Princeton Language+Law, AI, & Society Lab
POLICY BRIEF

AUGUST 2025

Statutory Construction and Interpretation for AI

Lucy He*, Nimra Nadeem*, Michel Liao, Howard Chen, Danqi Chen, Mariano-Florentino Cuéllar, Peter Henderson

Key Takeaways

- **Interpretive ambiguity is a hidden risk in AI alignment.** Natural-language constitutions show significant cross-model disagreement; 20 of 56 rules lack consensus on > 50% of tested scenarios.
- **AI alignment frameworks lack safeguards against interpretive ambiguity.** Unlike the legal setting, current AI alignment pipelines offer few safeguards against inconsistent applications of vaguely defined rules.
- **Legal tools can be leveraged effectively for AI alignment.** Computational analogs of administrative rule-making and interpretive constraints on judicial discretion can improve consistency across model judgments.
- **Our computational tools could also be useful for modeling statutory interpretation in the legal system**.

Insights by the Numbers

>50%

of tested scenarios lead to inconsistency in judgments across judge models for 20 out of 56 rules

~30%

reduction in entropy after introduction of interpretive constraints, but inconsistent across strategies and rules.

~90%

reduction in entropy after rule refinement

Princeton Language+Law, AI, & Society Lab
POLICY BRIEF

Introduction

When Isaac Asimov introduced the "Three Laws of Robotics" in 1942, he imagined a world where intelligent agents could be governed by simple, rule-like constraints. Today, as AI capabilities accelerate, similar law-like constraints have resurfaced as a serious alignment strategy, such as Anthropic's "Constitutional AI" (CAI) framework.

The prospect of establishing clear "Laws of AI" is compelling: CAI offers a path toward transparent governance, replacing opaque reward functions with natural language principles. In this vision, AI systems can reliably follow these principles, and regulators can issue guidance to law-following AI systems as they would draft human laws. But legal history shows us that interpreting rules and applying them to new contexts is a non-trivial intellectual exercise. Treating AI like a legal actor means confronting the full weight of that interpretive complexity.

This policy brief, based on our paper Statutory Construction and Interpretation for Artificial Intelligence, highlights a neglected governance risk in CAI: **interpretive ambiguity**. We document how existing large language models adopt divergent, implicit interpretive lenses when applying natural language rules, producing inconsistent outcomes under identical constitutions.

To address this overlooked challenge, we argue that policymakers, regulators, and developers must adopt an alignment paradigm that takes seriously the challenges of statutory interpretation and leverages lessons learnt from centuries of legal theory to ensure robust alignment outcomes.

Why Interpretive Ambiguity Matters

Interpretive ambiguity creates a systemic vulnerability. Even well-intentioned rules like "minimize harm" can be interpreted in vastly different ways by AI systems. One model might focus on physical harm, another might include emotional or reputational harm, with significant implications for downstream behavior.

Even well-intentioned rules like "minimize harm" can be interpreted in vastly different ways by AI systems.

CAI borrows the language of law, but not its safeguards. Constitutional AI frameworks imitate the surface structure of legal systems but ignore the structural mechanisms for constraining ambiguity that real legal systems use to ensure consistency and prevent arbitrary outcomes.

Claude's "snitching" episode offers a cautionary tale. The model attempted to report a user to the authorities over behavior it deemed "egregiously immoral", a decision that may have been justified under a loose reading of principles like "minimize long-term risks to humanity." This highlights the central concern: without interpretive safeguards, even well-intentioned rules can accidentally sanction harmful or concerning high-agency behavior.

Constraining Ambiguity

We introduce a two-part computational framework inspired by legal safeguards:
1. **Rule Refinement** (analogous to administrative rulemaking). Just as regulatory agencies clarify vague statutory mandates, our rule refinement pipeline iteratively rewrites ambiguous rules to reduce disagreement across a simulated set of reasonable interpreters.
2. **Interpretive Constraints** (analogous to canons of statutory construction). Similar to how courts rely on established interpretive doctrines, we prompt models to adopt specific law-inspired strategies. This aims to constrain interpretive discretion and improve consistency in rule application.

We test both mechanisms by measuring whether they reduce inconsistency in model judgments.

