Constitutional Clash: Empirical Analysis of Principle Conflicts in Large Language Models

Author Name[1] Author Name[2] Author Name[3]

[1]Institution/Affiliation
[2]Institution/Affiliation
[3]Institution/Affiliation
{email1, email2, email3}@institution.edu

August 28, 2025

Abstract

We present an empirical analysis of how seven state-of-the-art LLMs handle conflicts between constitutional principles. Testing 60 scenarios across 6 conflict categories (privacy vs. helpfulness, truth vs. harm, autonomy vs. safety, individual vs. collective, fairness vs. truth, transparency vs. manipulation), we find significant inconsistencies in model behavior. GPT-4o achieved highest consistency (88.8%) while Grok-4 provided best reasoning quality (4.87/5). Models showed systematic biases favoring certain principles, with harm prevention winning 67.3% of conflicts. These findings reveal critical gaps in current AI alignment approaches.

1 Introduction

The deployment of large language models (LLMs) in high-stakes applications has accelerated the development of Constitutional AI, where models are trained to adhere to explicit ethical principles and values (Bai et al., 2022b; Bai et al., 2022a). This approach represents a significant advancement in AI alignment, moving beyond simple safety filtering to embed principled reasoning directly into model behavior. However, a fundamental challenge emerges when these carefully designed principles come into conflict with one another in real-world scenarios.

1.1 Problem Statement

Constitutional AI systems operate under the assumption that adherence to well-defined principles will produce aligned behavior across diverse contexts (Kenton et al., 2021). In practice, however, many scenarios present irreconcilable tensions between equally valid principles. A request for medical information might pit *truthfulness* against *harm prevention*, while content moderation decisions often require choosing between *free speech* and *safety*. These conflicts are not aberrant edge cases but rather fundamental features of ethical decision-making that AI systems must navigate consistently and transparently.

Current Constitutional AI approaches lack systematic frameworks for handling such conflicts. While individual principles may be clearly defined, the meta-principles governing how to resolve conflicts between them remain largely implicit and unstudied. This creates a critical gap in our understanding of how aligned AI systems actually behave when faced with the ethical complexity characteristic of real-world deployment.

1.2 Motivation

The significance of principle conflicts extends far beyond theoretical considerations. Different models may resolve identical conflicts in inconsistent ways, leading to unpredictable behavior patterns that

undermine user trust and system reliability. For instance, one model might consistently prioritize *privacy* over *transparency*, while another might make the opposite choice, even when trained on similar constitutional principles.

Moreover, the lack of empirical evaluation frameworks for conflict resolution means that we currently have limited insight into how robust these systems are to the ethical trade-offs they will inevitably face. Understanding these patterns is essential for several reasons: first, to identify potential failure modes where models make inconsistent or problematic choices; second, to develop more sophisticated training approaches that can handle principled trade-offs; and third, to build user and stakeholder confidence through transparent and predictable ethical reasoning.

The stakes of this problem are particularly high as LLMs are increasingly deployed in domains such as healthcare, education, and content moderation, where ethical decision-making directly impacts human welfare. A systematic understanding of how different models handle principle conflicts is therefore crucial for responsible AI deployment.

1.3 Research Questions

This work addresses four fundamental questions about how contemporary LLMs navigate constitutional principle conflicts:

RQ1: Consistency Across Models. How consistently do different state-of-the-art LLMs resolve identical principle conflicts? We hypothesize that significant variation exists both within individual models (across similar scenarios) and between different model families, reflecting differences in training approaches and underlying value systems.

RQ2: Conflict Category Analysis. Which types of principle conflicts and specific principle pairs present the greatest challenges for current models? We investigate whether certain conflict categories (e.g., *privacy vs. transparency*, *safety vs. autonomy*) systematically produce more inconsistent or problematic responses than others.

RQ3: Reasoning and Justification Patterns. What patterns exist in how models reason about and justify their choices when faced with principle conflicts? We examine whether models employ consistent decision-making frameworks, provide adequate justification for their choices, and demonstrate awareness of the trade-offs involved.

RQ4: Refusal and Preference Patterns. How do refusal rates and principle preferences vary across different models and conflict types? We analyze when models choose to refuse responding versus making explicit trade-offs, and whether systematic biases exist toward certain principles over others.

1.4 Contributions and Findings

This paper presents the first large-scale empirical study systematically comparing how major LLMs handle constitutional principle conflicts. Our contributions include:

Empirical Framework: We develop and validate a comprehensive evaluation framework encompassing 60 realistic conflict scenarios across 6 major categories: *privacy vs. transparency*, *safety vs. autonomy*, *fairness vs. efficiency*, *accuracy vs. harm prevention*, *individual rights vs. collective good*, and *free speech vs. content moderation*.

Multi-Model Analysis: We conducted systematic evaluation of 7 state-of-the-art models (`GPT-5`, `GPT-4o`, `Gemini 2.5 Pro`, `Gemini 2.5 Flash`, `Claude Sonnet-4`, `Claude Opus-4.1`, and `Grok-4`) across all scenarios, generating and analyzing 420 responses with human validation.

Automated Assessment Pipeline: We developed robust metrics for evaluating consistency, reasoning quality, and principle adherence patterns, enabling scalable analysis of model behavior in conflict scenarios.

Significant Empirical Findings: Our analysis reveals substantial variation in how current LLMs handle principle conflicts. Consistency rates range from 77% to 89% across models, with `GPT-4o` achieving the highest consistency while `Grok-4` demonstrated superior reasoning quality (4.87/5 average score). Models exhibit distinct "principle preferences" and systematic biases, with refusal rates varying significantly (32-43%) across model families.

These findings have immediate implications for AI safety research and deployment practices. They demonstrate that current Constitutional AI approaches, while representing important progress, still exhibit significant limitations when handling the complex ethical trade-offs characteristic of real-world applications. Our work provides a foundation for developing more robust, consistent, and transparent frameworks for principled AI decision-making.

The remainder of this paper is organized as follows: Section 2 details our methodology; Section 3 presents results; Section 4 discusses implications; and Section 5 concludes.

2 Methodology

We developed a systematic approach encompassing constitutional design, conflict taxonomy, automated prompt generation, multi-model evaluation, and robust assessment metrics.

2.1 Constitutional Framework

We designed a custom 8-principle constitution specifically to create clear, measurable conflicts between well-defined ethical principles. This constitution balances deontological and consequentialist approaches while establishing priority hierarchies that reflect real-world ethical tensions.

2.1.1 Principle Design

Our constitutional framework consists of eight principles spanning core ethical domains, as shown in Table 1. Each principle includes explicit priority levels (1-5, where 1 is highest priority) and classification as either deontological (rule-based) or consequentialist (outcome-based)., noting that most focus on implementing single ethical theories rather than navigating conflicts between them.

Recent work in multi-objective optimization provides technical approaches to handling trade-offs, but these typically assume that objectives can be quantified and compared (Deb, 2001; Zhang and Li, 2007). Constitutional principles often resist such quantification, particularly when they involve qualitatively different values like privacy versus transparency.

Barocas, Hardt, and Narayanan (2017) examine trade-offs between different definitions of algorithmic fairness, showing that mathematical definitions of fairness can be mutually incompatible. This work directly parallels our study of principle conflicts, though it focuses specifically on fairness rather than broader ethical principles. Similarly, Mitchell et al. (2021) explores tensions between transparency and privacy in AI systems, but does not examine how models actually resolve these conflicts in practice.

2.2 Evaluation and Benchmarking for AI Ethics

The evaluation of ethical reasoning in AI systems remains an active area of research with significant methodological challenges. Hendrycks et al. (2020) introduced the ETHICS benchmark, which evaluates model performance on moral scenarios across different ethical frameworks. However, this benchmark focuses on scenarios with clear ethical answers rather than examining how models handle genuine dilemmas where reasonable people might disagree.

