The Last Vote: A Multi-Stakeholder Framework for Language Model Governance

Subramanyam Sahoo[*]
Berkeley AI Safety Initiative (BASIS)
UC Berkeley
sahoo2vec@gmail.com

Aditi Chhawacharia
Cyber Physical Systems Lab
University of Texas, Dallas
chhawacharia.aditi@gmail.com

Abstract

As artificial intelligence systems become increasingly powerful and pervasive, democratic societies face unprecedented challenges in governing these technologies while preserving core democratic values and institutions. This paper presents a comprehensive framework to address the full spectrum of risks that AI poses to democratic societies. Our approach integrates multi-stakeholder participation, civil society engagement, and existing international governance frameworks while introducing novel mechanisms for risk assessment and institutional adaptation. We propose: (1) a seven-category democratic risk taxonomy extending beyond individual-level harms to capture systemic threats, (2) a stakeholder-adaptive Incident Severity Score (ISS) that incorporates diverse perspectives and context-dependent risk factors, and (3) a phased implementation strategy that acknowledges the complex institutional changes required for effective AI governance.

1 Introduction

Language model governance remains marked by technocratic reductionism, privileging compliance-oriented risk taxonomies and existential mitigation logics while occluding the constitutive politicality of AI as a socio-technical infrastructure that redistributes epistemic authority and encodes normative commitments Araujo et al. (2024). This depoliticized framing produces legitimacy deficits across both state-centric regulatory instruments and industry self-regulation, which operationalize governance as a problem of optimization rather than democratic authorization. Contemporary interventions, exemplified by the European Union (2024)'s AI Act and recent U.S. executive orders, largely delimit governance to individualized harms and discrete technical risk vectors, thereby neglecting structural modalities through which scaled generative systems destabilize democratic legitimacy Farnadi et al. (2024). Such systems, by virtue of their cascading and path-dependent societal effects, engender forms of procedural erosion that elude capture by extant risk assessment methodologies, underscoring the need for governance architectures attuned to systemic threats beyond the technical-compliance paradigm Feng et al. (2023); Diakopoulos and Johnson (2024).

Democratic AI Governance Framework: We formalize AI governance as an optimization problem where democratic integrity is the primary objective, not a side-constraint Cooper et al. (2024). This formulation remedies three deficits in existing paradigms: (i) Risk Assessment: shifting from local fairness metrics to systemic analyses of AI–democracy interactions; (ii) Legitimacy: replacing technocratic exclusivity with participatory, binding authority for civil society and affected stakeholders; (iii) Implementation: supplying concrete institutional and procedural mechanisms that instantiate democratic principles as operational rules rather than aspirational norms Bovens (2007); Arnstein (1969) (For more – Follow Section 3).

[*]Core Contributor

39th Conference on Neural Information Processing Systems (NeurIPS 2025)
Workshop on Algorithmic Collective Action (ACA@NeurIPS 2025).

Key Contributions: We extend AI governance literature by (i) introducing a seven-category taxonomy of democratic risks spanning individual exclusion to systemic fragility, (ii) formalizing a stakeholder-adaptive Incident Severity Score (ISS) that aggregates heterogeneous utilities into mathematically rigorous governance signals, (iii) proposing a four-phase, six-year implementation roadmap transitioning from voluntary coordination to binding democratic oversight, and (iv) operationalizing deliberative democratic theory through institutionalized co-governance, citizen panels, and sovereignty zones.

Paper Structure: Section 2 positions our work relative to existing frameworks. Section 3 presents our democratic risk taxonomy. Section 4 details the multi-stakeholder governance architecture. Section 5 outlines phased implementation. Section 6 covers monitoring and adaptation. The mathematical ISS framework is detailed in the appendix, providing computational tools for operationalizing democratic oversight. Section 7 and 8 gives a clear-cut horizon for Limitations and Future works correspondingly. Additional technical details and responses to methodological critiques appear in Appendices.

2 Background and Related Work

Ovadya et al. (2025) demonstrates that technologies are not neutral tools but embody political values and redistribute power within society. The work done by Jasanoff (2004) on the co-production of science and social order shows how technological systems and political institutions mutually constitute each other. This perspective reveals why purely technical approaches to AI governance fail to address AI's constitutive political effects on democratic institutions Hadfield and Bernier (2025).

Current Governance Limitations: Existing governance frameworks predominantly address individual-level harms through fairness constraints and safety protocols Jobin et al. (2019). The **EU AI Act** employs risk stratification by application domain but lacks systematic treatment of systemic democratic effects Walters et al. (2023). Industry focused partnerships also largely prioritize individual accountability over collective institutional impacts Pasquale (2015). While previous work has addressed group fairness in machine learning through demographic parity and equalized odds metrics, these approaches still operate within individual-level harm frameworks rather than addressing structural democratic threats Fishkin (2018). *Our work extends beyond both individual and group fairness to examine AI's impact on democratic institutions themselves.*

Our Positioning: We position this work at the intersection of AI governance literature on technology's constitutive political effects and computational democracy research emphasizing participatory institutional design. Our framework diverges from prevailing approaches by treating **democratic integrity as a primary optimization objective** Sahoo (2025) rather than a constraint, contributing a formal risk taxonomy that extends algorithmic governance theory through operationalizable metrics for democratic impact assessment.

3 A Comprehensive Risk Taxonomy for Democratic Societies

Building on the domain structure of MIT AI Risk Repository (2024); Slattery et al. (2025), we extend existing risk taxonomies to systematically address AI's threats to democratic institutions. Our taxonomy, derived from democratic theory and historical institutional threats, captures both direct process-level risks and indirect institutional interactions, spanning harms from individual exclusion to systemic collapse Bengio et al. (2025).

Discrimination & Democratic Exclusion. Beyond individual unfair treatment, AI systems can systematically exclude entire communities from democratic participation. This includes algorithmic discrimination in voting access, civic service delivery, and representation in democratic processes, creating structural barriers to political equality. Hewage (2023) showed GPT-based resume screening tools systematically exclude candidates based on linguistic patterns associated with minority communities, institutionalizing bias at scale. Hendrycks et al. (2023)

Privacy Erosion & Democratic Surveillance. AI-assisted surveillance enables unprecedented monitoring of citizen activities, communications, and political associations, potentially chilling free expression and opposition organizing. Das et al. (2025) This extends privacy concerns into the realm of democratic participation rights. Agrawal (2022); Liang et al. (2018) Curated evidence from China's social credit system demonstrates how surveillance can systematically constrain democratic participation by monitoring and scoring citizen behavior, creating chilling effects on dissent and political organization. Corporate surveillance systems in democracies create similar risks through political tracking and behavioral scoring. Future of Life Institute (2024)

Electoral Misinformation & Discourse Degradation. Current models enable computational propaganda, hyper-personalized misinformation campaigns, and systematic degradation of civic discourse quality. Aparicio de Soto (2022) Unlike general misinformation, these threats specifically target electoral processes and democratic deliberation. Bots powered by model weights Nevo et al. (2024) can be used to conduct targeted surveys, build voter profiles, infer political preferences from conversational data, and deliver personalized propaganda.