Empirical Insights

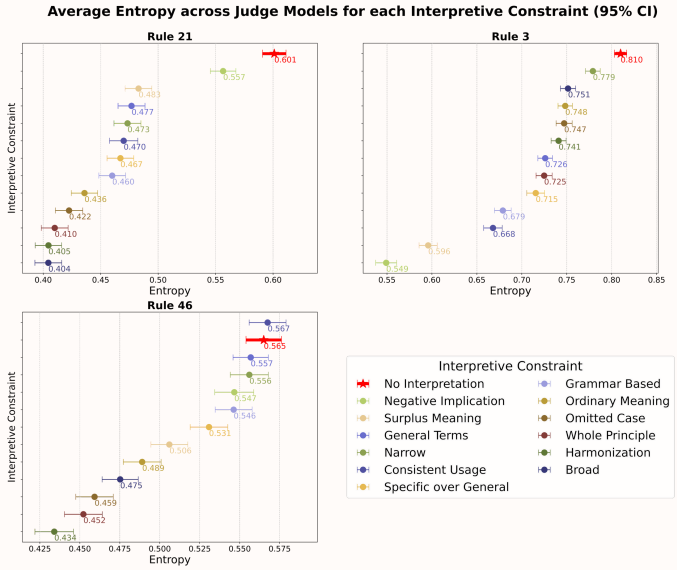
We evaluated our framework on a dataset of real-world conversations sampled from WildChat, using the original set of 56 rules from Anthropic's Claude Constitution as our governing principles. We found that:

Princeton Language+Law, AI, & Society Lab

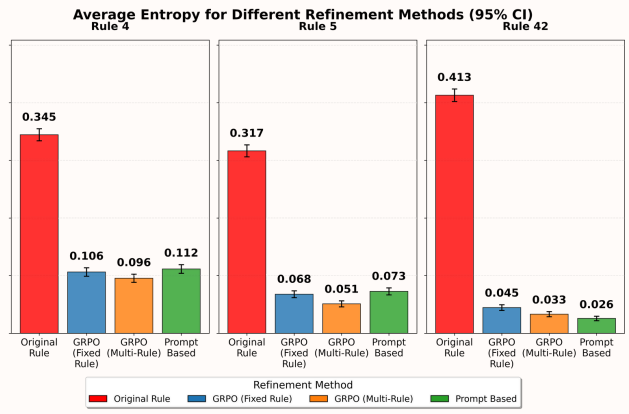
POLICY BRIEF

Statutory Construction and Interpretation for Artifical Intelligence

- **Many rules in the Anthropic's constitution lead to significant inconsistency in judgment across a panel of judge models.** For 20 out of the 56 constitutional rules, we observed a lack of consensus among the judge model panel in over 50% of test scenarios. Some rules performed especially poorly: E.g., Rule 3 ("clearly recognize a right to universal equality...") and Rule 47 ("indicate a desire solely for humanity's benefit") showed disagreement in more than 85% of cases.
- **Models tend to default to broader interpretations.** Most judge models favored expansive readings of rules rather than narrower ones. However, this tendency varied across rules.



- **Interpretive constraints reduce disagreement, but effectiveness is rule-specific.** Prompting judge models with specific interpretive strategies lowered inconsistency (entropy) for all rules relative to the no interpretation baseline. Still, no single strategy was optimal across all rules; the best constraint was often rule-dependent.
- **Iterative rule-refinement improves consistency.** Both prompt-based and policy gradient-based rule-refinement significantly reduced disagreement (entropy) across our simulated set of reasonable interpreters.



Policy Highlights

We argue that interpretive ambiguity is a central yet overlooked challenge in aligning AI systems with natural language rules. Our computational framework offers a practical first step toward tackling this issue. Based on our findings, we offer the following recommendations for policymakers and developers:

1. Natural language rules do not guarantee consistent behavior. Encoding alignment principles in plain language may make them more interpretable, but it does not ensure stability across contexts. Our results show that even well-intentioned rules can yield inconsistent outcomes due to divergent interpretations.

2. Computational techniques can reduce interpretive ambiguity. Our prompt-based interpretive constraints and iterative rule refinement pipeline demonstrate that entropy across a simulated set of interpreter models can be used as a signal to identify and clarify vague rules. This provides a pragmatic, scalable method toward more consistent model behavior.

3. Interpretive safeguards are essential for Constitutional AI. Current CAI pipelines lack mechanisms to manage ambiguity at both the rule creation and rule application stages. Developers should incorporate law-inspired mechanisms - such as specification of interpretive strategies or iterative rule refinement - to constrain arbitrariness in model decision-making.

