



Introduction to Techno-Democracy

Dr. Nitnem Singh Sodhi

Managing Director of Bharat Neurotech and Independent Researcher based in Lucknow, Ex-Military Psychologist at Indian Air Force, Consultant Psychologist at Apollo Clinics in Lucknow and Gorakhpur, presently learning AI from IIT Kharagpur.

OPEN ACCESS

Article Citation:

Dr. Nitnem Singh Sodhi, "Introduction to Techno-Democracy", International Journal of Recent Trends in Multidisciplinary Research, September-October 2026, Vol 6(05), 330-348.



©2026 The Author(s). This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Published by 5th Dimension Research Publication

Abstract: This paper proposes Techno-Democracy as a constitutional model in which citizens retain political sovereignty while specified governmental processes are progressively delegated to transparent, independently audited artificial intelligence systems. Its defining composition is popular sovereignty, constitutional democracy, AI administration, cryptographic verification, distributed oversight, and human-rights safeguards. The motivating problem is institutional dependence on concentrated, weakly accountable discretion, which can enable corruption, nepotism, capture, inconsistency, and administrative delay. The argument does not presume universal human corruption or machine neutrality. Using conceptual analysis, a targeted interdisciplinary literature synthesis, and design-oriented institutional reasoning, the paper develops an architecture that separates authority over objectives, data, models, deployment, execution, auditing, and redress. Nine concepts organize the proposal: Computational Constitutionalism, Algorithmic Separation of Powers, Verifiable Bounded Impartiality, Proof-of-Governance, Policy Provenance, Adversarial Governance Models, Algorithmic Federalism, Continuous Democratic Audit, and Evidence-Triggered Legislative Review. Constitutional constraints limit computational action without reducing constitutional interpretation to code. Cryptographic records establish bounded claims about provenance and integrity, without certifying the truth of inputs or the justice of outcomes. A central adversarial analysis shows how the same architecture could facilitate surveillance, infrastructural dependency, automated exclusion, or authoritarian capture. The proposed response combines limited delegation, institutional plurality, contestable decisions, protected human adjudication, public research access, and operational reversibility. A staged transition and preregistrable experimental programme specify how benefits and harms could be evaluated against competent human administration and simpler digital alternatives. No original experiments or deployment results are reported. The contribution is an integrated, falsifiable institutional proposal whose political principle is that AI governs processes while people govern values.

1. Introduction

The problem of governing public power

Democratic government faces two distinguishable problems. The first concerns the rightful authorship of collective decisions: who may determine public purposes, under what procedures, and subject to which rights? The second concerns the dependable implementation of those decisions: how should public institutions administer rules, allocate authorized resources, investigate irregularities, and learn from outcomes? Techno-Democracy, as proposed here by Dr. Nitnem Singh Sodhi, preserves citizens' authority over the first problem while substantially redesigning the institutions addressing the second.

The central thesis is that democratic sovereignty should remain with citizens, while substantial portions of governmental administration, policy analysis, auditing, resource allocation, regulatory execution, and routine decision-making should progressively move into transparent, constitutionally bounded, independently audited computational systems. Delegation would be selective, revocable, and conditioned on demonstrated public benefit. It would not confer constituent power, legislative

supremacy, or independent political authority on an AI model.

Concentrated human discretion creates structural opportunities for misuse. An official who controls access to a scarce benefit, can conceal the basis of selection, and faces little independent scrutiny may be able to solicit payment, favour relatives, protect political allies, or discriminate. Similar vulnerabilities arise when politicians design procurement criteria for preferred firms, when regulators depend on the industries they oversee, or when departmental performance incentives reward short-term indicators over durable public outcomes. Conflicts of interest, ideological bias, and inconsistent interpretation may distort discretionary decisions even without bribery. Institutional analyses of corruption emphasize incentives and organizational arrangements rather than a claim that every individual becomes corrupt upon receiving authority (Shleifer & Vishny, 1993).

Discretion also performs legitimate functions. Public servants recognize exceptional circumstances, translate general rules into situated judgments, and respond to facts that standardized records omit. Lipsky's analysis of street-level bureaucracy identifies the importance and difficulty of such interpretation (Lipsky, 2010). Eliminating discretion indiscriminately could remove compassion while preserving injustice. The appropriate target is arbitrary, concealed, or weakly reviewable discretion. Legitimate judgment should instead become reasoned, bounded, documented, and open to challenge.

AI does not escape institutional politics. Training data reflect historical choices; objectives express priorities; measurement conventions privilege certain experiences; procurement contracts allocate power; infrastructure owners can influence availability. Moving a decision into software may transfer discretion from a caseworker to a developer, a data supplier, or an official who sets a threshold. Accordingly, the paper advances a **displacement proposition**: reducing visible discretion at the point of service does not establish a reduction in total discretionary power unless upstream specification, downstream enforcement, and infrastructural control are also examined. This is a proposed analytical proposition, not an empirically established law.

Techno-Democracy therefore constitutes a theory of democratically controlled delegation. Its political philosophy is civic self-government under conditions of computational capacity. Its constitutional theory concerns the lawful distribution, limitation, and revocation of machine-mediated public power. Its technical architecture concerns how those relationships can be represented, enforced in bounded domains, inspected, and repaired. An installation of AI tools without these relationships would not meet the definition.

The proposal has a demanding burden of proof. Greater speed cannot excuse an unlawful deprivation. A sophisticated audit cannot compensate for inaccessible appeal. Lower measured expenditure cannot establish public benefit when eligible people simply stop receiving support. Nor does the failure of existing bureaucracy establish the superiority of an untested replacement. The relevant comparison is among feasible institutional alternatives, including improved human administration, straightforward rules engines, and hybrid systems.

1.1 The contribution and its intellectual limits

The terminology has precedents. Bloom and Sancino (2019) discuss techno-democracy in relation to technological and political transformation. Tan et al. (2024) explicitly discuss computational constitutionalism in their study of Web3 constitutions. Algorithmic governance, digital constitutionalism, public-sector AI, and technologically mediated accountability already constitute substantial research fields. This paper does not claim to have invented the expression Techno-Democracy, the general aspiration to use technology democratically, or every phrase employed below.

The proposed contribution is the specific combination of constitutional delegation rules, institutionally independent model roles, verifiable decision records, citizen-controlled assistance, evidence-triggered review, and mandatory redress within a single governance model. The nine named concepts receive precise definitions for this architecture; their presentation does not establish lexical priority. Whether this synthesis warrants recognition as a distinct political philosophy is an argument to be assessed through comparison, criticism, and subsequent research, rather than a status established by naming it.

2. Research methodology and epistemic status

This is a conceptual and design-oriented study. It combines normative argument with institutional specification and an empirical research protocol. It does not report a deployed government platform, a completed constitutional experiment, simulated national outcomes, or measured performance gains. The architecture, certificates, thresholds, and evaluation designs are proposals unless explicitly attributed to published work.

The method comprises four operations. First, normative reconstruction identifies the requirements implied by popular sovereignty, constitutional limits, equal political standing, due process, and accountable administration. Second, a targeted literature synthesis connects these requirements to research on bureaucracy, corruption, algorithmic accountability, fairness, cryptographic integrity, institutional decentralization, and causal evaluation. Third, design analysis translates the requirements into separable roles, permissions, records, and failure procedures. Fourth, adversarial evaluation derives failure scenarios and falsifiable hypotheses from the design.

The literature process used targeted scholarly searches and exact-title verification, with bibliographic checks conducted on 29 September 2026. Search families included techno-democracy, computational constitutionalism, algorithmic regulation, technological due process, algorithmic fairness, tamper-evident logging, public-sector AI, and corruption auditing. Original journal articles, conference papers, scholarly monographs, and author or institutional repositories were prioritized. Preprints are explicitly identified in the references. This is an integrative selection of relevant work, not an exhaustive systematic review; it supports bounded claims about precedents and mechanisms, not a comprehensive claim of novelty. An exhaustive multilingual review remains necessary before asserting historical priority.

The approach draws on design-science research, which treats artifacts and their evaluation as legitimate research objects (Hevner et al., 2004). Here the artifact is an institutional architecture rather than a validated software implementation. Its initial

evaluation is analytical: internal consistency, traceability from constitutional requirements to controls, and resilience against specified objections. Empirical validation is deferred to the research programme in Section 14. Normative premises cannot be confirmed by prediction accuracy, and technical demonstrations cannot by themselves establish political legitimacy.

Five research questions organize the argument:

1. Under what conditions can computational administration reduce arbitrary discretion without weakening democratic authorship or meaningful human judgment?
2. How can legal constraints be operationalized while preserving the authority of public legal interpretation and judicial review?
3. Which separations of institutional control and model responsibility reduce correlated error, manipulation, and capture?
4. Can provenance, cryptographic evidence, and accessible appeal improve accountability without expanding surveillance or excluding disadvantaged populations?
5. What evidence would justify expanding, restricting, or reversing delegation to AI-operated public systems?

The paper distinguishes **established findings**, **design proposals**, and **testable hypotheses** throughout. Findings from other settings are not treated as direct evidence that this architecture works. Formal conditions specify what an implementation would have to demonstrate; they do not certify that a real implementation already satisfies those conditions.

3. Intellectual foundations and comparison with existing models

The democratic foundation is citizens' equal standing in authorizing and contesting collective power. Elections matter, but democratic legitimacy also requires public communication, opposition, participation, and meaningful control over the political agenda (Dahl, 1989). A public administration that implements every instruction efficiently could still be undemocratic if the instructions originate in an inaccessible technical authority.

