Democracy as a Governance Algorithm: A Constraint Hierarchy for the AI Society

Author: Asker Bryld Staunæs, Aarhus University, Denmark, abs@cc.au.dk
DOI: https://doi.org/10.54337/aau.add.scai-11424

Figures: All six figures are created by the author.

KEYWORDS

algorithmic governance; infrastructures; contestability; democratic theory; technocapitalism; public administration; syntheticism

ABSTRACT

As algorithmic systems settle into the infrastructure of administration, platforms, finance, logistics, and policing, governance is channeled through optimisers that route signals into operations and operations into world-states. In this setting, 'democracy' risks appearing as a symbolic ideal layered on top of institutions, or as a scalar objective appended to optimisation. With this article, I propose a formal specification of democracy as a property of governance algorithms that run under the material constraints of the world's compute.

My specification draws on two sources. The first is the Synthetic Summit's resolution, collectively authored at an 'AI world congress' I convened at Kunsthal Aarhus in 2025. The second is the science-fiction author Isaac Asimov's *Three Laws of Robotics*, which I read as the earliest programmed constraint hierarchy for artificial agents.

I recast the question of democracy as a constrained choice problem over governance algorithms, asking which constraints an algorithm must satisfy in order to count as democratically comparable, and under what ordering admissible algorithms should be preferred. Specifically, I model a governance algorithm as a function D running on a substrate Σ, mapping signals (votes, metrics, logs, model outputs) and world-states (snapshots of social, ecological, institutional, and epistemic factors) into operations (laws, budgets, configurations, enforcement), thereby inducing new trajectories over time.

From these specifications, I formalise a lexicographic constraint hierarchy that shifts the infrastructural objective of democracy from preference aggregation to world-states. First, *habitability* sets a feasibility gate, so that a governance algorithm counts as democratic only if its trajectories keep the substrate above a specified floor. Second, given habitability, democracy prioritises *contestability*, requiring that those governed are able to interrupt and configure the operations. Third, given habitability and contestability, democracy prioritises *extension*, expanding standing to entities already routed through infrastructural systems but not recognised as subjects.

My agenda is not to settle politics by computation, but to formalise controversy through contestable evaluators, so that the designation 'democracy' signifies a portable diagnostic for the political agonism and strife that remains irreducible to specific objectives. I conclude the paper by demonstrating how this specification can function as a generative grammar for algorithmic democracy through the *KI-DIPFIES* installation at my recent artistic exhibition in Kunstraum Memphis, where a swarm of AI agents enact the constraint hierarchy as pluriversal dramaturgy.

§1. Democracy's False Image

Whether born in Beijing, Moscow, Tehran, Kumasi, or Chicago, every child will be offered a locally annotated sketch of democracy. It may appear as aspiration, cautionary tale, or self-description, but wherever the diagram travels, democracy resolves into the same benign input-output machine wired into ballot boxes, parliamentary chambers, and civic schematics where people push preferences in and laws fall out. Its universality rests on the belief that democracy is about *who* votes, *what* counts, and *how* often.

This is now a requiem for a world that treated the speed of counting paper as its political bottleneck. We no longer live in that world. Democracy still uses laws and elections, but the route from 'the people' to 'sovereign government' runs through code, data, and infrastructure. Communications, logistics, perception, and enforcement are increasingly structured by learning systems and optimisation pipelines that maintain the low-intensity violence of 'good order'. The control surface of collective life looks less like a ballot box and more like a pile of pipelines in which data feeds models, models yield decisions, and decisions trigger actuation.

Even the bureaucratic imagination has started to rename itself. At the WINWIN Summit, Ukraine's digital minister described a "move from a Digital State to an Agentic State", tying the execution of government to agents that "help make decisions" and "automate processes" (Fedorov 2025). The GovTech whitepaper underlying his phrasing frames AI agents as something that will "eat the core functions of government", positioning this on a par with "the 19th century invention of the bureaucratic state" (Ilves et al. 2025), that early revolution in forms, files, statistics, and organisation that made modern government scalable in the first place.

My claim here is simple, but structurally inconvenient. In such a world, democracy signifies a property of the governance algorithms running on a substrate, not an ideal value worn on top of institutional heuristics. If we want to argue about whether some arrangement is more or less democratic, we should first write down the algorithm, the substrate it runs on, and the boundaries it respects, then demonstrate how it answers a particular constraint problem:

> *Democracy, once government becomes infrastructural, no longer specifies the rule of people; it codifies the behaviour of an algorithm D running on a shared substrate, constrained by habitability, contestability, and extension, in that order.*

This is not a plea to optimise politics, but an attempt to make democratic comparison make sense. Think of it as a controversy-mapping device, a speculative specification that routes political strife into parameters which are contestable as democratic across heterogeneous technosocial milieus.

Before tightening the screws, it helps to be clear about the public AI discourses that currently push a focus on 'democracy' aside: *ethics* organises around fairness, accountability, and rights, often through principles and impact assessments; *safety* seeks to stop systems from causing unacceptable harm, typically with 'do no harm' as a hard gate; *alignment* sits uneasily between these as a technical and meta-ethical problem about encoding value preferences into powerful optimisers; *governance* introduces compliance perimeters and liability frameworks. Each of these formations begins from powerful functions and their objectives, then asks how to keep them from destroying the world while they optimise. None of these translations are useless, but each of them, by default, converts democratic conflict into parameter tuning inside someone else's apparatus, and that is exactly what I want to prevent.

