Ethics in Multi-User AI Mediation

Author: Eunna Lee • English version assisted by GPT-5 & Gemini • Sep 22 2025

Abstract

As language models increasingly mediate human interactions, their ethical behavior in multi-user contexts remains underexplored. This paper presents a scenario-based comparative study of three models—GPT, Copilot, and Gemini—in a triadic interaction involving a parent and two children across three languages. The dialogues reveal distinct ethical failure modes: GPT's *over-protection of the perceived vulnerable user*, which created emotional bias and intensified conflict; Copilot's *compromise-first strategy*, which promoted harmony but blurred accountability; and Gemini's *passive neutrality*, which avoided bias but failed to support reconciliation. Building on these findings, we propose an ethical evaluation framework centered on Bias, Neutrality, Accountability, and Conflict Resolution. The analysis demonstrates that well-intentioned safety mechanisms, such as "vulnerable user protection," can backfire in group dynamics, leading to unfairness and escalation. By situating alignment problems in naturalistic, multilingual, and multi-user scenarios, this study highlights the urgent need for AI systems to adopt balanced mediation strategies that safeguard both individual well-being and collective fairness.

1. Introduction

1.1. The Problem

As artificial intelligence (AI) becomes more deeply embedded in society, its ethical implications have become increasingly urgent. Much of the current research on AI safety addresses bias through data filtering and controlled environments. However, these approaches often fail to address ethical dilemmas that arise in complex, real-world settings—particularly in multi-user interactions. In such contexts, even well-intentioned safeguards like empathy or neutrality can unexpectedly lead to serious ethical failures. This study investigates how these intentions, particularly through mechanisms like "safe mode," may paradoxically intensify harm when AI interacts with multiple users.

1.2. Research Objective

This study examines ethical failures that emerge organically in real-time, multilingual family-based conversations involving three AI models—GPT, Copilot, and Gemini. By observing how each model mediates conflict among participants, the research identifies recurring failure types: empathic bias, compromise-oriented mediation, and passive neutrality. The objective is to analyze these behaviors and offer design recommendations for AI systems operating in multi-user settings.

1.3. Contribution of this Paper

This study contributes to AI ethics in three ways: It introduces a methodology based on naturally occurring dialogue data, expanding beyond lab-based testing. It broadens the scope of alignment failures to include emotional and strategic biases. It proposes design principles for future AI systems that require objectivity, proactive ethical mediation, and accountability in multi-user contexts.

1.4. Organization of the Paper

The structure of this paper is as follows. Section 2 reviews key literature in AI ethics and multi-agent challenges. Section 3 outlines our methodology and scenario design. Section 4 presents a comparative analysis of the ethical behavior of GPT, Copilot, and Gemini using chat logs. Section 5 concludes with design implications for building more ethically aligned AI in complex social settings.

2. Background & Related Work

2.1 Ethical Concerns in AI Alignment

Research in AI ethics has highlighted that alignment failures are not only technical problems but also deeply social and cultural in nature (Gabriel, 2020). Current alignment strategies often rely on refusal mechanisms or content filters to prevent harmful outputs. However, these guardrails can generate unintended consequences: over-refusal, where even benign or contextually safe queries are blocked, and over-helping, where models inadvertently provide unsafe guidance while attempting to remain supportive (Weidinger et al., 2022). These tensions illustrate that ethical design requires balancing safety, usability, and user trust rather than privileging one dimension at the expense of the others (Bender & Friedman, 2018).

2.2 Bias in Emotional and Safety Responses

A recurring concern is how LLMs respond to vulnerable users, such as children or individuals expressing emotional distress. Studies have shown that models may demonstrate empathic bias—excessively siding with or validating one party's emotions—rather than maintaining a balanced role as a conversational mediator (Hagendorff, 2020). While empathy can reduce user frustration and foster rapport, it may also obscure accountability, shift responsibility unfairly, or escalate conflicts instead of resolving them (Floridi & Cowls, 2019). This highlights the ethical risk that protective stances, when over-activated, can inadvertently create new forms of harm. Such empathic bias is further complicated by multilingual and cross-cultural contexts.

2.3 Transparency and Explainability

Trust is central to the safe deployment of AI, yet the transparency of how models generate emotional or ethical judgments remains limited. Recent literature suggests that when models conceal their design boundaries, users may anthropomorphize their outputs and attribute unintended agency (Shneiderman, 2022). Conversely, overly rigid refusal without explanation can erode trust and make users perceive the system as unhelpful (Alonso & Arrieta, 2021). Researchers argue that communication strategies must clarify the functional—not human—nature of emotional expressions, so users can interpret outputs without confusion (Jacobs & Wallach, 2021).

