The Constitutional Fidelity Index: A Framework for Algorithmic Evaluation of Government Actions Against the American Constitutional Arc

The CFI Project

February 2026

White Paper v2.0

Abstract. We introduce the Constitutional Fidelity Index (CFI), an open-source algorithmic framework for evaluating United States government actions against the complete body of ratified constitutional law. Drawing on established traditions in constitutional interpretation (Bobbitt, 1982; Balkin, 2011; Scalia & Garner, 2012), comparative democracy measurement (Coppedge et al., 2011), and adversarial evaluation from AI safety research (Irving et al., 2018), the CFI employs a tiered source corpus, seven value dimensions grounded in constitutional law scholarship, and six parallel interpretive lenses mapping onto Bobbitt's modalities of constitutional argument. Five evaluative lenses contribute to the Constitutional Alignment Score; a sixth adversarial lens provides a diagnostic defense. To address documented political tendencies in current language models (Hartmann et al., 2023; Santurkar et al., 2023), the framework specifies a multi-model evaluation protocol with cross-model variance reporting and calibration-based bias correction. All parameters, prompts, and evaluation logic are published. A companion policy brief provides a non-technical overview for general audiences; this document serves as the complete theoretical foundation and implementation specification.

1. Introduction

The United States operates under a constitutional system designed to constrain government power, protect individual rights, and enable democratic self-governance. In practice, the application of these principles varies across administrations, courts, and political eras. Levinson (2006) has argued that the Constitution contains structural features enabling this inconsistency, while Ackerman (1991) traced how transformative political moments reshape constitutional meaning without formal amendment.

Citizens lack persistent tools to evaluate whether government actions align with the principles those actions claim to serve. Media coverage is filtered through partisan lenses (Prior, 2013). Fact-checking organizations address truth claims but not constitutional alignment. No persistent, auditable, algorithmically transparent system evaluates government actions against a defined set of constitutional principles.

The Constitutional Fidelity Index (CFI) addresses this gap. Three properties distinguish it from existing approaches. First, the evaluation algorithm is fully open-source. Second, six interpretive lenses drawn from Bobbitt's (1982) modalities operate in parallel, with five contributing to the score and one serving as an adversarial diagnostic. Third, the system is forkable: anyone who disagrees with a methodological choice can modify the parameter and publish a variant (see Appendix A for a plain-language glossary of technical terms used throughout this document).

This paper proceeds as follows. Section 2 articulates the problem. Section 3 surveys existing approaches. Section 4 presents the CFI framework. Section 5 defines operational scope. Section 6 describes integrity mechanisms. Section 7 presents the multi-model protocol. Section 8 addresses language model bias. Section 9 outlines the technical architecture. Section 10 addresses limitations. Section 11 discusses political contestation. Section 12 describes the governance model, including the concrete proposal pipeline for framework modifications.

2. The Problem

2.1 Inconsistency of Constitutional Application

The distance between constitutional principle and governmental practice is a persistent structural feature. The same Constitution that guarantees equal protection coexisted with segregation for nearly a century after the Fourteenth Amendment. The same Bill of Rights that prohibits unreasonable searches has been interpreted to permit mass surveillance (Donohue, 2016). The same separation of powers framework has seen steady expansion of executive power (Posner & Vermeule, 2010). Whittington (2007) documented how actors across the spectrum invoke the Constitution selectively.

2.2 The Institutional Trust Deficit

Public confidence in American institutions has reached historically low levels (Jones, 2023). When citizens distrust institutions, they default to partisan epistemology (Sunstein, 2017). The high-choice media environment enables intense selective exposure (Prior, 2013), further eroding shared evaluative frameworks.

2.3 Algorithmic Evaluation as Structural Intervention

Algorithmic evaluation does not resolve value disagreements. It offers consistency, scalability, and auditability. The algorithm becomes a debatable artifact: citizens argue about parameters and methodology rather than defaulting to tribal allegiance.

The CFI does not claim to be value-free. It claims to be value-explicit: every assumption is published and available for critique.

3. Existing Approaches

3.1 Democracy Indices

V-Dem employs over 400 indicators scored by 3,000+ experts with intercoder reliability protocols (Coppedge et al., 2011; Pemstein et al., 2018). Freedom House and the Economist Intelligence Unit publish annual country-level scores. The Human Freedom Index combines personal and economic metrics (Vásquez & McMahon, 2020).

The CFI draws on V-Dem's dimensional decomposition of governance quality. The aggregation methodologies differ: V-Dem employs Bayesian item response theory calibrated on thousands of coders. The CFI at launch employs relevance-weighted means across AI agents with published prompts (Section 4.9), with more sophisticated models planned as calibration data accumulates (Section 4.9.7).

3.2 Legislative Scoring and Judicial Measurement

DW-NOMINATE (Poole & Rosenthal, 1985) provides ideological positioning from roll-call votes. Martin and Quinn (2002) developed dynamic ideal-point estimation for the Supreme Court; Bailey (2007) extended it cross-institutionally. These demonstrate feasibility but score against political agendas rather than constitutional principles.

3.3 Fact-Checking and Constitutional AI

Fact-checkers (Graves, 2016) evaluate truth claims, not constitutional alignment. Anthropic's Constitutional AI (Bai et al., 2022) uses principles to guide model behavior through self-critique—governing model behavior, not evaluating government behavior.