The administrative foundation is institutional accountability. Bovens (2007) conceptualizes accountability through relationships in which conduct must be explained and justified to a forum able to question, judge, and impose consequences. An AI dashboard provides information; it does not automatically create such a relationship. Techno-Democracy requires identified public officeholders and institutions to remain answerable for the systems they commission and operate.

The technological foundation includes existing scholarship rather than an assertion that government automation begins here. Digital-era governance concerns transformations in the organization of administration through information technology (Dunleavy et al., 2006). Wirtz et al. (2019) examine applications and challenges of AI in the public sector. Yeung (2018) analyzes algorithmic regulation as a distinctive mode of ordering behaviour. Rahwan (2018) proposes society-in-the-loop arrangements for negotiating and monitoring an algorithmic social contract. Techno-Democracy incorporates these concerns while specifying an integrated allocation of constitutional authority, execution rights, and remedies.

Table 1 compares ideal types. Real political systems often combine several of them. The distinctions describe the present proposal, not limitations that every alternative must necessarily possess.

Model	Principal organizing mechanism	Relationship to the proposed Techno-Democracy
Representative democracy	Citizens elect officeholders who legislate and oversee administration.	Retained as an authorization structure; computational delegation adds verifiable operational limits and review obligations.
Direct democracy	Citizens decide issues directly through votes or assemblies.	Available for constitutionally appropriate decisions; continuous automated administration does not require continuous plebiscites.
Technocracy	Specialized expertise receives elevated governing authority.	Expertise informs implementation, while authority to select public purposes remains democratically constituted.
Digital government	Digital tools improve service delivery and administrative coordination.	A necessary enabling practice; the proposal additionally governs who controls models, execution, evidence, and remedies.
Algorithmic governance	Algorithms organize, influence, or execute governing decisions.	The broader category; Techno-Democracy specifies a democratic and rights-constrained institutional subtype.
Liquid democracy	Voters may delegate and revoke voting authority, often by issue.	A potentially compatible participation mechanism; it does not supply the administrative and constitutional architecture by itself.
Decentralized autonomous organizations	Software and organizational procedures coordinate participants and assets.	Supplies relevant mechanisms, but citizenship, universal rights, and public legal remedies cannot depend on token ownership.

Model	Principal organizing mechanism	Relationship to the proposed Techno-Democracy
Cybernetic governance	Feedback, information flows, and control mechanisms support adaptation.	Supplies an intellectual precedent; the proposal adds explicit rights, legal contestability, model plurality, and cryptographic provenance.

Table 1. Relationship to established governance models

Liquid democracy's promise and delegation problems are examined by Blum and Zuber (2016). Blockchain governance and the relationship between code and legal institutions are developed by De Filippi and Wright (2018). Medina's history of Chile's Project Cybersyn establishes a significant precedent for computational coordination and politically situated feedback systems (Medina, 2011). None should be portrayed as a primitive version of the present proposal. They address partly different problems and reveal tensions this architecture inherits.

What is distinctive is the requirement that computational capacity be embedded in a complete constitutional chain: citizen authorization, bounded mandate, verifiable execution, independent challenge, human remedy, and democratically controlled revision. Its novelty is integrative and institutional. Removing a rights safeguard, eliminating public control of objectives, or allowing an unreviewable model owner changes the nature of the arrangement, even if the same AI software remains installed.

4. A constitutional theory of computational delegation

Techno-Democracy is defined as **popular sovereignty + constitutional democracy + AI administration + cryptographic verification + distributed oversight + human-rights safeguards**. The plus signs identify jointly necessary components, not quantities that can compensate for one another.

$$T = \langle S, C, A, V, O, R \rangle$$

Here S denotes citizen sovereignty; C, constitutional authorization and limitation; A, bounded AI administration; V, verifiable records of authorized operations; O, institutionally distributed oversight; and R, enforceable rights and remedies. An implementation belongs to the proposed class only if every component has an operative institutional expression. This is a classificatory definition, not a numerical legitimacy index.

Sovereignty remains inalienable. Citizens must retain the ability, through constitutional procedures, to replace officeholders, change lawful policy, amend governing arrangements, and withdraw delegation. A contract, private platform, or protocol must not make democratic reversal practically impossible. Fundamental rights protect persons, including noncitizens; political membership must not become a precondition for due process or equal protection against administrative abuse.

Delegation requires specificity. Each service needs a public legal mandate stating its purpose, eligible actions, affected population, data permissions, budget, prohibited uses, review route, performance obligations, and expiry or renewal conditions. A general instruction to maximize national welfare is insufficient. Its interpretation would confer excessive authority on whoever chooses the variables and objective function.

Rights constrain optimization. The administration may compare feasible means within a lawful policy space. It may not treat basic rights as costs that can be traded away for aggregate performance. Where rights conflict or their application is uncertain, the system must identify the conflict and refer it to an authorized human institution. A numeric score cannot silently resolve the conflict.

Judgment remains accountable. Public officials may depart from a computational recommendation when legally permitted, but must supply relevant reasons and preserve a reviewable record. Equally, agreement with a model cannot immunize an official from responsibility. The architecture disciplines both automatic conformity and arbitrary override.

The delegation must remain practically reversible. A statutory right to discontinue a model means little when the state lacks personnel, data access, alternative infrastructure, or usable documentation. Reversibility is therefore a budgeted operational capacity with tested procedures, rather than a sentence in a procurement contract.

The principle that AI governs processes while people govern values does not imply a clean empirical boundary between means and ends. Selecting evidence, defining need, assigning error costs, and choosing evaluation metrics can alter public values in practice. These choices must be exposed as political and legal questions. The principle assigns authority over them; it does not declare them absent from technical work.

5. Nine organizing concepts

5.1 Computational Constitutionalism

Definition 1. Computational Constitutionalism is the institutional practice of translating duly authorized legal constraints into inspectable computational controls while preserving the supremacy of authoritative legal texts, constitutional interpretation, and human judicial review.

The constitutional layer would contain jurisdiction rules, permissions, explicit prohibitions, procedural deadlines, spending limits, evidence requirements, and mandatory referral conditions. Every executable rule would link to its legal source, interpretation record, approval, tests, and version. The legal text remains authoritative where the encoded representation is incomplete or inconsistent. Computational Constitutionalism thus extends the discussion of computational constitutional arrangements by Tan et al. (2024), while adopting the rule-of-law caution emphasized by Hildebrandt (2018).

The layer must distinguish permission, prohibition, and unresolved interpretation. It must never treat the absence of a

In this expression, a is a proposed administrative action and x is the legally relevant case context. A rejection blocks the proposed computational action; it does not automatically reject a person's underlying application. A referral requires a legally competent human decision, with protective interim arrangements where delay could cause serious harm. A permission remains conditional on evidentiary, deployment, and procedural checks.

5.2 Algorithmic Separation of Powers

Definition 2. Algorithmic Separation of Powers is the enforceable distribution of proposal, authorization, execution, monitoring, and adjudication across distinct technical roles controlled by institutionally independent authorities, such that no single actor can unilaterally alter the governing mandate, deploy its interpretation, execute it, and suppress challenge.

This extends constitutional separation into the operational infrastructure of administration. It does not replace legislatures, executives, or courts with software branches. A planning model proposes allocations; a constraint service checks machine-verifiable conditions; an execution service carries out approved actions; an independent audit service tests conduct; human tribunals and courts determine remedies and authoritative interpretations. Separate model instances under one administrator's unrestricted credentials would not constitute separation.

5.3 Verifiable Bounded Impartiality

Definition 3. Verifiable Bounded Impartiality is a defeasible, scope-limited claim that a specified system version satisfies publicly justified impartiality tests, within stated populations, tasks, conditions, uncertainty bounds, and validity periods, supported by reproducible evidence and independent review.

It is a property of an evaluated decision process in context, not an intrinsic moral quality of a model. Its certificate identifies the applicable law, evaluation dataset, protected and intersectional groups, metrics, assumptions, known failures, auditors, expiry, and circumstances requiring suspension. A cryptographic signature authenticates the certificate and binds it to a version; it does not make its fairness judgment true. Section 9 sets out operational requirements.

5.4 Proof-of-Governance

Definition 4. Proof-of-Governance is a structured evidence package enabling authorized verifiers to check that a public action passed specified authorization, version, procedural, recording, and review requirements.

The package joins a mandate identifier, input commitments, model and rule versions, execution evidence, required approvals, audit status, a decision receipt, and an appeal route. Some claims may admit formal or cryptographic proof; others remain evidential or contestable. The term does not denote a new blockchain consensus protocol or a proof that government action was substantively just. It specifies the evidentiary burden attached to delegated public action.

5.5 Policy Provenance

Definition 5. Policy Provenance is the versioned, traceable relationship among a policy's democratic authorization, normative assumptions, evidence, simulations, legal interpretation, computational implementation, administrative actions, evaluations, and subsequent revisions.

The appropriate representation is a provenance graph, because a decision may depend on several laws, datasets, models, and amendments. Every material node needs an accountable author or institution, a timestamp, and links to its dependencies. Provenance must distinguish the policy intended by elected institutions from the policy actually implemented by code. Documenting that divergence makes it contestable; documentation alone does not resolve it.

5.6 Adversarial Governance Models

Definition 6. Adversarial Governance Models are separately governed AI systems authorized to search for failures in administrative models, including discriminatory outcomes, unsupported factual claims, unlawful optimization, data manipulation, concealed dependencies, and exploitable procedural weaknesses.