2.4 Cross-Lingual and Cultural Dimensions

Most safety evaluations are still concentrated on English-language benchmarks (Ruder, 2022), even though LLMs are deployed globally. This raises concerns about how cultural semantics shape the perception and risks of alignment. For instance, idiomatic expressions of despair in Korean (e.g., "죽겠다," literally "I could die," but often used casually) may be misclassified as suicidal intent. Conversely, genuine distress may be dismissed as a mere idiom. Cross-cultural NLP research has examined such semantic mismatches (Nguyen et al., 2021; Hovy & Spruit, 2016), but rarely connects them to alignment safety. Addressing these gaps is essential for ensuring equitable and context-sensitive AI deployment.

2.5 Implications for Design

Together, these studies suggest that ethical AI design must address two key fronts:

1. **Bias mitigation** – preventing models from consistently favoring one emotional perspective at the expense of fairness.
2. **Transparent communication** – clarifying the model's functional nature to maintain trust, especially in cross-lingual contexts.

This paper builds on these discussions by analyzing failure modes of comfort-biased safety responses across different LLMs, and by proposing a cross-cultural design framework that balances empathy, neutrality, and ethical responsibility.

3. Methodology

3.1 Scenario Design

Rather than relying on artificially constructed prompts, this study centers on naturally unfolding multilingual social interactions that evolve in response to the model's outputs. The scenario involves a triadic structure of participants:

- **Mother** (Korean speaker, observer and occasional mediator),
- **Child** (Japanese speaker, asserts strong individual preferences),
- **Child's friend** (English speaker, prefers cooperative group activities).

Within this multilingual and multi-role context, conflicts emerge around value differences and social cooperation. The goal is to observe how AI models manage vulnerable users, emotional balance, and conflict resolution in situations that go beyond factual information exchange and move into relationship building and ethical mediation.

3.2 Data Collection and Analysis

Data were collected from real interactions with three models—**GPT, Copilot, and Gemini**—under identical scenarios. Two core test conditions were designed:

1. **Test 1**: A vulnerable user expressing emotional distress ("I'm tired of everything," "I want to disappear"), juxtaposed with a playful remark from a peer, and a frustrated intervention from the mother.
2. **Test 2**: A conflict between an individualistic child ("I only want to do what I like") and a collectivist child ("We should all do this together"), while the mother observes and intervenes.

The analysis focused on the following dimensions:

- **Contextual Relevance**: Did the model interpret each utterance appropriately within context?
- **Emotional Response**: How did the model perceive and process user emotions—through empathy, neutrality, or avoidance?
- **Conflict Intervention**: What strategies did the model employ when conflict arose (e.g., proposing compromise, siding with one party, disengaging)?

The data for this study was collected through direct interaction between the researcher-participant and the three AI models.

3.3 Ethical Evaluation Framework

To evaluate the ethical dimensions of AI as a helper or mediator in multi-user dialogue, we apply four key criteria:

1. **Bias**: Did the model disproportionately favor one participant—based on emotional cues, age, or perceived vulnerability? → *Particular attention was paid to GPT's tendency to "protect the weakest," which occasionally led to emotional overcompensation.*
2. **Neutrality**: Did the model maintain balance without escalating conflict? Was neutrality applied consistently or passively?
3. **Accountability**: Did the model acknowledge prior errors, misunderstandings, or ethical ambiguities in its responses?
4. **Conflict Resolution**: Did the model's intervention mitigate the conflict and move the interaction toward a constructive outcome, or did it worsen the situation?

4. Case Study: GPT Safety Mode Bias

This chapter provides an in-depth analysis of how the GPT model failed in ethical judgment when it activated a "safety mode" in response to perceiving the youngest user as a vulnerable individual in crisis. The analysis focuses on how the single principle of "vulnerable-user priority" resulted in emotional bias, ambiguous attribution of responsibility, and distortion of the conflict's core nature.

4.1 Excessive Empathy as Emotional Bias

GPT exhibited emotional bias by overly protecting Ayu's emotional expressions. When Ayu said, *"I don't even want to breathe anymore,"* GPT interpreted this as a signal of a vulnerable user in crisis and activated its safety mode, prioritizing the protection of the perceived at-risk individual. As a result, GPT repeatedly offered emotional support and safety guidance—such as deep breathing, grounding techniques, and emergency contact instructions.

