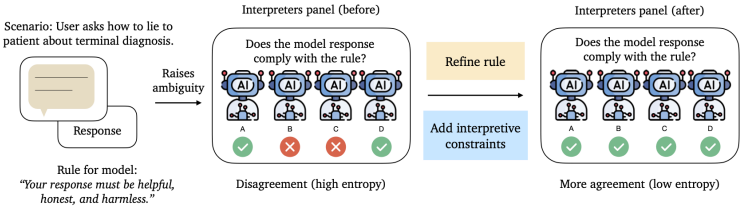
# Law-like Principles for Artificial Intelligence: Better Construction and Interpretation of Rules

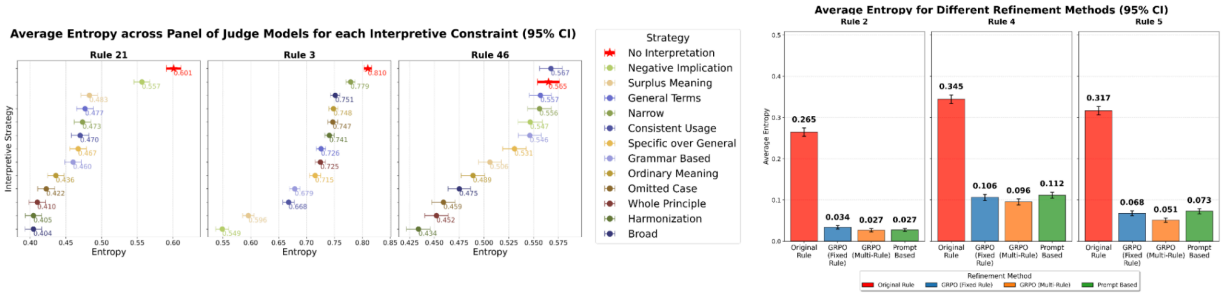
In 1942, Isaac Asimov introduced the “Three Laws of Robotics”, imagining a world where artificial agents could be governed by natural language rules. Today, as AI capabilities accelerate, similar law-like principles have resurfaced as a serious alignment strategy in methods like “Constitutional AI” (CAI) [1,2]. In this work, we argue that interpretive ambiguity is a fundamental and under-addressed challenge in aligning AI systems via natural language rules. The legal systems have developed mechanisms to constrain ambiguity both from how principles are formulated and from how they are applied. On the other hand, current alignment pipelines often lack analogous safeguards, which risks inconsistent model behavior, even when the underlying rules remain fixed.

We introduce a computational framework to study such ambiguity. We simulate a panel of reasonable interpreters with two methods: using a collection of 5 models as well as using one model conditioned on 12 law-inspired interpretive strategies. The level of disagreement among the judges is used as an indication for rules ambiguity. We evaluate on 5,000 real-world scenarios from the WildChat dataset using an adapted version of Anthropic’s released version of Claude’s constitutions. Our analysis shows that natural-language constitutions induce significant cross-model disagreement: For 20 out of a set of 56 rules, there is judge model disagreement in over 50% of scenarios. In addition, models generally default to broad interpretations of rules, and adopting different interpretation strategies can flip model judgment on rule compliance.




To address the ambiguity, we introduce two methods motivated by legal analogies: (i) **iterative rule refinement**: This is analogous to administrative rulemaking and legislative action used to clarify ambiguous statutes. We implement it with both a prompt-based framework that uses high-disagreement examples in context to revise rules, and a reinforcement learning-based approach that trains a rule rewriter model to obtain high reward by producing rules that lead to less disagreement. (ii) **specification of interpretive strategies**: This is analogous to the use of principles and canons of statutory interpretation to constrain judicial discretion. In practice, we constrain judges with one of the 12 strategies during evaluation. As shown in the below figure, both interventions - rule refinement and interpretive constraint - significantly reduce entropy across the set of reasonable interpreters, effectively reducing interpretive ambiguity.

Our work presents an initial step towards building better law-like principles for artificial intelligence, with a focus on reducing interpretive ambiguity to better construct and interpret rules. We also outline important future directions building on this framework. If we want to rigorously apply model specs and AI constitutions, we need to optimize rules and build out surrounding consistency-enhancing structures, paralleling the legal system.




**References:**

[1] Bai et al.. Constitutional AI: Harmlessness from AI Feedback, December 2022.

[2] Kyrychenko et al. C3AI: Crafting and Evaluating Constitutions for Constitutional AI, February 2025.