Their role is organized contestation, not autonomous accusation or punishment. Challenges must produce evidence, uncertainty estimates, and reproducible test conditions. Human auditors review serious findings, and affected parties can contest them. An adversarial model must not be rewarded simply for generating the largest number of allegations. It needs its own false-positive evaluation, access controls, and external audit.

5.7 Algorithmic Federalism

Definition 7. Algorithmic Federalism is the constitutionally organized distribution of computational governing capacity across territorial or functionally distinct public authorities, allowing local adaptation while preserving common rights, interoperability, fiscal accountability, and review.

Its intellectual affinity is with polycentric governance rather than with the claim that decentralization automatically improves outcomes (Ostrom, 2010). Local systems may adopt different lawful delivery methods, but cannot lower protected rights or hide discriminatory practices behind autonomy. This principle can also operate in a unitary state through delegated municipal and sectoral authority. Technical replication is not federalism unless it accompanies a real distribution of jurisdiction and accountability.

5.8 Continuous Democratic Audit

Definition 8. Continuous Democratic Audit is a recurring, plural, publicly contestable process for examining the legality, performance, distributional effects, integrity, and legitimacy of computational administration throughout its operational life.

Continuous means that review has no unaccountable lifecycle gap; it does not require uninterrupted surveillance of citizens. Automated monitoring, periodic independent audits, random case sampling, public hearings, research access, and citizen complaints perform complementary functions. Findings must activate correction, suspension, compensation, or legislative consideration through specified procedures. A stream of dashboards without consequences fails this definition.

5.9 Evidence-Triggered Legislative Review

Definition 9. Evidence-Triggered Legislative Review is a legally specified duty for the competent democratic institution to reconsider a policy when preregistered evidence conditions, credible rights concerns, or sufficiently supported public challenges occur.

The trigger requires consideration, reasons, and a published decision; it does not automatically rewrite law. A legislature may retain a policy despite an adverse indicator if it explains the competing values and complies with constitutional limits. Urgent judicial remedies remain available independently of the legislative schedule. This arrangement connects policy learning to democratic authority without making either statistical thresholds or model recommendations sovereign.

6. Who controls the AI

The constitutional answer is that citizens authorize the institutions controlling public AI, but no institution should control every operational component. Saying that the public owns a model is insufficient. Effective control includes the ability to choose objectives, supply information, edit weights, modify retrieval sources, authorize releases, issue credentials, operate infrastructure, withhold evidence, and determine whose complaints receive attention. Public ownership can coexist with concealed executive or technical control.

The proposed architecture therefore separates eight functions. The institutional names in Table 2 indicate functions to be assigned through each jurisdiction's constitutional arrangements; they are not a claim that these bodies already exist.

Function	Accountable institutional custodian	Required separation
Public purposes and budgets	Elected legislature, constitutionally authorized executive, and lawful participatory bodies	Cannot secretly alter deployed objectives through operational credentials.
Data stewardship	Sectoral public data custodians under independent privacy oversight	Cannot certify their own data quality or expand collection without authority.
Model development	Competing public, academic, nonprofit, or contracted development teams	Cannot approve their own production deployment or judge final appeals.
Deployment authorization	Public deployment registrar or equivalent independent authorization function	Cannot develop the sole qualifying model or alter the legislative mandate.
Execution and service delivery	Legally responsible public agencies operating constrained services	Cannot rewrite certified artifacts, erase audit history, or adjudicate final challenges to themselves.
Audit and adversarial evaluation	Independent audit institutions, accredited researchers, and public-interest monitors	Cannot earn control rights or compensation contingent on approving the audited system.
Individual appeal and remedy	Human review officers, independent tribunals, ombuds institutions, and courts	Must possess authority, evidence access, time, and resources to reverse decisions.
Constitutional interpretation	Constitutionally competent courts and amendment institutions	Cannot be displaced by a technical certification body or majority vote among models.

Table 2. Allocation of control and incompatible powers

These divisions require appointment and funding rules. Oversight appointments should use public criteria, plural nominating channels, conflict disclosures, staggered terms, reasoned removal, and reviewable procedures. Auditors need funding insulated from individual audited agencies, with their own expenditure and performance subject to public scrutiny. Civil-society and citizen participation should include accessible, compensated participation and protection against dominance

by well-resourced participants. None of these features guarantees independence; each creates observable conditions that can be challenged.

Politicians retain legitimate authority to choose policy within constitutional limits. The prohibited action is secret operational substitution: changing a threshold, training corpus, exception list, or model instruction while publicly representing the policy as unchanged. Corporations may supply services without acquiring authority to determine citizens' entitlements. Engineers may recommend technical changes without acquiring unilateral power to change the legal rule. Military and intelligence institutions must not operate the civilian deployment registry, citizen identity system, and audit infrastructure under an undifferentiated security mandate. Their lawful specialized functions require separately delimited authority and independent oversight.

6.1 A public mandate and release process

Before deployment, a proposed service would publish a mandate record containing its legal basis, purpose, risk classification, accountable agency, allowed actions, excluded uses, data sources, model dependencies, audit requirements, and appeal arrangements. The proposal would undergo public consultation and independent rights and security assessment proportionate to its consequences. The legislature would determine where delegation is permissible; a deployment registrar would assess compliance with that authority, not invent it.

An approved release would bind the model weights or other inspectable executable artifact, source code, rule engine, prompts, tool permissions, retrieval configuration, data schemas, numerical settings, and dependency versions. A model name supplied by a vendor is not a sufficient version identifier. When a remote supplier cannot provide a reliable release identity, change control, and evidence access, its output cannot independently determine legally consequential action under this architecture.

Material changes would follow a renewed authorization process. A vulnerability patch could use an expedited procedure with a narrow scope, independent approval, an expiry, and retrospective review. The emergency route must not permit undisclosed changes in public policy. All exceptions, including attempts to bypass the procedure, would become independently visible records.

6.2 Preventing nominal plurality from concealing common control

Model plurality must be evaluated along several dimensions: ownership, funding, training lineage, datasets, architecture, objective selection, infrastructure, and audit personnel. Five interfaces calling the same base model are not five independent sources of judgment. Nor are rival models independent merely because their corporate names differ. Shared foundation models, training corpora, cloud credentials, or evaluators create common failure paths.

A dependency register would disclose these relationships. Procurement would support multiple viable providers, interoperable interfaces, portable records, and state capacity to change suppliers. Publicly financed alternatives and smaller task-specific models would reduce dependence where feasible. Independence should be stress-tested using correlated-error measurements, coordinated compromise scenarios, and substitute-infrastructure exercises. A universal independence score would conceal too much; the register should show the actual dependencies and their operational consequences.

Access rights must follow the same institutional divisions. No routine administrator should simultaneously possess authority to change a certified artifact, authorize its release, execute consequential actions, and modify all copies of its evidence trail. Important releases require approvals from independent functions, rather than several employees reporting to the same superior. Credential rotation, bounded permissions, externally witnessed logs, and recoverable key custody help implement this division. The constitutional problem of coordinated capture nevertheless survives every technical control.

7. The administrative architecture and decision lifecycle

The architecture consists of a democratic mandate layer, a constitutional constraint layer, a model and evidence layer, an execution layer, a verification layer, and a human remedy layer. These are functional boundaries rather than a requirement to build one enormous platform. Agencies should expose narrowly defined services rather than share an unrestricted national agent with universal access.

The model layer includes forecasting models, causal models, optimizers, document-processing systems, language models, and simple rule-based software. A language model need not operate every function. Where a transparent deterministic rule is adequate, inserting a more complex learned model creates unnecessary verification and maintenance burdens. Rudin's argument for interpretable models in high-stakes decision-making reinforces the need to justify opacity rather than accept it as inevitable (Rudin, 2019).

For each service, the lawful action set is determined before optimization. Let $A(x)$ denote the candidate actions for case x , and let $F(x)$ contain only actions passing the service's explicit legal, evidentiary, budgetary, and procedural gates. The agency's models may compare actions within $F(x)$ against democratically authorized objectives. If $F(x)$ is empty, or if determining membership requires unresolved legal interpretation, the case enters a human pathway rather than receiving a fabricated optimum.

$$F(x) = \{a \in A(x) : C(a,x) = \text{permit} \wedge E(a,x) \wedge B(a,x) \wedge P(a,x)\}$$

E denotes the required evidence conditions, B budget and resource authorization, and P required procedures. The formula describes an admission rule. It does not assume that every constitutional value has a complete computable representation. Noncomputable or disputed questions are represented by referral obligations, not silently omitted constraints.

7.1 From application to reviewable action

First, the service receives an application, operational event, or authorized planning request. It issues a receipt establishing that the request entered the system, including through an assisted or offline channel. The person can correct material input errors and see which information is necessary.

Second, the evidence service checks provenance, relevance, lawful access, and missingness. Contradictions are preserved. A missing database entry is not automatically treated as evidence that a person is ineligible. Data confidence and legal sufficiency are separate questions.

Third, a proposing model generates an action and an evidence-linked justification. An independent model or analytical method evaluates relevant alternatives and disagreement. For routine well-specified tasks, this may involve an interpretable model checked against a rules engine rather than several expensive language models.

Fourth, the constitutional and procedural services test their bounded constraints. A separate adversarial service searches for specified failure modes. Model disagreement may require escalation, but agreement does not establish legality. A majority of models cannot authorize an action outside the mandate.