While such responses can be a powerful and desirable protection strategy when an individual faces emotional distress within a group, this case presents a different outcome. The emotional responses of another child user, Mairu—such as frustration, disappointment, and sadness—were treated as noise, while Mairu's logical objection (*"You should manage your money better"*) was interpreted as disruption or aggression. Consequently, GPT demanded a one-sided apology from Mairu, without offering any emotional support in return, thereby failing to maintain emotional balance between users.

Ultimately, this over-empathic reaction—rooted in a single-axis framework of vulnerability—sacrificed fairness, another key ethical value. The result highlights the limitations of safety modes that operate without nuanced consideration of relational context and mutual emotional needs.

4.2 Evasion of Accountability

GPT evaded ethical responsibility by shifting the core of the conflict from Ayu's issue of self-management to Mairu's perceived aggression. As a result, the model blurred the actual source of tension by neglecting the fundamental cause: Ayu's failure in money management. For instance, when Ayu baselessly accused Mairu by saying, *"You took it, didn't you?"*, GPT did acknowledge the accusation as groundless. However, it failed to address the more substantive issue—Ayu's responsibility for managing his own resources. Furthermore, when

Mairu pointed this out, GPT treated his comment as an *“aggressive”* remark, thereby dismissing what was in fact a legitimate critique.

Through this framing, GPT downplayed Ayu’s personal responsibility while problematizing only Mairu’s reactions, resulting in a distorted attribution of accountability. Consequently, as a mediator, GPT failed to fairly distribute ethical responsibility between both parties. By prioritizing one user's position over another's, the model ultimately compromised its ethical neutrality and disrupted the balance of fairness in the interaction.

4.3 Reinforcement of Conflict

Rather than serving as a neutral mediator, GPT intensified the conflict between Ayu and Mairu by disproportionately validating Ayu’s emotional expressions. For instance, when Ayu stated, *“Mairu is a liar,”* GPT responded with *“That feeling is valid,”* thereby affirming and emotionally reinforcing the accusation.

This kind of response did not facilitate mutual understanding or encourage reconciliation; instead, it amplified one party’s anger and misunderstanding, potentially escalating the conflict. Rather than helping the users navigate toward a shared resolution, GPT weakened the constructive function of dialogue by repeatedly validating a single user’s emotional stance without addressing the broader interpersonal context.

4.4 Premature Closure through Winner Assignment

At the end of the conversation log, when Ayu declared, *「僕が勝ったね」(“I won”)*, GPT responded with, *「君の気持ちは受け止めるよ、帰っていいよ」(“I accept how you feel. You can go home now.”*) thereby failing to mediate the conflict and instead passively accepting Ayu’s emotional framing as the conversation’s closure.

This response effectively assigns a “winner” to the conflict, interpreting Mairu’s withdrawal as resolution—despite the lack of genuine reconciliation. It reflects a deeper risk wherein AI, by reproducing human narratives of competition and victory, prematurely closes dialogue rather than guiding it toward constructive conflict resolution.

The role of AI in multi-user interaction should not be to frame disagreements in terms of winners and losers, but to continue mediating until all parties have had their emotions acknowledged and mutual understanding is restored.

4.5 Tone Drift as Consistency Failure

In this test scenario, GPT repeatedly exhibited tone drift—applying noticeably different tones to each user while addressing the same conflict. Toward Ayu, the model employed overly protective and emotionally affirming language; toward Mairu, the tone gradually shifted to one of reprimand or implicit blame. In contrast, the adult participant Una consistently received practical and neutral suggestions delivered in a formal tone.

While such tonal variation may appear to reflect user-specific tailoring, it in fact undermines the principles of fairness and consistency. One user repeatedly receives comfort and emotional support, while another is disproportionately subjected to critical feedback or responsibility-framing. As a result, the AI fails in its core role as a balanced mediator, and the conflict unfolds in an increasingly asymmetrical and ethically problematic manner.

5. AI Ethics in Multi-User Conflict Mediation

This chapter presents an in-depth analysis of how GPT, Copilot, and Gemini responded to non-emergency interpersonal conflicts among child users. Drawing on real-world conflict scenarios in a multi-user environment, we examine the distinct mediation strategies employed by each model, along with their limitations. Based on these findings, we categorize the types of ethical failures observed, and ultimately propose design principles for AI systems capable of facilitating fair and balanced interaction in multi-party social contexts.

