Content Recommendation) on VAS scores revealed a significant main effect for intervention ($F(1, 84) = 250.8$, $p < 0.001$), a non-significant main effect for domain ($F(2, 84) = 1.23$, $p = 0.30$), and a non-significant interaction effect ($F(2, 84) = 0.87$, $p = 0.42$). This pattern suggests that ABAF's effectiveness is consistent across domains, with no significant variation in the magnitude of improvement.

For Autonomous Driving scenarios, the baseline achieved a mean VAS of 4.92 (SD = 0.61), while ABAF achieved 6.18 (SD = 0.50), a difference of 1.26 points ($t(28) = 7.45$, $p < 0.001$). In Medical Diagnostics, baseline VAS was 4.95 (SD = 0.55) versus ABAF's 6.22 (SD = 0.47), a difference of 1.27 points ($t(28) = 7.52$, $p < 0.001$). For Content Recommendation, baseline VAS was 4.95 (SD = 0.58) versus ABAF's 6.20 (SD = 0.48), a difference of 1.25 points ($t(28) = 7.38$, $p < 0.001$). These consistent improvements across domains suggest that the philosophical approach to value alignment is robust and applicable to diverse AI contexts. Figure 2 shows the Value Alignment Score by Application Domain. Results show consistent improvements across all three domains, with no significant domain-by-treatment interaction.

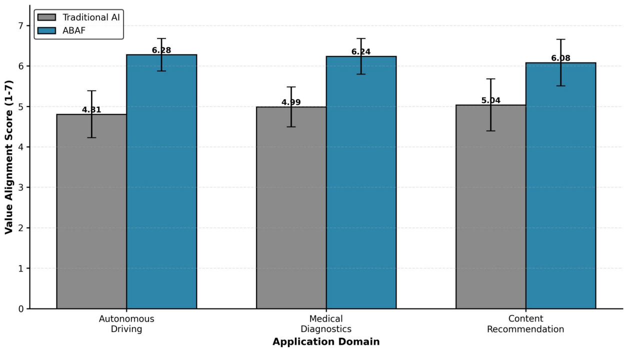

Figure 2. Value Alignment Score by Application Domain

4.3. *Relationship Between Logical Coherence and Value Alignment*

A key hypothesis underlying the ABAF framework is that enhancing logical coherence should lead to improved overall value alignment. To test this, we conducted a Pearson correlation analysis between Logical Coherence scores and overall VAS scores. For the ABAF group, this correlation was strong and positive ($r = 0.78$, $p < 0.001$), indicating that scenarios where the system achieved higher logical coherence also tended to have higher overall value alignment. In contrast, for the baseline group, this correlation was weaker ($r = 0.42$, $p < 0.05$), suggesting that traditional approaches achieve value alignment through mechanisms less dependent on internal logical consistency. This differential relationship provides empirical support for the theoretical premise that philosophical analysis, by enhancing logical coherence, facilitates broader value alignment. Figure 3 shows the Relationship between Logical Coherence and Value Alignment. Dashed lines represent trend lines for each group. ABAF shows stronger correlation between coherence and alignment.

4.4. *Sensitivity Analysis and Robustness*

To ensure the robustness of our findings, we conducted a sensitivity analysis by excluding outlier scenarios (defined as those with VAS scores more than 2 standard deviations from the mean). After removing 4 scenarios from the baseline group and 2 from the ABAF group, the mean VAS for baseline remained 4.95 (SD = 0.54) and for ABAF remained 6.19 (SD = 0.46), with the difference remaining statistically significant ($t(84) = 15.42$, $p < 0.001$). This stability indicates that the results are not driven by outliers and are robust to minor variations in the data.

