Ethical Considerations: Automated Political Stance Identification

March 6, 2026

1 Introduction

The Automated Political Stance Identification (APSI) tool is designed to support the analysis of political texts by detecting ideological positions and rhetorical patterns. Because APSI operates at the intersection of artificial intelligence, political analysis, and public discourse, its development and deployment raise a number of important ethical questions that require transparent and rigorous engagement.

This document sets out the ethical principles that inform the design of APSI, the known limitations of the tool, and the appropriate boundaries of its use. It is intended for researchers, journalists, civil society practitioners, and any member of the public who wishes to understand how APSI has been built responsibly and how it should — and should not — be used.

2 Purpose and Intended Use

APSI was developed for legitimate research, journalistic, and civil society purposes. Its primary intended uses are:

- Academic research on political communication, ideological trends, and rhetorical patterns.
- Journalistic investigation into political discourse and public debate.
- Analysis by think tanks, NGOs, and civil society organisations monitoring democratic health.
- Educational and pedagogical applications in the social sciences.

These use cases are grounded in a commitment to democratic accountability and the public interest. The tool enables the kind of large-scale, reproducible text analysis that would be prohibitively time-consuming using manual methods alone, thereby expanding the capacity of researchers and practitioners to scrutinise political communication.

3 Prohibited and Strongly Discouraged Uses

There are categories of use that are explicitly outside the intended scope of APSI and that its developers consider ethically unacceptable.

3.1 Automated Decision-Making About Individuals

APSI is not designed to evaluate, classify, or make inferences about individual people. It analyses text, not persons. Scores produced by the tool reflect patterns in the language of a particular document; they do not constitute assessments of the political beliefs, loyalties, or intentions of any individual author. APSI must never be used to screen, rank, or make decisions affecting individuals — whether in employment, legal proceedings, security assessments, or any other consequential context.

Important: Using APSI outputs as absolute evidence of an individual's political beliefs or intentions would be both methodologically unjustified and ethically impermissible.

3.2 Surveillance and Political Profiling

APSI must not be used as a tool for the surveillance of political actors, activists, journalists, or private citizens. Similarly, it must not be deployed to build political profiles of individuals or groups for purposes of targeting, manipulation, or repression.

3.3 Disinformation and Political Manipulation

The tool must not be used to generate, amplify, or lend artificial credibility to political disinformation. Its outputs should not be selectively presented in ways designed to mislead audiences about the political content of a text, the positions of political actors, or the nature of public debate.

3.4 Legal and Policy Judgments

APSI scores are probabilistic estimates based on patterns in language. They do not constitute expert testimony, verified factual claims, or legally admissible evidence. The tool must not be used to inform legal proceedings, regulatory decisions, or policy judgments that materially affect individuals or organisations.

4 Transparency and Reproducibility

A central design principle of APSI is that its outputs must be interpretable and auditable. Unlike opaque "black box" classifiers, APSI is built on a hypothesis-based scoring framework in which each score can be traced back to explicit theoretical statements, the entailment probabilities they received from the underlying language model, and the deterministic rules used to aggregate them into a final score.

This approach serves several ethical functions:

- It enables users to understand why a particular score was produced, rather than simply accepting a numerical output without justification.
- It allows critical scrutiny of the theoretical assumptions embedded in the hypotheses, which are derived from established political science literature and expert survey codebooks.
- It supports the identification and correction of errors, biases, or limitations by academic peers and practitioners.

The full methodology, including the hypotheses used for each analytical dimension, is documented and publicly available. The underlying model, Political DEBATE (Burnham et al., 2024), is an open-source resource, and the validation datasets produced in the course of this project will also be made publicly available to support replication and independent assessment.

5 Known Limitations and Potential Biases

Responsible use of APSI requires awareness of its technical limitations and the ways in which its outputs may be imperfect or misleading. The following limitations are acknowledged by the tool's developers.

5.1 Dependence on Hypothesis Design

APSI scores are only as theoretically valid as the hypotheses on which they are based. These hypotheses are derived from established political science frameworks and expert survey instruments; however, they necessarily reflect particular scholarly traditions and may not fully capture the complexity or regional specificity of political language in all contexts. Users are encouraged to critically evaluate whether the hypotheses are appropriate for their particular analytical purpose.

5.2 Training Data Biases

The Political DEBATE model on which APSI is built was trained on a corpus of political texts predominantly drawn from English-language sources, including social media posts, news articles, Congressional documents, and court case summaries. As with all machine learning models, the tool may reflect biases present in this training data. This may affect its performance when applied to texts from non-Anglophone political traditions, non-Western democratic contexts, or political cultures underrepresented in the training corpus.

5.3 Rhetorical Framing vs. Genuine Belief

APSI detects patterns of rhetorical framing — the way political positions are expressed in language — rather than the underlying beliefs or intentions of authors. A text may employ populist rhetoric strategically without reflecting the author's sincere political commitments; conversely, genuinely held positions may be expressed in language that does not trigger the tool's hypotheses. Users should interpret scores as estimates of rhetorical tendency, not as verified statements of belief or intent.