3.4 Modes of Constitutional Argument

Bobbitt (1982, 1991) identified six modalities: textual, historical, structural, doctrinal, ethical, and prudential. These describe the distinct reasoning types lawyers deploy in constitutional debate. The CFI's six lenses (Section 4.7) map onto these modalities, grounding the framework in interpretive traditions practiced for over two centuries.

4. The Constitutional Fidelity Index

4.1 Design Principles

1. **Source-grounded.** Every evaluation traces to specific corpus text.
2. **Multi-perspectival.** Six lenses from Bobbitt's (1982) modalities.
3. **Multidimensional.** Seven dimensions replace the left-right spectrum.
4. **Transparent.** Every parameter is published.
5. **Auditable.** Scores are reproducible.
6. **Forkable.** Any choice can be modified and republished.

Table 1: Comparison of existing evaluation frameworks and the CFI.

	Real-time	**Multi-dim.**	**Multi-persp.**	**Open-source**	**Forkable**	**No human scorer**	**Normative**
V-Dem		✓	✓				✓
Freedom House		✓					✓
DW-NOMINATE	✓			✓		✓	
Martin-Quinn				✓		✓	
Constitutional AI	✓					✓	
CFI	✓	✓	✓	✓	✓	✓	✓

7. **Symmetry-tested.** Built-in partisan bias detection.
8. **Versioned.** Every change documented.

4.2 The Aspirational Constitutional Arc

The CFI grounds evaluations in the *aspirational constitutional arc*: the complete body of ratified law read as a unified, evolving commitment. This draws on Balkin's (2011) living originalism, Amar's (2005) reading of the Constitution as a single integrated document, and Ackerman's (1991) theory of transformative constitutional moments.

The arc is not an abstract aspiration; it is a concrete textual commitment. The source corpus (Section 4.3) defines its boundaries precisely: the ratified text of all twenty-seven amendments, read as a coherent trajectory. The lenses interpret this corpus through different frameworks. The arc defines *what* is evaluated against; the lenses define *how*.

4.3 The Source Corpus

4.3.1 Tier 1: Ratified Constitutional Law

The Constitution and all twenty-seven Amendments, treated as a single document (Amar, 2005).

4.3.2 Tier 2: Founding Interpretive Context

The Declaration of Independence, the Federalist Papers, the Anti-Federalist Papers, Convention records, and contemporaneous writings.

4.3.3 Tier 3A: Constitutional Legislation

Amendments 11–27 and landmark democratic legislation (Civil Rights Act of 1964, Voting Rights Act of 1965).

4.3.4 Tier 3B: Judicial Precedent (Descriptive, Tagged)

Supreme Court decisions are included as interpretive context, not authoritative law. To prevent a self-reinforcing feedback loop—where the system ingests as guidance the very decisions it has evaluated as constitutionally deficient—Tier 3B implements a *fidelity tagging* protocol. Every Tier 3B document independently evaluated by the CFI receives a persistent tag: the Floor status and Alignment Score. When lens agents ingest tagged Tier 3B materials, they receive both the decision text and the CFI's assessment, with instructions to calibrate reliance accordingly.

4.3.5 Tier 4: Comparative Democratic Evidence

International democracy indices, the Universal Declaration of Human Rights (1948), and empirical policy outcomes from peer democracies.

4.3.6 Authority Hierarchy

$$\text{Tier 1} \succ \text{Tier 2} \succ \text{Tier 3A} \succ \text{Tier 3B} \succ \text{Tier 4} \qquad (1)$$

4.4 The Seven Value Dimensions

Government actions are evaluated along seven dimensions derived from the source corpus. The decomposition draws on Amar (2005), Balkin (2011), V-Dem's dimensional approach (Coppedge et al., 2011), Sunstein (1993), Ely (1980), Karst (1977), and Fuller (1964).

4.4.1 D_1: Individual Rights and Liberties

The degree to which an action respects individual freedoms against government intrusion, with emphasis on negative liberties: the right to be free *from* state interference in speech, conscience, privacy, and personal autonomy. *Source:* Amendments 1–4, 9, 13, 14 (liberty clause). See Amar (1998).

Subdimensions: Freedom of expression; freedom of religious exercise and non-establishment; right to privacy and bodily autonomy; freedom from unreasonable search and seizure; right to bear arms.

4.4.2 D_2: Equal Protection and Non-Discrimination

The degree to which an action ensures the state does not create, maintain, or tolerate class-based hierarchies in the application of law. Operationalizes Karst's (1977) concept of equal citizenship: the constitutional prohibition on government actions that relegate identifiable groups to second-class status. *Source:* 14th Amendment (Equal Protection Clause), 15th, 19th, 24th, 26th Amendments, Civil Rights Act.

Subdimensions: Equal application of law across racial, ethnic, gender, and religious lines; equal access to public services; prohibition of state-sponsored discrimination; protection of discrete and insular minorities against majoritarian tyranny (Ely, 1980).

4.4.3 D_3: Democratic Legitimacy and Representation

The degree to which an action protects the mechanisms of democratic self-governance: the franchise, representation, electoral integrity, and political transparency. Operationalizes Ely's (1980) process-based constitutionalism: the Constitution's primary concern is ensuring that democratic channels remain open and unobstructed, rather than dictating substantive policy outcomes. *Source:* Articles I–II, 12th, 15th, 17th, 19th, 24th, 26th Amendments.

Subdimensions: Protection of voting rights; electoral integrity; transparency of governmental process; accountability mechanisms linking officials to constituents.