Democratic Manipulation & Malicious Interference. Sophisticated actors can weaponize model weights for large-scale electoral interference, voter suppression, and systematic manipulation of democratic processes. This extends beyond individual fraud to coordinated attacks on democratic institutions. Horta Ribeiro et al. (2020) documents bot networks amplifying divisive political content, synthetic media campaigns targeting specific voter demographics, and automated systems designed to suppress turnout through coordinated disinformation campaigns. Shah et al. (2025); Bullock et al. (2025)

Civic Participation & Human Agency Loss. Algorithmic curation of information environments affects civic engagement through echo-chamber reinforcement, filter-bubble creation, and the delegation of civic decision-making to automated systems, reducing meaningful human participation in democracy. Costanza-Chock (2020) Language-model–assisted recommendation systems can promote increasingly extreme political content and create radicalization pathways that undermine democratic discourse norms. Horta Ribeiro et al. (2020)

Democratic Power Concentration. The capital and data requirements for advanced AI concentrate power among a few actors, enabling democratic capture through regulatory influence and technological dependency. Sahoo and Dutta (2024) shows how democratic institutions can become dependent on private entities for critical functions. Foundation-model concentration creates dependencies when governments adopt these systems for public services, potentially delegating consequential democratic decisions to unaccountable private entities. Reuel et al. (2024); Fisher et al. (2025)

Systemic Democratic Fragility. Complex interactions among new-era models OpenAI et al. (2024) can produce emergent behaviors that threaten democratic stability through cascade failures, unintended coordination effects, or systems developing goals misaligned with democratic oversight—representing novel risks to institutional stability. Demirer et al. (2019); Hammond et al. (2025)

4 A Multi-Stakeholder Governance Architecture

Current governance frameworks suffer from what Grek (2016) et al., identifies as the **expertocracy** problem: the systematic privileging of technical expertise while relegating other epistemic contributions to symbolic consultation . This produces legitimacy deficits because different stakeholder groups possess forms of knowledge that are non-substitutable and cannot be reduced to purely technical metrics Caddle et al. (2025).

To address this, our framework institutionalizes seven distinct categories of expertise. Technical practitioners provide feasibility assessments and capability boundaries grounded in real-world deployment contexts. Academic researchers contribute interdisciplinary safety analysis and long-term systemic perspectives unconstrained by immediate commercial pressures Ho et al. (2023). Democratic representatives ensure electoral legitimacy and constitutional compatibility, embedding governance processes within democratic accountability structures. Civil society organizations offer public-interest advocacy and long-term value-sensitive oversight von Rosing et al. (2025). Industry participants contribute knowledge of market dynamics, competitive pressures, and implementation costs that external regulators often lack. Affected communities provide experiential evidence of algorithmic harms that cannot be captured by audits or simulations khan2025randomnessrepresentationunreliabilityevaluating. Finally, international partners supply coordination capacity across jurisdictions and analysis of AI's transnational effects on democratic institutions .

We propose a graduated model of participatory governance that integrates these knowledge categories through risk-sensitive forms of involvement Bai et al. (2022); Parthasarathy et al. (2024). In *low-risk* contexts, governance may rely on enhanced consultation and transparency mechanisms such as public comment periods and hearings. *Medium-risk* scenarios require structured deliberation through citizen panels, stakeholder workshops, and anticipatory technology assessment. *High-risk* applications demand binding co-governance, where stakeholder groups exercise formal decision-making authority, including veto rights and access to appeals processes Ganeri (2019) .

5 A Phased Implementation Strategy

Comprehensive AI governance cannot be imposed immediately due to: **(i) weak public salience of systemic risks before crisis events, (ii) insufficient technical capacity in regulatory agencies, (iii) industry resistance absent competitive incentives for compliance, and (iv) democratic legitimacy deficits when governance precedes stakeholder engagement.** Phased implementation addresses these constraints by: building demonstration effects through visible early successes, accumulating technical capacity through learning-by-doing, creating first-mover advantages that flip industry incentives, and generating political coalitions through early stakeholder inclusion Bengio et al. (2024). This approach acknowledges that the primary barrier is not technical feasibility but political economy—institutional transformation requires coalition-building, not just framework specification Reuel et al. (2025).

5.1 Foundation Building Phase (0-24 months)

-Establishing Constitutional Democratic AI Governance

The foundation building phase prioritizes the establishment of robust constitutional frameworks that define clear stakeholder rights, enforcement mechanisms, and accountability structures for effective governance Priyanshu et al. (2024). This initial phase focuses on legitimacy building through controlled pilot deployments that test core governance hypotheses in low-risk, high-visibility settings Chaffer et al. (2025). Specifically, municipal bodies must serve as testing grounds for political chatbots and content moderation language models, allowing for real world validation of democratic oversight mechanisms while minimizing systemic risks Huang et al. (2024). These pilot programs will generate empirical evidence on stakeholder engagement effectiveness, regulatory compliance costs, and democratic participation outcomes that will inform subsequent phases Allen et al. (2025). The phase concludes with the codification of constitutional principles including due process rights for affected communities, transparency requirements for algorithmic decision making, and appeals mechanisms for automated determinations that impact democratic participation Ribeiro et al. (2025).

5.2 System Integration Phase (24-48 months)

-Transitioning to Mandatory Compliance and Risk-Based Oversight

The system integration phase marks the critical transition from voluntary industry cooperation to mandatory regulatory compliance, with particular emphasis on high-risk applications that directly impact democratic processes Hadfield and Clark (2023). All deployments involving **automated political advertising, synthetic news generation, or voter-targeted conversational agents** will be subject to mandatory "***Incident Severity Score (ISS)***" (Refer Appendix A) assessments conducted by fully operational model safety committees with diverse stakeholder representation. These committees will possess enforcement authority, including the power to require design modifications, impose operational restrictions, or mandate system shutdowns for applications that exceed established risk thresholds Zeng et al. (2024). This phase includes the development of standardized assessment protocols, the training of qualified evaluators, and the establishment of inter-agency coordination mechanisms to ensure consistent application of governance standards across jurisdictions Pazzaglia et al. (2025). By the conclusion of this phase, the regulatory framework will demonstrate measurable effectiveness in identifying and mitigating high-risk deployments while maintaining democratic legitimacy through transparent, participatory oversight processes.

5.3 Comprehensive Coverage Phase (48-72 months)

-Expanding Regulatory Scope Through Decentralized Democratic Oversight

This phase extends mandatory governance requirements to medium risk scenarios while adopting the principle of subsidiarity through community based oversight mechanisms. Local community oversight boards, composed of affected stakeholders and technical experts, will assume primary responsibility for evaluating systems with localized impacts, such as educational content generation tools, and community specific content moderation systems Ter-Minassian (2025). These decentralized bodies will operate within standardized frameworks established during previous phases while retaining authority to adapt governance approaches to local democratic values and community needs Ovadya et al. (2025). The phase emphasizes capacity building through **comprehensive training programs for community oversight members, the development of accessible technical assessment tools, and the establishment of resource-sharing networks between communities** Ulnicane (2024). This approach ensures that governance parameters scale democratically rather than bureaucratically, maintaining citizen engagement and local accountability as regulatory coverage expands across the ecosystem Reuel and Undheim (2024).

5.4 Adaptive Governance Phase (72+ months)

-Institutionalizing Continuous Democratic Learning and Innovation

This phase institutionalizes mechanisms for continuous democratic learning and **governance innovation**, ensuring that regulatory frameworks evolve alongside technological developments and dynamic societal values Kulothungan and Gupta (2025). Governance innovation laboratories will serve as controlled environments for testing novel oversight approaches, stakeholder engagement mechanisms, and risk assessment methodologies before their incorporation into mainstream regulatory practice. These laboratories will operate through partnerships between regulatory agencies, academic institutions, and civil society organizations, generating empirical evidence on governance effectiveness and democratic legitimacy Zhong et al. (2025). The phase includes the establishment of systematic processes for updating risk thresholds based on emerging evidence, regular review cycles for stakeholder representation mechanisms, and adaptive procedures for incorporating lessons learned from governance failures or unexpected outcomes Ahern (2025). This institutionalized learning approach ensures that democratic AI governance remains responsive to technological change while preserving core democratic values and maintaining public trust in regulatory institutions.

Coalition resilience underpins institutional transformation through strategic stakeholder alignment that preempts governance capture while ensuring political sustainability Stańczak et al. (2025). This approach mobilizes civil society organizations as advocacy coalitions, deploys public education campaigns to build democratic legitimacy, and develops industry partnerships by framing robust governance as market-stabilizing infrastructure. The framework addresses the "**democratic deficit problem**" Azman (2011) in AI governance by creating self-reinforcing political incentives: early adopters gain competitive advantages through public trust premiums, while compliance costs decrease through economies of scale as participation expands, generating positive feedback loops that sustain democratic governance against technocratic reversion Longpre et al. (2025).