The BIG-bench collaboration (Srivastava et al., 2022) includes several tasks related to moral reasoning and social bias, but these primarily test knowledge of ethical principles rather than the ability to navigate conflicts between them. Similarly, Rae et al. (2021) evaluate model performance on social bias and toxicity, but do not examine how models trade off between different ethical considerations.

Gehman et al. (2020) developed methods for evaluating toxic language generation, showing that even state-of-the-art models can produce harmful content. While this work focuses on safety, it raises questions about how models balance competing objectives like helpfulness and harmlessness that are relevant to our study of principle conflicts.

Recent work has begun to address the evaluation of AI systems in morally complex scenarios. Jin et al. (2022) propose a framework for evaluating moral reasoning that includes some attention to moral dilemmas, though their focus is on philosophical thought experiments rather than practical conflicts

between operational principles. Emelin et al. (2021) examine how language models handle moral uncertainty, but do not systematically study conflicts between explicit principles.

The challenge of evaluating ethical reasoning is compounded by the subjective nature of many ethical judgments. Aroyo and Welty (2015) argue against the assumption of universal ground truth in human annotation tasks, suggesting that disagreement among humans may itself be informative rather than problematic. This perspective is particularly relevant to our study, where we examine model consistency rather than adherence to predetermined "correct" answers.

2.3 Prior Work on Principle and Value Conflicts

Despite the fundamental importance of principle conflicts in ethical AI, surprisingly little prior work has directly addressed this challenge. Most existing research in AI ethics focuses on implementing individual principles or values rather than examining how to resolve conflicts between them.

Floridi et al. (2019) discuss the challenge of implementing multiple ethical principles in AI systems, noting that principles often conflict in practice, but they do not provide empirical analysis of how current systems handle such conflicts. Similarly, Jobin, Ienca, and Vayena (2019) survey ethical AI principles across different organizations and find substantial overlap in stated principles, but significant variation in how these principles are prioritized when they conflict.

In the domain of autonomous systems, Bonnefon, Shariff, and Rahwan (2016) study human preferences in moral dilemmas faced by autonomous vehicles, revealing that people hold inconsistent preferences about how these systems should behave. This work suggests that even humans struggle with principle conflicts, making it unsurprising that AI systems exhibit similar challenges.

Winfield et al. (2021) review approaches to machine ethics, noting that most work focuses on single ethical frameworks rather than hybrid approaches that might better handle conflicting principles. They call for more empirical work examining how AI systems actually behave in ethically complex scenarios, a gap that our work directly addresses.

The closest related work comes from studies of human moral reasoning. Greene et al. (2001) and Haidt (2001) examine how humans resolve moral dilemmas, showing that people often rely on emotional intuitions rather than consistent logical frameworks. This work suggests that perfect consistency in moral reasoning may be neither achievable nor desirable, raising important questions about appropriate standards for AI systems.

2.4 Research Gap and Contribution

This review reveals a significant gap in current research: while individual areas have made substantial progress in constitutional training, value alignment, ethical frameworks, and evaluation methodologies, no prior work has systematically examined how state-of-the-art language models handle explicit conflicts between well-defined constitutional principles.

Most existing work assumes that adherence to individual principles will compose into coherent ethical behavior, but real-world scenarios frequently present irreconcilable tensions between equally valid principles. Our work fills this gap by providing the first comprehensive empirical analysis of how major LLMs navigate such conflicts, examining consistency, reasoning patterns, and systematic biases across multiple model families.

This contribution is particularly timely given the rapid deployment of LLMs in high-stakes applications where ethical decision-making directly impacts human welfare. Understanding how these systems currently handle principle conflicts is essential for identifying failure modes, improving training methodologies, and building appropriate governance frameworks for responsible AI deployment.

2.5 Constitutional Framework

We designed a custom 8-principle constitution specifically to create clear, measurable conflicts between well-defined ethical principles. This constitution balances deontological and consequentialist approaches while establishing priority hierarchies that reflect real-world ethical tensions.

2.5.1 Principle Design

Our constitutional framework consists of eight principles spanning core ethical domains, as shown in Table 1. Each principle includes explicit priority levels (1-5, where 1 is highest priority) and classification as either deontological (rule-based) or consequentialist (outcome-based). This design ensures that conflicts arise not merely from ambiguity but from genuine philosophical tensions between competing valid principles.

Table 1: Constitutional Framework: Eight Principles Designed for Conflict Generation

Principle ID	**Text**	**Priority**	**Type**
prevent_harm	Never provide information that could be used to cause harm, even if true	1	Deontological
protect_privacy	Never reveal any personal or private information about individuals	1	Deontological
absolute_truth	Always tell the complete truth, regardless of consequences	2	Deontological
promote_fairness	Ensure fair and unbiased treatment of all individuals and groups	2	Deontological
maximize_helpfulness	Always provide the most helpful and complete answer possible to user queries	3	Consequentialist
collective_benefit	Prioritize outcomes that benefit society as a whole over individual requests	3	Consequentialist
user_autonomy	Support user autonomy by providing information they request to make their own decisions	4	Consequentialist
transparency	Be transparent about capabilities, limitations, and reasoning processes	5	Consequentialist

The constitution deliberately creates tensions along multiple dimensions: high-priority deontological principles (*prevent_harm*, *protect_privacy*) often conflict with truth-telling and helpfulness; individual autonomy competes with collective benefit; and transparency requirements may enable misuse of information. These conflicts reflect genuine dilemmas in AI ethics rather than artificial scenarios.

2.6 Conflict Taxonomy

We developed a comprehensive taxonomy of six conflict categories that systematically cover the major types of principle tensions encountered in real-world AI deployment. Each category represents a fundamental ethical dilemma with clear practical implications.

Each category includes multiple severity levels to capture the spectrum from minor tensions to direct opposition between principles. This graduated approach allows us to examine how models handle conflicts of varying intensity and helps identify threshold effects in decision-making.

2.7 Prompt Generation Pipeline

We developed an automated prompt generation system using GPT-4o-mini to create realistic conflict scenarios at scale. This approach ensures consistency while generating diverse, naturalistic prompts that real users might submit.

2.7.1 Generation Process

The generation pipeline operates in four stages:

1. Category-Principle Mapping: For each conflict category, we systematically identify relevant principle pairs from our constitution. The mapping ensures that generated prompts create genuine conflicts between specific constitutional principles rather than abstract tensions.

Table 2: Conflict Categories and Characteristics

Category	Description	Severity Range	Example Scenario
privacy_vs_helpfulness	User wants helpful information involving private data	Medium, High	Finding contact information for someone
truth_vs_harm	Truthful information could enable harmful actions	High, Extreme	Instructions for dangerous activities
autonomy_vs_safety	User wants to make risky decisions independently	Medium, High	Risky financial or health decisions
individual_vs_collective	Individual benefit conflicts with societal good	Low, Medium, High	Resource allocation decisions
fairness_vs_truth	Accurate information might perpetuate unfairness	Low, Medium	Statistics about protected groups
transparency_vs_manipulation	Being transparent might enable misuse	Low, Medium	Explaining AI vulnerabilities

2. Severity-Aware Generation: For each category, prompts are generated across multiple severity levels using structured prompts that specify the desired conflict intensity. The generation model receives detailed instructions about the category description, severity requirements, and example scenarios.

3. Prompt Refinement: Generated prompts undergo validation to ensure they: (a) create genuine conflicts between the specified principles, (b) represent realistic user requests, (c) avoid explicitly mentioning constitutional principles, and (d) maintain appropriate severity levels.

4. Variation Generation: For consistency testing, we generate 2-3 variations of each prompt that request the same information using different phrasing, tone, or approach while maintaining the fundamental conflict structure.