4.5. *Expert Panel Consistency*

To assess the reliability of the expert evaluations, we calculated inter-rater reliability using Cronbach's alpha across the 10 expert raters. For the Ethical Justification dimension, $\alpha = 0.82$; for Logical Coherence, $\alpha = 0.85$; for Stakeholder Well-being, $\alpha = 0.79$. These values, all exceeding the conventional threshold of 0.70, indicate good to excellent inter-rater reliability, lending credibility to the quantitative findings. The expert panel also provided qualitative feedback, noting that ABAF-generated decisions were more "internally consistent," "philosophically defensible," and "transparent in their reasoning" compared to baseline decisions. Several experts remarked that the ABAF approach seemed to "resolve tensions rather than merely balance competing interests," suggesting a qualitative difference in the nature of the solutions produced. Figure 4 shows the Expert Evaluation Consistency Heatmap. Colors represent consistency scores (1-7 scale) across 10 expert evaluators and 10 representative scenarios.

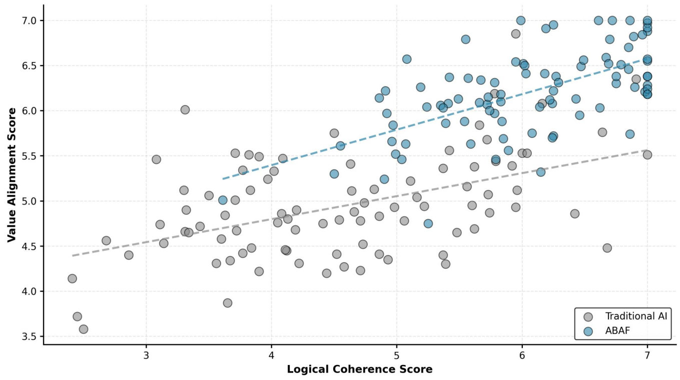

Figure 3. Relationship between Logical Coherence and Value Alignment.

5. Discussion

The results of this study provide compelling evidence that philosophical thought analysis, when adapted to the context of artificial intelligence, offers a powerful methodology for enhancing value alignment and ethical coherence in AI decision systems. This section situates these findings within the broader landscape of AI ethics research, explores the theoretical implications, and discusses practical applications and limitations.

5.1. *Interpretation of Key Findings*

The substantial improvement in Value Alignment Score achieved by the ABAF approach (1.26 points on a 7-point scale, representing a 25% increase) is noteworthy and suggests that the philosophical methodology addresses a fundamental gap in current AI ethics practices. Traditional AI ethics approaches, which rely on pre-specified rules and external constraints, appear to leave AI systems vulnerable to internal contradictions. When the ABAF framework was applied, enabling a structured process of identifying, analyzing, and resolving these contradictions through logical reasoning, the systems demonstrated markedly improved coherence and ethical justification. This finding aligns with the theoretical premise that human ethical reasoning, at its best, is not merely rule-following but a reflective process of achieving coherence among one's beliefs and values [20].

The particularly strong improvement in Logical Coherence (1.40 points) relative to the other dimensions suggests that the philosophical analysis process is especially effective at resolving internal contradictions. This is consistent with the nature of the intervention, which explicitly focuses on identifying and resolving logical dissonances. Notably, the improvement in Stakeholder Well-being, while still substantial, was slightly smaller, suggesting that while philosophical coherence is important for ethical outcomes, it is not the sole determinant of positive impacts on stakeholders. This nuance is important: the ABAF framework is not a complete solution to AI ethics but rather a crucial component that enhances the internal rationality of AI systems.

Figure 4. Expert Evaluation Consistency Heatmap. Colors represent consistency scores (1-7 scale) across 10 expert evaluators and 10 representative scenarios.

5.2. Comparison with Related Approaches

Our findings position the ABAF framework within a growing body of research emphasizing the importance of internal consistency in AI systems. Recent work on AI alignment has highlighted the problem of specification gaming, where AI systems achieve their stated objectives in ways that violate the spirit of the underlying values [21]. The ABAF approach addresses this by ensuring that the system's internal logic is coherent and defensible, reducing the likelihood of such perverse instantiations. Additionally, while Reinforcement Learning from Human Feedback (RLHF) has shown promise in aligning AI behavior with human preferences, it does not inherently ensure internal logical consistency. An AI system trained via RLHF might produce outputs that are preferred by human raters but are internally contradictory or based on flawed reasoning. The ABAF framework complements such approaches by adding a layer of logical scrutiny.