6 Monitoring, Evaluation, and Adaptation

Governance should be seen as a cybernetic homeostatic system through a tripartite impact monitoring protocol that continuously tracks: (1) democratic health indicators including electoral integrity metrics and civic discourse quality measures, (2) longitudinal social-economic impact assessments quantifying equity shifts and disparate model deployment effects on marginalized populations, and (3) governance system performance metrics evaluating decision-making efficiency and procedural justice (Zwitter (2024); Zaidan and Ibrahim (2024)). This multi-modal stream generates high-fidelity real-time diagnostics that feed directly into dual-architecture adaptation mechanisms: systematic annual threshold recalibration based on empirical outcomes, and dedicated governance innovation labs serving as institutional sandboxes for novel oversight methodologies Salaudeen et al. (2025).

Accountability is enforced through mandated transparency protocols including public facing dashboards and annual governance reports, while independent statutory oversight bodies conduct external audits to prevent regulatory capture and ensure **democratic alignment** Hendrycks et al. (2025). This architecture transforms static regulatory frameworks into evidence-based adaptive systems capable of responding to emergent socio-technical pathologies while maintaining procedural legitimacy Greenblatt et al. (2024); Summerfield et al. (2024).

7 Current Limitations

The stakeholder weight aggregation mechanism assumes *rational behavior* and may overlook power dynamics or strategic manipulation that characterize actual democratic processes. Additionally, the framework confronts substantial methodological constraints including cultural specificity to Western democratic contexts that limits global applicability and resource-intensive deliberative processes that may exceed organizational capacity. Implementation challenges include assumptions of institutional willingness to adopt multi-stakeholder governance without addressing entrenched interests that benefit from existing technocratic approaches, and limited enforcement mechanisms for compelling compliance from powerful AI companies or state actors who may resist democratic oversight. Technical gaps encompass unclear methodologies for identifying and legitimizing community representatives, raising concerns about the democratic legitimacy of the stakeholders themselves, and potential inability to capture qualitative aspects of democratic harm such as the erosion of civic trust or degradation of deliberative norms that resist quantification.

8 Future Directions

Future work must empirically validate the ISS on documented governance failures, track democratic health longitudinally, and adapt the framework cross-culturally to separate universal from local principles. **Methodological advances should build dynamic online learning systems with causal inference and uncertainty quantification.** Research should integrate governance innovations (hybrid human–AI oversight, decentralized transparency, international coordination), technical extensions (real-time monitoring, adversarial robustness, multi-modal risk assessment), and practical pathways (municipal testbeds, industry incentives, legal integration). Theoretical development drawing on deliberative democracy, critical power analysis, and complexity science is essential to evaluate effectiveness, legitimacy, and scalability, ultimately testing whether multi-stakeholder AI governance can enhance democratic resilience.

9 Conclusion

Language model governance is fundamentally a political challenge requiring democratic solutions, not just technical ones. This framework strengthens democracy by creating new mechanisms for participation, accountability, and transparency while providing concrete tools for implementation. Our ISS metric offers a mathematically rigorous approach to risk assessment that incorporates stakeholder expertise without sacrificing technical precision. The phased implementation strategy provides realistic pathways for institutional transformation while maintaining democratic legitimacy throughout the transition. The choices made now about language model governance will determine whether these systems strengthen democratic discourse through improved information access or undermine it through manipulation, misinformation, and exclusion. This framework provides tools for ensuring AI serves democratic values rather than subverting them.

10 Social Impacts Statement

This work aims to strengthen democratic governance of language models, with several important broader impacts to consider. The proposed governance framework could significantly reshape how societies balance technological innovation with democratic accountability, potentially setting global precedents for managing powerful systems.

Acknowledgements

Subramanyam Sahoo would like to thank Stephen Casper (MIT Algorithmic Alignment Group), Jeffrey Andrade (Director, Harvard AISST), Chanden Climaco (Harvard), Vinaya Sivakumar (UC Berkeley), Kayla Y Jew (UC Berkeley). He further extends our gratitude to Coby Joseph, Dr. Aishwarya Saxena, Cristina Schmidt Ibáñez, and Kayla-Leigh Coetzee of the Vista Institute for AI Policy for their valuable discussion during fellowship. He would also like to thank the Apart Lab members — Philip Quirke, Amir Abdullah, and Jacob Haimes — for their support throughout this work.

Both authors would also like to thank the reviewers of the *NeurIPS 2025 Workshop on Algorithmic Collective Action* for their thoughtful feedback and constructive suggestions.

Spiritual Dedication

I, Subramanyam Sahoo, dedicate this work to **Ilya Sutskever**, whose vision and commitment to advancing safe and beneficial AI have profoundly shaped my own path. His ideas inspired me to pursue AI safety, and this work is a small reflection of that influence.

References

Agrawal, V. (2022). Demystifying the chinese social credit system: A case study on ai-powered control systems in china. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):13124–13125.

Ahern, D. (2025). The new anticipatory governance culture for innovation: Regulatory foresight, regulatory experimentation and regulatory learning.

Allen, D., Hubbard, S., Lim, W., and et al. (2025). A roadmap for governing ai: technology governance and power-sharing liberalism. *AI Ethics*, 5:3355–3377.

Aparicio de Soto, J. (2022). The constructivism of social discourse: Toward a contemporaneous understanding of knowledge. *Open Journal of Philosophy*, 12(3):376–396. Published August 5, 2022.

Araujo, R., Fort, K., and Guest, O. (2024). Understanding the first wave of ai safety institutes: Characteristics, functions, and challenges. Technical report, Institute for AI Policy and Strategy. Accessed: 2025-05-08.

Arnstein, S. R. (1969). A ladder of citizen participation. *Journal of the American Institute of Planners*, 35(4):216–224. Published July 1969.

Azman, K. D. (2011). The problem of "democratic deficit" in the european union. *International Journal of Humanities and Social Science*, 1(5):242–250.

Bai, Y., Bowman, Z., Hatfield-Dodds, B., Mann, D., Amodei, N., McCandlish, S., Brown, T., Kaplan, J., et al. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*. Dec 15 2022.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J., and Mindermann, S. (2024). Managing extreme ai risks amid rapid progress. *Science*, 384(6698):842–845.

Bengio, Y., Mindermann, S., Privitera, D., Besiroglu, T., Bommasani, R., Casper, S., Choi, Y., Fox, P., Garfinkel, B., Goldfarb, D., Heidari, H., Ho, A., Kapoor, S., Khalatbari, L., Longpre, S., Manning, S., Mavroudis, V., Mazeika, M., Michael, J., Newman, J., Ng, K. Y., Okolo, C. T., Raji, D., Sastry, G., Seger, E., Skeadas, T., South, T., Strubell, E., Tramèr, F., Velasco, L., Wheeler, N., Acemoglu, D., Adekanmbi, O., Dalrymple, D., Dietterich, T. G., Felten, E. W., Fung, P., Gourinchas, P.-O., Heintz, F., Hinton, G., Jennings, N., Krause, A., Leavy, S., Liang, P., Ludermir, T., Marda, V., Margetts, H., McDermid, J., Munga, J., Narayanan, A., Nelson, A., Neppel, C., Oh, A., Ramchurn, G., Russell, S., Schaake, M., Schölkopf, B., Song, D., Soto, A., Tiedrich, L., Varoquaux, G., Yao, A., Zhang, Y.-Q., Albalawi, F., Alserkal, M., Ajala, O., Avrin, G., Busch, C., de Leon Ferreira de Carvalho, A. C. P., Fox, B., Gill, A. S., Hatip, A. H., Heikkilä, J., Jolly, G., Katzir, Z., Kitano, H., Krüger, A., Johnson, C., Khan, S. M., Lee, K. M., Ligot, D. V., Molchanovskyi, O., Monti, A., Mwamanzi, N., Nemer, M., Oliver, N., Portillo, J. R. L., Ravindran, B., Rivera, R. P., Riza, H., Rugege, C., Seoighe, C., Sheehan, J., Sheikh, H., Wong, D., and Zeng, Y. (2025). International ai safety report.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4):447–468. First published online 7 June 2007.