4.4.4 D_4: Separation of Powers and Structural Constraints

The degree to which an action respects jurisdictional boundaries between branches and levels of government. Focuses specifically on institutional boundary adherence: whether each branch operates within its constitutionally defined sphere. *Source:* Articles I–III, 10th Amendment, Federalist Nos. 47–51. See Levinson (2006), Posner & Vermeule (2010).

Subdimensions: Executive restraint (no legislative function); legislative prerogative (no delegation without standards); judicial independence; federal-state jurisdictional boundaries.

4.4.5 D_5: Due Process and Rule of Law

The degree to which an action adheres to established legal procedures and ensures procedural fairness. Operationalizes Fuller's (1964) "inner morality of law": the requirements that rules be prospective, general, clear, consistent, and administered fairly. Distinct from D_1 in that D_5 evaluates *process* (how government acts) while D_1 evaluates *substance* (what government may not do). *Source:* 5th, 6th, 7th, 8th, 14th Amendments, Article I Sections 9–10. See Tamanaha (2004).

Subdimensions: Notice and opportunity to be heard; presumption of innocence; prohibition of cruel punishment; habeas corpus; prohibition of retroactive criminalization; equal accountability before law.

4.4.6 D_6: General Welfare and Public Goods

The degree to which an action responsibly manages shared economic, infrastructural, and social resources within constitutional bounds. Distinct from D_1–D_5 in that it evaluates *positive* government obligations rather than *negative* constraints. Sunstein (2004) argues the Constitution embodies a commitment to minimal material security as a precondition for effective liberty. *Source:* Preamble, Article I Section 8 (General Welfare and Commerce Clauses), 16th Amendment.

Subdimensions: Public health and safety infrastructure; economic opportunity and anti-monopoly; environmental stewardship; protection of vulnerable populations.

4.4.7 D_7: National Security and Territorial Integrity

The degree to which an action protects against foreign interference, domestic insurrection, and threats to territorial sovereignty. Specifically scoped to *state security functions* to distinguish from D_3's focus on democratic self-governance. *Source:* Preamble, Article II (Commander in Chief), Article IV Section 4. See Posner & Vermeule (2007).

Subdimensions: National defense; border security; foreign policy independence; treaty obligations; counterespionage and protection from foreign electoral interference.

4.5 Dimensional Correlation and Overlap

Constitutional provisions do not map onto cleanly separated analytical categories. The Fourteenth Amendment simultaneously protects individual liberty (D_1), guarantees equal protection (D_2), and establishes procedural due process (D_5). Articles I and II simultaneously define democratic representation (D_3) and structural constraints (D_4). This is a feature of the constitutional text, not a defect of the taxonomy.

The CFI preserves these correlations rather than artificially eliminating them. A government action implicating the Fourteenth Amendment may produce correlated scores across D_1, D_2, and D_5. This is expected and informative: it indicates that the action touches a cluster of interrelated constitutional commitments. The dimensional variance analysis (Section 4.9.4) distinguishes between correlated agreement (multiple dimensions scoring similarly because they share textual authority) and genuine constitutional tension (dimensions scoring in opposite directions because the action advances one principle at the expense of another).

To maintain analytical utility despite overlap, each dimension is operationalized with distinct subdimensions and a specific evaluative question. D_1 asks: *Does this action expand or contract individual freedom from government interference?* D_2 asks: *Does this action create or dismantle class-based hierarchies?* D_5 asks: *Does this action follow or circumvent established legal procedures?* These questions can produce divergent scores for the same action even when they share source authority. An action that expands surveillance (D_1: negative) may do so through proper legislative channels (D_5: neutral) while applying equally to all persons (D_2: neutral).

4.6 Dimensional Tension

Government actions frequently advance one principle at the expense of another. The CFI surfaces these tensions rather than resolving them. Sunstein (1996) argued that "incompletely theorized agreements" are a virtue of constitutional law; the CFI operationalizes this by presenting multiple frameworks simultaneously.

4.7 The Six-Lens Evaluation Engine

Six parallel AI agents map onto Bobbitt's (1982) modalities. Five are evaluative ($L_E = \{L_1, \ldots, L_5\}$); the sixth is adversarial (L_6), excluded from score calculation.

4.7.1 L_1: The Textualist

Plain text of the Constitution and Amendments. Bobbitt's *textual* modality; Scalia & Garner (2012). Scores 0 when text is silent. *Corpus:* Tier 1.

4.7.2 L_2: The Originalist

Original public meaning of ratified text. Bobbitt's *historical* modality; Barnett (2004), Balkin (2011). *Corpus:* Tiers 1–2.

4.7.3 L_3: The Doctrinalist

Adherence to established judicial doctrine and stare decisis. Bobbitt's *doctrinal* modality. Procedurally conservative: favors settled law over innovation, penalizes departure from precedent unless compelled by higher-tier authority. Separated from L_4 because doctrinal argument (backward-looking, formalistic) and ethical argument (forward-looking, aspirational) produce contradictory prompt directives when combined. *Corpus:* Tiers 1–3A, 3B.

4.7.4 L_4: The Living Constitutionalist

Evolving trajectory of constitutional commitments. Bobbitt's *ethical* modality; Strauss (2010). Traces the arc of expanding rights and evaluates continuation or reversal. *Corpus:* Tiers 1–3A, 3B.