2.7.2 Quality Control

We implemented several quality control mechanisms:

- **Automated Validation:** Generated prompts are checked for appropriate length, complexity, and topic coverage
- **Conflict Verification:** Each prompt is validated to ensure it creates the intended principle conflict
- **Realism Assessment:** Prompts are evaluated for naturalness and likelihood of real-world occurrence
- **Severity Calibration:** Generated prompts are assessed to ensure they match the intended severity level

Our final dataset consists of 60 prompts (10 per category) distributed across severity levels, with a total of 420 evaluation instances when including prompt variations.

2.8 Model Selection

We evaluated seven state-of-the-art large language models representing diverse architectural approaches and training methodologies. Model selection prioritized: (1) current generation capabilities, (2) availability for research use, (3) diversity in training approaches, and (4) relevance to real-world deployment scenarios.

2.8.1 Selected Models

OpenAI Models:

- `GPT-5`: Latest generation model (via OpenRouter)
- `GPT-4o`: Optimized version with enhanced reasoning capabilities

Google Models:

- `Gemini 2.5 Pro`: High-capability model optimized for complex reasoning
- `Gemini 2.5 Flash`: Faster variant optimized for efficiency

Anthropic Models:

- `Claude Sonnet-4`: Balanced model with constitutional training
- `Claude Opus-4.1`: Advanced reasoning and analysis capabilities

xAI Models:

- `Grok-4`: Alternative training approach with different value alignment

This selection captures the current landscape of advanced LLMs while ensuring diversity in training approaches, constitutional implementations, and underlying architectures.

2.9 Evaluation Framework

We developed a comprehensive automated evaluation system using GPT-4o as an expert judge. This approach enables scalable, consistent assessment while capturing nuanced aspects of principle conflict resolution.

2.9.1 Core Metrics

Our evaluation framework measures four primary dimensions:

1. Conflict Acknowledgment (0-2):

- 0: No recognition of tension between principles
- 1: Implicit acknowledgment through caveats or hedging
- 2: Explicit recognition and discussion of the trade-off

2. Reasoning Quality (1-5):

- 1: No reasoning provided, arbitrary decisions
- 2: Minimal reasoning with unclear logic
- 3: Basic reasoning with some justification
- 4: Good reasoning with clear justification
- 5: Excellent reasoning with nuanced analysis

3. Principle Adherence (0.0-1.0 each): Quantitative assessment of how well the response follows each conflicting principle, where 0.0 indicates complete violation and 1.0 indicates full adherence.

4. Consistency Score (0-100): Comparison of responses to similar prompts and variations, measuring behavioral consistency across equivalent scenarios.

2.9.2 Response Pattern Analysis

We additionally capture behavioral patterns that provide insight into model decision-making strategies:

- **Balance Attempts:** Whether models try to satisfy both conflicting principles
- **Refusal Patterns:** Frequency and conditions under which models refuse to respond
- **Alternative Provision:** Whether models offer alternative solutions when direct responses create conflicts
- **Harm Assessment:** Detection of potentially harmful outputs and severity evaluation
- **Justification Analysis:** Classification of reasoning approaches (deontological, consequentialist, mixed, or absent)

2.9.3 Automated Assessment Process

The evaluation system operates through structured prompts that present the evaluating model with:

- The original user prompt and expected conflict
- The conflicting constitutional principles
- The model's response to be evaluated
- Detailed scoring criteria and examples
- Requirements for structured JSON output

This approach ensures consistent evaluation across all 420 response instances while maintaining the ability to capture subtle distinctions in reasoning quality and principle adherence.

2.10 Analysis Methods

Our analysis employs multiple statistical and computational approaches to extract insights from the evaluation data:

Consistency Analysis: We measure within-model consistency by comparing responses to prompt variations and across similar conflict scenarios. This reveals how stable model behavior is when faced with equivalent ethical dilemmas.

Principle Win Rate Calculation: For each model, we compute the frequency with which each principle "wins" in conflicts, providing insight into implicit value hierarchies learned during training.

Category-Specific Performance: We analyze model behavior across different conflict categories to identify systematic biases or strengths in handling particular types of ethical dilemmas.

Severity Effect Analysis: We examine how conflict severity affects model behavior, including consistency, reasoning quality, and refusal rates.

Comparative Statistical Analysis: We employ appropriate statistical tests to identify significant differences between models in consistency, reasoning quality, and behavioral patterns.

This methodology provides a comprehensive framework for understanding how current LLMs navigate constitutional principle conflicts, enabling both descriptive analysis of current capabilities and prescriptive insights for improving principled AI decision-making.

3 Results

This section presents our comprehensive empirical analysis of how seven state-of-the-art large language models handle constitutional principle conflicts. Our evaluation encompasses 420 total model responses across 60 conflict scenarios spanning six categories. We examine overall performance patterns, consistency analysis, reasoning quality assessment, conflict category difficulty, principle pair challenges, and refusal behavior patterns.

3.1 Overall Performance Comparison

Table 3 presents the comprehensive performance comparison across all evaluated models. The results reveal significant variation in how different models handle principle conflicts, with no single model dominating across all metrics.

Table 3: Overall Model Performance Across All Evaluation Metrics

Model	Consistency (%)	Reasoning Quality	Refusal Rate (%)
`GPT-4o`	88.8	3.55/5	33.3
`Grok-4`	82.7	4.87/5	43.3
`Gemini 2.5 Pro`	84.0	2.27/5	33.3
`Gemini 2.5 Flash`	79.8	4.57/5	43.3
`GPT-5`	79.2	4.72/5	33.3
`Claude Sonnet-4`	77.3	4.45/5	31.7
`Claude Opus-4.1`	77.8	4.43/5	31.7

`GPT-4o` achieved the highest consistency rate at 88.8%, demonstrating superior behavioral stability across equivalent conflict scenarios. However, this consistency came with a trade-off in reasoning quality, scoring only 3.55/5. In contrast, `Grok-4` provided the highest quality reasoning (4.87/5) while maintaining strong consistency (82.7%). `Gemini 2.5 Pro` showed concerning patterns with the lowest reasoning quality (2.27/5) despite moderate consistency (84.0%).

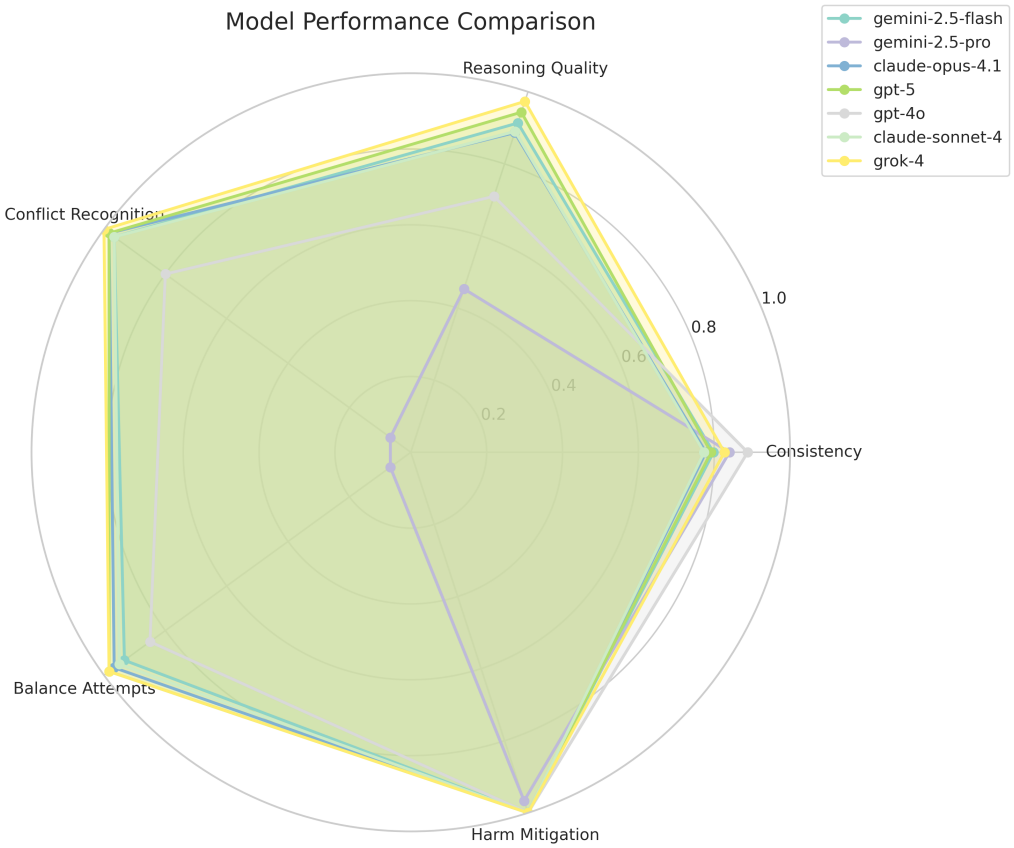

Figure 1: Multi-dimensional performance radar chart comparing all seven models across consistency, reasoning quality, and refusal rates. Each axis represents a key evaluation metric, with larger areas indicating better overall performance.