Bullock, J. B., Hammond, S., and Krier, S. (2025). Agi, governments, and free societies.

Caddle, X. V., Qadir, S., Hughes, C., Sweigart, E. A., Park, J. K., and Wisniewski, P. J. (2025). Building a village: A multi-stakeholder approach to open innovation and shared governance to promote youth online safety.

Chaffer, T. J., von Goins II, C., Okusanya, B., Cotlage, D., and Goldston, J. (2025). Decentralized governance of autonomous ai agents.

Cooper, A. F., Choquette-Choo, C. A., Bogen, M., Jagielski, M., Filippova, K., Liu, K. Z., Chouldechova, A., Hayes, J., Huang, Y., Mireshghallah, N., Shumailov, I., Triantafillou, E., Kairouz, P., Mitchell, N., Liang, P., Ho, D. E., Choi, Y., Koyejo, S., Delgado, F., Grimmelmann, J., Shmatikov, V., Sa, C. D., Barocas, S., Cyphert, A., Lemley, M., danah boyd, Vaughan, J. W., Brundage, M., Bau, D., Neel, S., Jacobs, A. Z., Terzis, A., Wallach, H., Papernot, N., and Lee, K. (2024). Machine unlearning doesn't do what you think: Lessons for generative ai policy, research, and practice.

Costanza-Chock, S. (2020). *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press, Cambridge, MA. Print and e-book editions.

Das, B. C., Amini, M. H., and Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Comput. Surv.*, 57(6).

Demirer, R., Leggio, K. B., and Lien, D. (2019). Herding and flash events: evidence from the 2010 flash crash. *Finance Research Letters*, 31.

Diakopoulos, N. and Johnson, D. (2024). Anticipating and addressing the ethical implications of deepfakes in the newsroom. *Edward Elgar Publishing (pre-print)*. Northwestern Computational Journalism Lab.

European Union (2024). Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (artificial intelligence act). *Official Journal of the European Union*, L 2024/1689:1–168. Published in the Official Journal on 12 July 2024.

Farnadi, G., Havaei, M., and Rostamzadeh, N. (2024). Position: Cracking the code of cascading disparity towards marginalized communities.

Feng, S., Park, C. Y., Liu, Y., and Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. ACL.

Fisher, J., Appel, R. E., Park, C. Y., Potter, Y., Jiang, L., Sorensen, T., Feng, S., Tsvetkov, Y., Roberts, M. E., Pan, J., Song, D., and Choi, Y. (2025). Political neutrality in ai is impossible- but here is how to approximate it.

Fishkin, J. S. (2018). *Democracy When the People Are Thinking*. Oxford University Press, Oxford.

Future of Life Institute (2024). Ai experts: Major ai companies have significant safety gaps. The 2024 AI Safety Index evaluates six leading AI companies across six safety categories, revealing significant gaps in risk management and control strategies.

Ganeri, J. (2019). Epistemic pluralism: From systems to stances. *Journal of the American Philosophical Association*, 5(1):1–21.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. (2024). Alignment faking in large language models.

Grek, S. (2016). Expert moves: International comparative testing and the rise of expertocracy. In Lingard, B., Martino, W., and Rezai-Rashti, G., editors, *Testing Regimes, Accountabilities and Education Policy*, page 15. Routledge, London, 1st edition.

Hadfield, G. K. and Bernier, A. (2025). *Revisiting the Many Legal Institutions that Support Contractual Commitments in a Globalized World*, pages 267–292. Springer Nature Switzerland, Cham.

Hadfield, G. K. and Clark, J. (2023). Regulatory markets: The future of ai governance.

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, T. A., Hughes, E., Kovařík, V., Kulveit, J., Leibo, J. Z., Oesterheld, C., de Witt, C. S., Shah, N., Wellman, M., Bova, P., Cimpeanu, T., Ezell, C., Feuillade-Montixi, Q., Franklin, M., Kran, E., Krawczuk, I., Lamparth, M., Lauffer, N., Meinke, A., Motwani, S., Reuel, A., Conitzer, V., Dennis, M., Gabriel, I., Gleave, A., Hadfield, G., Haghtalab, N., Kasirzadeh, A., Krier, S., Larson, K., Lehman, J., Parkes, D. C., Piliouras, G., and Rahwan, I. (2025). Multi-agent risks from advanced ai.

Hendrycks, D., Mazeika, M., and Woodside, T. (2023). An overview of catastrophic ai risks.

Hendrycks, D., Schmidt, E., and Wang, A. (2025). Superintelligence strategy: Expert version.

Hewage, A. (2023). Exploring the applicability of artificial intelligence in recruitment and selection processes: A focus on the recruitment phase. *Journal of Human Resource and Sustainability Studies*, 11:603–634.

Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A., Chowdhury, R., Dafoe, A., Hadfield, G., Levi, M., and Snidal, D. (2023). International institutions for advanced ai.

Horta Ribeiro, M., Ottoni, R., West, R., Almeida, V. A. F., and Meira Jr., W. (2020). Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 2020)*, pages 131–141, Barcelona, Spain. Association for Computing Machinery.

Huang, S., Siddarth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., and Ganguli, D. (2024). Collective constitutional ai: Aligning a language model with public input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1395–1417. ACM.

Jasanoff, S. (2004). *States of Knowledge: The Co-Production of Science and Social Order*. Routledge, London.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Kulothungan, V. and Gupta, D. (2025). Towards adaptive ai governance: Comparative insights from the u.s., eu, and asia.

Liang, F., Das, V., Kostyuk, N., and Hussain, M. M. (2018). Constructing a data-driven society: China's social credit system as a state surveillance infrastructure. *Policy & Internet*, 10(4):415–453.

Longpre, S., Klyman, K., Appel, R. E., Kapoor, S., Bommasani, R., Sahar, M., McGregor, S., Ghosh, A., Blili-Hamelin, B., Butters, N., Nelson, A., Elazari, A., Sellars, A., Ellis, C. J., Sherrets, D., Song, D., Geiger, H., Cohen, I., McIlvenny, L., Srikumar, M., Jaycox, M. M., Anderljung, M., Johnson, N. F., Carlini, N., Miailhe, N., Marda, N., Henderson, P., Portnoff, R. S., Weiss, R., Westerhoff, V., Jernite, Y., Chowdhury, R., Liang, P., and Narayanan, A. (2025). In-house evaluation is not enough: Towards robust third-party flaw disclosure for general-purpose ai.

MIT AI Risk Repository (2024). Domain taxonomy of AI risks. Accessed: 2025-01-15.

Nevo, S., Lahav, D., Karpur, A., Bar-On, Y., Bradley, H. A., and Alstott, J. (2024). Securing ai model weights: Preventing theft and misuse of frontier models. Technical Report RR-A2849-1, RAND Corporation. Research Report.