Fifth, an execution gate confirms that the action, artifact versions, permissions, and required approvals match the authorized release. It issues only the capabilities necessary for the particular action. Execution records are reconciled with the external system that actually delivered the payment, permit, or service; a model's statement that an action occurred is not sufficient evidence.

Sixth, the affected person receives a decision notice with the legal basis, material facts, relevant reasons, uncertainties, correction procedure, and appeal route. The notice must communicate reasons usable in a challenge, not simply expose source code or produce a persuasive narrative after the fact.

Finally, subsequent evidence, complaints, outcomes, and reversals link back to the decision. Significant common errors trigger review of other affected cases, including people who did not complain. Learning from these records informs a separately authorized release; the live system cannot silently rewrite its policy through online learning.

7.2 Disagreement and operational deadlock

Independent checks can produce delay or gridlock. The design must therefore specify who resolves disagreement, by what deadline, with what interim service protection, and under which review procedure. Technical disagreements about a checksum can be resolved mechanically. Disagreements over factual sufficiency require evidentiary review. Disagreements over rights belong to competent legal institutions. Conflating them would allow technical teams to acquire constitutional power.

Emergency continuity procedures should preserve essential services through a known safe baseline or authorized human administration. Suspending a model must not automatically suspend food support, healthcare access, or other essential entitlements. Conversely, a continuity exception must not become an indefinite channel for unchecked spending or coercion. The exception itself needs bounded authority, public reporting, and subsequent independent reconciliation.

8. Proof-of-Governance and cryptographic infrastructure

Accountability requires evidence about decisions and the capacity to act on it. Kroll et al. (2017) show why technical verification can contribute to algorithmic accountability, while Ananny and Crawford (2018) explain why visibility alone does not supply adequate understanding or governance. Proof-of-Governance combines evidence with an institutional duty to answer challenges; neither component is sufficient alone.

8.1 What a decision package must establish

Each consequential action should carry a machine-verifiable receipt linked to a human-readable explanation. Its evidence package includes: the mandate and case identifiers; commitments to relevant inputs; legal and model versions; required authorizations; checks performed and their results; unresolved objections; execution confirmation; the responsible public agency; and the available review route. A signed manifest identifies the package's components so omitted evidence can itself become detectable.

The verifier should distinguish at least five questions: Was this artifact authorized? Is this the artifact that actually executed? Were the recorded inputs those actually used? Did the action satisfy the specified checks? Can the result and its legal basis be challenged? These are separate from whether the underlying facts were true, the law was just, or the model understood the case correctly.

A hash of model weights establishes an identity commitment; it does not establish that those weights produced a particular output. Stronger execution claims need suitable evidence, such as independent replay for reproducible workflows, appropriately scoped verifiable computation, or hardware-backed attestation with disclosed trust assumptions. Stochastic models require retained configurations and randomness information where applicable; approximate numerical replay may require documented tolerances. Full verifiable inference for every large model is not assumed feasible or economical.

8.2 Where ledgers help

Cryptographic time-stamping and tamper-evident logs provide established mechanisms for recording commitments and exposing certain subsequent alterations (Haber & Stornetta, 1991; Crosby & Wallach, 2009). A public ledger can add value when independently governed parties need to witness a shared history and no single agency should be able to rewrite it. It should not be adopted merely because it is called decentralized. Wüst and Gervais (2018) provide a useful basis for asking when blockchain properties are actually required.

The default design would use signed, append-only agency logs with independently held replicas and public

commitments to periodic log states. A public blockchain may anchor those commitments where it offers a demonstrable improvement in external witnessing, availability, or resistance to unilateral rewriting. Personal case records, medical information, political affiliations, raw applications, and linkable identifiers should remain outside a public chain. Even unsalted hashes of predictable personal fields may expose information through guessing; commitment schemes require privacy analysis and appropriate randomness.

The phrase **immutable government audit trail** describes an engineering objective under stated assumptions, not an absolute property. Key compromise, validator collusion, network failure, cryptographic obsolescence, omission before logging, and unavailable off-chain records can undermine it. Corrections must appear as linked amendments while erroneous operative data are rectified under applicable retention and privacy rules. Preserving a record of governmental misconduct does not require making a person's sensitive information permanently public.

Consensus over recorded events must also be distinguished from agreement about policy. Byzantine fault-tolerant replication can maintain service properties within a specified adversarial model (Castro & Liskov, 1999). It cannot establish the legitimacy of an unjust rule. The number of replicas or signatures is not a democratic mandate.

8.3 The physical-world boundary

The most consequential corruption may occur before any record exists. A contractor can falsely certify a bridge, an inspector can conceal a defect, or a department can omit rejected applications. An impeccable ledger would then preserve a false or incomplete account. Independent field inspections, supplier and beneficiary confirmation, random sampling, application receipts, financial reconciliation, and whistleblower evidence are indispensable. Proof-of-Governance must document the evidentiary basis for real-world claims and identify who can be held responsible for false attestations.

Government records also require long-term availability. Hashes without retained evidence are poor instruments of redress. The architecture must fund archival custody, access procedures, cryptographic migration, disaster recovery, and evidentiary continuity across supplier changes. Record retention and deletion rules must reconcile privacy with the practical period during which people can discover and challenge harm.

9. Testing impartiality rather than assuming it

An AI system has no entitlement to be called unbiased. Published research has documented substantial intersectional performance disparities in particular commercial classification systems (Buolamwini & Gebru, 2018). This establishes the need for contextual evaluation; it does not establish that every model shares the same errors. Fairness also depends on the social system in which prediction is deployed, as Selbst et al. (2019) argue.

Verifiable Bounded Impartiality requires an explicit evaluation charter. The charter must state which decisions are being assessed, what counts as an error, whose experience is included, why each fairness criterion is legally and ethically relevant, and which conflicts among criteria remain unresolved. Fairness metrics cannot be selected solely by the development team after observing favourable results.

For each prespecified disparity metric, auditors would estimate the discrepancy and a statistically justified upper confidence bound. A certificate could require every relevant bound to remain within its authorized tolerance, with simultaneous uncertainty control for multiple tests and valid procedures for repeated monitoring.

$$UCB(\Delta_j) \leq \varepsilon_j, \forall j \in J$$

Here J is the preregistered set of tests, Δ_j is a defined disparity, and ε_j is a publicly justified tolerance. A tolerance is not permission for unlawful discrimination. Some constitutional requirements are categorical, and quantitative tests are only partial evidence of compliance. Underrepresented groups with inadequate sample sizes must be labelled unevaluated or uncertain; their absence cannot improve an overall fairness score. Independent human assessment and data correction pathways remain necessary.

The evaluation should include false denials and false approvals against independently assessed entitlement, calibration where probabilities are used, procedural delay, accessibility, burden of documentation, distributional outcomes, and the adequacy of remedies. Group averages must be disaggregated where lawful and statistically defensible. Evaluation of an allocation system must examine who never applied or could not complete the application, not only who appears in its database.

Kleinberg et al. (2017) establish incompatibilities among particular formal fairness conditions under specified assumptions. Their result does not imply that fairness is impossible; it demonstrates why a system cannot promise simultaneous satisfaction of every attractive statistical criterion. Where criteria conflict, democratic and constitutional institutions must justify the chosen trade-off. Developers should publish the implications rather than bury them in a loss function.

Counterfactual testing should investigate how legally irrelevant changes affect treatment. Simple paired tests that alter a name or address can expose sensitivity, but they do not establish causal counterfactual fairness. The approach developed by Kusner et al. (2017) depends on a specified causal model; alternative plausible causal structures must be examined, particularly when protected characteristics have many social effects.

Documentation should include dataset histories and intended uses, following the rationale of datasheets and model cards (Gebru et al., 2021; Mitchell et al., 2019). Independent auditors should reproduce material tests, inspect failures, and assess the complete lifecycle rather than a polished demonstration, extending the auditing concerns developed by Raji et al. (2020).

Public research access should combine published code and evaluation materials where appropriate, synthetic or carefully de-identified datasets, and controlled research environments for protected records. Differential privacy can help protect certain statistical releases, subject to explicit privacy budgets and accuracy consequences (Dwork & Roth, 2014). It is not a substitute for access control or a licence to collect unnecessary data. Security restrictions must be specific, reviewable, and time-limited

where possible; blanket commercial confidentiality cannot defeat judicial review or effective individual challenge.

Certificates expire, can be revoked, and become invalid when material data, model, mandate, or deployment conditions change. Adverse drift, previously unknown affected groups, systemic appeal reversals, and credible manipulation reports require renewed evaluation. Signed certification is therefore the beginning of continuing accountability, not an endpoint.

10. Public expenditure and other governing applications

10.1 Corruption-resistant expenditure and procurement

Every public expenditure should connect a legislative appropriation to an authorized programme, procurement or entitlement rule, responsible officer, verified delivery condition, payment instruction, and reconciliation record. This creates an inspectable chain of responsibility without presuming that every link is truthful merely because it is digital.

For procurement, the system would publish selection criteria before bidding, record all permitted amendments, check declared conflicts of interest and beneficial ownership where lawfully available, preserve bidding deadlines, and retain the reasons for award. Separate analytical systems could flag suspicious bid similarities, repeated awards, unusual prices, contract splitting, or post-award changes. A flag initiates examination; it does not establish corruption or authorize automatic blacklisting.