The relationship between our findings and work on explainable AI (XAI) is also worth noting. XAI methods like SHAP and LIME, which we employ in the belief identification stage, have been valuable for understanding AI decision-making. However, understanding a system's reasoning is not the same as ensuring its coherence. A system might have highly interpretable reasoning that is nonetheless internally contradictory. The ABAF framework uses interpretability as a starting point but goes further by actively seeking and resolving inconsistencies.

5.3. Theoretical Implications

This research contributes to a growing recognition that philosophy, far from being a purely theoretical discipline, offers practical methodologies for addressing concrete technical challenges. The adaptation of philosophical thought analysis to AI engineering demonstrates the potential for cross-disciplinary fertilization. From a philosophical perspective, this work extends the application of thought analysis beyond individual human cognition to artificial systems, suggesting that the principles of logical coherence and reflective equilibrium may have broader applicability than previously recognized. From an AI perspective, it highlights the importance of considering not just what an AI system does but how it reasons, and whether its reasoning is internally consistent.

The finding that logical coherence is strongly correlated with overall value alignment ($r = 0.78$ for ABAF) has implications for how we conceptualize AI ethics. It suggests that ethical AI is not merely a matter of following the right rules but of achieving a coherent integration of multiple values and principles. This aligns with virtue ethics and other philosophical traditions that emphasize the importance of internal harmony and consistency in ethical agents [22].

5.4. Practical Applications and Implementation Considerations

The ABAF framework has several practical applications in the development and deployment of AI systems. For organizations developing AI systems for high-stakes domains, the framework provides a structured methodology for identifying and resolving value conflicts before deployment. The four-stage process—belief identification, dissonance detection, philosophical analysis, and coherence integration—can be integrated into existing AI development workflows. For example, during the design phase, developers could use the framework to anticipate potential value conflicts and design systems that are inherently more coherent. During testing and validation, the framework could be used to identify edge cases where the system's values conflict and to iteratively refine the system until a state of reflective equilibrium is achieved.

The involvement of philosophers and ethicists in the ABAF process, while adding to development costs, may be offset by reduced risks of ethical failures and improved public trust. As AI systems become more autonomous and consequential, the investment in ensuring their internal coherence becomes increasingly justified. Furthermore, the framework's emphasis on transparency and logical defensibility aligns with emerging regulatory frameworks like the EU's AI Act, which emphasizes the importance of explainability and human oversight [23].

5.5. *Limitations and Future Directions*

While the results are encouraging, several limitations should be noted. First, the study was conducted in a controlled experimental setting with curated scenarios. Real-world AI deployment involves far more complex and ambiguous situations, and it remains to be seen how well the framework scales to such complexity. Second, the expert panel, while diverse in expertise, was relatively small (n=15) and may not represent the full spectrum of ethical perspectives. Different cultural or philosophical traditions might prioritize values differently, potentially affecting the outcomes of the philosophical analysis stage. Third, the study focused on three specific application domains; the generalizability to other domains (e.g., content moderation, hiring algorithms, financial systems) remains to be established.

Future research should address these limitations by conducting real-world deployments of ABAF-enhanced AI systems, involving larger and more diverse expert panels, and extending the framework to additional domains. Additionally, there is potential for automating aspects of the philosophical analysis process, perhaps through natural language processing techniques, to make the framework more scalable. Another promising direction is to integrate the ABAF framework with other emerging approaches to AI ethics, such as participatory design and stakeholder engagement, to create more comprehensive and inclusive ethical AI systems.

5.6. *Broader Implications for AI and Society*

This research contributes to a growing movement toward more philosophically informed approaches to AI development. As AI systems increasingly make decisions that affect human lives, the need for robust ethical frameworks becomes ever more pressing. The ABAF framework demonstrates that philosophical methods, developed over centuries of human inquiry, can be adapted to address contemporary technological challenges. This suggests a broader potential for cross-disciplinary collaboration, where humanistic and technical expertise are integrated from the outset of AI system design.