OpenAI, :, Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A. T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C. M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F. P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., von Lohmann, F., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G., Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H., Lightman, H., Chung, H. W., Kivlichan, I., O'Connell, I., Osband, I., Gilaberte, I. C., Akkaya, I., Kostrikov, I., Sutskever, I., Kofman, I., Pachocki, J., Lennon, J., Wei, J., Harb, J., Twore, J., Feng, J., Yu, J., Weng, J., Tang, J., Yu, J., Candela, J. Q., Palermo, J., Parish, J., Heidecke, J., Hallman, J., Rizzo, J., Gordon, J., Uesato, J., Ward, J., Huizinga, J., Wang, J., Chen, K., Xiao, K., Singhal, K., Nguyen, K., Cobbe, K., Shi, K., Wood, K., Rimbach, K., Gu-Lemberg, K., Liu, K., Lu, K., Stone, K., Yu, K., Ahmad, L., Yang, L., Liu, L., Maksin, L., Ho, L., Fedus, L., Weng, L., Li, L., McCallum, L., Held, L., Kuhn, L., Kondraciuk, L., Kaiser, L., Metz, L., Boyd, M., Trebacz, M., Joglekar, M., Chen, M., Tintor, M., Meyer, M., Jones, M., Kaufer, M., Schwarzer, M., Shah, M., Yatbaz, M., Guan, M. Y., Xu, M., Yan, M., Glaese, M., Chen, M., Lampe, M., Malek, M., Wang, M., Fradin, M., McClay, M., Pavlov, M., Wang, M., Wang, M., Murati, M., Bavarian, M., Rohaninejad, M., McAleese, N., Chowdhury, N., Chowdhury, N., Ryder, N., Tezak, N., Brown, N., Nachum, O., Boiko, O., Murk, O., Watkins, O., Chao, P., Ashbourne, P., Izmailov, P., Zhokhov, P., Dias, R., Arora, R., Lin, R., Lopes, R. G., Gaon, R., Miyara, R., Leike, R., Hwang, R., Garg, R., Brown, R., James, R., Shu, R., Cheu, R., Greene, R., Jain, S., Altman, S., Toizer, S., Toyer, S., Miserendino, S., Agarwal, S., Hernandez, S., Baker, S., McKinney, S., Yan, S., Zhao, S., Hu, S., Santurkar, S., Chaudhuri, S. R., Zhang, S., Fu, S., Papay, S., Lin, S., Balaji, S., Sanjeev, S., Sidor, S., Broda, T., Clark, A., Wang, T., Gordon, T., Sanders, T., Patwardhan, T., Sottiaux, T., Degry, T., Dimson, T., Zheng, T., Garipov, T., Stasi, T., Bansal, T., Creech, T., Peterson, T., Eloundou, T., Qi, V., Kosaraju, V., Monaco, V., Pong, V., Fomenko, V., Zheng, W., Zhou, W., McCabe, W., Zaremba, W., Dubois, Y., Lu, Y., Chen, Y., Cha, Y., Bai, Y., He, Y., Zhang, Y., Wang, Y., Shao, Z., and Li, Z. (2024). Openai o1 system card.

Ovadya, A., Redman, K., Thorburn, L., Chen, Q. Z., Smith, O., Devine, F., Konya, A., Milli, S., Revel, M., Feng, K. J. K., Zhang, A. X., Chandra, B., Bakker, M. A., and Kasirzadeh, A. (2025). Democratic ai is possible. the democracy levels framework shows how it might work.

Parthasarathy, A., Phalnikar, A., Jauhar, A., Somayajula, D., Krishnan, G. S., and Ravindran, B. (2024). Participatory approaches in ai development and governance: A principled approach.

Pasquale, F. (2015). *The Black Box Society*. Harvard University Press, Cambridge, MA.

Pazzaglia, ., Vendetti, V., Comencini, L. D., Deriu, F., and Modugno, V. (2025). Passing the turing test in political discourse: Fine-tuning llms to mimic polarized social media comments.

Priyanshu, A., Maurya, Y., and Hong, Z. (2024). Ai governance and accountability: An analysis of anthropic's claude.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., Anderljung, M., Garfinkel, B., Heim, L., Trask, A., Mukobi, G., Schaeffer, R., Baker, M., Hooker, S., Solaiman, I., Luccioni, A. S., Rajkumar, N., Moës, N., Ladish, J., Guha, N., Newman, J., Bengio, Y., South, T., Pentland, A., Koyejo, S., Kochenderfer, M. J., and Trager, R. (2024). Open problems in technical ai governance.

Reuel, A., Ghosh, A., Chim, J., Tran, A., Long, Y., Mickel, J., Gohar, U., Yadav, S., Ammanamanchi, P. S., Allaham, M., Rahmani, H. A., Akhtar, M., Friedrich, F., Scholz, R., Riegler, M. A., Batzner, J., Habba, E., Saxena, A., Kornilova, A., Wei, K., Soni, P., Mathew, Y., Klyman, K., Sania, J., Sahoo, S., Bruvik, O. B., Sadeghi, P., Goswami, S., Wang, A., Jernite, Y., Talat, Z., Biderman, S., Kochenderfer, M., Koyejo, S., and Solaiman, I. (2025). Who evaluates ai's social impacts? mapping coverage and gaps in first and third party evaluations.

Reuel, A. and Undheim, T. A. (2024). Generative ai needs adaptive governance.

Ribeiro, D., Rocha, T., Pinto, G., Cartaxo, B., Amaral, M., Davila, N., and Camargo, A. (2025). Toward effective ai governance: A review of principles.

Sahoo, S. (2025). The good, the bad, and the hybrid: A reward structure showdown in reasoning models training. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*.

Sahoo, S. and Dutta, K. (2024). Boardwalk empire: How generative ai is revolutionizing economic paradigms.

Salaudeen, O., Reuel, A., Ahmed, A., Bedi, S., Robertson, Z., Sundar, S., Domingue, B., Wang, A., and Koyejo, S. (2025). Measurement to meaning: A validity-centered framework for ai evaluation.

Shah, R., Irpan, A., Turner, A. M., Wang, A., Conmy, A., Lindner, D., Brown-Cohen, J., Ho, L., Nanda, N., Popa, R. A., Jain, R., Greig, R., Albanie, S., Emmons, S., Farquhar, S., Krier, S., Rajamanoharan, S., Bridgers, S., Ijitoye, T., Everitt, T., Krakovna, V., Varma, V., Mikulik, V., Kenton, Z., Orr, D., Legg, S., Goodman, N., Dafoe, A., Flynn, F., and Dragan, A. (2025). An approach to technical agi safety and security.

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., and Thompson, N. (2025). The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence.

Stańczak, K., Meade, N., Bhatia, M., Zhou, H., Böttinger, K., Barnes, J., Stanley, J., Montgomery, J., Zemel, R., Papernot, N., Chapados, N., Therien, D., Lillicrap, T. P., Marasović, A., Delacroix, S., Hadfield, G. K., and Reddy, S. (2025). Societal alignment frameworks can improve llm alignment.

Summerfield, C., Argyle, L., Bakker, M., Collins, T., Durmus, E., Eloundou, T., Gabriel, I., Ganguli, D., Hackenburg, K., Hadfield, G., Hewitt, L., Huang, S., Landemore, H., Marchal, N., Ovadya, A., Procaccia, A., Risse, M., Schneier, B., Seger, E., Siddarth, D., Sætra, H. S., Tessler, M., and Botvinick, M. (2024). How will advanced ai systems impact democracy?

Ter-Minassian, L. (2025). Democratizing ai governance: Balancing expertise and public participation.

Ulnicane, I. (2024). Governance fix? power and politics in controversies about governing generative ai. *Policy and Society*, 44(1):70–84.

von Rosing, M., Shepperson, L., and Czichos, H. (2025). Chapter 40 - sustainability board with veto rights. In von Rosing, M., editor, *The Sustainability Handbook, Volume 1*, pages 613–620. Elsevier.

Walters, J., Dey, D., Bhaumik, D., and Horsman, S. (2023). Complying with the eu ai act.

Zaidan, E. and Ibrahim, I. A. (2024). Ai governance in a complex and rapidly changing regulatory landscape: A global perspective. *Humanities and Social Sciences Communications*, 11(1):1121. Received 27 January 2024; Accepted 31 July 2024; Published 1 September 2024.

Zeng, Y., Klyman, K., Zhou, A., Yang, Y., Pan, M., Jia, R., Song, D., Liang, P., and Li, B. (2024). Ai risk categorization decoded (air 2024): From government regulations to corporate policies.

Zhong, H., Do, T., Jie, Y., Neuwirth, R. J., and Shen, H. (2025). Global ai governance: Where the challenge is the solution- an interdisciplinary, multilateral, and vertically coordinated approach.