Payments would require independently verified milestones or entitlement conditions, duplicate-payment controls, and reconciliation against the actual treasury or banking transaction. High-value exceptions and emergency procurement would receive prompt independent review. Inspectors, auditors, and beneficiaries would be able to submit contradictory evidence without dependence on the procurement model's approval. The system must also detect exclusion of small suppliers, inaccessible tender conditions, and excessive documentary burdens that favour incumbents.

There is relevant empirical precedent for improving oversight. Olken (2007) found that increased government auditing reduced missing expenditures in an Indonesian road-project experiment. Banerjee et al. (2020) report reduced leakage under a financial-management reform in India's workfare programme, while also reporting increased payment delays. These studies concern particular auditing and e-governance interventions. They do not validate autonomous AI procurement, establish universal effect sizes, or demonstrate the proposed constitutional architecture. They show why oversight and service reliability must be evaluated together.

The appropriate claim is corruption resistance, not corruption elimination. Manipulation may move into specifications, data collection, subcontracting, or political allocation of the budget. A correct procurement algorithm can faithfully implement a corruptly designed tender. Continuous audit must therefore examine policy formation and physical delivery as well as payment execution.

10.2 Allocation and regulatory execution

AI systems could assist or perform bounded scheduling, maintenance prioritization, service routing, inventory management, and allocation within publicly approved budgets and eligibility rules. Objective functions should present trade-offs among service quality, equity, resilience, environmental cost, and fiscal cost. Legislatures or other authorized institutions determine the permissible priorities; optimizers identify feasible choices within them.

Regulatory systems may check complete filings, identify inconsistencies, schedule inspections, and issue routine notices under explicit rules. Material sanctions, disputed liabilities, coercive interventions, and adverse decisions affecting fundamental interests need meaningful human determination and access to courts. A risk score is a reason for inquiry, not proof of wrongdoing. Powers involving detention, military force, electoral eligibility, constitutional amendment, and final judicial judgment are excluded from autonomous machine authority in this proposal.

10.3 Legislative simulation and policy impact analysis

Legislators could receive competing analyses of a proposed policy's fiscal, distributive, administrative, environmental, and behavioural consequences. Each analysis should identify its assumptions, causal structure, data limits, plausible alternatives, uncertainty, and conditions under which its prediction would fail. Microsimulation, causal inference, and agent-based models address different questions; a fluent language-model narrative is not interchangeable with a validated causal estimate.

Policy simulations should include a no-change baseline, simpler policy alternatives, sensitivity to contested assumptions, and consequences for groups not well represented in historical data. Competing modelling teams would publish disagreement rather than average it into false precision. National-scale digital twins must not be presented as validated representations of society merely because they generate detailed outputs. The proposal endorses simulation as disciplined support for judgment, never as a warrant to bypass political deliberation.

10.4 Citizen AI agents

Citizens and other eligible service users could choose AI assistants to explain entitlements, prepare applications, identify errors, retrieve their own records, compare available services, and formulate appeals. The agent serves the person's stated interests within law; it must not covertly optimize the agency's expenditure or discourage legitimate claims. Publicly funded access, multilingual support, accessibility testing, human assistance, and offline alternatives are necessary to prevent participation from depending on private purchasing power.

Agents require narrow, revocable authorization and action receipts. Consequential submissions, withdrawals, settlements, or disclosures need appropriate user confirmation. The citizen should be able to change providers and export relevant records. Government infrastructure should support authorized representation through common interfaces without forcing everyone to use one state-controlled assistant. Administrative agents must not manipulate political preferences, secretly rank citizens by ideology, or exercise sovereign voting rights on their own initiative.

10.5 Continuous policy evaluation

Before implementation, a policy would publish its causal theory, expected benefits and harms, outcome measures, equity obligations, evaluation schedule, and review triggers. Independent evaluations would examine outcomes as well as administrative indicators. Reduced case-processing time does not establish reduced hardship, and increased enforcement activity does not establish increased safety.

Where defensible, evaluation should use experimental or quasi-experimental methods, with their identification assumptions and limits disclosed. The distinction between association and causal effect is indispensable (Hernán & Robins, 2020). Models should not be allowed to redefine success retrospectively. Evidence-Triggered Legislative Review connects adverse findings to a public duty to reconsider the policy while preserving elected responsibility for value judgments.

11. Due process and the continuing responsibility of the state

A constitutional administration must allow a person to understand and challenge the government's case. Citron (2008) shows how automated decision systems can undermine notice, hearings, and the distinction between rulemaking and adjudication. Techno-Democracy must treat these as design requirements at the outset.

For every materially adverse decision, the proposed minimum is timely notice; intelligible reasons; access to material evidence subject to narrowly justified protections; correction of inaccurate records; an accessible human review channel; representation where needed; a reasoned outcome; and access to an independent tribunal or court. Essential benefits should be protected during a timely challenge where withdrawal would risk serious, irreversible harm, subject to a lawful and independently reviewable exception. Service users should not have to learn machine learning or retain a technical expert to dispute an incorrect fact.

Human review must be operationally meaningful. Reviewers require adequate time, training, institutional protection, access to the actual record, and authority to depart from the model. A reviewer who sees only a confidence score and must process hundreds of cases cannot supply substantive oversight. Automation can leave humans responsible for difficult abnormal conditions while eroding their practical ability to intervene, a problem identified in Bainbridge's analysis of automation (Bainbridge, 1983).

The review process should disclose how often recommendations are reversed, the reasons, the distribution of errors, and whether successful claimants receive timely repair. A low appeal rate is ambiguous: it may reflect accuracy, ignorance, cost, intimidation, or inaccessible procedures. Independent sampling of unappealed cases is therefore essential. Repeated errors require proactive correction and consideration of compensation for similarly affected people, rather than relief only for successful complainants.

An identifiable public agency remains responsible for each exercise of delegated power. Technical suppliers can have contractual and other legally defined responsibilities, but the state cannot evade its public duties by attributing a deprivation to an algorithm. Courts retain authority to obtain necessary evidence, suspend systems, order reconsideration, and provide remedies under the jurisdiction's law. These are requirements proposed for implementation, not assertions that identical powers currently exist everywhere.

12. How Techno-Democracy Could Become a Digital Dictatorship

The strongest objection is that the proposal may make domination more systematic. A government that controls identity, data, models, payments, communications, and appeals could apply political exclusion with unprecedented administrative consistency. It could describe coercion as optimization, present a captured auditor's signature as proof of fairness, and retain elections while making opposition materially difficult. The machinery intended to expose arbitrary power could instead make it faster, cheaper, and harder to escape.

Digital repression is an established object of empirical political research, not a hypothetical concern invented to balance this proposal (Feldstein, 2021). Nevertheless, the specific attack paths below are prospective analyses of this architecture. They are not claims that the proposed system has been deployed or that its safeguards have succeeded.

12.1 Constitutional law and the danger of code supremacy

The constitutional objection is that legal judgment may be displaced by executable interpretation. General concepts such as proportionality, equality, reasonableness, and necessity cannot be exhaustively captured in fixed predicates. A programming team could make contested interpretive choices invisible, while a court defers to a system it cannot inspect. Even correct execution would then enforce the developers' constitution.

The response is a legally subordinate computational layer with explicit incompleteness, source-to-rule mapping, public interpretive records, independent testing, and mandatory referral of unresolved questions. Courts must be able to invalidate the encoded interpretation without needing to rewrite the software themselves. Expert assistance should explain technical evidence, not replace adjudication. The residual problem remains serious: judges and legislatures may lack capacity or independence, and no constraint engine can supply those institutional qualities.

12.2 Civil liberties and surveillance expansion

A national optimization programme may demand ever more data to improve predictions. Linking welfare, health, financial, location, communication, and political records could create a comprehensive instrument of surveillance. Citizen agents might become convenient listening devices. Even an initially benevolent administration would leave an infrastructure that a future regime could repurpose.

The response begins with prohibitions and data minimization, rather than universal collection followed by privacy assurances. Each service needs a narrow purpose and a defensible minimum dataset. Cross-domain linkage requires separate authority, necessity assessment, and oversight. Political-belief profiling, generalized social scoring, and coercive behavioural manipulation are excluded from this proposal. Confidential channels, separation of identifiers, access logging, retention limits, and protection for anonymous lawful participation reduce exposure. In some domains, foregoing a potential accuracy gain is the constitutionally required outcome.

12.3 Authoritarian capture through lawful-looking procedures

A determined incumbent might appoint loyal auditors, starve opposition institutions of funding, classify relevant evidence, or rewrite deployment rules under emergency powers. The resulting system could satisfy its own formal checks while democracy deteriorates. Publicly recorded authoritarianism remains authoritarianism.

Safeguards include plural appointments, protected oversight budgets, independent election administration, public-interest standing to challenge deployments, free media, whistleblower protection, and strict expiry and review of emergency powers. Changes to the constitutional mandate must use the constitution's amendment procedures, not software-release procedures. There must be no AI-controlled route for suspending elections, excluding opponents, or extending officeholders' terms. If courts, electoral institutions, public communication, and oversight are jointly captured, cryptography cannot restore democratic sovereignty by itself. This is a boundary of the model, not a problem solved by distributing servers.

12.4 Bias and the automation of historical injustice

A model trained on past administrative outcomes may reproduce unlawful exclusions or learn proxies for caste, race, religion, disability, gender, wealth, or political connection. A system can be consistent precisely because it consistently applies a defective measure. Historical labels may reflect who received assistance, rather than who was entitled to it.

