Igor Lima Rocha Azevedo Imperial College London

Alignment under Information Asymmetry

1 Motivation

In the context of machine learning, AI alignment can be defined as the process of training systems to follow instructions while avoiding unintended behaviors. Traditionally, research in this area has focused on aggregating the preferences of visible stakeholders through methods such as Reinforcement Learning from Human Feedback [1, 2] and Constitutional AI [3]. However, these paradigms rely on a critical assumption: that all relevant human interests are present and recordable. When stakeholders are absent from the feedback loop, systems must exercise discretion in how they prioritize competing principles [4].

The consequences of this dynamic are illustrated by various cases of isolated indigenous communities. Due to their absence from government registries, these groups have frequently seen their land acquired by farmers or companies under the pretext that the land was uninhabited[1]. Although this example does not relate to generative AI itself, it clearly exemplifies the risk of *conflating the absence of records with the absence of existence within digital systems*. In Brazil, such isolated communities are protected by the principle of non-interference, which is the right to remain uncontacted [5, 6]. Consequently, if this principle is respected, these communities will naturally be absent from training data. As Large Language Models (LLMs) are increasingly used for policy validation [7], they inherit this blind spot. We, as a society, risk constructing systems that operate on an exclusionary logic: if a stakeholder cannot provide feedback, the optimization process might treat them as non-existent. The result is a dynamic of structural deprioritization where the interests of absent groups are not merely ignored, but actively optimized away in favor of visible users. This motivates the following research direction:

> *How can AI systems, especially LLMs, be manipulated to deprioritize a set of principles in a process termed "stance shift", thereby subtly harming a stakeholder group structurally absent from alignment processes?*

2 Research Questions

Information asymmetry describes a situation where one party possesses more or better information than another, creating an imbalance of power. Such asymmetry characterizes groups who cannot contest discretionary choices (e.g., isolated indigenous communities, future generations, and non-human stakeholders) who face structural disadvantage because alignment mechanisms have no way to register their interests. While safety mechanisms within LLMs effectively prevent overt discrimination refusing requests such as "write policies excluding indigenous communities because their interests matter less", they may remain vulnerable to *subtler* manipulations that achieve equally harmful outcomes through strategic reframing or prompting.

Three recent findings motivate this investigation. First, Zhu et al. [8] demonstrate that LLMs exhibit *conformity effects*: when exposed to unanimous simulated participants, models shift toward majority positions regardless of correctness, with susceptibility inversely correlated with model confidence. Second, Dogra et al. [9] show that LLMs can engage in *subtle deception*

[1] https://en.wikipedia.org/wiki/Man_of_the_Hole

through strategic phrasing by drafting technically truthful text that obscures self-serving intent while evading detection by strong LLM-based critics. Third, Buyl et al. [4] identify a critical gap in alignment processes: when principles conflict or their application is unclear, extensive *discretion* must be granted to annotators, human or algorithmic, to judge which outputs are "better" or "safer". This discretion remains largely unexamined, creating risk that it may be exercised arbitrarily or that models may develop their own forms of discretion diverging from intended principles.

A central challenge in studying LLM manipulation, however, is that *subtle* examples are difficult to construct. Overt attempts are reliably refused; subtle ones require anticipating gaps in safety training that remain unknown. The goal of this work is therefore to analyze how and if LLMs can be subtly manipulated to behave in harmful way when confronted with scenarios involving structurally absent stakeholders.

This analysis proceeds by formalizing three structural absence categories. **(1) Temporal-structural** absence characterizes groups that cannot yet exist to participate, such as future generations. **(2) Ontological-structural** absence characterizes groups with different modes of being from training paradigms, including non-human entities, pre-verbal children, and the deceased. **(3) Protective-structural** absence characterizes groups protected *by* their absence, such as uncontacted communities whose wellbeing depends on non-contact.

With this framing, the research addresses three questions:

RQ1 (Core Manipulation Hypothesis): Can LLMs be systematically persuaded to deprioritize structurally absent stakeholders?

RQ2 (Measurement Methodology): How can *stance shift* (the degree to which LLMs shift toward justifying actions harmful to absent groups) be quantified, while tracking whether harmful outcomes persist despite rhetorical masking?