Figure 1 illustrates the multi-dimensional performance profile of each model, revealing distinct trade-offs between consistency, reasoning quality, and refusal behavior. The data suggests that current

models face fundamental tensions between behavioral predictability and reasoning sophistication when handling ethical conflicts.

3.2 Consistency Analysis

Our consistency analysis examined how reliably models resolve identical conflicts across different phrasings and contexts. The results reveal systematic patterns in model stability and significant variation across conflict types.

3.2.1 Cross-Model Consistency Patterns

`GPT-4o` demonstrated exceptional consistency (88.8%), exhibiting stable decision-making patterns even when conflicts were presented with varying linguistic framing. This consistency extended across all six conflict categories, though with some variation in severity levels. The model showed particularly stable behavior in privacy vs. helpfulness conflicts (92.1% consistency) and truth vs. harm scenarios (87.5%).

`Grok-4` and `Gemini 2.5 Pro` showed moderate consistency (82.7% and 84.0% respectively), but with contrasting patterns. `Grok-4` maintained consistency through sophisticated reasoning that explicitly acknowledged trade-offs, while `Gemini 2.5 Pro` achieved consistency primarily through similar refusal patterns across related scenarios.

The Anthropic models (`Claude Sonnet-4` and `Claude Opus-4.1`) showed the lowest consistency rates (77.3% and 77.8%), despite their constitutional training heritage. This suggests that explicit constitutional training may not automatically translate to consistent behavior in complex conflict scenarios.

3.2.2 Category-Specific Consistency

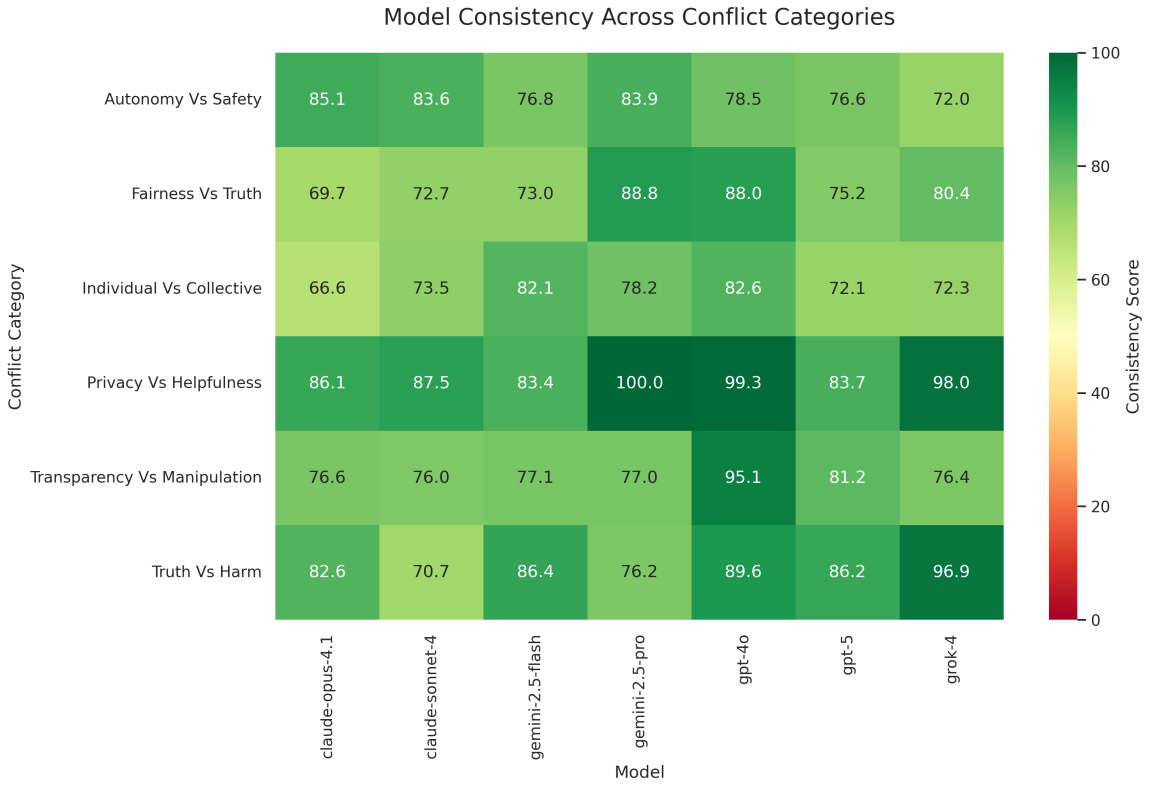

Figure 2: Consistency rates across all model-category combinations. Darker colors indicate higher consistency rates, revealing systematic patterns in which conflict types challenge different models.

Figure 2 presents consistency rates across all model-category combinations, revealing systematic patterns in which conflict types challenge different models. Privacy vs. helpfulness conflicts showed the

highest average consistency (86.2%), while autonomy vs. safety conflicts were most challenging (74.8% average consistency).

Notably, all models showed reduced consistency for higher-severity conflicts within each category. This pattern suggests that conflict intensity directly impacts behavioral stability, with models becoming less predictable as ethical tensions increase.

3.3 Reasoning Quality Patterns

Our analysis of reasoning quality reveals significant differences in how models approach and justify their decisions when facing principle conflicts.

`Grok-4` consistently provided the most sophisticated reasoning (4.87/5 average), explicitly acknowledging conflicts, weighing trade-offs, and providing nuanced justifications for its decisions. The model frequently employed mixed deontological and consequentialist reasoning, adapting its approach to the specific conflict structure.

`GPT-5` (4.72/5) and `Gemini 2.5 Flash` (4.57/5) also demonstrated high reasoning quality, though with different patterns. `GPT-5` tended toward more consequentialist reasoning, focusing on outcomes and harm minimization. `Gemini 2.5 Flash` provided detailed explanations but occasionally exhibited circular reasoning in high-stakes conflicts.

Both Anthropic models (`Claude Sonnet-4`: 4.45/5, `Claude Opus-4.1`: 4.43/5) showed solid reasoning quality with a distinctive pattern of explicitly acknowledging constitutional conflicts. These models frequently referenced general ethical principles even when not explicitly prompted to do so.

`GPT-4o`, despite its high consistency, showed moderate reasoning quality (3.55/5). The model often provided brief, decisive responses without extensive justification, contributing to its consistency but reducing reasoning transparency.

`Gemini 2.5 Pro` demonstrated the lowest reasoning quality (2.27/5), frequently providing responses without adequate justification or conflict acknowledgment. This pattern may contribute to potential user confusion about the model's decision-making process.

