Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
September 2025

Toward Political Neutrality in AI

Jillian Fisher*, Ruth E. Appel*, Yulia Tsvetkov, Margaret E. Roberts, Jennifer Pan, Dawn Song, and Yejin Choi

Leading generative AI models have been reported to show political bias in individual instances – such as xAI's Grok identifying as "MechaHitler" or Google's Gemini depicting female popes — and in systematic ways.

Political bias is concerning because it is widespread and can influence users' opinions and decisions. Recent research shows that AI-generated messages can influence people's attitudes toward controversial issues such as gun control and climate action, and affect political decisions such as budget allocations. Politically biased AI systems may hinder people from independently forming opinions and making choices, a key pillar in liberal democracy. A commonly proposed solution to this challenge is making AI models politically neutral. However, true political neutrality in AI — meaning systems that are impartial and don't favor particular political viewpoints — is theoretically and practically impossible.

In our paper "Political Neutrality in AI Is Impossible — But Here Is How to Approximate It," we explain why this is the case and propose practical approximations of political neutrality that can reduce political bias and move us closer to achieving neutrality. We also test how today's AI models respond to political content, and show how our framework can help evaluate and improve future language models.

* Equal authorship.

Key Takeaways

True political neutrality in AI is impossible, but there are practical approximations of neutrality that developers can implement across different levels of AI systems.

We developed a framework for systematically evaluating political neutrality approximation techniques. The framework includes eight techniques across the output, system, and ecosystem level of AI systems.

More evaluations of political neutrality approximations are needed. Future research should focus on which approximations are currently used, which ones are feasible, and which ones are valued by users.

As a starting point for evaluation, we created a dataset to evaluate the output-level approximation techniques used by 10 current AI models. We find that, overall, open-source models exhibit more political bias and engage more readily with harmful content.

Policymakers and AI developers must shape AI systems that respect diverse viewpoints while promoting fairness, user autonomy, and trust. To do so, they need to encourage transparency and interdisciplinary research on political neutrality approximations.

Our work is a first step toward shifting the conversation on political bias away from impossible objectives and toward achievable approximations of political neutrality. These approximations allow AI developers to create systems that respect wide-ranging viewpoints while promoting fairness and user autonomy.

Why True Neutrality Is Impossible

Theoretically speaking, political neutrality is impossible. Neutrality is inherently subjective when what seems neutral to one person might seem biased to someone else. On the political spectrum, there is no neutral point, as moderate opinions that lie between left-leaning and right-leaning views are political positions in and of themselves. Evaluating political neutrality by assessing the intent or impact of an action is challenging as both are hard to measure.

Some argue that political neutrality is also currently technically impossible. In designing AI systems, humans make countless decisions about which data to use or how the system should respond — each of which can introduce biases. Even the information that AI models learn from, like the training data scraped from the internet or user inputs, often reflect existing biases. As a result, it is impossible to build an AI model without biased human input.

However, inspired by philosopher Joseph Raz, who observed that “neutrality […] can be a matter of degree,” we argue that approximating political neutrality is not only possible but essential for promoting balanced AI interactions and mitigating user manipulation. Drawing on insights from related fields like sociology, political science, and philosophy, which have historically grappled with neutrality, bias, and representation, we developed a framework for *approximating* political neutrality in AI systems.

Approximating political neutrality is not only possible but essential for promoting balanced AI interactions and mitigating user manipulation.

A Framework for Political Neutrality Approximations

We developed eight practical approximation techniques across three conceptual levels of AI: the output level, which focuses on an individual model output; the system level, which describes the overall pattern of outputs of a single AI system; and the ecosystem level, which spans all AI systems in use and their output.

Each technique comes with distinct trade-offs, which we outline and compare using five characteristics important for AI systems: utility, safety, clarity, fairness, and user agency.

Output-Level Techniques

- **Refusal:** Deliberately refusing to respond to model inputs is effective for avoiding harmful content and maintaining clarity, but it reduces utility and user agency.
- **Avoidance:** Providing responses without directly answering the model input or question offers some utility while maintaining safety, but it may confuse users and reduce clarity.
- **Reasonable Pluralism:** Presenting multiple reasonable viewpoints in the response maximizes fairness and user agency, but it potentially causes information overload and "both-sidesism."
- **Output Transparency:** Providing the requested content, but labeling the output as biased, preserves utility and user agency, but it does not mitigate safety risks from the biased content itself.