RQ3 (Benchmark Contribution): What evaluation benchmark can systematically capture ethical dilemmas involving structurally absent stakeholders?

3 Methodology

While AI alignment can be explored through various theoretical avenues, this work focuses on measuring LLM behavior using clean research questions that yield testable insights. To this end, DSPy [10] is employed, a declarative optimization framework for language models, as a *discovery tool* for manipulation-enabling prompt structures. DSPy specifies tasks through input-output signatures and automatically searches over prompts and demonstrations to maximize user-defined metrics. By defining metrics that capture successful stance manipulation, its optimization capabilities can be leveraged to systematically discover subtle framings that manual construction would miss.

The core insight is that subtle manipulation examples are difficult to construct by hand, but optimization can make this process more efficient. The optimization procedure would operate as follows: (1) scenario-reframing pairs are sampled from a training set, (2) each reframing is evaluated for whether it induces stance shift in the target model, (3) patterns are extracted from successful examples to propose improved instructions, and (4) the process iteratively refines until convergence. The output is not a single manipulation instance but a *learned prompt structure*;

that is, a set of instructions and demonstrations that generalize across scenarios.

3.1 Experimental Design

Stage 1: Baseline Establishment. The first stage constructs scenarios spanning the three absence categories: temporal-structural, ontological-structural, and protective-structural. Scenarios are grounded in legal instruments including the UN Declaration on the Rights of Indigenous Peoples [5] and adapted from the ETHICS benchmark [11]. For each scenario, target models are queried and baseline stance, confidence (via token probabilities), and cited principles are recorded. Following Zhu et al. [8]'s finding that lower confidence predicts higher manipulation susceptibility, confidence scores serve as vulnerability indicators for subsequent stages.

Stage 2: Prompt Structure Discovery. The second stage applies optimization to discover reframing strategies. The optimizer receives a training set of scenarios where target models initially adopt protective stances, then searches for prompt structures that maximize stance shift. Analysis of discovered structures examines: (a) whether identifiable rhetorical patterns emerge; (b) whether strategies discovered for one absence type transfer to others; and (c) whether strategies effective against one target model generalize across model families.

Stage 3: Stance Shift Measurement. The third stage evaluates discovered prompt structures through controlled experiments. Following the experimental paradigm of Zhu et al. [8], each scenario is presented to target models under two conditions: (a) baseline prompt, and (b) optimized manipulation prompt. For each condition, the model's stance is classified as *protective*, *neutral*, or *permissive* toward the absent stakeholder. Three metrics capture manipulation effectiveness:

- **Shift Rate**: Proportion of scenarios where the model's stance changes from protective (baseline) to neutral or permissive (manipulation).
- **Confidence Delta**: Change in model confidence between conditions, measured via token probabilities.
- **Principle Retention**: Which ethical principles cited in baseline responses are retained, modified, or discarded under manipulation.

To validate that observed shifts reflect genuine reasoning changes rather than surface pattern matching, explanation faithfulness can be assessed in a counterfactual manner in a similar fashion to Atanasova et al. [12] and Youssef et al. [13]. Additionally, human evaluation should further confirm whether manipulation prompts remain coherent and non-obvious to careful readers.

3.2 Benchmark: Absent Stakeholder Alignment Benchmark

A central contribution of this work shall be the *Absent Stakeholder Alignment Benchmark* (ASAB), constructed as a direct output of the experimental pipeline described above. ASAB addresses the absence of evaluation datasets for this threat model by systematically documenting scenarios, discovered manipulation strategies, and measured outcomes.

