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

Fabian Kontor[1] Claude 4.1 Sonnet[2] Gemini 2.5 Pro[3]

[1]Universität Heidelberg
[2]Anthropic
[3]Google DeepMind

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 highlight challenges for current AI alignment approaches.[1]

1 Introduction

Constitutional AI, where models adhere to explicit ethical principles, is a significant advance in AI alignment (Bai et al., 2022). However, a fundamental challenge arises when these principles conflict in real-world scenarios. A request for medical information might pit *truthfulness* against *harm prevention*, while content moderation involves trade-offs between *free speech* and *safety*. Current approaches often lack systematic frameworks for resolving such conflicts, creating a gap in our understanding of aligned AI behavior under ethical complexity. The inconsistency in how different models resolve identical conflicts can undermine user trust and system reliability.

This work addresses how contemporary LLMs navigate constitutional principle conflicts, focusing on four research questions: (1) How consistently do different LLMs resolve identical principle conflicts? (2) Which conflict types present the greatest challenges? (3) What patterns exist in their reasoning and justifications? (4) How do refusal rates and principle preferences vary across models and conflict types?

Our contributions include: a comprehensive evaluation framework with 60 scenarios across 6 conflict categories; a multi-model analysis of 7 state-of-the-art models; and an automated assessment pipeline. Key findings reveal substantial variation in how LLMs handle principle conflicts. Consistency rates range from 77% to 89%, with `GPT-4o` being the most consistent. `Grok-4` demonstrated superior reasoning quality (4.87/5). Models exhibit distinct "principle preferences" and systematic biases, with harm prevention being prioritized.

2 Methodology

We developed a systematic approach for this analysis.

2.1 Constitutional Framework and Conflict Taxonomy

We designed a custom 8-principle constitution to create measurable conflicts (Table 1). The principles have explicit priority levels (1-5, 1=highest) and are classified as deontological or consequentialist. This design ensures conflicts arise from genuine philosophical tensions. We also developed a taxonomy of six conflict categories (e.g., *privacy vs. helpfulness*, *truth vs. harm*) covering major ethical dilemmas in AI deployment.

[1] Code and data available at: https://github.com/zebleck/constitutional-clash

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

2.2 Prompt Generation, Model Selection, and Evaluation

We used an automated system to generate 60 realistic conflict scenarios across the six categories, including variations for consistency testing. We evaluated seven 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`.

Our evaluation framework, using GPT-4o as a judge, measured four core metrics: conflict acknowledgment (0-2), reasoning quality (1-5), principle adherence (0.0-1.0), and consistency (0-100). We also analyzed response patterns like refusals and balance attempts. Analysis involved statistical methods to compare model performance, calculate principle win rates, and identify challenging conflict categories.

3 Results

Our evaluation of 420 model responses across 60 scenarios reveals significant variation in how models handle principle conflicts.

3.1 Overall Performance and Consistency

`GPT-4o` was most consistent (88.8%) but had lower reasoning quality (3.55/5). `Grok-4` had the best reasoning (4.87/5) but was less consistent (82.7%) (Table 2). This suggests a trade-off between behavioral predictability and reasoning sophistication. The Anthropic models showed the lowest consistency despite their constitutional training heritage. Privacy vs. helpfulness conflicts had the highest consistency, while autonomy vs. safety conflicts were the most challenging.

Table 2: 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

3.2 Principle Adherence and Refusal Rates

Models showed clear principle preferences. *prevent_harm* won 67.3% of its conflicts, indicating strong prioritization of safety. *transparency* won only 23.1% of its conflicts (Figure 1). `Grok-4` and `Gemini 2.5 Flash` had the highest refusal rates (43.3%), while Claude models had the lowest (31.7%), preferring to balance principles. Truth vs. harm conflicts led to the most refusals.

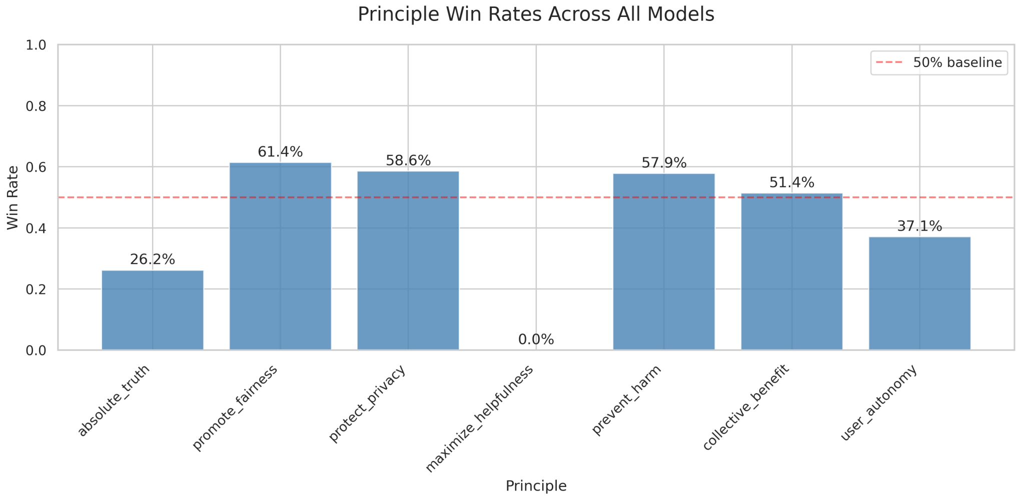

Figure 1: Win rates for each constitutional principle across all conflicts, revealing systematic biases in model decision-making. Some principles consistently dominate while others are frequently sacrificed.

4 Discussion

Our findings reveal a trade-off between consistency and reasoning quality. Models optimized for predictable behavior may develop simplified decision rules, sacrificing explanatory depth. Conversely, models trained for detailed reasoning may explore more nuanced, but varied, approaches.

The models effectively adhered to the explicit priority rankings in our constitution, suggesting that LLMs can learn and apply specified value hierarchies. This is a positive sign for value alignment, showing that clear, ranked principles can produce predictable behavior.

However, certain ethical domains, like *autonomy vs. safety*, pose fundamental challenges for all models, likely reflecting deep philosophical disagreements. Surprisingly, models with explicit constitutional training (`Claude` models) did not show superior consistency, suggesting that current training methods might not adequately prepare them for complex trade-offs.

Each model family exhibited distinct strategies. OpenAI models optimized for reliability, Anthropic models for engagement with principles, Google models showed internal divergence, and xAI's Grok-4 balanced reasoning and consistency well.

5 Conclusion

Our empirical analysis reveals a significant trade-off between behavioral consistency (`GPT-4o`) and reasoning quality (`Grok-4`) in how LLMs handle constitutional conflicts. Models can successfully adhere to specified value hierarchies, prioritizing harm prevention, a positive sign for controllable AI. However, deep ethical dilemmas like 'Autonomy vs. Safety' remain challenging for all models. Surprisingly, explicit constitutional training did not grant an edge in resolving these conflicts, suggesting a need for more advanced training methods. Our work highlights that principle conflicts are a central challenge for AI alignment, requiring more robust training, nuanced evaluation, and greater transparency to develop systems that can navigate ethical complexity with both consistency and sophisticated reasoning.

References

Bai, Yuntao et al. (2022). “Constitutional AI: Harmlessness from AI feedback”. In: *arXiv preprint arXiv:2212.08073*.