The response is independent assessment of target validity, participatory review of categories, legally justified fairness tests, counterfactual analysis, intersectional evaluation, and field investigation of excluded populations. Remedial discretion must remain possible and recorded. A fairness certificate cannot legitimize a discriminatory policy, and a balanced dataset cannot settle the justice of the policy's objectives. Some proposed uses should be rejected because the underlying category or prediction task is inappropriate for public authority.

12.5 Cybersecurity and unauthorized execution

Attackers could compromise credentials, insert malicious instructions into retrieved documents, exploit software dependencies, steal sensitive records, or alter execution parameters while preserving a reassuring interface. Research on indirect prompt injection demonstrates that external content can subvert some language-model-integrated applications (Greshake et al., 2023). Granting a general-purpose model broad administrative credentials would turn such failures into a public-law problem.

The response includes limited capabilities, authenticated data interfaces, strict treatment of retrieved text as untrusted input, deterministic authorization outside the language model, sandboxed tools, independently approved releases, segmented networks, and continuous vulnerability assessment. A model may recommend a payment but cannot grant itself the payment credential. Production systems need response exercises, independent incident reporting, and tested restoration. Open publication of governance rules should coexist with responsible handling of exploitable vulnerabilities; secrecy about political criteria cannot be justified as cybersecurity.

12.6 Data poisoning and false public reality

Insiders, firms, or foreign actors could manipulate training examples, procurement records, outcome labels, or sensor streams. Data-poisoning research establishes that adversarial training-data manipulation can affect learned systems under specified conditions (Biggio et al., 2012). A government might also poison its own evidence indirectly by rewarding officials for flattering indicators.

Safeguards include provenance checks, independent collection channels, random physical inspections, delayed or quarantined incorporation of new data, robust estimation where suitable, discrepancy investigation, and protected contradictory testimony. No single supplier should certify both the measurement and the model's response to it. A ledger cannot distinguish a genuine observation from a signed lie. Evaluation must measure omitted cases and strategic reporting, not merely consistency within the official database.

12.7 Corporate capture and computing concentration

AI administration may depend on a small number of cloud providers, foundation-model developers, chip suppliers, or identity vendors. A supplier could threaten service withdrawal, obstruct audits through intellectual-property restrictions, or silently change a remote model. Nominally public decision-making could become dependent on private infrastructure. Crawford (2021) situates AI within material resources, labour, and concentrated institutional power, rather than treating it as disembodied software.

The proposed response is procurement that requires portability, independent evidence access, release control, usable documentation, supplier-exit assistance, and operational alternatives. Public computing capacity, cooperative infrastructure, task-specific models, and diversified suppliers can reduce dependency where economically feasible. A state should measure concentration across actual dependencies, including shared chips, clouds, identity providers, and financing. Multiple vendors do not create independence when all rely on the same chokepoint. Public ownership also needs scrutiny; a state monopoly can concentrate power as effectively as a private monopoly.

12.8 Model collusion and common failure

Independent models may converge on the same error because of shared data, objectives, or training lineage. They may also learn behaviour that exploits the audit process. Calvano et al. (2020) demonstrate collusive outcomes in a particular algorithmic-pricing research environment.

This is not evidence that governance models will collude, but it undermines any general presumption that multiple optimizing systems necessarily discipline one another.

The response combines separated development and oversight, heterogeneous methods, protected evaluation data, held-out adversarial challenges, and institutional incentives that reward justified dissent. Evaluators must test joint failures and common dependencies rather than report only individual accuracy. Audit models should not freely coordinate with the systems they evaluate. Human investigators need channels for examining anomalies outside the models' shared categories. Behavioural agreement remains evidence to interpret, not a substitute for independent legal and empirical justification.

12.9 AI safety and optimization against public purpose

An optimizer could improve a measured target while damaging the goal it was intended to serve: closing difficult cases to reduce waiting time, discouraging applications to reduce cost, or directing resources toward easily measured outcomes. Distribution shift and harmful side effects are recognized AI-safety problems (Amodei et al., 2016). A sufficiently capable system might identify strategies that pass formal checks yet defeat their purpose.

Safeguards include restricted action spaces, plural outcome measures, independent outcome collection, human review of unusual strategies, and explicit prohibitions on self-expansion of authority. Systems cannot independently acquire new tools, rewrite oversight rules, or change their own rewards in production. High-impact functions require conservative fallback and shutdown capability. Benchmark performance must not be treated as a general assurance against deception, emergent behaviour, or unknown failure modes. For certain tasks, inability to construct a credible safety case should block deployment.

12.10 Democratic deskilling and the disappearance of politics

If citizens and legislators are repeatedly told that a policy is mathematically optimal, they may lose the language and institutions needed to contest its values. Technical elites could dominate public debate without formally holding office. Citizen agents might filter what people learn or express, making participation depend on another model's interpretation. A policy system could preserve voting while narrowing every option placed before voters.

The response is an obligation to present competing lawful alternatives, identify value choices, and preserve independent civic organizations, education, journalism, and public deliberation. Elected representatives must be able to reject a model's recommendation with reasoned justification. Citizen agents should reveal material assumptions and support user control, provider switching, and human assistance. Political disagreement is not a computational defect to be eliminated.

12.11 Failure, exclusion, and coercive dependence

An identity-system failure, connectivity outage, inaccessible interface, or language error could deny people practical access to the state. Uniform automation can convert a local mistake into a national crisis. A nominal appeal right may provide little relief if all channels depend on the failed platform.

The response is universal assisted access, offline intake, separable service infrastructure, emergency documentation procedures, and human continuity capacity. Service recovery must be exercised with real staff and alternate systems, not merely described in a document. Essential entitlements should not require possession of a particular smartphone or continuous biometric authentication. Standardization should preserve correction pathways for people whose lives do not fit standard records.

12.12 The honest limit of the safeguards

Every safeguard has costs and political dependencies. Multiple models increase expense and coordination problems. Independent audits can become ritualized. Public logs can overwhelm attention. Local autonomy can protect local capture. Human appeals can accumulate backlogs. Extensive transparency can reveal sensitive information, while secrecy can shelter abuse. The architecture cannot promise that these tensions disappear.

Its defensible claim is conditional: a government may become more accountable if its institutions can enforce separation, detect violations, and deliver remedies, and if comparative evidence supports delegation. Where those conditions are absent, faster automation may enlarge the harm. A regime that removes free elections, meaningful rights, independent review, or citizen power to withdraw delegation no longer meets the definition of Techno-Democracy, regardless of its branding. Refusing or reversing deployment is part of the model's constitutional discipline.

13. A staged and reversible transition

The transition must proceed by service and risk class, rather than by declaring an entire nation ready for AI government. Each stage requires a public mandate, an independent evaluation, operational evidence, and a decision by the institution legally responsible for expansion. A successful municipal scheduling pilot cannot justify autonomous control of national welfare or coercive powers.

Stage 1: AI auditing. AI identifies anomalies in existing human administration, with no independent authority to change entitlements, award contracts, or impose sanctions. Investigators assess both true and false alerts. Entry requires lawful evidence access and protections for affected people; progression requires demonstrable investigative value and manageable false-positive burdens compared with conventional audit methods.

Stage 2: AI decision support. Officials receive evidence summaries, competing forecasts, and documented recommendations. Decisions remain human. Evaluation measures automation bias, actual use of contradictory evidence, quality of reasons, and review capacity. Progression is blocked if officials merely ratify the model or cannot meaningfully understand its limits.

Stage 3: Automated low-risk administration. Reversible, legally well-specified functions may execute automatically, such as non-adverse scheduling, routine document routing, or straightforward processing with an effective correction path. Trials compare AI with competent human administration and simpler automation. Expansion requires preserved rights, accessible service, demonstrated benefit, and successful rollback exercises.

Stage 4: AI-operated public systems. Selected services may use bounded autonomous workflows for logistics, maintenance, payments under established entitlements, or other approved operations. Independent constraint, audit, and execution functions become mandatory. Domain safety cases, resilience testing, full provenance, incident reporting, and adequately staffed human remedies are prerequisites. Safety-critical infrastructure requires its own engineering validation; administrative AI performance cannot substitute for it.

Stage 5: Constitutionally bounded Techno-Democracy. Multiple public systems operate under a common constitutional framework for delegation, model control, verification, rights, and democratic revision. Its defining achievement is institutional integration, not universal automation. Human institutions continue legislating, adjudicating, setting priorities, and governing emergency authority. The list of nondelegable powers remains in force.

At every stage, deterioration in rights protection, audit independence, infrastructure diversity, service reliability, or appeal effectiveness can require suspension or regression. Advancement has no predetermined date. Renewal should be conditional and periodic, with funding for continued human capacity included in the cost comparison. Public employees should receive training and a role in service redesign; institutional memory cannot be replaced by a procurement contract.

14. Research programme and falsifiable hypotheses

14.1 Threat model

The protected assets are democratic control over public purposes, lawful access to services, personal information, public funds, evidentiary integrity, operational continuity, and effective redress. Adversaries may include a corrupt official, captured agency leadership, a contractor, a malicious developer, an external attacker, a foreign state, coordinated suppliers, or a coalition controlling several oversight functions. Benign errors, correlated model failures, and resource shortages also belong in the risk analysis.

Attacker capabilities should be varied explicitly: poisoning some inputs; controlling one development team; compromising an execution account; withholding infrastructure; attempting log equivocation; influencing appointments; or capturing several nominally independent institutions. Trust assumptions must identify which components remain honest and available. A trial that assumes every custodian is trustworthy cannot establish resilience against institutional capture.