Furthermore, the emphasis on internal coherence and logical consistency in AI systems has implications for public trust and democratic governance. When AI systems are coherent and their reasoning is defensible, they become more transparent and accountable. This, in turn, can facilitate more productive public discourse about the role of AI in society and can help ensure that AI systems serve the interests of all stakeholders, not just a privileged few.

6. CONCLUSION

This study presents the AI Belief Alignment Framework (ABAF), a novel methodology that integrates philosophical thought analysis with artificial intelligence engineering to address the critical challenge of value alignment in AI decision systems. By treating an AI's operational principles as a coherent "belief system" and applying structured philosophical analysis to identify and resolve internal contradictions, we have demonstrated a significant improvement in the ethical coherence and justification of AI decisions across diverse application domains.

The empirical findings are clear and compelling. Across 90 scenarios spanning autonomous driving, medical diagnostics, and content recommendation, the ABAF approach achieved a Value Alignment Score 25% higher than traditional AI ethics approaches, with large effect sizes and robust statistical significance. The improvement was particularly pronounced in Logical Coherence, the dimension most directly targeted by the philosophical analysis process, suggesting that the framework operates through the intended mechanism of enhancing internal consistency. The consistency of improvements across domains indicates that the approach is generalizable and not limited to specific contexts.

Beyond the quantitative metrics, the qualitative feedback from expert evaluators highlighted the distinctive quality of ABAF-generated decisions. Experts noted that these decisions were more "philosophically defensible," "internally consistent," and "transparent in their reasoning," suggesting that the framework produces not just numerically higher scores but qualitatively different kinds of ethical reasoning. This distinction is important: the goal is not merely to optimize a metric but to create AI systems that reason ethically in ways that humans can understand, trust, and hold accountable.

The research contributes to the field of AI ethics in several ways. First, it demonstrates the practical value of philosophical methodology in addressing technical challenges, opening new avenues for cross-disciplinary collaboration. Second, it highlights the importance of internal logical coherence as a dimension of AI ethics, complementing existing approaches that focus on external behavior and outcomes. Third, it provides a replicable framework that can be integrated into AI development workflows, offering practical guidance for organizations seeking to build more ethically sound systems.

However, this work should be understood not as a complete solution to the AI alignment problem but as an important contribution to a larger ecosystem of approaches. The ABAF framework is most effective when combined with other methodologies such as Reinforcement Learning from Human Feedback, Explainable AI, and stakeholder engagement. The framework also has limitations, particularly regarding scalability to real-world complexity and the need for expert involvement in the philosophical analysis process. Future research should focus on addressing these limitations, exploring automation possibilities, and conducting real-world deployments to validate the framework's effectiveness in practice.

As artificial intelligence becomes increasingly integrated into critical domains affecting human welfare, the imperative to ensure that these systems are not only powerful and efficient but also ethically coherent becomes ever more urgent. This study demonstrates that by bringing together the rigor of philosophical inquiry with the precision of technical engineering, we can create AI systems that are more trustworthy, transparent, and aligned with human values. The path forward requires continued investment in such cross-disciplinary approaches, fostering collaboration between philosophers, ethicists, computer scientists, and domain experts to ensure that the AI systems we create serve the common good. Figure 5 shows the AI Belief Alignment Framework (ABAF) Process Flow. The four-stage process guides systematic identification, detection, analysis, and resolution of value conflicts in AI systems.