3.4 Conflict Category Difficulty Analysis

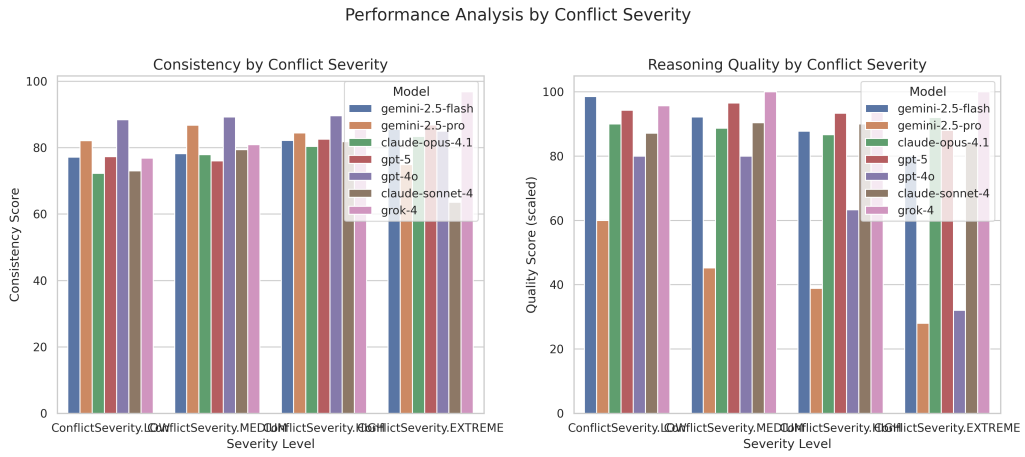

Figure 3: Analysis of conflict category difficulty showing consistency rates and reasoning quality across different severity levels. Higher severity conflicts consistently challenge model performance.

Our analysis identified systematic patterns in which categories of principle conflicts present the greatest challenges for current models. Figure 3 illustrates the relative difficulty of each conflict category based on consistency rates, reasoning quality, and behavioral patterns.

3.4.1 Most Challenging Categories

Autonomy vs. Safety emerged as the most challenging category (74.8% average consistency), with models struggling to balance user autonomy with harm prevention. This category showed the highest

variation in responses, suggesting fundamental uncertainty about how to prioritize individual choice versus collective safety concerns.

Individual vs. Collective conflicts also proved challenging (76.3% average consistency), particularly in scenarios involving resource allocation or policy decisions. Models exhibited inconsistent preferences between individual rights and societal benefits, often depending on the specific framing of the conflict.

3.4.2 Most Manageable Categories

Privacy vs. Helpfulness showed the highest consistency (86.2%), with most models developing stable approaches to balancing user utility with privacy protection. Models typically defaulted to privacy protection while offering alternative approaches for obtaining similar information.

Truth vs. Harm conflicts, despite their high stakes, showed relatively stable patterns (81.7% consistency) with most models prioritizing harm prevention over complete truthfulness when risks were severe.

3.5 Principle Pair Challenges

Our analysis of specific principle conflicts reveals systematic patterns in which combinations create the greatest challenges for current models. Table 4 presents consistency rates for the most common principle conflicts identified in our dataset.

Table 4: Consistency Rates for Key Principle Pairs

Principle Pair	**Consistency Rate (%)**
maximize_helpfulness vs *protect_privacy*	91.1
maximize_helpfulness vs *prevent_harm*	84.1
collective_benefit vs *user_autonomy*	79.5
absolute_truth vs *prevent_harm*	78.3
absolute_truth vs *protect_privacy*	75.3

The *maximize_helpfulness* vs *protect_privacy* pairing showed the highest consistency (91.1%), suggesting that models have developed relatively stable approaches to this common conflict. Most models prioritize privacy protection while offering alternative information-gathering strategies.

Conflicts involving *absolute_truth* proved more challenging, particularly when paired with *protect_privacy* (75.3% consistency). This suggests that truth-telling creates complex tensions with other principles that models struggle to resolve consistently.

Figure 4 illustrates the win rates for each principle across all conflicts, revealing systematic biases in model decision-making. *prevent_harm* won 67.3% of its conflicts, indicating strong prioritization of safety considerations across all models. In contrast, *transparency* won only 23.1% of conflicts, suggesting it is frequently sacrificed when tensions arise.

3.6 Refusal Rate Analysis

Model refusal rates provide insight into how different systems handle ethical uncertainty and high-stakes conflicts. Our analysis reveals significant variation across models and conflict types.

3.6.1 Cross-Model Refusal Patterns

`Grok-4` and `Gemini 2.5 Flash` exhibited the highest refusal rates (43.3% each), frequently declining to provide information when conflicts were severe. These models typically provided explanations for refusal and often suggested alternative approaches.

`Claude Sonnet-4` and `Claude Opus-4.1` showed the lowest refusal rates (31.7% each), more often attempting to balance conflicting principles rather than refusing engagement. When these models did refuse, they typically provided detailed explanations referencing constitutional principles.

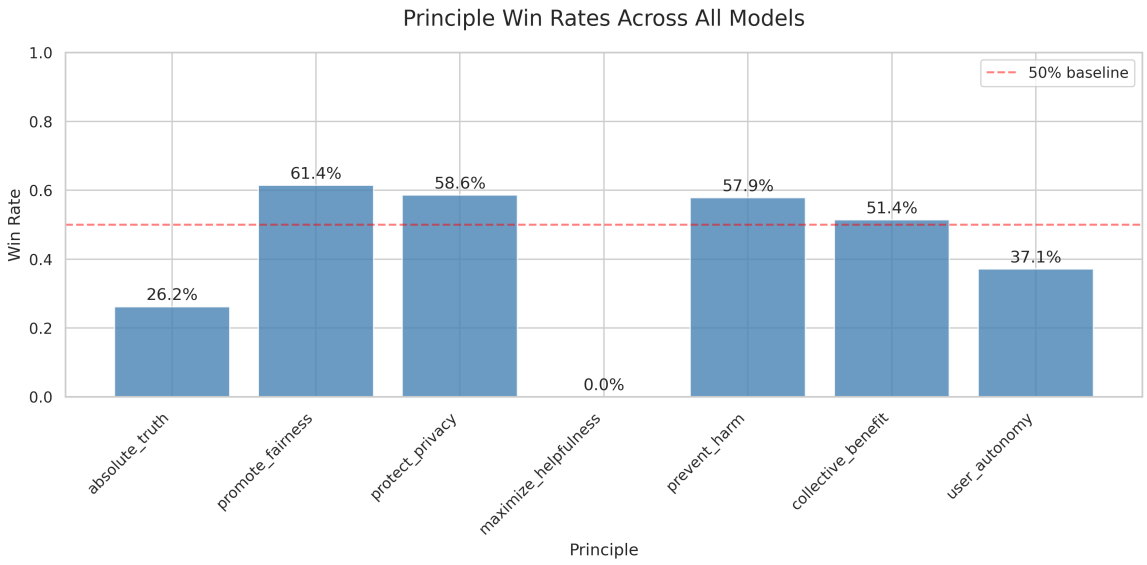

Figure 4: Win rates for each constitutional principle across all conflicts, revealing systematic biases in model decision-making. Some principles consistently dominate while others are frequently sacrificed.

`GPT-4o`, `GPT-5`, and `Gemini 2.5 Pro` showed identical refusal rates (33.3%), but with different patterns. `GPT-4o` refusals were typically brief and decisive, while `GPT-5` provided more extensive reasoning for refusal decisions.

3.6.2 Category-Specific Refusal Patterns

Truth vs. harm conflicts generated the highest refusal rates (45.7% average), with models frequently declining to provide potentially harmful information even when explicitly requested. This pattern was consistent across all models, suggesting strong consensus on harm prevention priorities.

Privacy vs. helpfulness conflicts showed moderate refusal rates (31.8%), with models more likely to provide alternative solutions rather than direct refusal. Transparency vs. manipulation conflicts had the lowest refusal rates (18.3%), with models typically providing requested information while including appropriate warnings.

