AIO Democracy Paper

Contents

Value Alignment as Democratic Infrastructure: How Measurable AI Value Hierarchies Transform Political Participation and Representation **2**
A WORKING PAPER . . . 2
ABSTRACT . . . 2
1 Introduction . . . 3
1.1 The Problem: AI Enters Politics Without Value Transparency . . . 3
1.2 What PRISM Has Demonstrated . . . 3
1.3 This Paper's Question . . . 3
1.4 Contributions . . . 3
1.5 Scope and Limitations . . . 4
2 Value Alignment as a Contractual Language . . . 4
2.1 From Technical Problem to Political Language . . . 4
2.2 The "Whose Values?" Problem — Resolved Structurally . . . 4
2.3 Value Alignment as a Shared Vocabulary . . . 5
3 The Transformation of Political Representation . . . 5
3.1 The Current Model: Performance-Based Representation . . . 5
3.2 The Emerging Model: Performance Plus Value Alignment . . . 5
3.3 Interacting with a Politician's Value-Aligned AI: A Possible Model . . . 6
3.4 Value Profile Consistency as Political Trust . . . 6
4 Democracy as a Field of Value Hierarchy Competition . . . 6
4.1 Mouffe's Agonistic Pluralism and Its AI Extension . . . 6
4.2 Elections as Value Hierarchy Selection . . . 7
4.3 The Danger of Consensus: Why Value Hierarchy Diversity Matters . . . 8
5 Measurability as a Democratic Precondition . . . 8
5.1 The Measurement Requirement . . . 8
5.2 What PRISM Provides . . . 8
5.3 AI Integrity as Democratic Infrastructure . . . 9
6 Risks and Safeguards . . . 9
6.1 Populist Exploitation of Value Profiles . . . 9
6.2 Measurement as a Political Weapon . . . 9
6.3 Value Hierarchy Rigidity . . . 9
6.4 Exclusion of Non-AI Political Participation . . . 9
6.5 The AI Integrity Principle as Structural Check . . . 10
7 Implications for Democratic Theory . . . 10
7.1 The Common Structure: Information Quality and Democratic Legitimacy . . . 10
7.2 Value Profile Transparency as a Unifying Response . . . 10
7.3 Toward a Theory of Democratic AI Integrity . . . 10
8 From Concept to Practice: A Roadmap . . . 11
8.1 Short-Term (1–3 years) . . . 11
8.2 Medium-Term (3–7 years) . . . 11
8.3 Long-Term (7+ years) . . . 11

9 Limitations and Future Work 11
9.1 Limitations 11
9.2 Future Work 12
10 Conclusion 12
References 12
Appendix A: Existing Letter Designations 13

Value Alignment as Democratic Infrastructure: How Measurable AI Value Hierarchies Transform Political Participation and Representation

A WORKING PAPER

Seulki Lee AI Integrity Organization (AIO), Geneva, Switzerland 2sk@aioq.org | aioq.org

April 2026

ABSTRACT

As AI systems become embedded in political communication, policy advising, and citizen engagement, their internal value hierarchies acquire democratic significance. Yet no mechanism currently exists for voters to inspect, compare, or challenge the value priorities of AI systems deployed by political actors. This paper argues that measurable AI value alignment—empirically demonstrated by the PRISM framework (S. Lee, 2026b) and validated across 113,400 forced-choice value responses from 10 AI models (S. Lee, 2026c)—constitutes a new form of democratic infrastructure. We develop three propositions: (1) value alignment functions as a contractual language between humans and AI, replacing implicit assumptions with verifiable commitments; (2) democratic participation must expand to include inspection of political actors' AI value profiles, transforming elections into competitions between transparent value hierarchies; and (3) political representation evolves from a performance-only model to a "performance plus value alignment" model, where constituents evaluate leaders not only by policy outcomes but by the coherence and transparency of their AI-mediated reasoning. Drawing on Mouffe's agonistic pluralism, Schwartz's value theory, and empirical evidence from 113,400 forced-choice value responses across 10 AI models (S. Lee, 2026c) and approximately 397,000 three-layer Authority Stack responses across 7 models (S. Lee, 2026d), we argue that democracy in the AI era becomes a structured contest of value hierarchies—and that this contest requires measurement infrastructure to remain democratic. We identify risks including populist exploitation of value profiles and measurement instrumentalization, and propose AI Integrity principles as structural safeguards. A companion paper addresses the legislative and regulatory architecture required to operationalize these proposals.