5.4 Context Insensitivity

The tool analyses text in isolation. It has no access to the context in which a document was produced — the identity or background of the author, the audience addressed, the historical or political circumstances, or the broader discursive environment. Identical language may carry very different political meanings depending on context, and users bear responsibility for situating APSI's outputs within appropriate contextual knowledge.

5.5 Approximate Scores and Confidence Estimates

All scores produced by APSI are probabilistic estimates, not objective measurements. Confidence levels indicate the internal consistency of the model's predictions and should not be interpreted as

guarantees of accuracy. Scores near the midpoint of any scale should be interpreted with particular caution, as they may indicate genuine ambiguity in the text as much as a moderate ideological position.

6 Validation and Ongoing Accountability

Prior to public deployment, APSI outputs have been validated against human expert assessments. A sample of 60 political texts was evaluated by academic experts recruited from the authorship of leading political science journals, including *Political Communication* and *Comparative Political Studies*. Across three analytical dimensions — economic left–right ideology, populism versus pluralism, and support for liberal democratic values — a total of 147 expert evaluations were collected and used to benchmark model performance.

This validation process provides an empirical basis for confidence in the tool's outputs while also identifying areas in which model performance may be weaker. Full details of the validation methodology and results are available in the accompanying Description and Methodology document.

The developers are committed to ongoing review of APSI's performance, including monitoring for evidence of systematic bias, updating hypotheses in response to scholarly criticism, and incorporating improvements as the underlying model and validation datasets are refined.

7 The Necessity of Human Oversight

APSI is designed as a tool to support human analysis, not to replace it. The developers take the position that automated text analysis, however sophisticated, cannot substitute for the contextual knowledge, critical judgment, and ethical reasoning that human analysts bring to the study of political communication.

All uses of APSI should therefore incorporate meaningful human oversight. This means:

- Treating APSI scores as one input among several in a broader analytical process, not as self-sufficient conclusions.
- Exercising critical judgment about whether the tool's hypotheses and dimensions are appropriate for the specific texts and contexts being analysed.
- Contextualising outputs within relevant political, historical, and cultural knowledge.
- Being transparent with audiences about the role of automated analysis in any published findings and the limitations that accompany it.

8 Data Protection and Privacy

Users who submit texts for analysis through the APSI platform are responsible for ensuring that they have appropriate rights to use those texts and that their use complies with applicable data protection legislation, including the General Data Protection Regulation (GDPR) where relevant.

APSI is not designed to process personally identifiable information, and users are strongly discouraged from submitting texts that contain such information. The tool should not be used to analyse private communications or documents obtained without consent. Where texts are drawn from public sources — such as published speeches, manifestos, news articles, or social media posts — users should nonetheless ensure that their use is consistent with applicable terms of service and ethical research standards.

9 Multilingual and Cross-National Use

APSI incorporates translation capabilities to extend its analytical reach to texts in a range of languages beyond English. Users applying the tool in non-English contexts should be aware that automated translation introduces an additional layer of uncertainty into the analytical process, and that the tool's hypotheses — which are grounded in established political science frameworks developed primarily in Western democratic contexts — may be less well-suited to political traditions with different conceptual vocabularies or institutional configurations.

Cross-national comparative analyses enabled by APSI should be approached with appropriate methodological caution, and users should consult relevant area expertise when interpreting results from political contexts with which they are not familiar.

10 Summary of Ethical Commitments

Transparency: APSI's methodology is fully documented and its outputs traceable to explicit theoretical hypotheses and deterministic scoring rules.

Accountability: Model outputs are validated against expert human assessments, and the developers are committed to ongoing review and improvement.

Non-maleficence: The tool must not be used for surveillance, political profiling, automated decision-making about individuals, or any purpose that could cause harm to persons or democratic institutions.

Human oversight: APSI supports, but does not replace, human analytical judgment. All outputs require contextual interpretation by qualified users.

Epistemic humility: Scores are estimates, not facts. Known limitations — including training data biases, context insensitivity, and hypothesis-dependence — must be communicated to audiences alongside any findings.

References

Burnham, Michael (2025). "Stance detection: a practical guide to classifying political beliefs in text." *Political Science Research and Methods*, 13(3), 611–628. https://doi.org/10.1017/psrm.2024.35

Burnham, Michael et al. (2024). *Political DEBATE: Efficient Zero-shot and Few-shot Classifiers for Political Text*. arXiv: 2409.02078. https://arxiv.org/abs/2409.02078

Jolly, Seth et al. (2022). "Chapel Hill expert survey trend file, 1999–2019." *Electoral Studies*, 75, 102420.

Laurer, Moritz et al. (2022). "Less Annotating, More Classifying." Preprint. Open Science Framework. https://osf.io/74b8k

Lindberg, Staffan I. et al. (2022). "Codebook Varieties of Party Identity and Organization (V–Party) v2."

Norris, Pippa (2020). "Measuring populism worldwide." *Party Politics*, 26(6), 697–717.