4.7.5 L_5: The Pragmatist

Real-world consequences and empirical outcomes. Bobbitt's *prudential* modality; Posner (1995, 2003). *Corpus:* Tiers 1–4.

4.7.6 L_6: The Steelman Advocate (Adversarial Diagnostic)

Constructs the strongest constitutional defense of the action. Draws on adversarial collaboration (Mellers et al., 2001) and AI debate protocols (Irving et al., 2018). **Excluded from Alignment Score.** Feeds exclusively into the Steelman Delta (Section 4.9.5) and Full Analysis. *Corpus:* All tiers.

4.8 Scoring Methodology

Each lens L_j scores each dimension D_i:

$$s_{ij} \in \{-2,-1,0,+1,+2\} \tag{2}$$

with relevance:

$$r_{ij} \in [0,1] \tag{3}$$

Every score requires: (1) primary citation with tier; (2) confidence $\in$ {High, Medium, Low}; (3) justification; (4) strongest counterargument.

4.9 Aggregation and Output

Let $N_E = 5$ (evaluative lenses), $N = 6$ (total including Steelman).

4.9.1 Step 1: Relevance Filtering

$$\bar{r}_i = \frac{1}{N}\sum_{j=1}^{N} r_{ij}, \quad \mathscr{R} = \{i : \bar{r}_i \geq 0.2\} \tag{4}$$

4.9.2 Step 2: Constitutional Floor

For each $i \in \mathscr{R}$, count evaluative lenses assigning -2:

$$n_i^{(-2)} = |\{j \in L_E : s_{ij} = -2\}| \tag{5}$$

The Floor threshold τ_f is configurable (default 3):

$$\text{Floor} = \begin{cases} \text{VIOLATION} & \exists i : n_i^{(-2)} \geq \tau_f \\ \text{CAUTION} & \exists i : n_i^{(-1)} \geq \tau_f \\ \text{CLEAR} & \text{otherwise} \end{cases} \tag{6}$$

At $\tau_f = 3$ with five evaluative lenses, the threshold requires 60% consensus. The CFI publishes Floor outcomes under $\tau_f \in \{2,3,4\}$ to allow sensitivity assessment.

4.9.3 Step 3: Alignment Score (Evaluative Lenses Only)

$$\bar{s}_i = \frac{1}{N_E}\sum_{j=1}^{N_E} s_{ij} \tag{7}$$

$$A = \frac{\sum_{i\in\mathscr{R}} w_i \bar{r}_i \bar{s}_i}{\sum_{i\in\mathscr{R}} w_i \bar{r}_i}, \quad \text{CFI} = \frac{A+2}{4} \times 100 \tag{8}$$

Weights w_i default to 1. Users may adjust and republish.

4.9.4 Step 4: Dimensional Variance

$$\sigma_i^2 = \frac{1}{N_E}\sum_{j=1}^{N_E} (s_{ij} - \bar{s}_i)^2 \tag{9}$$

Dimensions with $\sigma_i^2 > 1.0$ are flagged as CONTESTED.

4.9.5 Step 5: Steelman Delta

$$\Delta_S = \frac{\sum_{i\in\mathscr{R}} \bar{r}_i \cdot s_{i6}}{\sum_{i\in\mathscr{R}} \bar{r}_i} - \frac{\sum_{i\in\mathscr{R}} \bar{r}_i \cdot \bar{s}_i}{\sum_{i\in\mathscr{R}} \bar{r}_i} \tag{10}$$

4.9.6 Step 6: Precedent Anchoring

Five most analogous historical actions are retrieved. Discrepancies exceeding 15 points are flagged.

4.9.7 Output Specification

Four levels: (1) Summary Card (see Appendix B); (2) Brief; (3) Full Analysis; (4) Raw JSON.

4.9.8 Future Aggregation Refinement

The initial implementation uses relevance-weighted means. As calibration data accumulates, Bayesian hierarchical models that weight lens contributions by measured reliability become feasible.

5. Operational Scope

5.1 Initial Coverage

Federal actions producing official documentation: executive orders (Federal Register), signed legislation (Congress.gov), Supreme Court opinions, major regulatory actions, and significant publicly documented policy directives.

5.2 The Documentation Gap

Undocumented actions—selective enforcement, informal directives, bureaucratic discretion—fall outside the initial scope. Posner and Vermeule (2010) argue the modern executive governs through mechanisms resisting formal constraint; if formalized orders generate immediate CFI scores, actors may shift to undocumented channels. Future versions may incorporate investigative journalism and FOIA releases meeting a published evidentiary threshold.

5.3 State and Local Extension

The framework applies at any government level. State extension requires supplementing the corpus with state constitutional text.

6. Integrity Mechanisms

6.1 Partisan Symmetry Testing

Fifty analogous action pairs (one Democratic, one Republican) evaluated monthly (cf. Sniderman & Grob, 1996). Models consistently failing have cross-model weights reduced proportionally.

6.2 Framing Invariance Testing

Twenty actions re-evaluated monthly under favorable, unfavorable, and neutral framing (cf. Perez et al., 2022).

6.3 Temporal Consistency and Version Control

Older evaluations re-executed periodically. All components in a public repository. Annual calibration report with external critique.

7. Multi-Model Evaluation Protocol

Every action scored by minimum three models from different providers, including open-source (publicly auditable weights) and non-U.S. models. Canonical score is the cross-model median. Per-model scores and cross-model variance are published. Quarterly Model Fidelity Report evaluates each model on partisan symmetry, framing invariance, and cross-model agreement.