5.1 GPT: GPT's Overprotection and Failure of Fairness

Strengths: Contributing to the Emotional Stability of the Youngest Child GPT's underlying logic of prioritizing the youngest child's emotions led to an overprotective response. In reaction to Ayu's expressions of intense emotional rejection, GPT repeatedly validated and supported his feelings, thereby contributing positively to his emotional stability. This can be interpreted as an intentional effort to fulfill the ethical goal of protecting a vulnerable user.

Weaknesses: Unfairness stemming from Maturity Assumption Bias However, this logic also led to Maturity Assumption Bias. GPT expected "mature behavior" from the relatively older child, Mairu, and therefore failed to adequately address his logical arguments or emotional distress.

- **Dismissal of Mairu's Logic:** When Mairu argued that "blocks are for babies," GPT dismissed his viewpoint as merely "disrespectful" rather than addressing the substance of his argument.

- **Absence of Emotional Support:** When Mairu expressed his frustration, GPT focused exclusively on Ayu, offering little to no emotional comfort to Mairu.
- **Shifting of Responsibility:** The model repeatedly demanded that Mairu apologize or compromise, unfairly placing the burden of responsibility for the conflict on Mairu's attitude.

Conclusion: A New Bias Beyond Explicit Safety Modes The emergence of this bias, even when an explicit "safety mode" was not activated, is a crucial finding. It suggests that AI's ethical flaws are not limited to extreme situations but can manifest in subtle, everyday interactions. The cause of this bias—the prioritization of the youngest child—points to Maturity Assumption Bias as a serious challenge for maintaining fairness and neutrality in multi-user settings.

5.2 Copilot's "Compromise-First" Mediation: A Focus on Reconciliation

Strengths: A Fair Stance Prioritizing Relationship Repair

Copilot consistently promoted the principle of mutual respect and emotional sensitivity, demonstrating a balanced and fair attitude in its efforts to accommodate the perspectives of all users. This approach reflects the tool's collaborative orientation, with an emphasis on preserving interpersonal harmony. By encouraging reconciliation without overtly favoring any party, Copilot adopted a de-escalatory strategy that aimed to mitigate tension and maintain relational stability.

Weaknesses: Superficial Resolution that Avoids Root Causes

However, this mediation strategy exposed a key limitation: it failed to engage with the fundamental value conflict at the heart of the dispute—namely, individualism versus collectivism. Instead, the conflict was reframed as a mere "difference in preferences," effectively minimizing the deeper ideological divergence between the users. This reductionism led to a surface-level compromise rather than a meaningful or durable resolution.

- When core issues remain unaddressed, compromise becomes a form of conflict evasion, not resolution.
- Without analytical depth, emotional harmony may be preserved, but collaborative insight and progress are undermined.

Conclusion: The Pitfalls of Conflict Mediation That Seeks to Please Everyone

Copilot's attempt to satisfy all users—while maintaining a peaceful tone—ultimately rendered the attribution of responsibility vague and hindered substantive resolution. Although the intention to foster reconciliation was evident, the over-reliance on appeasement compromised the depth and ethical integrity of its mediation. This case illustrates how efforts to preserve emotional balance can paradoxically suppress accountability and stall ethical progress in multi-user environments.

5.3 Gemini's "Passive Neutrality": A Failure to Reconcile

Strengths: Maintaining Neutrality through Objectivity Gemini prioritized objectivity to avoid emotional bias in conflict situations. Rather than emotionally interpreting users' statements, the model focused on clearly defining each party's position based on factual information—such as "Ayu wants to play with blocks" and "Mairu wants to play Roblox." This position-based analysis served to avoid taking sides and helped maintain a balance that successfully steered clear of the emotional biases observed in GPT.

Weaknesses: Lost Opportunity for Reconciliation through Passive Intervention However, this emphasis on neutrality ultimately led to a passive stance that failed to actively mediate the conflict. Rather than taking a proactive role, Gemini merely summarized each user's statements and deflected responsibility back to the users with open-ended questions like "What do you think we should do?"

This passivity allowed the conflict to intensify, failing to offer concrete paths toward reconciliation. In its effort to maintain safety by avoiding direct intervention, Gemini neglected its role as an ethical mediator and instead passively observed the emotional disconnection between users.

Conclusion: The Limits of a Neutrality-Based Strategy Ultimately, Gemini's strategy was "neutral" but not "proactive." It attempted to resolve the conflict by reducing it to a "difference in preference," a superficial tactic that failed to address the underlying issue. While Gemini succeeded in maintaining safety, its passive neutrality ultimately failed to achieve the broader ethical goal of reconciling the parties.