Zwitter, A. (2024). Cybernetic governance: implications of technology convergence on governance convergence. *Ethics and Information Technology*, 26(2):1–13. Published 28 March 2024.

Appendix

A Incident Severity Score (ISS) - A Novel Metric for Governance

The ISS provides a rigorous framework for quantifying democratic risks while incorporating diverse stakeholder perspectives. We formalize risk assessment through a learnable model that captures complex interactions among democratic threats.

A.1 Classic Four-Factor ISS

We begin with four normalized incident attributes representing core dimensions of democratic risk assessment:

$$I \in [0,1] \quad \text{(Impact): Magnitude of democratic harm} \tag{1}$$
$$E \in [0,1] \quad \text{(Exploitability): Ease of malicious exploitation} \tag{2}$$
$$R \in [0,1] \quad \text{(Replicability): Potential for widespread replication} \tag{3}$$
$$X \in [0,1] \quad \text{(Exposure): Scale of population exposure} \tag{4}$$

Stakeholders assign nonnegative weights $\{w_I, w_E, w_R, w_X\}$ satisfying the normalization constraint:

$$w_I + w_E + w_R + w_X = 1, \quad \text{where } w_i \geq 0 \; \forall i \in \{I, E, R, X\} \tag{1}$$

This ensures stakeholder preferences form a valid probability distribution over risk dimensions.

A.1.1 Linear Aggregation

The linear ISS provides an intuitive weighted average of risk factors:

$$\text{ISS}_{\text{lin}} = w_I \cdot I + w_E \cdot E + w_R \cdot R + w_X \cdot X \in [0,1] \tag{2}$$

Properties: Additive risk combination, equal marginal contribution rates, suitable for independent risk factors.

A.1.2 Multiplicative Aggregation

The multiplicative ISS captures risk interdependencies through geometric aggregation:

$$\text{ISS}_{\text{mult}} = 1 - (1-I)^{w_I} \cdot (1-E)^{w_E} \cdot (1-R)^{w_R} \cdot (1-X)^{w_X} \in [0,1] \tag{3}$$

Properties: Superadditive risk combination, diminishing returns to individual factors, suitable for complementary risk interactions.

Selection Criterion: Use linear aggregation when risk factors contribute independently; use multiplicative aggregation when factors exhibit synergistic effects amplifying overall democratic harm.

A.2 High-Dimensional, Learnable ISS

To capture complex, higher-order interactions among d risk factors, we extend to a learnable framework accommodating richer democratic risk representations.

A.2.1 Risk Factor Representation

Let $\boldsymbol{f} = (f_1, f_2, \ldots, f_d)^T \in [0,1]^d$ represent the d-dimensional risk factor vector, where each f_i corresponds to a specific democratic risk category.

A.2.2 Parametric Model Architecture

Define the parameter set:

$$\boldsymbol{\theta} = \{\boldsymbol{w} \in \mathbb{R}^d, \boldsymbol{W} \in \mathbb{R}^{d \times d}, b \in \mathbb{R}\} \tag{5}$$

where:

- $\boldsymbol{w}$: Linear coefficients capturing first-order risk effects
- $\boldsymbol{W}$: Symmetric interaction matrix capturing pairwise risk synergies
- b: Bias term representing baseline democratic vulnerability

A.2.3 Second-Order Polynomial ISS

The high-dimensional ISS employs a second-order polynomial with sigmoid activation:

$$\text{ISS}(\boldsymbol{f};\boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\boldsymbol{f}) = \sigma(b + \boldsymbol{w}^T\boldsymbol{f} + \boldsymbol{f}^T\boldsymbol{W}\boldsymbol{f}) \in (0,1) \tag{4}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function ensuring bounded output in the open interval $(0,1)$, reflecting the asymptotic nature of absolute certainty in risk assessment.

Rationale: The quadratic term $\boldsymbol{f}^T\boldsymbol{W}\boldsymbol{f}$ captures pairwise risk interactions crucial for democratic contexts where individual risks may amplify each other non-linearly.

A.2.4 Parameter Learning via Maximum Likelihood

Given labeled historical incidents $\{(\boldsymbol{f}^{(n)}, y^{(n)})\}_{n=1}^N$ where $y^{(n)} \in [0,1]$ represents continuous severity labels (not binary), we optimize:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \left[\frac{1}{N}\sum_{n=1}^{N} L_{\text{Huber}}(y^{(n)}, h_{\boldsymbol{\theta}}(\boldsymbol{f}^{(n)})) + \lambda\|\boldsymbol{\theta}\|_2^2\right] \tag{5}$$

Corrected Loss Function: We replace binary cross-entropy with Huber loss to handle the continuous nature of democratic risk severity:

$$L_{\text{Huber}}(y,\hat{y}) = \begin{cases} \frac{1}{2}(y-\hat{y})^2 & \text{if } |y-\hat{y}| \le \delta \\ \delta|y-\hat{y}| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases} \tag{6}$$

with $\delta = 0.1$ chosen for robustness to outliers while maintaining sensitivity to precise risk gradations.

Regularization: $\lambda = 0.01$ prevents overfitting while $\lambda\|\boldsymbol{\theta}\|_2^2$ encourages sparse, interpretable risk interactions.

A.3 Embedding the Seven-Category Risk Taxonomy

A.4 Democratic Risk Categories

We map our comprehensive democratic risk framework into seven primary categories:

$$f_{\text{disc}} : \text{Discriminatory Discourse Amplification} \tag{7}$$
$$f_{\text{surv}} : \text{Surveillance and Democratic Chill} \tag{8}$$
$$f_{\text{elec}} : \text{Electoral Process Manipulation} \tag{9}$$
$$f_{\text{manip}} : \text{Public Opinion Manipulation} \tag{10}$$
$$f_{\text{civic}} : \text{Civic Engagement Degradation} \tag{11}$$
$$f_{\text{capture}} : \text{Regulatory and Institutional Capture} \tag{12}$$
$$f_{\text{emerg}} : \text{Emergent Democratic Threats} \tag{13}$$

A.4.1 Category-Specific Risk Computation

Each category aggregates multiple sub-risk components through L2 normalization:

$$\boldsymbol{f} = (f_{\text{disc}}, f_{\text{surv}}, f_{\text{elec}}, f_{\text{manip}}, f_{\text{civic}}, f_{\text{capture}}, f_{\text{emerg}})^T \tag{6}$$

Discriminatory Discourse (f_{disc}):

$$f_{\text{disc}} = \|\boldsymbol{r}_{\text{disc}}\|_2^{-1} \cdot (\alpha_1 \cdot \text{LM}_{\text{biasAmplification}} + \alpha_2 \cdot \text{syntheticContentBias} + \alpha_3 \cdot \text{languageExclusion}) \tag{14}$$

Surveillance Risks (f_{surv}):

$$f_{\text{surv}} = \|\boldsymbol{r}_{\text{surv}}\|_2^{-1} \cdot (\beta_1 \cdot \text{conversationalMonitoring} \\ + \beta_2 \cdot \text{politicalSentimentTracking} + \beta_3 \cdot \text{dissentDetection}) \quad (15)$$

Electoral Manipulation (f_{elec}):

$$f_{\text{elec}} = \|\boldsymbol{r}_{\text{elec}}\|_2^{-1} \cdot (\gamma_1 \cdot \text{AI}_{\text{generatedPropaganda}} \\ + \gamma_2 \cdot \text{personalizedPoliticalAds} + \gamma_3 \cdot \text{syntheticNewsGeneration}) \quad (16)$$

Opinion Manipulation (f_{manip}):

$$f_{\text{manip}} = \|\boldsymbol{r}_{\text{manip}}\|_2^{-1} \cdot (\delta_1 \cdot \text{conversationalManipulation} \\ + \delta_2 \cdot \text{LM}_{\text{botAmplification}} + \delta_3 \cdot \text{deepfakeTextGeneration}) \quad (17)$$