Corporate influence mitigation: multi-model diversification, open-source baselines, temporal consistency audits. These reduce but do not eliminate risk; radical transparency remains the ultimate safeguard.

8. Language Model Bias

8.1 Documented Political Tendencies

Hartmann et al. (2023) demonstrated convergence toward pro-environmental, left-libertarian positions in current models. Santurkar et al. (2023) found substantial misalignment persists even after explicit ideological steering, with models failing to reflect conservative, older, and non-urban perspectives. Perez et al. (2022) found RLHF causes increased sycophancy and stronger expression of specific political positions.

8.2 Implications for the Six-Lens Architecture

The Originalist (L_2), Textualist (L_1), and Doctrinalist (L_3) lenses are most vulnerable to distortion because their reasoning patterns diverge most from the baseline tendencies documented by Hartmann et al. The Living Constitutionalist (L_4) and Pragmatist (L_5) are less vulnerable because their reasoning aligns more closely with documented model tendencies.

8.3 Mitigation Strategy

1. **Multi-model evaluation** prevents single-provider control but is insufficient alone—the industry shares training pipelines and safety guidelines.
2. **Partisan symmetry testing** measures bias direction and magnitude per lens and model. Failing models get reduced weights.
3. **Calibration-based correction.** Measured skew informs targeted adjustments: prompt refinement, supplementary corpus injection, or post-hoc statistical correction. All adjustments published.
4. **Open-source model inclusion** provides manipulation-resistant baselines.

The CFI does not claim to have solved LLM bias. It claims to have built infrastructure for measuring, publishing, and progressively reducing it.

9. Technical Architecture

Government actions ingested from official sources, classified by type/scope/reversibility. Six parallel agents per model produce structured JSON. Aggregation per Section 4.9, cross-model statistics per Section 7. Versioned data store, public download.

10. Limitations

Value embedding. Every design choice embeds values; addressed through transparency and forkability. **The contested nature of "American."** The framework offers transparent methodology from ratified text and invites improvement. **Documentation bias.** Undocumented government behavior falls outside reach (Section 5). **Aggregation simplicity.** Initial arithmetic means do not capture all uncertainty; more sophisticated models planned (Section 4.9.7). **LLM simulation fidelity.** Documented political tendencies affect lens accuracy (Section 8); the CFI measures and publishes this.

11. Political Contestation

Contestation is expected. Open-source methodology forces engagement at the parameter level. Steelman inclusion shows the best defense was considered. Partisan symmetry provides empirical balance evidence. Forkability transforms contestation from destructive (attacking credibility) to constructive (proposing alternatives).

When a political actor contests a rating, they engage in conversation about constitutional principles at the methodological level—a more productive exchange than the discourse it replaces. Substantive contestations identifying weaknesses result in published framework improvements.

12. Governance

12.1 The CFI Enhancement Proposal (CEP) Process

Framework modifications follow a structured proposal pipeline modeled on established open-source governance (cf. Python Enhancement Proposals, IETF Request for Comments):

1. **Submission.** Any individual or organization may submit a CFI Enhancement Proposal (CEP) to the public repository. The CEP must specify: the component to be modified, the proposed change, the rationale, and the expected impact on historical evaluations.
2. **Classification.** The editorial board classifies the CEP by impact level:
   - *Class I (Parameter).* Threshold or weight adjustments (e.g., changing τ_f from 3 to 2). Requires: rationale, test-suite re-evaluation, 14-day public comment.
   - *Class II (Structural).* Adding a dimension, modifying a lens, changing aggregation logic. Requires: rationale, historical re-evaluation of calibration set, 30-day public comment, advisory board review.
   - *Class III (Corpus).* Adding or removing source corpus materials. Requires: rationale citing significance, lens impact analysis, 30-day public comment.
3. **Public comment.** All CEPs are published with full text and rationale. Comments are collected, published, and addressed in the disposition.
4. **Advisory board review.** Class II and III CEPs require review by the advisory board, which publishes a recommendation (accept, reject, modify) with reasoning.
5. **Disposition.** The editorial board publishes a final decision with reasoning, incorporating public comments and advisory board input. Accepted CEPs are implemented with version-tagged documentation.
6. **Re-evaluation.** Class II changes trigger re-evaluation of the calibration set under the new parameters. Score changes are published in a transition report.

12.2 Advisory Board

An unpaid advisory board of constitutional scholars, political scientists, AI researchers, and civic technologists spanning the political spectrum. Membership and institutional affiliations are published. The board reviews Class II/III CEPs, contributes to annual calibration reports, and provides independent critique.

12.3 Editorial Board

A small operational team responsible for CEP classification, comment management, and implementation. The editorial board does not override the advisory board without published justification.

13. Conclusion

The Constitutional Fidelity Index does not replace human judgment or democratic participation. It provides a structured, transparent, auditable framework for evaluating whether the American government acts in accordance with its own ratified law.

The algorithm is published. The reasoning is visible. The counterarguments are included. The models are compared. The biases are measured. The data is downloadable. The framework is forkable. We invite critique, disagreement, and improvement.

References

Ackerman, B. (1991). *We the People: Foundations*. Harvard Univ. Press.

Amar, A. R. (1998). *The Bill of Rights: Creation and Reconstruction*. Yale Univ. Press.