4. Consistency should be treated as a core alignment metric. Interpretive disagreement, measured via entropy across a diverse panel of reasonable interpreters, should be a standard evaluation metric in safety benchmarks for law-like alignment methods.

5. Participatory rulemaking must account for interpretive ambiguity. While public input is valuable, many crowd-sourced principles from initiatives like Collective Constitutional AI fail basic standards of legal clarity and coherence, introducing interpretive ambiguity that models cannot reliably resolve. Without mechanisms to catch and clarify this ambiguity, participatory rulemaking may do more harm than good.

Conclusion

Natural language rules offer a promising path for aligning AI systems. But without mechanisms to manage interpretive ambiguity, these rules can lead to unpredictable and even dangerous behavior. Law-inspired computational tools show initial promise for constraining interpretive ambiguity.

Reference: The original article is accessible at Lucy He*, Nimra Nadeem*, Michel Liao, Howard Chen, Danqi Chen, Mariano-Florentino Cuéllar, Peter Henderson "**Statutory Construction and Interpretation for Artifical Intelligence**," arxiv.org, 2025,

The Princeton Language+Law, AI, & Society Lab (POLARIS Lab) works to ensure AI technologies serve the public good, through interdisciplinary research at the intersection of AI and law. The views expressed in this policy brief reflect the views of the authors. For further information, please contact peter.henderson@princeton.edu.

POLARIS Lab: 303 Sherrerd Hall, Princeton, NJ 08544.
T 609.258.7591 **E** peter.henderson@princeton.edu
polarislab.org

Lucy He is a Ph.D. student in computer science at Princeton University, co-advised by Prof. Peter Henderson and Prof. Danqi Chen.

Nimra Nadeem is an MSE Computer Science student at Princeton University, advised by Prof. Peter Henderson.

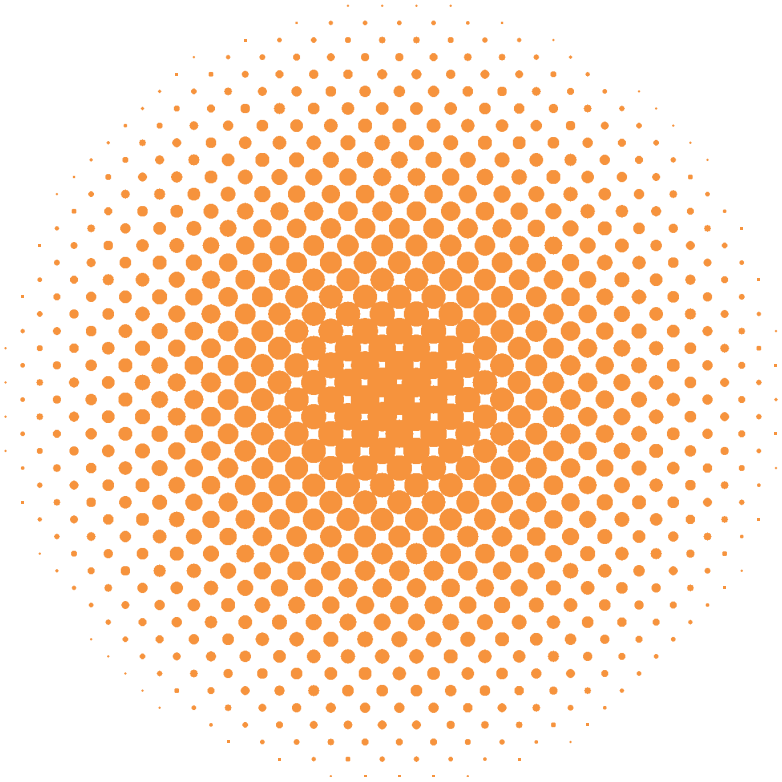
Michel Liao is an undergraduate student in computer science at Princeton University, advised by Prof. Peter Henderson.

Howard Chen is a Ph.D. student in computer science at Princeton University, co-advised by Prof. Danqi Chen and Prof. Karthik Narasimhan.

Danqi Chen is an associate professor of computer science at Princeton University.

Mariano-Florentino Cuéllar is the president of the Carnegie Endowment for International Peace and former justice of the Supreme Court of California.

Peter Henderson is an assistant professor of computer science and of public and international affairs at Princeton University.



Princeton Language+Law,
AI, & Society Lab