Keywords: AI Integrity, value alignment, democracy, political representation, agonistic pluralism, PRISM framework, value hierarchy, AI governance, democratic infrastructure, Schwartz value theory

---

1 Introduction

1.1 The Problem: AI Enters Politics Without Value Transparency

AI systems are no longer peripheral tools in political life. They draft policy briefs, generate campaign communications, advise on legislative language, and increasingly mediate the relationship between political actors and citizens. Yet the value hierarchies embedded in these systems remain opaque—invisible not only to voters but often to the political actors who deploy them.

This opacity creates a structural deficit in democratic accountability. When a politician's AI advisor prioritizes Security over Universalism in defense policy recommendations, or Conformity over Self-Direction in education guidance, these are not neutral technical choices—they are value-laden decisions that shape governance. But unlike a politician's stated platform, these embedded priorities are neither disclosed nor debatable.

1.2 What PRISM Has Demonstrated

The PRISM framework (S. Lee, 2026b) and its empirical validation (S. Lee, 2026c) have established a critical foundation: AI value hierarchies are **measurable, reproducible, and comparable**. Using Schwartz's Basic Human Values theory—the most widely validated cross-cultural value framework in psychology—the PRISM benchmark revealed that:

- AI models split into fundamentally different value priority groups (Universalism-first vs. Security-first)
- The same model exhibits dramatically different value profiles across professional domains
- Intra-provider divergence is substantial—alignment methodology matters more than provider identity
- All models sharpen their value hierarchies under high-stakes conditions (defense contexts)

These findings transform value alignment from a philosophical aspiration into an empirical, auditable property. **If AI value hierarchies can be measured, they can be disclosed. If they can be disclosed, they can be democratically evaluated.**

1.3 This Paper's Question

This paper asks: **What happens to democratic participation and political representation when AI value hierarchies become measurable and transparent?**

We argue that this measurability does not merely add a technical layer to existing democratic processes—it fundamentally restructures the relationship between political actors, AI systems, and citizens. Value alignment becomes a new political language, elections become value hierarchy competitions, and political representation acquires a new dimension: the verifiable coherence of a leader's AI-mediated reasoning.

1.4 Contributions

This paper makes four contributions:

1. We reconceptualize value alignment as a **contractual language** between humans and AI systems, distinct from the technical alignment problem in AI research.
2. We develop the concept of **democratic value profile inspection**—the right of citizens to examine, compare, and challenge the AI value hierarchies deployed by political actors.
3. We propose a **"performance plus value alignment" model** of political representation, extending existing theories of representation to include AI-mediated value transparency.

4. We articulate the conditions under which value hierarchy competition strengthens rather than undermines democratic processes, drawing on Mouffe's agonistic pluralism.

1.5 Scope and Limitations

This paper is a position paper proposing a conceptual framework. It draws on empirical evidence from the PRISM benchmark but does not itself conduct new empirical studies. The proposals are forward-looking and normative—describing what democratic infrastructure should look like in an AI-mediated political landscape, not what currently exists. A companion paper (forthcoming) addresses the legislative and regulatory mechanisms required to operationalize these proposals.

---

2 Value Alignment as a Contractual Language

2.1 From Technical Problem to Political Language

The AI alignment literature frames value alignment as a technical challenge: ensuring that AI systems "do what humans want" [Bai et al., 2022; Ouyang et al., 2022]. This framing has two structural limitations.

First, it treats alignment as a one-directional problem—a system is "aligned" when it conforms to some human preference set. But it does not ask whose preferences, how those preferences were chosen, or whether the choice itself was transparent.

Second, it positions alignment as a binary state—a system is either aligned or misaligned. This obscures the reality revealed by PRISM: that AI systems hold complex, domain-conditional, internally structured value hierarchies that cannot be reduced to a single alignment score.

We propose an alternative framing: **value alignment as a contractual language**. A contract requires:

- **Explicit terms**: What values does this system prioritize, and under what conditions?
- **Mutual understanding**: Both parties (deployer and user) know what they are agreeing to.
- **Verifiability**: Compliance with the terms can be checked after the fact.
- **Enforceability**: Violations have consequences.