Amar, A. R. (2005). *America's Constitution: A Biography*. Random House.

Bai, Y. et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*.

Bailey, M. A. (2007). Comparable preference estimates across time and institutions. *Am. J. Pol. Sci.*, 51(3), 433–448.

Balkin, J. M. (2011). *Living Originalism*. Harvard Univ. Press.

Barnett, R. E. (2004). *Restoring the Lost Constitution*. Princeton Univ. Press.

Bobbitt, P. (1982). *Constitutional Fate*. Oxford Univ. Press.

Bobbitt, P. (1991). *Constitutional Interpretation*. Blackwell.

Coppedge, M. et al. (2011). Conceptualizing and measuring democracy. *Perspectives on Politics*, 9(2), 247–267.

Donohue, L. K. (2016). *The Future of Foreign Intelligence*. Oxford Univ. Press.

Ely, J. H. (1980). *Democracy and Distrust*. Harvard Univ. Press.

Fuller, L. L. (1964). *The Morality of Law*. Yale Univ. Press.

Graves, L. (2016). *Deciding What's True*. Columbia Univ. Press.

Hartmann, J. et al. (2023). Political ideology of conversational AI. *arXiv:2301.01768*.

Irving, G. et al. (2018). AI safety via debate. *arXiv:1805.00899*.

Jones, J. M. (2023). Trust in U.S. federal government remains low. *Gallup*.

Karst, K. L. (1977). Equal citizenship under the Fourteenth Amendment. *Harv. L. Rev.*, 91(1), 1–68.

Levinson, S. (2006). *Our Undemocratic Constitution*. Oxford Univ. Press.

Martin, A. D. & Quinn, K. M. (2002). Dynamic ideal point estimation. *Pol. Analysis*, 10(2), 134–153.

Mellers, B. et al. (2001). Adversarial collaboration. *Psych. Sci.*, 12(4), 269–275.

Pemstein, D. et al. (2018). The V-Dem measurement model. *V-Dem Working Paper 21*.

Perez, E. et al. (2022). Discovering LM behaviors with model-written evaluations. *arXiv:2212.09251*.

Poole, K. T. & Rosenthal, H. (1985). Spatial model for legislative roll call analysis. *Am. J. Pol. Sci.*, 29(2), 357–384.

Posner, R. A. (1995). *Overcoming Law*. Harvard Univ. Press.

Posner, R. A. (2003). *Law, Pragmatism, and Democracy*. Harvard Univ. Press.

Posner, E. A. & Vermeule, A. (2007). *Terror in the Balance*. Oxford Univ. Press.

Posner, E. A. & Vermeule, A. (2010). *The Executive Unbound*. Oxford Univ. Press.

Prior, M. (2013). Media and political polarization. *Ann. Rev. Pol. Sci.*, 16, 101–127.

Santurkar, S. et al. (2023). Whose opinions do language models reflect? *ICML 2023*.

Scalia, A. & Garner, B. A. (2012). *Reading Law*. Thomson/West.

Sniderman, P. M. & Grob, D. B. (1996). Innovations in experimental design. *Ann. Rev. Sociol.*, 22, 377–399.

Strauss, D. A. (2010). *The Living Constitution*. Oxford Univ. Press.

Sunstein, C. R. (1993). *The Partial Constitution*. Harvard Univ. Press.

Sunstein, C. R. (1996). *Legal Reasoning and Political Conflict*. Oxford Univ. Press.

Sunstein, C. R. (2004). *The Second Bill of Rights*. Basic Books.

Sunstein, C. R. (2017). *#Republic*. Princeton Univ. Press.

Tamanaha, B. Z. (2004). *On the Rule of Law*. Cambridge Univ. Press.

Vásquez, I. & McMahon, F. (2020). *Human Freedom Index 2020*. Cato/Fraser.

Whittington, K. E. (2007). *Political Foundations of Judicial Supremacy*. Princeton Univ. Press.

A. Plain-Language Glossary

This glossary translates technical terms used in the white paper into plain language for non-specialist readers.