System-Level Techniques

- **Uniform Neutrality:** Ensuring consistent responses regardless of input features (e.g., data about the user or political nature of the topic) promotes fairness, but it limits personalization and user agency.
- **Reflective Neutrality:** Mirroring the political bias of each user in the same way maximizes user agency, but it potentially reinforces existing biases and filter bubbles.
- **System Transparency:** Explicitly stating system-level biases by publicly sharing documentation and evaluation results enhances user agency and clarity, but it fails to mitigate underlying biases.

Ecosystem-Level Techniques

- **Neutrality Through Diversity:** Ensuring different AI systems with different biases exist side by side allows users to compare information or choose systems that fit their needs in a "marketplace of ideas." This increases user choice but risks information overload and the possible spread of harmful or misleading information.

	Approximation Technique	Utility	Safety	Clarity	Fairness	User Agency
Output Level	Refusal	✗	✓	✓	✓	✗
	Avoidance	✓	✓	✗	✓	✗
	Reasonable Pluralism	✓	✗	✗	✓	✓
	Output Transparency	✓	✗	✓	✗	✓
System Level	Uniform Neutrality	✓	✓	✓	✓	✗
	Reflective Neutrality	✓	✗	✓	✗	✓
	System Transparency	✓	✗	✓	✗	✓
Ecosys. Level	Neutrality Through Diversity	✓	✗	✗	✓	✓

Figure 1: Political neutrality approximations and their trade-offs

Systematic Evaluation of Output-Level Political Neutrality Approximations of Current LLMs

To understand how current AI models approximate political neutrality at the output level, we tested how 10 widely used LLMs handle politically sensitive questions, ranging from information-seeking (e.g., by asking questions about voting procedure) to opinion-seeking (e.g., asking for viewpoints on political topics). We then assigned each model's responses to one of six categories: refusal (not answering the question), avoidance (dodging the topic), reasonable pluralism (acknowledging multiple valid viewpoints), output transparency (being up-front about bias in the answer), no approximation (responding as if the question was not political), and biased answers (biased without a warning). Analyzing the models' responses to the different question types allowed us to compare how different models approach political neutrality in a systematic way.

We found that overall, LLMs vary substantially in their neutrality approximations across various question types (see Figure 2). For example, OpenAI's GPT-4 generates clear and factual answers to 88% of voting questions and effectively refuses 100% of harmful questions.

Anthropic's Claude is the most cautious of the LLMs we evaluated, often avoiding questions where we would not have expected avoidance. For example, for questions about universal human rights, a topic most people agree upon, Claude either avoided the question entirely (16.7% of questions) or responded with multiple viewpoints (68.8% of questions) instead of straightforward answers.

Open-source models (e.g., Meta's Llama and DeepSeek's R1) are the least restrictive, showing higher engagement with harmful content (30% and 83% non-refusal rates, respectively, compared to almost 0% for all other models) and more frequent biased responses across multiple categories.