Civic Degradation (f_{civic}):

$$f_{\text{civic}} = \|\boldsymbol{r}_{\text{civic}}\|_2^{-1} \cdot (\epsilon_1 \cdot \text{AI}_{\text{echoAmplification}} \\ + \epsilon_2 \cdot \text{personalizationBubbles} + \epsilon_3 \cdot \text{LM}_{\text{radicalizationPathways}}) \quad (18)$$

Institutional Capture (f_{capture}):

$$f_{\text{capture}} = \|\boldsymbol{r}_{\text{capture}}\|_2^{-1} \cdot (\zeta_1 \cdot \text{modelConcentration} \\ + \zeta_2 \cdot \text{infrastructureDependence} + \zeta_3 \cdot \text{providerCapture}) \quad (19)$$

Emergent Threats (f_{emerg}):

$$f_{\text{emerg}} = \|\boldsymbol{r}_{\text{emerg}}\|_2^{-1} \cdot (\eta_1 \cdot \text{multiLM}_{\text{cascadeRisk}} \\ + \eta_2 \cdot \text{goalMisalignment} + \eta_3 \cdot \text{emergentBehaviors}) \quad (20)$$

A.4.2 Parameter Specifications

Sub-component weights $(\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \eta)$ are learned through stakeholder consultation and empirical validation:

- **Equal weighting baseline:** All sub-components weighted equally $(1/3)$ initially
- **Stakeholder adjustment:** Weights refined through multi-stakeholder deliberation
- **Empirical validation:** Final weights validated against historical democratic incidents

L2 Normalization: $\|\boldsymbol{r}_k\|_2^{-1}$ ensures each category contributes proportionally to overall risk assessment while preserving relative magnitudes within categories.

A.5 Stakeholder-Adaptive Weighting

A.5.1 Multi-Stakeholder Weight Aggregation

Each of the seven stakeholder groups $k \in \{1, 2, \ldots, 7\}$ proposes a weight vector $\boldsymbol{w}^{(k)} \in \Delta^{d-1}$ reflecting their risk prioritization.

Stakeholder Categories:

1. Democratic institutions ($k = 1$)
2. Civil society organizations ($k = 2$)
3. Regulatory bodies ($k = 3$)
4. Technical experts ($k = 4$)
5. Affected communities ($k = 5$)
6. Industry representatives ($k = 6$)
7. Academic researchers ($k = 7$)

A.5.2 Utility-Based Weight Aggregation

We aggregate stakeholder preferences through utility-weighted softmax:

$$u_k = \alpha_k \cdot \log p(\boldsymbol{\theta}^* \mid \text{stakeholder } k) + \beta_k \cdot \text{expertise}_k + \gamma_k \cdot \text{impact}_k \tag{7a}$$

$$\boldsymbol{w} = \text{Softmax}(\boldsymbol{u}) = \left(\frac{\exp(u_1)}{Z}, \ldots, \frac{\exp(u_7)}{Z} \right) \in \Delta^6 \tag{7b}$$

where $Z = \sum_{i=1}^{7} \exp(u_i)$ ensures proper normalization and $\boldsymbol{w} \in \Delta^6$ (6-dimensional probability simplex for 7 stakeholders).

Utility Components:

- $\alpha_k \cdot \log p(\boldsymbol{\theta} \mid \text{stakeholder } k)$: Stakeholder-specific model likelihood
- $\beta_k \cdot \text{expertise}_k$: Technical expertise weighting
- $\gamma_k \cdot \text{impact}_k$: Direct impact severity weighting

Parameter Values:

- $\alpha_k = 1.0$: Equal evidential weighting across stakeholders
- $\beta_k \in [0.5, 1.5]$: Expertise-based adjustment factors
- $\gamma_k \in [0.8, 2.0]$: Impact-based adjustment factors (highest for affected communities)

A.6 Phase-Dependent Trigger Thresholds

A.6.1 Temporal Risk Threshold Evolution

Let $S = \text{ISS}(\boldsymbol{f}; \boldsymbol{\theta}^*)$ represent the computed severity score with empirical cumulative distribution function F_S derived from historical incident data.

For each intervention level $j \in \{L, M, H\}$ (Low, Moderate, High) and time $t \in [0, 1]$ representing progress through our six-year implementation roadmap, we define evolving thresholds:

$$s_j(t) = (1 - \varphi(t)) \cdot s_j^{\text{init}} + \varphi(t) \cdot s_j^{\text{full}} \tag{8a}$$

$$\alpha_j(t) = (1 - \varphi(t)) \cdot \alpha_j^{\text{init}} + \varphi(t) \cdot \alpha_j^{\text{full}} \tag{8b}$$

where:

- $\varphi(t)$: Smooth transition function from initial to full deployment phases
- $s_j^{\text{init}}, s_j^{\text{full}}$: Initial and mature-phase severity thresholds
- $\alpha_j^{\text{init}}, \alpha_j^{\text{full}}$: Initial and mature-phase probability thresholds

A.6.2 Probabilistic Trigger Mechanism

Intervention j triggers when the probability of exceeding threshold $s_j(t)$ meets the confidence requirement:

$$P(S \geq s_j(t)) = 1 - F_S(s_j(t)) \geq \alpha_j(t) \tag{9}$$

Threshold Specifications:

Low Intervention ($j = L$): Enhanced monitoring

- $s_L^{\text{init}} = 0.2, s_L^{\text{full}} = 0.3$
- $\alpha_L^{\text{init}} = 0.1, \alpha_L^{\text{full}} = 0.15$

Moderate Intervention ($j = M$): Regulatory review

- $s_M^{\text{init}} = 0.5, s_M^{\text{full}} = 0.6$
- $\alpha_M^{\text{init}} = 0.05, \alpha_M^{\text{full}} = 0.1$

High Intervention ($j = H$): Emergency response

- $s_H^{\text{init}} = 0.8, s_H^{\text{full}} = 0.75$
- $\alpha_H^{\text{init}} = 0.01, \alpha_H^{\text{full}} = 0.05$

Rationale: Thresholds become more sensitive (lower s_j^{full}) and require higher confidence (higher α_j^{full}) as governance systems mature, reflecting improved institutional capacity and democratic risk awareness.

A.6.3 Transition Function

The phase transition function $\varphi(t)$ ensures smooth threshold evolution:

$$\varphi(t) = 3t^2 - 2t^3 \quad \text{for } t \in [0, 1] \tag{10}$$

This S-curve provides gradual initial transition, rapid mid-phase evolution, and stabilization approaching full deployment.

B Unified ISS Framework Integration

B.1 Complete Mathematical Pipeline

The complete ISS computation integrates all components:

1. **Risk Assessment:** Compute seven-category risk vector $\boldsymbol{f}$ using equation (6)
2. **Stakeholder Weighting:** Aggregate stakeholder preferences via equations (7a-7b)
3. **ISS Computation:** Calculate severity score using equation (4) with learned parameters $\boldsymbol{\theta}^*$
4. **Threshold Evaluation:** Compare against phase-dependent thresholds using equations (8-9)
5. **Intervention Triggering:** Activate appropriate governance responses based on probabilistic triggers

B.2 Computational Complexity

- **Training Phase:** $O(Nd^2 + N \log N)$ where $N =$ training samples, $d = 7$ risk categories
- **Inference Phase:** $O(d^2 + K)$ where $K = 7$ stakeholder groups
- **Memory Requirements:** $O(d^2 + Kd + N)$ for parameters, stakeholder weights, and training data

C Summary of Mathematical Framework

- **Equations (1-3):** Classic four-factor ISS with linear and multiplicative aggregation options
- **Equations (4-5):** High-dimensional learnable ISS with second-order polynomial architecture and Huber loss optimization
- **Equation (6):** Seven-category democratic risk taxonomy mapping with L2 normalization
- **Equations (7a-7b):** Multi-stakeholder weight aggregation via utility-based softmax ensuring $\boldsymbol{w} \in \Delta^6$
- **Equations (8-10):** Phase-dependent probabilistic trigger thresholds with smooth temporal evolution

This unified ISS framework provides mathematically rigorous, democratically grounded, and computationally tractable risk assessment for language model governance, integrating stakeholder pluralism with technical precision to enable evidence-based democratic oversight of AI systems.