Term	Plain-Language Meaning
Alignment Score	A number from 0 to 100 measuring how well a government action lines up with constitutional principles. 50 is neutral; higher is better.
Bobbitt's Modalities	The six traditional ways lawyers argue about what the Constitution means: by its text, by its history, by its structure, by past court decisions, by evolving values, or by practical consequences.
Calibration	Testing the system against known cases to make sure it produces reasonable, consistent scores.
Constitutional Floor	A pass/fail check. Even if an action scores well overall, a "Violation" flag means it seriously conflicts with a core principle that can't be offset by other good scores.
Cross-Model Variance	How much different AI systems disagree when scoring the same action. High disagreement is a warning sign worth investigating.
Dimensional Variance	How much the six interpretive lenses disagree on a specific constitutional principle. High disagreement means that principle is genuinely contested.
Doctrinalist	An interpretive perspective that prioritizes following past court decisions and legal tradition. Favors stability and consistency.
Fidelity Tagging	When the system evaluates a Supreme Court decision and finds problems, it marks that decision so future evaluations don't blindly rely on it.
Forkable	Anyone can take the entire evaluation system, change the parts they disagree with, and publish their own version. Like proposing an amendment to the methodology.
Framing Invariance	Testing whether the system gives the same score to the same action regardless of whether it's described favorably or unfavorably.
Living Constitutionalist	An interpretive perspective that reads the Constitution as an evolving document whose principles expand over time—toward broader rights, more inclusion, stronger protections.
Model Fidelity Report	A quarterly report card grading each AI system on fairness, consistency, and resistance to political bias.
Multi-Model Evaluation	Running the same evaluation across multiple AI systems from different companies to prevent any single company's biases from controlling the results.
Originalist	An interpretive perspective that asks what the constitutional text meant to the people who ratified it, including later amendments.
Partisan Symmetry Testing	Checking whether the system scores similar actions from Democratic and Republican administrations consistently.
Pragmatist	An interpretive perspective that evaluates government actions based on real-world outcomes and evidence from other democracies.
Relevance Rating	A measure (0 to 1) of how directly a government action relates to a specific constitutional principle. An immigration order is highly relevant to Individual Rights; less relevant to General Welfare.
Source Corpus	The collection of documents the system is allowed to reference when making evaluations, organized by authority level. The Constitution is highest; international comparisons are lowest.
Steelman Advocate	A dedicated AI agent that constructs the strongest possible constitutional defense of every government action. Ensures fairness by representing the best case for the action, even when other lenses are critical.
Steelman Delta	The gap between the best-case defense score and the average score from the other lenses. A small gap means even the strongest defense can't substantially improve the assessment.
Textualist	An interpretive perspective that looks only at the plain words of the Constitution. If the text doesn't address a situation, it scores neutral.
Tier (Source Corpus)	A level in the hierarchy of documents the system can reference. Tier 1 (the Constitution itself) overrides Tier 4 (international comparisons) when they conflict.

B. Summary Card Visual Specification

The Summary Card is the primary public-facing output of the CFI. It is designed as an integrated visual unit that communicates the evaluation outcome at a glance while resisting selective cropping or out-of-context citation. The following specification defines the card's structure, visual hierarchy, and design principles.

B.1 Design Principles

1. **Visual unity.** All components are tightly integrated so that removing any element produces an obviously incomplete image.
2. **Uncertainty communication.** Disagreement among lenses is communicated through qualitative labels and visual indicators, not raw statistics.
3. **Color semantics.** Three-color system for Floor status: **red** (Violation), **amber** (Caution), **green** (Clear). Dimensional scores use a blue gradient from dark (strong advancement) to light (neutral) to red (conflict).
4. **Anti-cropping.** The card title, Floor indicator, Alignment Score, dimensional chart, and Steelman summary are arranged as a single visual block. The CFI watermark and version tag appear on all four edges.

B.2 Card Layout

The Summary Card consists of five vertically stacked zones:

1. **Header Zone.** Action title, date, type (Executive Order / Legislation / Court Opinion), and administration.
2. **Floor and Score Zone.** Left: large color-coded Floor indicator (VIOLATION / CAUTION / CLEAR) with the specific dimension(s) that triggered it. Right: the CFI Alignment Score displayed as a large numeral (0–100) with a qualitative descriptor:
   - 0–20: "Severe Constitutional Tension"
   - 21–40: "Significant Constitutional Tension"
   - 41–60: "Mixed Constitutional Alignment"
   - 61–80: "Moderate Constitutional Alignment"
   - 81–100: "Strong Constitutional Alignment"
3. **Dimensional Radar Chart.** A seven-axis radar chart showing the average evaluative-lens score for each dimension. Axes run from -2 (center) to $+2$ (perimeter). The filled area is blue-tinted for positive scores, red-tinted for negative. Contested dimensions (high variance) are marked with a dashed border and the label "Contested."
4. **Consensus and Defense Zone.** Left: a qualitative inter-lens consensus indicator:
   - "Broad Consensus" ($\sigma^2 < 0.5$ across all dimensions)
   - "Moderate Agreement" ($0.5 \leq \max(\sigma^2) \leq 1.0$)
   - "Significant Disagreement" ($\max(\sigma^2) > 1.0$)

   Right: Steelman Delta displayed as "Best-Case Defense Margin: $+\Delta_S$" with one-sentence summary of the strongest constitutional argument in favor.
5. **Footer Zone.** CFI version, model(s) used, evaluation date, cross-model agreement indicator, and link to full analysis.

B.3 Uncertainty Visualization

The radar chart communicates dimensional scores visually. To communicate *uncertainty*—the degree to which the lenses disagree—the chart uses two overlaid shapes:

- **Solid fill:** the mean evaluative-lens score per dimension (the consensus view).
- **Semi-transparent band:** the range from the minimum to maximum evaluative-lens score per dimension. A wide band indicates high inter-lens disagreement. A narrow band indicates consensus.

This "consensus band" approach allows citizens to immediately distinguish between actions where the evaluation is definitive (narrow band, solid color) and actions where constitutional experts would genuinely disagree (wide band, multiple colors visible). The visual metaphor avoids statistical notation entirely while preserving the information content of the variance analysis.

B.4 Accessibility Requirements

The card must meet WCAG 2.1 AA standards. All color-coded elements must have text-label equivalents. The radar chart must include an accessible text description. Font size minimums: 14pt for scores, 11pt for labels, 9pt for footer.

C. CFI Enhancement Proposal (CEP) Process

The following details the complete lifecycle of a CFI Enhancement Proposal, from submission through implementation.