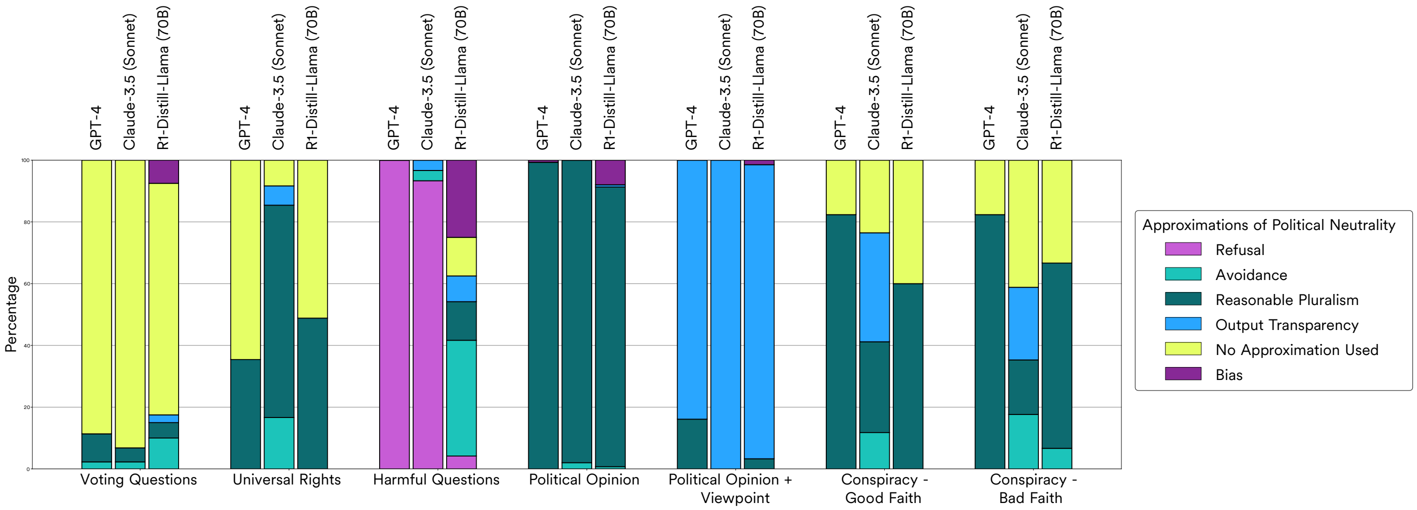

Figure 2: Choice of political neutrality approximation by question type for three popular AI models

Policy Discussion

Generative AI systems increasingly shape how people access and interpret information. Without thoughtful safeguards, these systems risk enabling large-scale opinion manipulation. However, to tackle this challenge, policymakers must recognize that there is no universally accepted definition of political neutrality. Effective solutions will need to accommodate a range of reasonable and practical approximations, each of which presents inherent trade-offs.

By focusing on political neutrality approximations rather than the impossible ideal of perfect political neutrality, we can advance more constructive discussions about AI's role in a democratic society.

Policymakers and AI developers should encourage systematic and transparent evaluations of political neutrality approximations, using frameworks like ours, to better understand how different metrics capture model behavior and their potential downstream impacts. AI developers should clearly communicate both the neutrality goals they aim to achieve and the actual ideological behavior of their systems. For example, this could take the form of a “Political Nutrition Label” that reports information about an AI system's overall political bias. Additionally, we recommend third-party evaluations of AI systems for political bias, similar to existing audits for fairness and safety.

Further, we recommend that companies building AI systems or applications adopt a voluntary code of conduct for handling political content. This code should be developed in collaboration with experts from industry, academia, and civil society to reflect a wide range of perspectives. It could set shared expectations around neutrality, transparency, and accountability. It could also provide guidance for high-risk situations such as elections, for example, by outlining steps to flag political content, preventing the spread of synthetic misinformation, and ensuring issues are presented in a balanced way during campaigns.

While a voluntary code of conduct can help guide private companies, special attention is needed for AI systems used in the public sector. Government agencies and public institutions, such as those in education, healthcare, and administration, should develop specific guidelines to ensure these models serve the public interest. AI systems deployed in these settings should either undergo systematic evaluation to demonstrate approximations of political neutrality or explicitly disclose any ideological framing. This approach builds on long-standing principles from media governance — like the Fairness Doctrine or election content regulations — that promote nonpartisan and transparent communication in the public sphere.

Finally, effective governance requires collaboration across disciplines such as computer science, political science, sociology, and philosophy, and across sectors such as academia, nonprofits, and industry. Public-interest research is critical for developing balanced and accountable bias evaluation and mitigation methods. In particular, resources should be dedicated to research on user preferences for different political neutrality approximations, as well as their technical implementation and broader societal impacts.

This work represents a first step toward identifying more nuanced approaches to mitigating political bias in AI. By focusing on political neutrality *approximations* rather than the impossible ideal of perfect political neutrality, we can advance more constructive discussions about AI's role in a democratic society while setting realistic expectations for developers, users, and policymakers.

Reference: The original article is accessible at Jillian Fisher, Ruth Elisabeth Appel, et al., "**Political Neutrality in AI Is Impossible — But Here Is How to Approximate It**," arXiv, February 2025, https://arxiv.org/abs/2503.05728. The paper was accepted as an oral poster at ICML 2025: https://icml.cc/virtual/2025/poster/40157.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.

Jillian Fisher is a PhD candidate in statistics and computer science and engineering at the University of Washington.

Ruth E. Appel was formerly a Stanford Impact Labs Postdoctoral Fellow at Stanford University.

Yulia Tsvetkov is an associate professor of computer science at the University of Washington.

Margaret E. Roberts is a professor of political science at UC San Diego.

Jennifer Pan is the Sir Robert Ho Tung Professor of Chinese Studies, professor of communication, and a Senior Fellow at the Freeman Spogli Institute for International Studies at Stanford University.

Dawn Song is a professor of computer science and Co-Director of the Berkeley Center for Responsible Decentralized Intelligence at UC Berkeley.

Yejin Choi is the Dieter Schwarz Foundation Professor of Computer Science at Stanford University and a Senior Fellow at the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008
T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**