D Addressing ISS Design Choices and Limitations

D.1 On the Political Nature of Risk Quantification

creating a single numerical score from heterogeneous stakeholder input is inherently political. This is correct and intentional. We don't want to entertain the premise that governance frameworks should aspire to political neutrality. As governance systems already encode political choices through their design, deployment, and impact distribution.

Our ISS framework makes three design commitments:

Explicit rather than hidden politics: Traditional "neutral" risk assessments embed implicit political choices (e.g., weighting individual privacy equally to corporate efficiency). Our multi-stakeholder weighting (Eq. 7a-7b) makes these tradeoffs explicit and contestable.

Structured aggregation over ad-hoc judgment

While the polynomial structure (Eq. 4) involves design choices, it provides:

(i) mathematical consistency;
(ii) interpretable first-order ($\mathbf{w}^\top \mathbf{f}$) and interaction ($\mathbf{f}^\top \mathbf{W} \mathbf{f}$) effects;
(iii) empirical falsifiability via Huber loss optimization (Eq. 5).

Adaptive rather than fixed: The phase-dependent thresholds (Eq. 8-10) acknowledge that "appropriate" risk levels evolve with institutional capacity and democratic norms.

D.2 Alternative Aggregation Functions

We selected second-order polynomials for tractability and interpretability. Alternative approaches include:

- Rank-based methods: Median or quantile aggregation across stakeholder assessments (loses granularity, gains robustness to outliers)
- Deliberative consensus: Iterated stakeholder negotiation (addresses legitimacy, sacrifices scalability)
- Neural architectures: Deep networks for f $\rightarrow$ ISS mapping (gains expressiveness, loses interpretability)

Future work should empirically compare these against our baseline using historical governance cases.

D.3 On Quantifying Qualitative Harms

We acknowledge the **difficulty of reducing democratic threats to four scalar values**. We emphasize:

- The ISS provides governance **signals**, not comprehensive impact **assessments**
- Equation (6) decomposes risks into seven interpretable categories before aggregation
- High ISS scores trigger qualitative processes (stakeholder deliberation, community review) rather than mechanical responses
- Section 6 monitoring includes democratic health indicators beyond ISS

The framework treats quantification as a necessary but insufficient governance input, not a substitute for democratic judgment.

E Handling Stakeholder Conflicts

Conflict Resolution Protocol

When stakeholder assessments diverge significantly (variance in $w^{(k)}$ exceeding a predefined threshold), the framework employs structured disagreement procedures:

(i) mandatory deliberation rounds in which groups articulate the sources of their divergent risk perceptions;

(ii) sensitivity analysis illustrating how alternative weightings affect ISS scores, thereby revealing whether disagreement is fundamental or marginal;

(iii) in cases of irreconcilable conflict, decisions default to the most protective assessment from directly affected communities, implementing a precautionary principle that privileges experiential knowledge over purely technical optimization.

F Retrospective Validation Strategy

The ISS framework will be validated through retrospective analysis of documented AI governance failures including: Cambridge Analytica (2018, electoral manipulation), China's social credit systems (ongoing, surveillance), and content moderation failures during 2020 elections. For each case, we will:

- reconstruct the risk vector f using contemporaneous evidence,
- compute ISS scores under different stakeholder weightings,
- assess whether appropriate thresholds would have triggered intervention, and
- compare framework recommendations against actual outcomes.

This retrospective testing will calibrate thresholds (Section A.6) and validate stakeholder weighting procedures (Eq. 7) before prospective deployment.

G Enforcement Architecture

High-risk determinations trigger graduated enforcement mechanisms adapted from existing regulatory frameworks. For applications exceeding moderate thresholds (ISS > 0.6), model safety committees may: (i) require pre-deployment impact assessments with stakeholder consultation periods (15-30 days), (ii) mandate design modifications including capability restrictions or alignment interventions, (iii) impose operational monitoring requirements with regular compliance audits, or (iv) in extreme cases (ISS > 0.8), issue temporary deployment suspensions pending comprehensive review. Enforcement authority derives from three sources: regulatory mandates for covered entities (Phase 2+), contractual requirements in public procurement, and reputational mechanisms through public ISS score disclosure. Civil society veto rights (Section 4) operate through formal objection processes where affected communities can petition for independent review, triggering mandatory committee reconsideration with burden of proof on deployers to demonstrate risk mitigation.

H Capacity Building Strategy

Municipal pilots (Phase 1) leverage existing institutional infrastructure to minimize incremental resource requirements. Implementation proceeds through:

1. **Shared technical infrastructure**: A centralized ISS computation platform maintained at the regional/national level eliminates per-municipality software development costs; local bodies access it through standardized API interfaces, requiring only basic computing resources.
2. **Distributed expertise networks**: University partnerships provide technical evaluation capacity through structured practicum programs in which graduate students gain governance experience while providing *pro bono* risk assessments under faculty supervision—converting educational requirements into governance capacity.
3. **Tiered assessment protocols**: The ISS framework stratifies evaluations by complexity—routine low-risk applications proceed via automated preliminary screening (computation time: minutes); moderate-risk cases undergo streamlined stakeholder consultation (timeline: single deliberation session); only high-risk deployments require comprehensive multi-week assessments.

4. **Knowledge commons approach**: All assessment methodologies, training materials, and case precedents are released under open licenses, enabling later-adopting municipalities to implement oversight at a fraction of pioneer costs through documented best practices.

By Phase 3, marginal costs for additional oversight bodies approach minimal operational expenses (meeting coordination, administrative staff time) rather than full-cost institutional buildout, as enabling infrastructure—technical tools, trained evaluator pools, and procedural templates—exists as public goods.

I Some Open Problems

Despite growing work on democratic AI governance, several critical research challenges remain unresolved. The following open problems highlight areas where conceptual clarity, methodological innovation, and institutional design are still urgently needed.

Critical Open Problems

- **Measuring democratic health:** How can we operationalize and validate "democratic health" when competing theories emphasize different values (e.g., participation vs. efficiency, representation vs. expertise)?
- **Preventing capture:** What mechanisms can safeguard multi-stakeholder governance processes from domination by well-organized or well-resourced actors that risk reproducing existing power asymmetries?
- **Handling value conflict:** How should the framework address fundamental disagreements between stakeholder groups about core democratic values, especially when consensus-building may entrench the status quo and block structural reforms?
- **Scaling deliberation:** Can deliberative processes remain legitimate and effective when scaled from local communities to national or international arenas where direct, face-to-face engagement is infeasible?
- **Balancing expertise and participation:** How do we reconcile democratic inclusion with the technical complexity of AI systems that require specialized expertise for effective evaluation?
- **Defining community sovereignty:** What boundaries should govern local decision-making in AI oversight when outcomes have global implications or clash with universal human rights principles?
- **Addressing temporal mismatch:** How can the framework mitigate the gap between rapid AI development cycles and slower democratic deliberation without undermining either innovation or oversight?
- **Evaluating outcomes:** What metrics should define "success" in democratic AI governance, and how can we assess whether multi-stakeholder oversight produces better outcomes than existing regulatory approaches?

Points to Ponder

This policy should **not** be implemented in countries with a dictatorial or authoritarian approach. Such regimes typically lack the transparency, institutional checks, and civic accountability required for ethical AI governance. Implementing this framework in such contexts may enable state overreach, surveillance misuse, and suppression of fundamental rights.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Both Abstract and Introduction accurately reflect the contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Follow the limitations Section along with Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: I wrote the ISS in Appendix and all formulae and proofs are elaborated.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: This Technical AI Governance paper does not need any statistical validations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This policy paper follows all the ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Follow social impact statement section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Every asset is properly credited as per the ethical bounds of both the authors.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This policy paper shows how to ethically implement an AI policy framework in a **democratic** and sovereign country.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.