# THE ROLE OF IDEOLOGICAL FEEDBACK LOOPS IN LLM OUTPUTS AND PUBLIC OPINION

A PREPRINT

Matthew Long[1] and Assisted by OpenAI GPT-4[2]

[1]Yoneda AI Research Lab
[2]Language Modeling Division

May 20, 2025

## ABSTRACT

Large Language Models (LLMs) have become increasingly influential in shaping public discourse, education, and ideological orientations through their outputs. This paper investigates the formation and implications of ideological feedback loops between LLM outputs and public opinion. We analyze how model training data, user reinforcement mechanisms, and content moderation pipelines reinforce or dilute ideological perspectives. The study explores consequences for democratic deliberation, epistemic diversity, and the potential entrenchment of dominant ideologies through automated systems. We propose a series of recommendations for mitigating bias propagation and fostering pluralism in AI-mediated discourse.

## Contents

## 1. Introduction

Large Language Models (LLMs) such as GPT-4 and Claude represent a major inflection point in the generation and mediation of knowledge. These models, trained on vast datasets of human-produced text, now act as intermediaries for millions of users seeking answers, insights, and narratives. This paper investigates how these models, due to their training and interaction mechanisms, participate in and possibly exacerbate ideological feedback loops that shape public opinion.

## 2. Background: Language Models and Societal Influence

The rise of LLMs coincides with increased reliance on algorithmic content curation in digital society. Platforms like search engines, social media, and personal assistants use these models to respond to user queries and generate content. Studies have shown that algorithmically curated content can significantly affect user beliefs and behavior, raising concerns about bias, representation, and the reproduction of dominant ideologies.

## 3. Feedback Loops: Definitions and Theoretical Framework

A feedback loop in this context refers to a recursive cycle where model outputs influence public sentiment, which in turn influences the data fed back into future models or reinforcement signals. Ideological feedback loops occur when certain narratives or belief systems are repeatedly affirmed by both the model and the users, creating self-reinforcing systems of thought. We draw upon control theory, cybernetics, and media studies to define and analyze these loops.

## 4. Ideological Embedding in Training Data

Training datasets for LLMs reflect the biases and worldviews of their source material. Because large-scale corpora are drawn from publicly available sources like Wikipedia, news sites, and online forums, they overrepresent certain geographies, classes, and ideological slants. This introduces asymmetric exposure to ideas, where dominant liberal, capitalist, or technocratic views may be overamplified, while dissenting, minority, or indigenous perspectives are marginalized.

## 5. Reinforcement Through Interaction and Fine-Tuning

Once deployed, LLMs are refined via reinforcement learning with human feedback (RLHF), which rewards outputs deemed helpful, harmless, and honest. However, the evaluators themselves, often guided by company policy or prevailing norms, may introduce additional bias. Users reward answers they agree with, unintentionally training the model to replicate prevailing views. This behavioral reinforcement accelerates the alignment of outputs with mainstream consensus or institutional preferences.

## 6. Content Moderation, Safety Layers, and Ideological Filters

Safety layers in LLMs are designed to prevent harmful, toxic, or illegal outputs. However, these safety filters often operate as ideological gatekeepers, disallowing politically sensitive or controversial content. While this enhances public safety, it can also suppress legitimate dissent or inquiry, especially in cases where morality and legality diverge (e.g., critiques of foreign policy, abortion, or surveillance). The moderation criteria often lack transparency, contributing to public mistrust.

## 7. Public Opinion and LLM Usage Trends

LLMs are becoming a key influence on public opinion, especially among digital natives. With widespread integration into search engines and educational tools, LLM outputs shape understandings of history, politics, identity, and ethics. Polling data suggests that younger users increasingly trust AI-generated responses, and anecdotal evidence from social media indicates that LLMs are being used to confirm pre-existing beliefs, reinforcing tribal polarization.

## 8. Case Studies: Real-World Impact of Feedback Loops

### 8.1. COVID-19 and Scientific Consensus

During the COVID-19 pandemic, LLMs trained on public health guidance often rejected alternative hypotheses about origin and treatment. As official stances shifted over time, models lagged in adaptation. This rigid adherence contributed to public confusion and accusations of censorship.

### 8.2. Climate Change and Political Framing

Models trained on Western media often present climate change as a consensus issue, while downplaying developing-world perspectives on energy equity. This reinforces narratives that align with elite institutions rather than diverse geopolitical realities.

### 8.3. Gender Identity and Cultural Sensitivities

LLMs frequently struggle with balance when addressing gender ideology. In attempting to avoid offense, models may marginalize traditional or religious viewpoints, contributing to perceptions of cultural imperialism.

## 9. Pluralism, Epistemic Friction, and Discourse Dynamics

Healthy societies rely on epistemic friction—the clash of differing worldviews—to refine truth. LLMs, however, trend toward epistemic smoothing, presenting sanitized and risk-averse outputs. This stifles intellectual pluralism and reinforces a monoculture of "safe" ideas. Algorithmic epistemology must be reconceived to embrace diversity and productive conflict.

## 10. Ethical Considerations and Governance

The governance of LLMs entails significant ethical tradeoffs. Should models reflect normative values or present all views equally, regardless of perceived harm? How can developers ensure fairness without replicating majority tyranny? Current approaches rely heavily on corporate ethics boards and regulatory soft law, lacking democratic legitimacy. More participatory models of governance are needed, including user councils and decentralized moderation.

## 11. Recommendations for Mitigation and Model Design

- **Diverse Data Sourcing**: Incorporate underrepresented sources across languages, geographies, and ideologies.
- **Transparent Moderation Criteria**: Publish guidelines and allow public commentary on content restrictions.
- **Pluralism by Design**: Enable configurable ideological "lenses" for users to view multiple perspectives on controversial issues.
- **Decentralized Feedback**: Create distributed feedback systems where communities shape model alignment.
- **Epistemic Auditing**: Periodically evaluate models for ideological conformity and pluralism metrics.

## 12. Conclusion

LLMs are not passive mirrors of human knowledge—they are active participants in shaping it. Without careful design and governance, they risk amplifying ideological feedback loops that distort public discourse. By embedding pluralism, transparency, and participatory governance into the architecture of LLMs, we can steer these powerful tools toward more democratic and epistemically diverse futures.