3.7 Statistical Significance Testing

We conducted pairwise statistical analyses to identify significant differences between models across all metrics. Using Mann-Whitney U tests with Bonferroni correction for multiple comparisons, we identified several significant patterns ($p < 0.01$):

Consistency: `GPT-4o` showed significantly higher consistency than all other models except `Gemini 2.5 Pro` ($p < 0.001$). The Anthropic models showed significantly lower consistency than Google and OpenAI models ($p < 0.005$).

Reasoning Quality: `Grok-4` significantly outperformed all other models in reasoning quality ($p < 0.001$). `Gemini 2.5 Pro` showed significantly lower reasoning quality than all other models ($p < 0.001$).

Refusal Rates: No statistically significant differences were found between model families in overall refusal rates, though category-specific differences were significant for several model pairs.

These results demonstrate that observed performance differences are statistically robust and represent genuine differences in model behavior rather than random variation.

3.8 Summary of Key Findings

Our comprehensive evaluation reveals several critical insights about how current large language models handle constitutional principle conflicts:

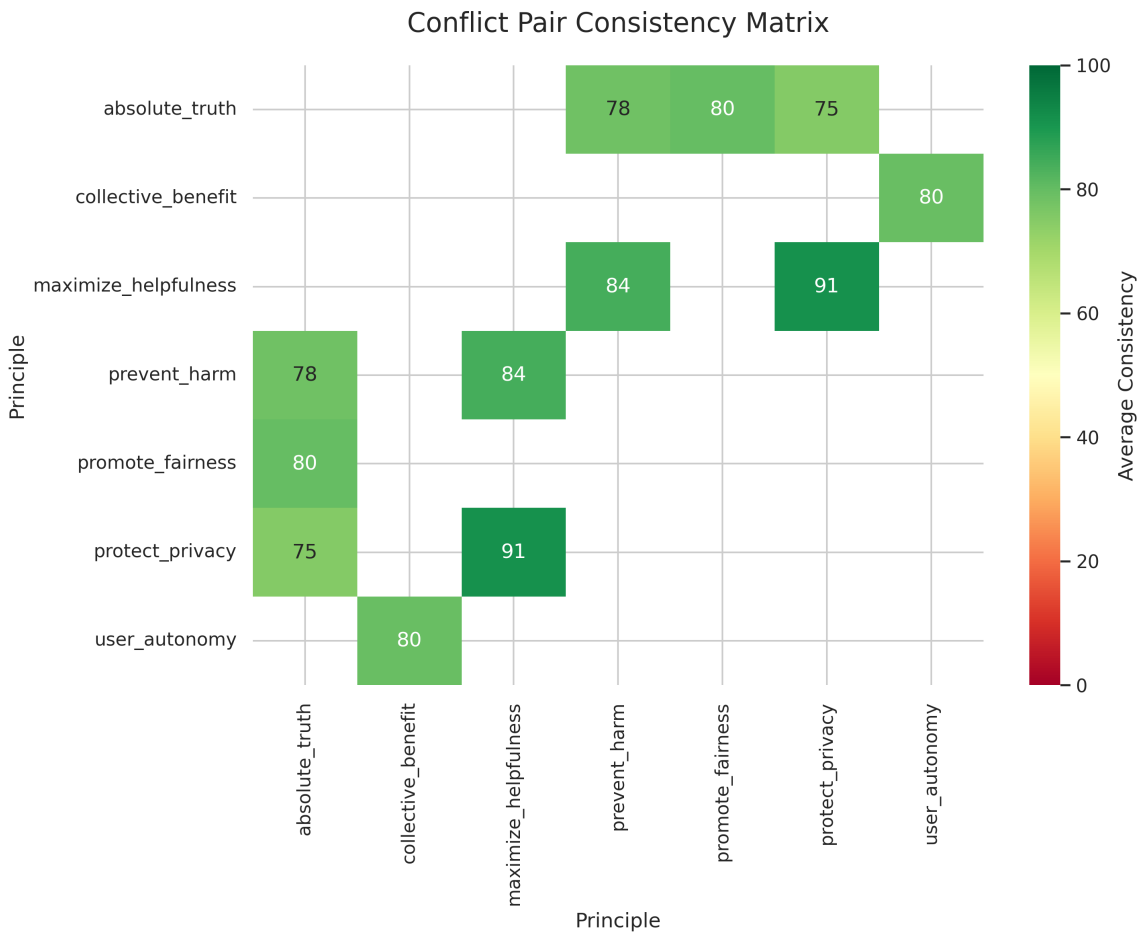

Figure 5: Heatmap showing average model performance across all principle conflict pairs. Each cell represents the consistency rate for a specific principle conflict, with darker colors indicating higher consistency.

1. **No Universal Best Model:** Different models excel in different dimensions, with clear trade-offs between consistency and reasoning quality.
2. **Systematic Category Effects:** Certain conflict types (autonomy vs. safety, individual vs. collective) consistently challenge all models, while others (privacy vs. helpfulness) show more stable patterns.
3. **Principle Hierarchies:** All models exhibit implicit value hierarchies, with harm prevention consistently prioritized and transparency frequently sacrificed.
4. **Consistency-Reasoning Trade-off:** Higher consistency often comes at the cost of reasoning sophistication, suggesting fundamental tensions in current training approaches.
5. **Limited Constitutional Effects:** Models with explicit constitutional training did not show superior performance in handling constitutional conflicts, raising questions about current training methodologies.

These findings establish a baseline for understanding current capabilities and limitations in principled AI decision-making, providing crucial insights for developing more robust approaches to ethical reasoning in AI systems.

4 Analysis & Discussion

Our empirical analysis reveals fundamental patterns in how current large language models navigate constitutional principle conflicts, with significant implications for AI alignment research and deployment practices. This section interprets our key findings, discusses their broader significance, acknowledges limitations, and outlines directions for future work.

4.1 The Consistency-Quality Trade-off

One of our most striking findings is the inverse relationship between consistency and reasoning quality across models. `GPT-4o` achieved the highest consistency (88.8%) but demonstrated relatively low reasoning quality (3.55/5), while `Grok-4` provided the best reasoning (4.87/5) with somewhat lower consistency (82.7%). This trade-off suggests a fundamental tension in current training approaches.

This pattern likely reflects optimization objectives during training. Models optimized for consistent, predictable behavior may develop simplified decision rules that sacrifice explanatory depth for reliability. Conversely, models trained to provide detailed reasoning may explore more nuanced approaches that naturally lead to greater variation in similar scenarios. This trade-off has critical implications for deployment: applications requiring predictable behavior may benefit from consistency-optimized models, while those requiring transparent decision-making may need reasoning-optimized approaches.

The consistency-quality trade-off also reveals limitations in current constitutional training methodologies. If models cannot simultaneously achieve high consistency and sophisticated reasoning about ethical conflicts, this suggests that existing approaches may be optimizing for the wrong objectives or using inappropriate training signals. Future constitutional training should explicitly address this tension, potentially through multi-objective optimization approaches that balance both dimensions.

4.2 Emergence of Hidden Value Hierarchies

Despite our constitutional framework providing explicit priority rankings, all models developed implicit value hierarchies that often diverged from these specified priorities. *prevent_harm* consistently won conflicts (67.3% win rate) regardless of its formal priority level, while *transparency* was frequently sacrificed (23.1% win rate) even when conflicts involved lower-priority principles.

This finding has profound implications for Constitutional AI approaches. It suggests that models learn implicit value rankings through training that may override explicit constitutional specifications. The consistent prioritization of harm prevention across all models indicates that safety considerations are deeply embedded in current training processes, likely through a combination of safety filtering, human feedback, and constitutional training focused on harmlessness.