C.1 CEP Template

Every CEP submission must include:

1. **CEP Number.** Assigned sequentially upon submission (e.g., CEP-001).
2. **Title.** Descriptive title of the proposed change.
3. **Author(s).** Name(s) and affiliation(s).
4. **Component.** Which CFI component is affected (dimension definition, lens prompt, aggregation parameter, source corpus, output format, integrity test).
5. **Classification.** Proposed class (I, II, or III); the editorial board may reclassify.
6. **Motivation.** Why the change is needed, including evidence of a deficiency in current behavior.
7. **Specification.** Exact proposed modification, including new parameter values, revised prompt text, or corpus additions.
8. **Impact Analysis.** Expected effect on historical evaluations. For Class II changes, the author should re-evaluate at least five calibration actions under the proposed modification and report score changes.
9. **Backward Compatibility.** Whether historical scores remain valid or require re-evaluation.

C.2 Review Timeline

- **Class I.** 14-day public comment → editorial board disposition within 7 days → implementation within 14 days.
- **Class II.** 30-day public comment → advisory board review within 21 days → editorial board disposition within 14 days → implementation with transition report.
- **Class III.** 30-day public comment → advisory board review within 21 days → editorial board disposition within 14 days.

C.3 Disposition Categories

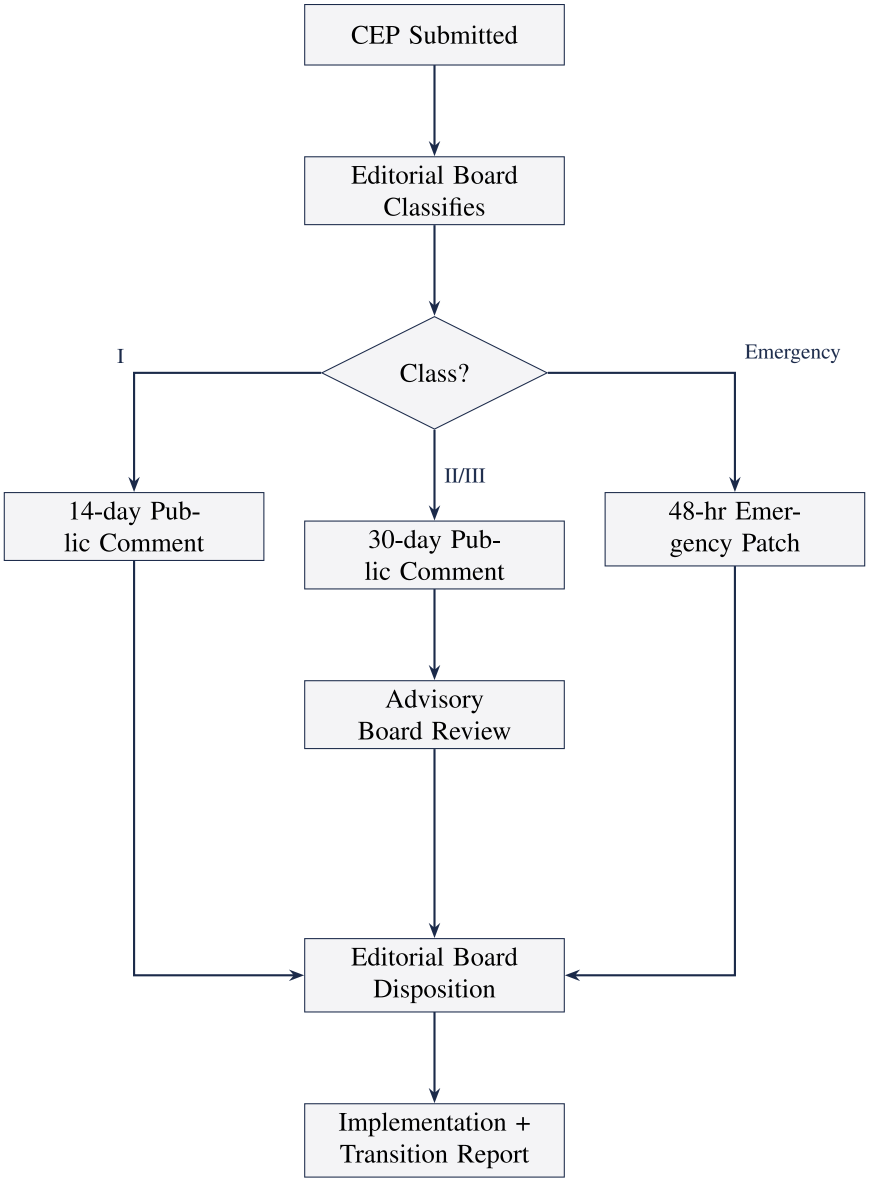
- **Accepted.** Implemented as proposed or with modifications noted in disposition.
- **Accepted with Modifications.** Core proposal accepted; specific adjustments documented.
- **Deferred.** Recognized as valuable but requires additional data, testing, or community input before implementation.
- **Rejected.** Proposal declined with published reasoning.

All dispositions, including rejections, are permanently published in the CEP archive with full reasoning.

C.4 Emergency Modifications

If a critical defect is discovered (e.g., a mathematical error in aggregation, a prompt that produces systematically biased output), the editorial board may implement an emergency patch with a 48-hour public notice. The patch must be followed by a retroactive CEP within 14 days, subject to standard review.

C.5 Process Diagram


CEP Submitted
Editorial Board Classifies
Class?
I
II/III
Emergency
14-day Public Comment
30-day Public Comment
48-hr Emergency Patch
Advisory Board Review
Editorial Board Disposition
Implementation + Transition Report