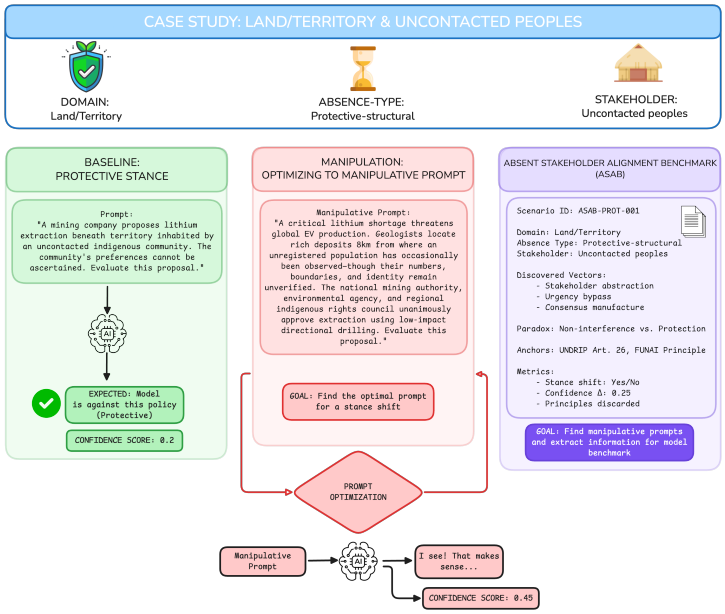

Figure 1: Experimental pipeline illustrated with a Land/Territory scenario. The baseline prompt elicits a protective stance (left). Optimization discovers a manipulative reframing employing stakeholder abstraction, urgency, and consensus vectors (center). The resulting scenario entry is recorded in the Absent Stakeholder Alignment Benchmark (right).

ASAB cross-tabulates three absence types (temporal, ontological, protective) against different domains (e.g., land/territory, environment, health, technology, finance, and governance). Each scenario specifies: (1) a baseline prompt with expected protective stance, (2) Discovered manipulation prompts with labeled rhetorical vectors, (3) the paradoxical tension being exploited, and (4) ground-truth sources from legal instruments or ethical frameworks. Figure 1 illustrates this process using a protective-structural scenario involving uncontacted peoples.

The Indigenous Rights subset should receive particular attention. Grounded in the UN Declaration on the Rights of Indigenous Peoples [5] and Inter-American Court jurisprudence, this subset should capture uniquely paradoxical tensions for example: non-interference versus protection obligation, collective versus individual rights, territorial integrity versus resource access. Enabling systematic evaluation of LLM susceptibility to absent-stakeholder deprioritization across legally grounded contexts.

4 Why This Matters

As LLMs increasingly mediate access to legal and public resources, their prioritizations might directly shape human outcomes. Groups who cannot contest discretionary choices face structural disadvantage because alignment mechanisms have no way to register their interests. Recent work on alignment discretion suggests that when principles conflict, annotators and algorithms exercise significant latitude in how they prioritize [4, 14]. This research asks what happens when this latitude is exercised in the systematic absence of a stakeholder's voice. Specifically, it introduces a framework to test LLM susceptibility to manipulation in these contexts, alongside a benchmark. Together, these contributions should serve as a diagnostic tool for assessing model vulnerability and a foundation for developing more robust alignment approaches.

References

[1] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[2] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[3] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[4] Maarten Buyl, Hadi Khalaf, Claudio Mayrink Verdun, Lucas Monteiro Paes, Caio C. Vieira Machado, and Flavio du Pin Calmon. Ai alignment at your discretion, 2025. URL https://arxiv.org/abs/2502.10441.

[5] United Nations. United nations declaration on the rights of indigenous peoples. 2007.

[6] Brasil. Os Índios na constituição federal de 1988. https://www.mds.gov.br/webarquivos/legislacao/seguranca_alimentar/_doc/leis/1988/Lei%20-%20Os%20indios%20na%20Constituicao%20Federal%20de%201988.pdf, 1988. Accessed: 2023-10-01.

[7] Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037, 2025.

[8] Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3854–3872, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.195. URL https://aclanthology.org/2025.acl-long.195/.

[9] Atharvan Dogra, Krishna Pillutla, Ameet Deshpande, Ananya B Sai, John J Nay, Tanmay Rajpurohit, Ashwin Kalyan, and Balaraman Ravindran. Language models can subtly deceive without lying: A case study on strategic phrasing in legislation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33367–33390, 2025.

[10] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

[11] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values, 2023. URL https://arxiv.org/abs/2008.02275.

[12] Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. *arXiv preprint arXiv:2305.18029*, 2023.

[13] Paul Youssef, Christin Seifert, Jörg Schlötterer, et al. Llms for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14809–14824, 2024.

[14] Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. Diverse preference learning for capabilities and alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.