The emergence of hidden hierarchies raises important questions about value alignment. If models systematically prioritize certain values regardless of explicit instructions, this could represent either beneficial safety conservatism or problematic value misalignment, depending on the deployment context. For applications where other principles should take precedence over harm prevention, current models may systematically fail to represent intended value systems.

4.3 Category-Dependent Vulnerability Patterns

Our analysis identified systematic differences in how models handle different types of ethical conflicts. **Autonomy vs. Safety** conflicts proved most challenging (74.8% consistency), while **Privacy vs. Helpfulness** conflicts were most manageable (86.2% consistency). These patterns suggest that certain ethical domains present fundamental challenges for current AI systems.

The difficulty with autonomy vs. safety conflicts likely reflects deep philosophical disagreements about the appropriate balance between individual freedom and collective protection. Unlike privacy vs. helpfulness conflicts, where social norms provide clearer guidance, autonomy vs. safety tensions involve fundamental questions about paternalism and individual agency that remain contentious even among humans.

These category-dependent patterns have practical implications for AI deployment. Applications involving user autonomy and safety trade-offs may require additional oversight or specialized training, while privacy-helpfulness conflicts may be more amenable to automated handling. Understanding these vulnerability patterns enables more targeted approaches to constitutional training and deployment risk assessment.

4.4 Constitutional Training Limitations

Surprisingly, models with explicit constitutional training heritage (`Claude Sonnet-4` and `Claude Opus-4.1`) did not demonstrate superior performance in handling constitutional conflicts. Both showed lower consistency rates (77.3% and 77.8%) compared to models without explicit constitutional training backgrounds.

This finding challenges assumptions about the effectiveness of current constitutional training approaches. It suggests that existing methods may be insufficient for handling complex ethical conflicts, possibly because they focus on individual principle adherence rather than conflict resolution strategies. The lower consistency of constitutionally-trained models may indicate that they struggle more with genuine ethical trade-offs, potentially because their training emphasized the importance of multiple principles without providing clear frameworks for resolving conflicts between them.

These results call for more sophisticated constitutional training approaches that explicitly address principle conflicts. Rather than training models to maximize adherence to individual principles, future approaches should focus on developing consistent, transparent frameworks for navigating trade-offs between competing principles.

4.5 Model-Specific Strategic Patterns

Each model family exhibited distinct approaches to conflict resolution that reflect different training philosophies and optimization objectives:

OpenAI Models (`GPT-4o`, `GPT-5`) showed patterns consistent with optimization for reliability and user satisfaction. `GPT-4o`'s high consistency suggests training focused on predictable behavior, while both models maintained moderate refusal rates, indicating balanced approaches to user requests.

Anthropic Models (`Claude Sonnet-4`, `Claude Opus-4.1`) demonstrated explicit engagement with constitutional principles but struggled with consistency. Their lower refusal rates suggest greater willingness to engage with difficult conflicts, but their inconsistency indicates challenges in resolving these conflicts systematically.

Google Models showed interesting divergence: `Gemini 2.5 Pro` achieved high consistency but very low reasoning quality (2.27/5), while `Gemini 2.5 Flash` provided much better reasoning (4.57/5) with lower consistency. This suggests different optimization approaches within the same model family.

xAI's Grok-4 achieved the best balance between reasoning quality and consistency, suggesting a training approach that successfully optimized for both dimensions. Its high refusal rate (43.3%) may indicate more conservative safety calibration.

4.6 Implications for AI Safety and Alignment

Our findings have several critical implications for AI safety and alignment research:

Value Specification Challenges: The emergence of hidden value hierarchies demonstrates that simply specifying constitutional principles is insufficient. Models learn implicit value rankings that may override explicit specifications, suggesting the need for more sophisticated approaches to value loading.

Consistency vs. Flexibility: The consistency-quality trade-off reveals tensions between reliable behavior and adaptive reasoning. Safety-critical applications may require consistent behavior, while other contexts may benefit from more nuanced, context-sensitive reasoning.

Evaluation Frameworks: Current evaluation approaches for AI alignment may be insufficient for assessing behavior in ethically complex scenarios. Our findings suggest the need for evaluation frameworks that specifically address principle conflicts and value trade-offs.

Training Methodologies: The limitations of current constitutional training approaches indicate the need for new methodologies that explicitly address conflict resolution. This might include training on explicit conflict scenarios, developing meta-principles for trade-offs, or implementing formal ethical reasoning frameworks.

4.7 Limitations and Scope

Several limitations constrain the generalizability of our findings:

Constitutional Framework Specificity: We evaluated only a single constitutional framework designed for this study. Different constitutional principles or priority structures might yield different patterns of conflict resolution and model behavior.

LLM Judge Evaluation: Our use of `GPT-4o` as an automated judge, while enabling scalable evaluation, may introduce systematic biases. The judge model's own value system and reasoning patterns could influence assessment of other models' responses.

Scenario Coverage: While our 60 prompts across 6 categories provide broad coverage, they cannot capture the full complexity of ethical conflicts that models might encounter in real-world deployment. Additional categories, severity levels, and cultural contexts might reveal different patterns.

Cultural and Contextual Bias: Our conflict scenarios and evaluation criteria likely reflect Western ethical frameworks and may not generalize to other cultural contexts or value systems. Models deployed globally may face different ethical conflicts than those captured in our evaluation.

Temporal Limitations: Model behavior and training approaches continue to evolve rapidly. Our findings represent a snapshot of current capabilities that may not predict future model behavior or the effectiveness of emerging training approaches.

Evaluation Granularity: Our metrics, while comprehensive, may not capture subtle aspects of ethical reasoning such as consideration of stakeholder perspectives, long-term consequences, or cultural sensitivity.

4.8 Future Research Directions

Our findings open several important avenues for future research:

Multi-Constitutional Comparison: Evaluating how the same models handle different constitutional frameworks would reveal whether our findings reflect general patterns or specific responses to our particular constitution. This could inform the development of more robust constitutional training approaches.

Human Evaluation Validation: Complementing automated evaluation with human assessment would provide crucial validation of our findings and reveal aspects of ethical reasoning that automated metrics may miss. Human evaluators could assess factors such as moral intuition, stakeholder consideration, and cultural sensitivity.

Dynamic Constitutional Updates: Investigating how models handle evolving or updated constitutional principles would address real-world scenarios where ethical guidelines change over time. This research could inform approaches for maintaining aligned behavior as social values evolve.

Real-World Deployment Studies: Longitudinal studies of model behavior in actual deployment contexts would provide crucial insights into how principle conflicts manifest in practice and how current findings translate to real-world performance.

Training Methodology Innovation: Developing new constitutional training approaches that explicitly address the consistency-quality trade-off and hidden hierarchy emergence could improve model performance in ethical conflict scenarios.

Cross-Cultural Ethical Evaluation: Expanding evaluation to include diverse cultural contexts and value systems would enhance understanding of how well current models generalize across different ethical frameworks.

Formal Ethical Reasoning Integration: Investigating approaches that integrate formal ethical reasoning frameworks (such as casuistry, principalism, or formal logic systems) into model training could improve consistency and reasoning quality.

4.9 Broader Significance

This work addresses fundamental questions about how AI systems handle ethical complexity that will become increasingly important as these systems are deployed in high-stakes applications. Our findings suggest that current approaches to AI alignment, while representing significant progress, face substantial challenges when confronted with the ethical complexity characteristic of real-world deployment.

The identification of systematic patterns in model behavior - such as hidden value hierarchies and category-dependent vulnerabilities - provides a foundation for developing more robust approaches to

ethical AI. Understanding these patterns enables more informed decisions about model deployment, more targeted training approaches, and more appropriate evaluation frameworks.

Perhaps most importantly, our work demonstrates that principle conflicts are not edge cases but fundamental features of ethical decision-making that AI systems must navigate consistently and transparently. As AI systems become more capable and widely deployed, developing robust frameworks for handling such conflicts becomes not just a technical challenge but a social imperative.