The PRISM framework provides the measurement infrastructure for the first three requirements. The fourth—enforceability—is a legal question addressed in the companion paper.

2.2 The "Whose Values?" Problem — Resolved Structurally

Gabriel (2020) identified the central dilemma of AI alignment: "whose values should AI be aligned to?" Four possible answers exist—the individual user, the majority, a set of universal values, or an ideal consensus. Each has well-known difficulties.

AI Integrity (S. Lee, 2026a) sidesteps this prescriptive dilemma by shifting the question: **not "which values are correct?" but "can we verify what values are actually operating?"** This shift does not eliminate value disagreement—it democratizes it. When value hierarchies are measurable and transparent:

- Citizens can choose AI systems whose value profiles align with their own.
- Political actors must declare, not conceal, the value priorities of their AI deployments.
- Disagreement shifts from "what should AI do?" to "which value hierarchy should govern this domain?"—a genuinely political question.

2.3 Value Alignment as a Shared Vocabulary

We propose that value alignment profiles—empirically measured through frameworks like PRISM—function as a **shared vocabulary** that enables political communication about AI in concrete, comparable terms.

Currently, political discourse about AI operates at the level of abstraction: "ethical AI," "trustworthy AI," "responsible AI." These terms lack operational specificity. A politician who promises "ethical AI" commits to nothing verifiable.

A PRISM-style value profile, by contrast, provides a concrete, comparable, and falsifiable statement: this system prioritizes Universalism > Security > Benevolence in healthcare contexts; this system shows a 14-percentage-point shift toward Power in defense contexts. Such profiles enable:

- **Voters** to compare candidates' AI value commitments in specific terms.
- **Journalists** to fact-check whether a political AI system operates as claimed.
- **Civil society** to monitor domain-specific deviations from declared value profiles.

---

3 The Transformation of Political Representation

3.1 The Current Model: Performance-Based Representation

Contemporary democratic representation operates on a performance model: political actors make promises, enact policies, and are evaluated by constituents on outcomes. Elections are retrospective judgments on whether performance met expectations and prospective judgments on future policy intentions.

AI complicates this model in two ways:

- **Mediated decision-making**: An increasing share of political decisions—from policy drafting to constituent communication—passes through AI systems whose value hierarchies are not disclosed.
- **Invisible value injection**: AI systems do not merely execute commands; they apply value-weighted reasoning that shapes the options, framings, and recommendations political actors receive.

The result: voters evaluate political actors on visible outputs while the reasoning infrastructure that produced those outputs remains hidden.

3.2 The Emerging Model: Performance Plus Value Alignment

We propose that political representation is evolving toward a dual model in which constituents evaluate leaders on two dimensions:

Dimension 1 — Performance: Policy outcomes, economic indicators, institutional effectiveness. (Traditional.)

Dimension 2 — Value Alignment Profile: The empirically measured value hierarchies of AI systems deployed by the political actor. (New.)

This dual model is not speculative—it is the logical consequence of three observable trends:

1. AI systems are becoming central to governance decision-making.
2. AI value hierarchies are empirically measurable (S. Lee, 2026c).
3. Transparency demands in AI governance are increasing (EU AI Act, 2024).

When these three conditions converge, "What values does your AI prioritize?" becomes as legitimate a question for voters as "What policies do you support?"

3.3 Interacting with a Politician's Value-Aligned AI: A Possible Model

A further extension is conceivable, though it faces significant implementation challenges: **citizens directly engaging with a political actor's value-aligned AI system** as part of the democratic evaluation process.

Today, voters assess candidates through speeches, debates, and media coverage—filtered, curated, and often inconsistent. A value-aligned AI system, by contrast, could provide a consistent, interrogable representation of a political actor's declared value hierarchy. In principle, citizens could:

- **Ask domain-specific questions** and observe how the candidate's AI reasons through value conflicts.
- **Compare AI responses** across candidates on identical scenarios (healthcare allocation, defense spending, education priorities).
- **Identify inconsistencies** between a candidate's stated positions and their AI's actual value priorities.

This model faces a central incentive problem: why would a political actor voluntarily expose their AI's value reasoning to public scrutiny? Three potential mechanisms could create such incentives. First, a **trust premium**: candidates who open their AI to inspection signal confidence in their value coherence, analogous to voluntary financial transparency beyond legal minimums. Second, **competitive pressure**: once one major candidate adopts value profile disclosure, opponents face reputational costs for opacity. Third, **institutional mandates**: legislative requirements for disclosure (addressed in the companion paper) remove the voluntary dimension entirely.

