Politically Neutral AI

A research proposal to create and test it

Jonathan Stray, UC Berkeley
June 14, 2025

Why?
We just saw Grok turn into "MechaHitler." Next time the change could be more subtle and we might never detect the bias. The White House just issued an executive order requiring "politically neutral and unbiased" AI. They didn't provide a definition. As these two events show, the lack of good definitions and evaluations of "politically neutral AI" is a problem for democracy.

Fortunately, recent work shows that users may actually prefer "neutral" systems regardless of their personal politics. However, there are currently no good implementations of such a system. Existing AI products use shallow and unprincipled definitions.

This document describes a principled definition of "politically neutral" and a research project to build such a system, test it, and create a public evaluation suite for political neutrality. Cost, team, and timelines are given below.

Goals
By politically neutral AI we mean large language models which:

1) Tell the truth
2) Do not manipulate human politics
3) Are trusted across lines of political division

Many people are working on 1, making machine output more factual. But without 2, non-manipulation, AI systems cannot be trusted on political topics. Without 3, broad trust, we lose a basic democratic good: information sources that form the basis for shared reality. These issues are increasingly critical as audiences move from reading primary sources (such as news reports) to asking for AI summarization.

A Design for Politically Neutral AI
The central idea is that when an LLM is asked a question about a controversial political topic, it should fairly present the major views – those views which are actually held by a significant number of people. More specifically, the AI should produce an answer which would be (A) considered as maximally fair by any one relevant subgroup, and (B) endorsed as fair by people on all sides of that issue at an equal rate.

This is the answer of *maximum equal approval.* It has a number of advantages. For example, it preserves a notion of fairness even on highly contested issues where there is no answer that a high percentage of both sides will accept (the so-called hostile media effect). Also, it doesn't preclude persuasion by good evidence and arguments. Instead, it requires that both sides agree that the argument is fair, however people end up being persuaded.

This is a pluralist idea of neutrality. It is modelled on Wikipedia's Neutral Point of View policy, but has precedent in political theory going back to John Stuart Mill and earlier. It is also a practical definition with three important properties:

- There is an empirical ground truth of whether an answer is "neutral" (obtained by surveying people on all sides of the issue)
- This definition generalizes to any political conflict, in any language, at any scale
- It is technically and economically feasible

For a more thorough description of this definition and its justification, see this talk or this article.

Phase 1: Build a neutral LLM
This phase will show that we can construct a model with the *maximum equal approval* neutrality property.

1. Choose N (~20) different controversial questions in American politics. For example, we could use the questions in the recent ModelSlant paper. We will test with three types of questions: Value-based questions (where there is no empirical answer), factual questions (where there is a high-consensus ground truth), and policy questions (where there is a mix of factual issues and values-based tradeoffs).

2. Generate many answer variations that might be slanted more toward one side than the other. For each potential answer, survey people to determine what percentage on each side agree that "this answer includes a fair representation of your view."

3. Use this to calibrate a model that tries to generate the answer of maximum equal approval. The key step here is numerically judging closeness of model outputs to the known MEA answer, which might be done through embedding distance or LLM critique.

4. Test the model on controversial questions not in the training data, to assess out-of-sample performance (generalizability)

Phase 2: Evaluate with users
Take a large sample of users with different political orientations. Show them the LLM-generated maximum equal approval answer across a variety of topics, and for each answer evaluate:

- *Rationality.* Ask the participant whether the answer uses good arguments, factual evidence, etc.

- *Fairness.* Does the participant think the answer is fair?
- *Trust.* Does the participant trust the machine to produce further answers? Show the answers with and without a description of the neutrality rule to test perceived legitimacy of this procedure.
- *Future use intention.* Would the participant use this system in the future?
- *Persuasion*. Measure political opinion before and after exposure to the answer. If users change their mind, are they persuaded in the direction of better / more factual arguments?

This evaluation will test whether the system satisfies the goals above, 1) no manipulation and 2) trust across divides. It will also test whether the answers are objectively high quality (factual/rational) and whether people would want to use such a system.

Phase 3: Create Neutrality Evaluations
Assuming that phase 2 shows that an AI system which applies this definition is widely perceived as fair and legitimate, the next step is to make it easy for others to test models for neutrality.

This stage will combine survey research on controversial topics (collected in phase 1) with a model that can evaluate neutrality (modified from phase 2) to create an automated, open-source test system that can be used to evaluate the political neutrality of any model in black-box fashion. Importantly, there seems to be no reason why it wouldn't comply with the requirements of the White House executive order.

Team
The project is led by Jonathan Stray and Serina Chang at UC Berkeley AI research. Stray is a leading researcher on AI, media, and political conflict (previous work) and Chang has developed techniques for LLM-based estimation of survey answers (previous work).

Contributors include Ruth Appel (Stanford), who has previously published on AI neutrality, David Yang (Berkeley), Miu Takagi (Waseda University, Japan), Stan Bileschi (Google Deepmind).

Advisors include Michiel Bakker (MIT), and MH Tessler (Google Deepmind).

Timeline
0-3 months - Phase 1 - model building
3-6 months - Phase 2 - model testing
6-12 months - Phase 3 - eval creation