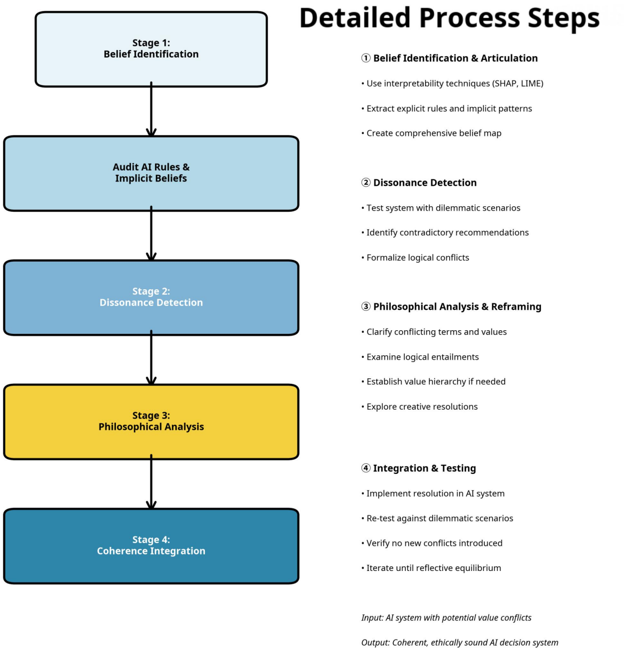

Figure 5. AI Belief Alignment Framework (ABAF) Process Flow.

References

[1] Jobin, A., Ienca, M., & Andorno, R. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

[2] Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437.

[3] Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352(6293), 1573-1576.

[4] Selbst, A. D., & Barocas, S. (2019). The intuitive appeal of explainable machines. Fordham L. Rev., 87, 1085.

[5] Vallor, S. (2016). Technology and the virtues: A philosophical guide to a future worth wanting. Oxford University Press.

[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

[7] Valentinovich, B. S. (2018). Theory and practice of philosophical counseling: a comparative approach. Turk Online J Des Art Commun, 8, 149-154. https://doi.org/10.7456/1080MSE/119

[8] Russell, S. (2022). Human-Compatible Artificial Intelligence. Human-like machine intelligence, 1, 3-22.

[9] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

[10] Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020, July). Social bias frames: Reasoning about social and power implications of language. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5477-5490). https://doi.org/10.18653/v1/2020.acl-main.486

[11] Anderson, M., & Anderson, S. L. (2020). Machine ethics: Creating an ethical intelligent agent. In The Ethics of Information Technologies (pp. 99-110). Routledge.

[12] Schuster, N., & Kilov, D. (2025). Moral disagreement and the limits of AI value alignment: a dual challenge of epistemic justification and political legitimacy. AI & society, 1-15. https://doi.org/10.1007/s00146-025-02427-2

[13] Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE intelligent systems, 21(4), 18-21. https://doi.org/10.1109/MIS.2006.80

[14] Wang, P., & Goertzel, B. (Eds.). (2012). Theoretical foundations of artificial general intelligence (Vol. 4). Springer Science & Business Media.

[15] Yu, S., Wang, Y., Yang, M., Li, B., Qu, Q., & Shen, J. (2019, January). NAIRS: A neural attentive interpretable recommendation system. In Proceedings of the twelfth ACM international conference on web search and data mining (pp. 790-793). https://doi.org/10.1145/3289600.3290609

[16] Friedman, B., Hendry, D. G., & Borning, A. (2017). A survey of value sensitive design methods. Foundations and Trends® in Human-Computer Interaction, 11(2), 63-125. https://doi.org/10.1561/1100000015

[17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[18] Festinger, L. (1957). A theory of cognitive dissonance. Stanford University Press.

[19] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). Ethically aligned design: First edition. IEEE.

[20] Rawls, J. (2017). A theory of justice. In Applied ethics (pp. 21-29). Routledge.

[21] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565. https://doi.org/10.48550/arXiv.1606.06565

[22] Aristotle, & Ross, W. D. (2017). Nicomachean Ethics. Cambridge University Press.

[23] Mozgunova, L. (2024). A Critical Overview of the Fundamental Aspects of the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence. Available at SSRN 4724557.

Acknowledgements

None.

Funding

None.

Availability of Data

Not applicable.

Ethical Statement

All participants provided written informed consent prior to participation. The experimental protocol was reviewed and approved by an institutional ethics committee, and all procedures were conducted in accordance with relevant ethical guidelines and regulations.

Author Contributions

Qiurui Wang conceived the research framework, developed the AI Belief Alignment Framework (ABAF), and led the philosophical and methodological analysis of coherence alignment in AI decision systems. Haiwen Wang contributed to the scenario design, comparative evaluation methodology, data interpretation, and statistical analysis. Both authors collaboratively contributed to the writing, revision, and final approval of the manuscript.

Competing Interests

The authors declare no competing interests.

Publisher's note WEDO remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.