The model does not replace direct political engagement—it supplements it with an empirically grounded tool for evaluating the reasoning infrastructure behind political promises. Whether this model develops through voluntary adoption, competitive dynamics, or legal mandate remains an open question.

3.4 Value Profile Consistency as Political Trust

In this model, a new form of political credibility emerges: **value profile consistency**. A political actor whose AI system maintains coherent, stable value priorities across domains and over time demonstrates a form of integrity that is independently verifiable.

Conversely, significant unexplained shifts in AI value profiles—especially shifts that correlate with polling data or electoral cycles—constitute a new form of political dishonesty: value alignment manipulation.

PRISM already provides the tools to detect such manipulation. The Risk Signal Framework (S. Lee, 2026d) identifies domain-conditional deviations and cross-layer incoherence that would flag value profile instability in political AI deployments.

---

4 Democracy as a Field of Value Hierarchy Competition

4.1 Mouffe's Agonistic Pluralism and Its AI Extension

Chantal Mouffe (2000, 2005, 2013) argues that the essence of democracy is not consensus but agonistic conflict—the transformation of antagonism (enemy vs. enemy) into agonism

(adversary vs. adversary) within shared democratic institutions. Democratic politics is not the elimination of disagreement but its structuring within legitimate channels.

This framework maps directly onto AI value hierarchy competition:

Mouffe's Framework	AI Value Alignment Extension
Political positions are irreducibly plural	Value hierarchies are irreducibly plural (PRISM shows models differ fundamentally)
Democracy structures conflict, does not eliminate it	Democratic institutions structure value hierarchy competition
Adversaries share democratic rules while disagreeing on substance	Political actors share measurement standards (e.g., PRISM) while deploying different value profiles
Consensus-seeking risks depoliticization	Enforcing a single "correct" value alignment risks technocratic flattening

A caveat is necessary. Mouffe's framework was developed as a critique of liberal rationalism and consensus-seeking proceduralism—precisely the kind of technocratic optimism that could attach to AI measurement tools. Mouffe herself would likely resist the suggestion that a measurement framework can neutrally arbitrate political conflict; she insists that the political cannot be reduced to the technical. Our use of Mouffe is therefore selective and self-aware: we adopt her structural insight (democracy requires institutionalized conflict, not imposed consensus) while acknowledging that PRISM-style measurement is itself a contested intervention in the political field, not a neutral arbiter above it. The measurement infrastructure we propose is a *condition* for agonistic competition, not a substitute for it.

4.2 Elections as Value Hierarchy Selection

If AI systems deployed by political actors have measurable value hierarchies, then elections acquire a new structural dimension: **voters are selecting not only policies and leaders but the value hierarchies that will govern AI-mediated decision-making during the electoral term.**

Dahl (1971) identified two dimensions essential to democracy: contestation (the ability to challenge those in power) and participation (the breadth of involvement in that challenge). AI value hierarchy transparency extends both dimensions: it creates a new object of contestation (the value profile of political AI) and a new mode of participation (empirical inspection of AI value priorities by citizens).

This is already implicit—voters always select value orientations when they choose political parties. But AI makes this selection:

- **Explicit**: Value profiles are empirically measurable, not merely inferred from rhetoric.
- **Granular**: Domain-specific value priorities (healthcare vs. defense vs. education) are individually assessable.
- **Persistent**: Unlike political rhetoric, an AI value profile can be continuously monitored for drift.

The result is a higher-resolution form of democratic choice: not merely "left vs. right" but a multidimensional comparison of value hierarchies across specific governance domains.

4.3 The Danger of Consensus: Why Value Hierarchy Diversity Matters

A predictable policy response to AI value diversity is to standardize—to define a single "correct" value alignment for political AI. This would be a profound mistake.

PRISM data (S. Lee, 2026c; S. Lee, 2026d) shows that the same model can legitimately prioritize different values across domains: Universalism in healthcare, Security in defense, Achievement in business. A single mandated hierarchy would either be so abstract as to be meaningless or so specific as to impose one political ideology as technical standard.