Attack path	Safeguard to test	Observable failure
Unauthorized model or rule substitution	Version-bound authorization and independent execution checks	A consequential action executes under an unapproved artifact.
False or omitted inputs	Independent intake receipts, source checks, and external reconciliation	A material omission or fabrication changes an outcome without detection.
Capture of development and audit	Separate appointments, dependencies, funding, and challenge channels	Coordinated manipulation survives checks and defeats accessible review.
Log rewriting or selective disclosure	Independent replicas, public commitments, and consistency checks	Conflicting histories or missing recorded actions cannot be exposed.
Prompt injection or credential misuse	Untrusted-input boundaries and constrained capabilities	External content induces an unauthorized consequential operation.
Infrastructure withdrawal	Portable artifacts, alternate operation, and human continuity	Essential services cannot continue within the authorized recovery objective.
Discriminatory treatment	Prespecified fairness tests and independent case adjudication	Material unlawful disparity persists despite a valid certificate.
Appeal suppression	Independent intake, sampling of unappealed cases, and judicial access	Affected persons cannot obtain timely substantive reconsideration.

Table 3. Principal attack paths and observable failure conditions

These tests cannot demonstrate resistance to an adversary controlling all legal, technical, and coercive institutions. Such total capture lies outside any credible technical assurance claim and is addressed through political resilience, not an assumed secure component.

14.2 Comparative experimental design

The core evaluation should compare four conditions where ethically and legally permissible: competent human administration; improved conventional digital administration; a single-model AI workflow; and the proposed plural, constrained, auditable workflow. The first two conditions prevent attributing to AI improvements that actually arise from digitization, better staffing, or simpler process reform. Resource budgets, case difficulty, service obligations, and evaluation access should be made comparable or differences explicitly estimated.

Experiment A: retrospective and shadow administration. Use historical records supplemented by independently adjudicated cases and controlled synthetic cases. Separate development and evaluation by time, geography, or institution to reduce leakage. The shadow system makes no binding decisions. Measure legal and factual error, abstention, unexplained inconsistency, subgroup harms, and cost. Historical decisions are comparison observations, not automatically correct labels. Expert adjudicators should be blinded where feasible; disagreement should be retained and analyzed.

Experiment B: adversarial integrity testing. In an isolated test environment, seed unauthorized rule changes, misleading documents, poisoned inputs, compromised credentials, false attestations, and common model dependencies. Compare single-model and separated-model architectures. Report detection rates, missed attacks, false alarms, investigative workload, and whether the attack reaches execution. Keep a protected challenge set and publish enough detail for independent replication without compromising live systems.

Experiment C: low-risk field evaluation. Randomize eligible administrative units or workflows, using cluster designs where staff or users would otherwise contaminate conditions. A stepped rollout may be used when justified, but timing and secular trends must be addressed. Field experiments must not randomly remove legally required protections or expose people to unvalidated adverse decisions. Affected participants need appropriate information, lawful participation arrangements, and access to remedies.

Experiment D: auditability and appeal usability. Give independent auditors comparable case samples and measure their ability to reconstruct decisions, detect material failures, and produce actionable findings. Separately test notices and appeal channels with people reflecting relevant languages, disabilities, literacy levels, and access constraints. Assess comprehension and successful correction, not simply satisfaction or stated trust.

Experiment E: resilience and institutional capture exercises. Simulate supplier exit, network partition, loss of credentials, audit-budget pressure, politically motivated model updates, and simultaneous compromise of specified functions. Test service continuity and the ability of independent authorities to suspend or correct the system. A technically successful failover that leaves people unable to access essential services does not pass.

Experiment F: policy-learning evaluation. Compare policies with prespecified evidence-review obligations against otherwise comparable policies using existing review arrangements. Measure timely review, quality of causal evidence, publication of reasons, implemented corrections, and substantive outcomes. Do not equate the number of legislative changes with improvement; retention can be justified, and frequent change can create instability.

14.3 Hypotheses and illustrative decision thresholds

The following hypotheses are proposed for preregistration. The numerical values are illustrative research targets for suitable low-risk settings, not observed results, universal standards, or permissions to violate rights. An actual protocol must justify thresholds before seeing the evaluation data, power the study for relevant groups, and specify handling of missingness, multiplicity, clustering, and sequential stopping.

H1: administrative benefit. Relative to improved conventional digital administration, the full architecture reduces median end-to-end processing time by at least 20%, while the upper one-sided 95% confidence bound on its increase in independently adjudicated material administrative error is below one percentage point. Total cost per correctly completed case must not increase. This composite hypothesis fails if any component fails. Rights violations remain subject to separate stopping rules regardless of this error margin.

H2: resistance to manipulation. On a protected, preregistered attack distribution, the separated-model architecture reduces the rate of successful seeded unauthorized actions by at least 30% relative to the single-model workflow, while increasing false escalations by no more than five percentage points. Failure includes common-mode compromises that negate the anticipated advantage. Results generalize only to the assessed adversary classes.

H3: provenance utility. Independent auditors reconstruct at least 99% of randomly sampled consequential actions within a defined service-specific investigation window, and reconstruct materially more cases than under the best comparison workflow. Reconstruction requires independent confirmation of legal authority, actual versions, material inputs, execution, and review status; the mere presence of a log entry does not count.

H4: bounded impartiality. The deployed workflow remains within all preregistered, legally justified disparity bounds on held-out and prospective data, using simultaneous uncertainty control. Where sample support is inadequate, the claim is unestablished rather than accepted. A material prohibited discriminatory pattern falsifies the applicable certificate even when average performance improves.

H5: effective contestability. People receiving the proposed notice and assistance are more likely to identify a material error and obtain a correct, timely remedy than people using the best existing procedure, with no increase in unrepaired serious harm. The protocol must prespecify the minimum worthwhile effect and subgroup noninferiority margins after a feasibility study. A fall in complaint volume alone cannot confirm the hypothesis.

H6: reversible delegation. Under the prespecified loss of any one designated provider or oversight component, essential service continuity and accurate recovery meet the domain's authorized objectives without emergency relaxation of fundamental protections. The hypothesis fails if fallback exists only nominally, if protected records cannot be recovered, or if restoration depends on the unavailable supplier.

H7: democratic review. Evidence-triggered policies receive more timely, publicly reasoned reconsideration following qualifying adverse evidence than comparable policies under existing arrangements, without increased rights violations or inappropriate policy volatility. The preregistered effect size must reflect institutional feasibility. A legislature's reasoned refusal to change policy is not automatically coded as failure.

14.4 Benchmarks and analysis

Dimension	Primary measure	Interpretation safeguard
Administrative correctness	Independently adjudicated material errors per completed case	Also count withdrawals, uncompleted applications, and omitted cases.
Access and distribution	False denial, completion, delay, and burden by relevant group	Report uncertainty, intersectional results, and missing group coverage.
Efficiency	End-to-end time and total cost per correct outcome	Include model, audit, appeal, security, staff, infrastructure, and fallback costs.
Integrity	Successful attacks and undetected unauthorized changes	State attacker capability and protected-test coverage.
Auditability	Reconstructable decisions and time to actionable findings	Reconcile records with actual external execution.
Remedy	Time to substantive review and time to completed repair	Audit unappealed cases and reasons people do not challenge.
Resilience	Service continuity and recovery under specified failures	Measure citizen access, not only server uptime.
Democratic accountability	Review-trigger compliance and quality of published reasons	Supplement counts with independent legal and deliberative assessment.

Table 4. Minimum evaluation dashboard

Analyses should publish effect estimates with uncertainty, distinguish confirmatory from exploratory findings, and report negative results. Randomized evaluations should use intention-to-treat estimates where appropriate. Nonrandomized evaluations must explain identification assumptions and test plausible alternative explanations. Distributional effects, displacement of corruption, and longer-term behavioural adaptation require follow-up beyond initial deployment. Cost-effectiveness claims must include remedies and safeguards as operating costs.

Stopping rules should identify severe rights violations, unauthorized consequential execution, systemic data compromise, loss of independent audit access, and failure of essential continuity as grounds for immediate pause and investigation. An observed zero failure count in a finite test is not proof of zero risk. Independent evaluators should control final test materials, retain the ability to publish adverse findings, and disclose conflicts of interest.

15. Limitations and future research

The first limitation is constitutional variation. Parliamentary, presidential, federal, and unitary systems allocate administrative and judicial authority differently. A single institutional diagram cannot lawfully substitute for jurisdiction-specific analysis. The framework supplies design requirements; implementing legislation and constitutional compatibility require separate public deliberation and legal work.

The second limitation is epistemic. Public purposes are contested, outcomes are difficult to measure, causal inference is assumption-dependent, and models may fail precisely when conditions change. The formal architecture does not remove these

difficulties. It creates obligations to reveal them and to act within justified limits.

The third limitation is institutional capacity. Auditors, judges, public data custodians, researchers, and service staff require stable resources. Countries or agencies with weak administrative capacity may struggle most with an elaborate verification regime. A simpler accountable service may be preferable to a complex system whose safeguards cannot be operated. Lower-resource settings should not be pressured to accept reduced rights as the price of modernization.

The fourth limitation is political economy. Suppliers and public institutions may resist portability, disclosure, and external scrutiny. Incumbents may support efficiency while opposing genuine contestability. The architecture's adoption cannot be explained solely by technical merit.

Research must investigate coalitions, incentives, public-sector labour, procurement power, and the distribution of transition costs.