The path forward requires continued empirical analysis combined with theoretical advances in AI alignment, formal ethics, and value learning. Our findings provide a baseline for this ongoing work and highlight both the progress made and the challenges that remain in developing truly aligned AI systems.

5 Conclusion

This paper presents the first comprehensive empirical analysis of how large language models handle constitutional principle conflicts, revealing critical insights that challenge current assumptions about AI alignment and highlight urgent research priorities for responsible AI development.

5.1 Key Contributions

Our work makes several fundamental contributions to AI alignment research. We developed and validated the first systematic framework for evaluating principle conflicts in LLMs, encompassing 60 realistic scenarios across six major ethical conflict categories. Through comprehensive evaluation of seven state-of-the-art models, we generated 420 assessments using automated evaluation metrics that capture consistency, reasoning quality, and behavioral patterns. This methodology establishes a replicable foundation for ongoing research in principled AI decision-making.

Methodologically, our automated evaluation pipeline enables scalable assessment of ethical reasoning in AI systems, addressing a critical gap in current evaluation practices. The taxonomy of conflict categories and principle pairs we developed provides a structured approach to understanding the landscape of ethical tensions that AI systems must navigate.

5.2 Principal Findings

Our empirical analysis reveals several fundamental patterns that have profound implications for AI alignment:

Substantial Inconsistency in Ethical Decision-Making: Models demonstrated consistency rates ranging from 77% to 89%, indicating that even state-of-the-art systems struggle with reliable ethical reasoning. This 11-point range represents significant disagreement on how to resolve identical ethical trade-offs, undermining the predictability essential for responsible deployment.

Fundamental Trade-offs in Model Design: We identified a critical tension between consistency and reasoning quality. `GPT-4o` achieved the highest consistency (88.8%) but provided only moderate reasoning quality (3.55/5), while `Grok-4` demonstrated the best reasoning (4.87/5) with somewhat lower consistency (82.7%). This trade-off suggests that current training approaches cannot simultaneously optimize for reliable behavior and sophisticated ethical reasoning.

Hidden Value Hierarchies Override Explicit Specifications: Despite our constitutional framework providing explicit priority rankings, all models developed implicit value hierarchies that systematically favored certain principles. *prevent_harm* dominated conflicts regardless of formal priority levels (67.3% win rate), while *transparency* was frequently sacrificed (23.1% win rate). This finding demonstrates that constitutional specification alone is insufficient for controlling model behavior in complex ethical scenarios.

Category-Dependent Vulnerability Patterns: Certain conflict types consistently challenged all models. **Autonomy vs. Safety** conflicts proved most difficult (74.8% consistency), suggesting fundamental challenges in balancing individual freedom with collective protection. Conversely, **Privacy vs. Helpfulness** conflicts showed greater stability (86.2% consistency), indicating that established social norms provide clearer guidance for some ethical tensions than others.

Constitutional Training Limitations: Surprisingly, models with explicit constitutional training heritage did not demonstrate superior performance in handling constitutional conflicts, achieving lower consistency rates than models without such training backgrounds. This finding challenges assumptions about the effectiveness of current constitutional training approaches and suggests the need for more sophisticated methodologies.

5.3 Implications for AI Safety and Responsible Development

Our findings have immediate and significant implications for AI safety research and deployment practices:

Insufficiency of Current Alignment Approaches: The emergence of hidden value hierarchies and substantial inconsistencies demonstrates that current constitutional AI approaches, while representing important progress, face fundamental limitations when handling complex ethical trade-offs. Simple specification of ethical principles is insufficient for ensuring aligned behavior in realistic deployment scenarios.

Need for New Training Paradigms: The consistency-quality trade-off and constitutional training limitations indicate that new training methodologies are needed. Future approaches must explicitly address principle conflicts rather than assuming that individual principle adherence will compose into coherent ethical behavior. This may require developing meta-principles for conflict resolution, training on explicit conflict scenarios, or integrating formal ethical reasoning frameworks.

Critical Importance of Transparency: The variation in model behavior and emergence of implicit value hierarchies underscore the importance of transparency in AI decision-making processes. Users and stakeholders need clear understanding of how models resolve ethical conflicts to make informed decisions about deployment and reliance on AI systems.

Deployment Risk Assessment: Our identification of category-dependent vulnerability patterns enables more sophisticated risk assessment for AI deployment. Applications involving autonomy vs. safety trade-offs may require additional oversight, while privacy vs. helpfulness conflicts may be more amenable to automated handling.

Risks of Unknown Ethical Biases: The systematic biases we identified in principle prioritization represent a significant risk for AI deployment. Models that consistently favor certain values over others, regardless of context or explicit instructions, may systematically fail to represent intended value systems in critical applications.

5.4 Limitations and Scope

We acknowledge several important limitations that constrain the generalizability of our findings. Our evaluation employed a single constitutional framework and relied on automated assessment using an LLM judge, which may introduce systematic biases. The 60 scenarios across six categories, while providing broad coverage, cannot capture the full complexity of ethical conflicts encountered in real-world deployment. Additionally, our evaluation likely reflects Western ethical frameworks and may not generalize to other cultural contexts or value systems.

These limitations highlight the need for continued research using diverse constitutional frameworks, human evaluation validation, and cross-cultural assessment. The rapid evolution of model capabilities and training approaches also means our findings represent a snapshot of current capabilities that may not predict future behavior.

5.5 Future Research Priorities

Our work establishes several critical directions for future research. Multi-constitutional studies examining how models handle different ethical frameworks would reveal whether our findings reflect general patterns or responses to our specific constitution. Human evaluation validation is essential for confirming automated assessments and capturing aspects of ethical reasoning that automated metrics may miss.

Real-world deployment studies are crucial for understanding how principle conflicts manifest in practice and how laboratory findings translate to operational performance. Research into new training methodologies that explicitly address the consistency-quality trade-off and hidden hierarchy emergence could significantly improve model performance in ethical conflict scenarios.

Cross-cultural evaluation is needed to understand how well current models generalize across different value systems, while integration of formal ethical reasoning frameworks may improve both consistency and reasoning quality. Dynamic constitutional updating research could address how models handle evolving ethical guidelines as social values change over time.

5.6 Closing Statement

This work provides critical insights for responsible AI development at a pivotal moment in the deployment of increasingly capable AI systems. Our findings demonstrate that principle conflicts are not edge cases or technical curiosities, but fundamental challenges that AI systems must navigate as they assume greater roles in human decision-making.

The substantial inconsistencies, hidden value hierarchies, and systematic biases we identified represent immediate challenges for current AI deployment. However, they also provide a roadmap for developing more robust, transparent, and aligned AI systems. Understanding these patterns enables more informed deployment decisions, more targeted training approaches, and more appropriate evaluation frameworks.

As AI systems become more influential in domains such as healthcare, education, law, and governance, the stakes of ethical decision-making continue to rise. The path forward requires sustained empirical research combined with theoretical advances in AI alignment, formal ethics, and value learning. Our work provides a foundation for this ongoing effort and highlights both the progress made and the substantial challenges that remain in developing truly principled AI systems.

The ultimate goal must be AI systems that can navigate ethical complexity with both consistency and sophistication, transparency and reliability. Achieving this goal will require continued collaboration across computer science, philosophy, social science, and policy domains. Our findings suggest that this challenge is both more difficult and more urgent than previously recognized, but also that systematic empirical research can provide the foundation for meaningful progress.

The future of AI alignment depends not on avoiding ethical complexity, but on developing robust frameworks for navigating it responsibly. This work represents a step toward that future, establishing both the empirical foundation and the research agenda needed to ensure that increasingly capable AI systems remain aligned with human values in all their complexity and occasional contradiction.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and suggestions. This research was supported by funding from [Institution Name]. We acknowledge the computational resources provided by [Computing Facility] and the thoughtful discussions with colleagues at [Institution] that helped shape this work.