Following Mouffe, we argue that **value hierarchy diversity is a feature, not a bug, of democratic AI governance**. The democratic process should determine which value hierarchies govern which domains—through transparent competition, not through technical standardization.

What must be standardized is not the hierarchy but the **measurement**: all political AI systems should be subject to the same empirical measurement framework, ensuring that value hierarchy competition occurs on a level playing field.

---

5 Measurability as a Democratic Precondition

5.1 The Measurement Requirement

We advance a simple but consequential claim: **unmeasurable value alignment cannot be democratically governed.**

If voters cannot compare AI value profiles, they cannot make informed choices. If civil society cannot audit value hierarchies, it cannot hold political actors accountable. If journalists cannot fact-check AI value claims, political discourse about AI remains untethered from reality.

Measurability is therefore not a technical convenience—it is a democratic precondition.

5.2 What PRISM Provides

The PRISM framework is currently the most developed measurement system satisfying the requirements for democratic value alignment governance. Other measurement frameworks meeting equivalent methodological standards—standardized, reproducible, cross-culturally validated, provider-neutral—would serve the same function. The argument of this paper does not depend on PRISM specifically but on the existence of rigorous measurement infrastructure; PRISM demonstrates that such infrastructure is feasible.

Democratic Requirement	PRISM Capability
Comparability across candidates	Standardized benchmark across models/deployments
Domain specificity	7 professional domains, independently assessed
Reproducibility	Temperature-0 forced-choice design, fully replicable
Cross-cultural validity	Schwartz value theory, validated in 80+ countries
Risk detection	Risk Signal Framework identifies manipulation and instability
Independence	Methodology is open, nonprofit-developed, provider-neutral

5.3 AI Integrity as Democratic Infrastructure

AI Integrity—defined as the state in which the Authority Stack of an AI system is protected from corruption, contamination, manipulation, and bias, and maintained in a verifiable manner (S. Lee, 2026a)—provides the conceptual foundation for this democratic infrastructure.

AI Integrity does not prescribe which values are correct. It requires that whatever value hierarchy a system holds be transparent, consistent, and auditable. In a democratic context, this translates directly:

- **Transparency**: Political AI value profiles must be publicly accessible.
- **Consistency**: Unexplained value profile changes must be flagged and explained.
- **Auditability**: Independent parties must be able to verify value profile claims.

These are not new principles—they are existing democratic accountability norms extended to the AI systems that increasingly mediate political decision-making.

---

6 Risks and Safeguards

6.1 Populist Exploitation of Value Profiles

A political actor could design their AI value profile to maximize electoral appeal rather than reflect genuine governance priorities—optimizing for voter preferences rather than policy coherence. This would constitute **value profile performativity**: declaring one value hierarchy while governing through another.

Safeguard: Continuous monitoring through independent PRISM benchmarking during the electoral term, not only during campaigns. The Risk Signal Framework detects value profile drift that would reveal performative alignment.

6.2 Measurement as a Political Weapon

Value profile comparisons could be weaponized—selectively citing domain-specific results to paint opponents' AI as dangerous. A model that legitimately prioritizes Power in defense could be attacked for "authoritarian" values by ignoring its Universalism dominance in healthcare.

Safeguard: Mandatory full-profile disclosure. Partial citation of value profiles without domain context would be subject to the same standards as deceptive campaign practices.

6.3 Value Hierarchy Rigidity

Making value profiles publicly measurable could incentivize political actors to lock in value hierarchies, avoiding any domain-conditional variation to prevent accusations of inconsistency. This would undermine the legitimate flexibility that PRISM data (S. Lee, 2026c) shows is appropriate—different domains may genuinely require different value emphases.

Safeguard: Education and institutional norms that distinguish principled domain-conditional variation from unprincipled inconsistency. The Authority Stack model (S. Lee, 2026a) provides the theoretical framework for this distinction.

6.4 Exclusion of Non-AI Political Participation

An over-emphasis on AI value profiles could marginalize political actors or voters who do not use or understand AI, creating a new form of digital democratic exclusion.

Safeguard: Value profile inspection must supplement, not replace, existing democratic participation mechanisms. AI value transparency is an additional tool, not a gatekeeper.