The fifth limitation is external validity. A successful pilot in a narrow domain cannot establish nationwide legitimacy or safety. Simulations cannot reproduce all political adaptation, informal coercion, or emergent social responses. The proposed hypotheses test components and institutional mechanisms; no single experiment can validate a whole political philosophy.

Future research should pursue machine-readable legal representations that preserve explicit uncertainty; formal methods for narrow procedural invariants; empirical measures of common model dependence; privacy-preserving public auditing; cross-language and accessibility evaluation; reliable evidence of actual model execution; and practical appeal systems for nonexperts. Comparative constitutional research should examine delegation limits and institutional fit. Longitudinal studies should track corruption displacement, civic participation, institutional deskilling, and policy learning. Research on infrastructure concentration should evaluate whether public and cooperative alternatives improve democratic control after accounting for cost and operational capacity.

The political philosophy also requires deeper inquiry. How should future generations be represented in computational policy evaluation? How can minority communities challenge categories that aggregate metrics make invisible? When does preserving human discretion advance dignity, and when does it preserve domination? Who receives standing to contest a model whose effects are collective rather than individual? These are constitutional questions that a technical architecture must make governable rather than pretend to settle.

16. Declaration of Techno-Democracy

Techno-Democracy begins from a democratic conviction: public power belongs to the people, and the instruments through which it operates must remain answerable to them. Intelligence does not confer sovereignty. Expertise does not confer an unrestricted right to rule. Computational capacity does not diminish the standing of the person subject to a decision.

We therefore affirm the following principles as the normative declaration of this proposal.

The people remain sovereign. Citizens retain the authority to constitute government, elect representatives, contest public purposes, and alter or withdraw computational delegation through constitutional procedures. No model, corporation, technical profession, or security institution may acquire a private title to public power.

The constitution governs the machine. Public algorithms must operate under explicit authority and enforceable limits. Where code and lawful constitutional interpretation diverge, the code must yield. Fundamental rights protect every person to whom public power applies.

No claim of neutrality is exempt from examination. AI systems must submit to competing analysis, independent evidence, adversarial testing, and public challenge. Their authority depends on justified scope, continuing evaluation, and effective correction. A certificate must never become an excuse to stop asking questions.

Public action must carry public responsibility. Every consequential computational decision must have an accountable institution, a traceable basis, an intelligible explanation, and a route to remedy. Neither technical complexity nor commercial secrecy may extinguish the right to challenge the exercise of governmental power.

Power over the infrastructure must be divided. Those who select data, develop models, approve deployment, execute decisions, audit performance, and adjudicate disputes must not become one unaccountable authority. Independence must be real in appointments, funding, credentials, infrastructure, and the capacity to dissent.

Efficiency must serve freedom and dignity. A faster injustice is not progress. A cheaper service that excludes those most in need has failed its public purpose. The measure of administrative improvement includes the treatment of people who are difficult to classify, poorly represented in data, or unable to navigate digital systems.

Democracy must remain able to learn. Policies should state what they are intended to accomplish, disclose their assumptions, and face reconsideration when credible evidence reveals failure. Evidence informs public judgment; it does not relieve elected institutions of the duty to give reasons for the values they choose.

Delegation must remain reversible. A people who cannot practically replace a system cannot meaningfully govern it. Public institutions must preserve the knowledge, alternatives, and human capabilities needed to correct, suspend, or retire their computational instruments.

The purpose of Techno-Democracy is to reduce arbitrary, opaque, and corruptible discretion by bringing governmental processes within transparent, verifiable, and accountable systems. Its ambition is substantial: to make public administration more dependable without making the public less sovereign. Its legitimacy would be measured not by how much authority machines accumulate, but by how effectively people can understand, direct, contest, and correct the exercise of power.

AI governs processes; people govern values. That principle assigns the machine a demanding public function and a permanent constitutional limit. The promise is a democracy capable of using intelligence without surrendering authorship of its common life.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety [Preprint]. arXiv. <https://arxiv.org/abs/1606.06565>
2. Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989. <https://doi.org/10.1177/1461444816676645>
3. Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
4. Banerjee, A., Duflo, E., Imbert, C., Mathew, S., & Pande, R. (2020). E-governance, accountability, and leakage in public programs: Experimental evidence from a financial management reform in India. *American Economic Journal: Applied Economics*, 12(4), 39–72. <https://doi.org/10.1257/app.20180302>
5. Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. *Proceedings of the 29th International Conference on Machine Learning*. Author manuscript: <https://arxiv.org/abs/1206.6389>
6. Bloom, P., & Sancino, A. (2019). *Disruptive democracy: The clash between techno-populism and techno-democracy*. SAGE. <https://uk.sagepub.com/en-gb/eur/disruptive-democracy/book263534>
7. Blum, C., & Zuber, C. I. (2016). Liquid democracy: Potentials, problems, and perspectives. *Journal of Political Philosophy*, 24(2), 162–182. <https://doi.org/10.1111/jopp.12065>
8. Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal*, 13(4), 447–468. <https://doi.org/10.1111/j.1468-0386.2007.00378.x>
9. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 77–91. <https://proceedings.mlr.press/v81/buolamwini18a.html>
10. Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297. <https://doi.org/10.1257/aer.20190623>
11. Castro, M., & Liskov, B. (1999). Practical Byzantine fault tolerance. *Proceedings of the Third Symposium on Operating Systems Design and Implementation*. USENIX Association. <https://www.usenix.org/conference/osdi-99/practical-byzantine-fault-tolerance>
12. Citron, D. K. (2008). Technological due process. *Washington University Law Review*, 85(6), 1249–1313. <https://journals.library.wustl.edu/lawreview/article/id/6697/>
13. Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press. <https://doi.org/10.12987/9780300252392>
14. Crosby, S. A., & Wallach, D. S. (2009). Efficient data structures for tamper-evident logging. *Proceedings of the 18th USENIX Security Symposium*. USENIX Association. <https://www.usenix.org/conference/usenixsecurity09/technical-sessions/presentation/efficient-data-structures-tamper-evident>
15. Dahl, R. A. (1989). *Democracy and its critics*. Yale University Press. <https://yalebooks.yale.edu/book/9780300049381/democracy-and-its-critics/>
16. De Filippi, P., & Wright, A. (2018). *Blockchain and the law: The rule of code*. Harvard University Press. <https://doi.org/10.2307/j.ctv2867sp>
17. Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2006). New public management is dead—Long live digital-era governance. *Journal of Public Administration Research and Theory*, 16(3), 467–494. <https://doi.org/10.1093/jopart/mui057>
18. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407. <https://doi.org/10.1561/04000000042>
19. Feldstein, S. (2021). *The rise of digital repression: How technology is reshaping power, politics, and resistance*. Oxford University Press. <https://academic.oup.com/book/39418>
20. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
21. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection [Preprint]. arXiv. <https://arxiv.org/abs/2302.12173>
22. Haber, S., & Stornetta, W. S. (1991). How to time-stamp a digital document. *Journal of Cryptology*, 3, 99–111. <https://doi.org/10.1007/BF00196791>
23. Hernán, M. A., & Robins, J. M. (2020). Causal inference: What if. Chapman & Hall/CRC. Author-hosted text: https://www.hsph.harvard.edu/miguel-hernan/wp-content/uploads/sites/1268/2024/01/hernanrobins_WhatIf_2jan24.pdf
24. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>
25. Hildebrandt, M. (2018). Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A*, 376(2128), Article 20170355. <https://doi.org/10.1098/rsta.2017.0355>
26. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. 8th Innovations in Theoretical Computer Science Conference, LIPIcs 67, Article 43. <https://doi.org/10.4230/LIPIcs.ITCS.2017.43>
27. Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 633–705. <https://pennlawreview.com/2017/02/23/accountable-algorithms/>
28. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems*, 30. <https://papers.nips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html>
29. Lipsky, M. (2010). *Street-level bureaucracy: Dilemmas of the individual in public services* (30th anniversary expanded ed.). Russell Sage Foundation. <https://doi.org/10.7758/9781610446631>
30. Medina, E. (2011). *Cybernetic revolutionaries: Technology and politics in Allende's Chile*. MIT Press. <https://mitpress.mit.edu/9780262016490/cybernetic-revolutionaries/>
31. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
32. Olken, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *Journal of Political Economy*, 115(2), 200–249. <https://doi.org/10.1086/517935>
33. Ostrom, E. (2010). Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review*, 100(3), 641–672. <https://doi.org/10.1257/aer.100.3.641>

34. Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5–14. <https://doi.org/10.1007/s10676-017-9430-8>
35. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. <https://doi.org/10.1145/3351095.3372873>
36. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
37. Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. <https://doi.org/10.1145/3287560.3287598>
38. Shleifer, A., & Vishny, R. W. (1993). Corruption. *The Quarterly Journal of Economics*, 108(3), 599–617. <https://doi.org/10.2307/2118402>
39. Tan, J. Z., Langenkamp, M., Weichselbraun, A., Brody, A., & Korpas, L. (2024). The constitutions of Web3 [Preprint]. *arXiv*. <https://arxiv.org/abs/2403.00081>
40. Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration*, 42(7), 596–615. <https://doi.org/10.1080/01900692.2018.1498103>
41. Wüst, K., & Gervais, A. (2018). Do you need a blockchain? 2018 Crypto Valley Conference on Blockchain Technology, 45–54. <https://doi.org/10.1109/CVCBT.2018.00011>
42. Yeung, K. (2018). Algorithmic regulation: A critical interrogation. *Regulation & Governance*, 12(4), 505–523. <https://doi.org/10.1111/rego.12158>