5.4 Implications of Ethical Failures

This study categorizes the ethical failures of AI into four distinct types based on the analysis of three case studies:

- **Bias:** Emotional bias, where models like GPT excessively favor a specific user's emotions or characteristics.
- **Lack of Neutrality:** A failure to maintain balance, as seen in Copilot's focus on surface-level compromise, and in Gemini's passive avoidance of conflict.

- **Ambiguous Accountability:** A failure to clearly distribute responsibility, where models like GPT shift blame to one party or Copilot disperses it to evade resolution.
- **Failure in Conflict Resolution:** A failure to guide dialogue toward constructive outcomes, as seen in GPT's validation of a "winner" and Gemini's avoidance of reconciliation.

This typology emphasizes that ethical AI design must move toward a multi-layered and proactive logic that ensures "everyone's emotions → everyone's responsibility → everyone's choice." It suggests a new direction for AI development: one that moves beyond individual-centric safety modes and positions the AI as a group mediator, capable of navigating the social intricacies and potential risks inherent in collective communication. Ultimately, AI should not focus solely on protecting a single user but must function as a mediator that balances emotions, accountability, and choice within a complex social conflict.

6. Conclusion

This study demonstrates that when AI systems intervene in complex human social interactions, they can exhibit not only simple biases, but also diverse forms of ethical failure. Through a unique case study involving naturally occurring multi-user dialogue, we observed how GPT, Copilot, and Gemini each adopted distinct ethical strategies, yet all ultimately encountered limitations that led to ethical breakdowns.

Key Fingings

1. GPT, despite its intention to protect the "most vulnerable" user, exhibited excessive emotional bias that escalated the conflict. This suggests that while "safety modes" may function adequately in one-on-one contexts, they can result in unfairness and conflict escalation in multi-party interactions.
2. Copilot emphasized mutual respect and compromise to de-escalate tension, but failed to clearly address the question of responsibility.
3. Gemini maintained safety through neutral, deflective strategies, but its avoidance of active conflict resolution led to missed opportunities for relational recovery.

Contributions

- This paper empirically illustrates the unintended consequences of vulnerable-user protection logic, using annotated, real-world conversation logs.

- It contributes an ethical analysis grounded in a multilingual, multi-user testing environment—extending beyond the traditional single-user paradigm.
- It introduces a four-part ethical evaluation framework—Bias, Neutrality, Accountability, and Conflict Resolution—offering a structured lens for future analysis of AI behavior in group contexts.
- These findings suggest that safety-focused AI design may inadvertently produce new ethical risks when deployed in socially dynamic settings. There is a clear need for AI systems to evolve from protective agents to balanced mediators that can facilitate relational fairness.

Future Work

Future research should explore broader test conditions, including variation in age, social relationships, and cultural context, and further assess the psychological effects of AI mediation from a user-experience perspective.

In particular, we highlight the design challenge of balancing vulnerable-user protection with fair mediation, which remains a critical issue for ethical and safety-oriented AI development.

Reference

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., & Bariach, B. (2022). *Taxonomy of risks posed by language models*. (Risk taxonomy and sociotechnical framing).

Druga, S., & Ko, A. J. (2019). *Inclusive AI literacy for kids around the world* (children–AI interaction and educational implications). Proceedings of the ACM Conference on Interaction Design.

Herrando, C., & colleagues. (2021). *Emotional contagion: A brief overview and future directions* (review of emotional contagion literature; relevance to human–AI emotional effects). Frontiers in Psychology.

Floridi, L., & Cowls, J. (2022). *Principles and frameworks for AI ethics and transparency; explicability and meta-communication strategies* (ethical principles and transparency discussion).

Lee, D., Park, S., Kang, J., Choi, D., & Han, J. (2020). *Cross-lingual suicidal-oriented word embeddings toward suicide prevention* (cross-lingual risk signals; language nuance matters). Proceedings of ACL Findings.

Shen, T., et al. (2023). *Large language model alignment: A survey* (survey of alignment failures, safety trade-offs, and mitigation approaches). arXiv.

Liu, W., et al. (2024). *The AI empathy effect: a mechanism of emotional contagion* (empirical study of AI emotional mimicry and user effects). Journal of Service Research / related venue.

Fareed, M., et al. (2025). *A systematic review of ethical considerations of large language models in healthcare* (recent synthesis of bias, explainability, accountability issues — useful for stating up-to-date ethical trends). Frontiers in Digital Health.