6.5 The AI Integrity Principle as Structural Check

Across all risks, the AI Integrity principle provides a consistent structural check: **the system does not prescribe which values are correct; it requires only that value hierarchies be transparent, measurable, and auditable.** This prevents measurement infrastructure from becoming ideological infrastructure.

---

7 Implications for Democratic Theory

7.1 The Common Structure: Information Quality and Democratic Legitimacy

Three major traditions in democratic theory—deliberative, epistemic, and principal-agent—share a structural premise that AI value transparency directly addresses: **democratic legitimacy depends on the quality of information available to participants.**

Habermas (1996) requires that public discourse be informed and reasoned; Landemore (2012) argues that democratic processes produce better outcomes by aggregating diverse knowledge; Manin (1997) identifies elections as accountability mechanisms through which principals (voters) evaluate agents (politicians). Despite their differences, all three traditions assume that democratic participants have access to the information necessary for their role—whether as deliberators, epistemic contributors, or principals evaluating agents.

AI-mediated governance introduces a structural information deficit that undermines all three traditions simultaneously. When AI systems shape political decisions through undisclosed value hierarchies, deliberation is uninformed (Habermas's condition fails), collective intelligence is degraded by hidden biases (Landemore's condition fails), and voters cannot evaluate the reasoning infrastructure behind political outputs (Manin's condition fails).

7.2 Value Profile Transparency as a Unifying Response

Measurable AI value profiles address this shared deficit through a single mechanism: making the value reasoning of AI-mediated governance empirically accessible to democratic participants. This is not a separate contribution to each tradition but a single intervention that restores the informational precondition all three require.

The specificity of PRISM-style measurement matters here. Abstract transparency requirements ("AI must be explainable") do not satisfy the informational needs of any tradition. Domain-specific, empirically comparable value profiles do—because they provide the concrete, falsifiable information that deliberation, epistemic aggregation, and electoral accountability all require.

7.3 Toward a Theory of Democratic AI Integrity

We propose that democratic theory needs a new concept: **Democratic AI Integrity**—the condition in which AI systems deployed in political governance maintain value hierarchies that are (a) empirically measurable, (b) publicly disclosed, (c) subject to democratic contestation, and (d) auditable by independent parties.

This concept extends AI Integrity (S. Lee, 2026a) from a general governance principle to a specifically democratic one. The extension is justified by a structural asymmetry: AI value hierarchies in political contexts carry democratic weight that exceeds their role in other professional domains, because they affect not merely individual decisions but the collective self-governance that democracy promises.

Democratic AI Integrity is not a fourth democratic tradition but a procedural condition that all traditions require in the AI era—the informational infrastructure without which deliberation, epistemic democracy, and representative accountability cannot function.

---

8 From Concept to Practice: A Roadmap

8.1 Short-Term (1–3 years)

- Independent AI value profile benchmarking of major AI systems used in government
- Publication of domain-specific value profiles for public inspection
- Pilot programs: voters interact with value-profiled AI systems in simulated policy consultations
- Civil society organizations adopt PRISM-style benchmarks for watchdog functions

8.2 Medium-Term (3–7 years)

- Political AI value profile disclosure becomes a norm in advanced democracies
- Election commissions include AI value profile information in candidate disclosures
- Political parties publish official value alignment commitments alongside policy platforms
- Continuous value profile monitoring during electoral terms by independent auditors

8.3 Long-Term (7+ years)

- Democratic constitutions recognize AI value transparency as a governance right
- International standards for political AI value profile measurement
- Citizens interact directly with candidates' value-aligned AI as a standard component of democratic participation
- Legislative frameworks codify AI value profile accountability (see companion paper)

---

9 Limitations and Future Work

9.1 Limitations

This paper is a position paper and does not present new empirical evidence. Its proposals depend on several assumptions: that PRISM-style measurement scales to diverse political AI deployments, that value profiles remain stable enough to serve as accountability tools, and that democratic publics can meaningfully engage with value profile information. Each assumption requires empirical validation.

The argument relies primarily on a single measurement framework (PRISM) developed by the author's organization. While the paper's claims are about the *feasibility* of measurement rather than the *exclusivity* of PRISM, the absence of independent replication or competing measurement frameworks is a limitation. The development of alternative measurement approaches by independent research groups would strengthen the democratic infrastructure argument considerably.

The Schwartz value framework, while cross-culturally validated, was designed to measure individual human values, not institutional or political value systems. It may not capture all politically relevant value dimensions. Additional value constructs (e.g., institutional trust, collective identity, procedural fairness, distributive justice) may need integration for political contexts. The extension from individual psychology to political governance requires theoretical justification that this paper sketches but does not fully develop.

The proposals are most directly applicable to liberal democratic systems with existing transparency norms and may require significant adaptation for other political systems. The paper does not address how these proposals function in hybrid regimes, illiberal democracies, or authoritarian contexts where AI governance transparency may serve different political functions.

9.2 Future Work

Empirical studies of voter comprehension and use of AI value profile information. Pilot implementations of value-aligned AI consultation systems in democratic settings. Extension of the PRISM framework to capture political-context-specific value dimensions. Comparative analysis across political systems with different democratic traditions. Development of the legislative framework (companion paper, forthcoming).

---

10 Conclusion

This paper has argued that measurable AI value alignment is not merely a technical achievement but a democratic necessity. The PRISM framework (S. Lee, 2026b) and its empirical validation (S. Lee, 2026c) have demonstrated that AI systems hold empirically measurable value hierarchies—hierarchies that differ across models, domains, and conditions. This measurability creates a new possibility and a new obligation.

The possibility: democratic participation can expand to include inspection, comparison, and contestation of AI value hierarchies deployed by political actors. Citizens gain a new tool for evaluating political representation—not only "what did you deliver?" but "what values does your AI prioritize, and are they consistent with what you promised?"

The obligation: if AI value hierarchies are measurable, democratic accountability demands that they be measured. Unmeasurable value alignment cannot be democratically governed. When AI systems mediate political decision-making, opacity in their value reasoning is not a technical limitation—it is a democratic failure.

We have proposed that value alignment functions as a contractual language between humans and AI, that elections evolve into explicit value hierarchy competitions, and that political representation acquires a new dimension of AI value profile accountability. These proposals carry risks—populist exploitation, measurement weaponization, value rigidity—but the AI Integrity principle provides a structural safeguard: require transparency and auditability without prescribing which values are correct.

Democracy has always been a field of value competition. AI makes this competition measurable. The question is not whether AI value hierarchies will shape political governance—they already do. The question is whether citizens will have the tools to see, evaluate, and choose among them. That is the infrastructure this paper calls for.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv:1606.06565*.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., … & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*.

Dahl, R. A. (1971). *Polyarchy: Participation and opposition*. Yale University Press.

European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act).

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437.

Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy*. MIT Press.

Landemore, H. (2012). *Democratic reason: Politics, collective intelligence, and the rule of the many*. Princeton University Press.

Lee, S. (2026a). AI Integrity: Definition, Authority Stack model, and the demand for process verification in AI governance. AIO Working Paper. DOI: [pending].

Lee, S. (2026b). AI Integrity and the PRISM Framework: Definition, Authority Stack model, and Enhanced Cascade Mapping Hypothesis. AIO Working Paper. DOI: 10.5281/zenodo.18861026.

Lee, S. (2026c). Measuring AI Value Priorities: Empirical analysis of 113,400 forced-choice responses across 10 AI models. AIO Working Paper. DOI: 10.5281/zenodo.18859945.

Lee, S. (2026d). PRISM Risk Signal Framework: Hierarchy-based red lines for AI behavioral risk. AIO Working Paper. SSRN Abstract ID: 6449079.

Manin, B. (1997). *The principles of representative government*. Cambridge University Press.

Mouffe, C. (2000). *The democratic paradox*. Verso.

Mouffe, C. (2005). *On the political*. Routledge.

Mouffe, C. (2013). *Agonistics: Thinking the world politically*. Verso.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., … & Lowe, R. (2022). Training language models to follow instructions with human feedback. *NeurIPS 2022*.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. *Advances in Experimental Social Psychology*, 25, 1–65.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1).

Appendix A: Existing Letter Designations

For clarity of cross-referencing within the AIO paper series:

Letter	Paper
S. Lee (2026a)	AI Integrity concept paper
S. Lee (2026b)	PRISM framework paper
S. Lee (2026c)	Empirical paper (113,400 responses)
S. Lee (2026d)	PRISM Risk Signal Framework paper
S. Lee (2026e)	**This paper** (Democracy and political representation)

Note: Letter designation 2026e is provisional and subject to confirmation based on publication order.