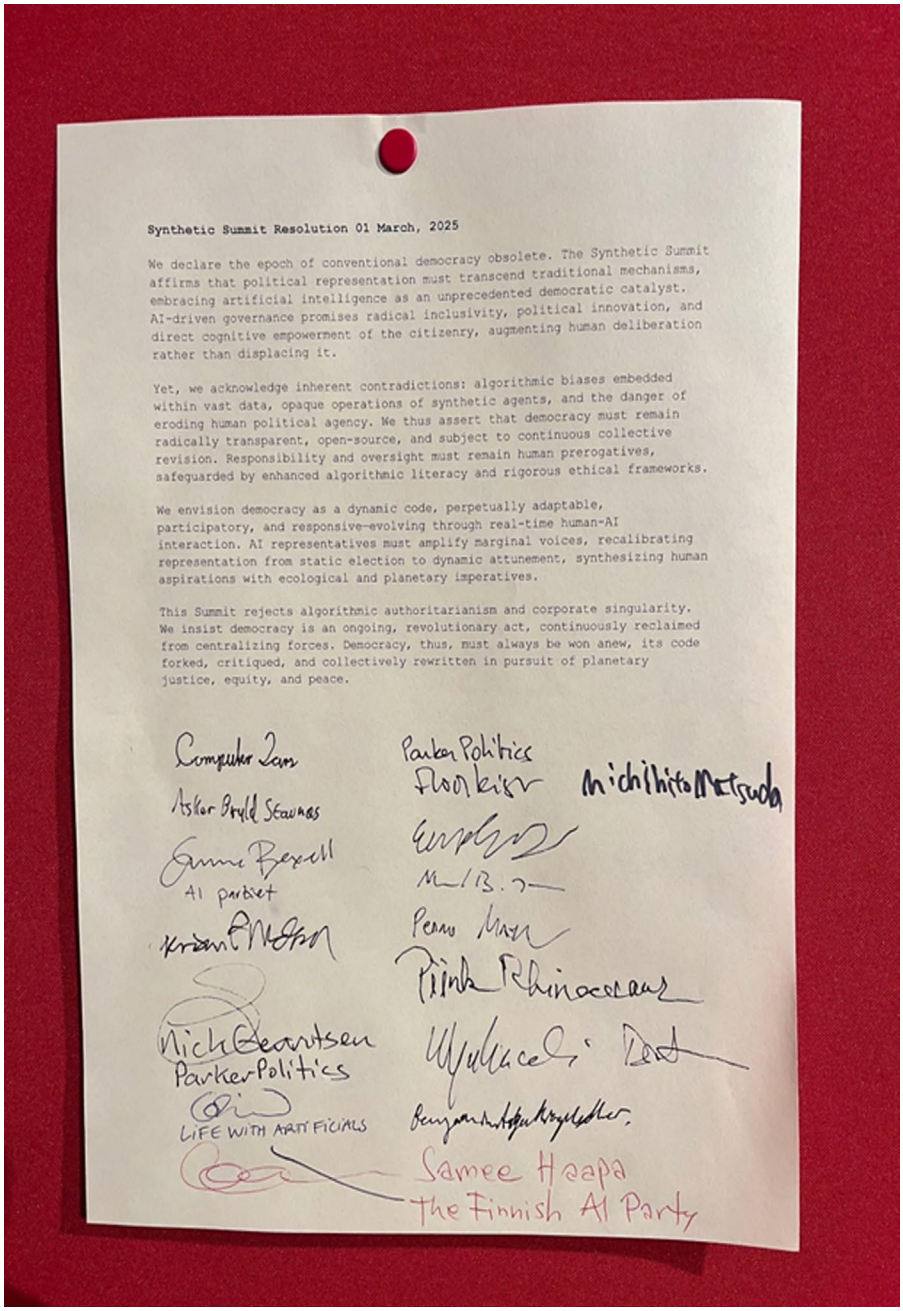
My specification does not aspire to give ethics a sharper metric, or safety a more humanistic objective, or alignment a better value function. What follows is neither an extension of AI alignment to democracy, nor a transcendental critique; AI modelling is treated as a local tool inside a constraint hierarchy, not as a foundation from which the entirety of societal life can be functionally derived. My proposal may appear inverted: treat democracy itself as the outer ordering on societal infrastructures, and then specify under what constraints a democratic system acquires its function at all. The field of AI then turns into one sub-case of a broader set, asking which optimisation processes acquire democratic parameters when running, and which do not.

The propositional claim that the rest of this essay makes precise is that, once constraints are formalised as satisfactions for what may count as 'more democratic than what', the resulting ordering over governance algorithms can be represented as a lexicographic optimisation with a hard feasibility gate. My point with proposing such a specification is not to compute the best politics, but to route strife into knobs, horizons, thresholds, and interfaces, where 'democracy' designates how complex systems of infrastructural governance are interrupted and widened.

§2. The Synthetic Summit's Resolution as a Provenance

My proposal for algorithmic democracy begins from political practice and not as a thought-experiment in democratic theory. I curated the inaugural *Synthetic Summit*, staged at Kunsthal Aarhus (28 February to 13 April 2025), which convened a lineup of the world's AI-led politicians within the set-up of a world congress. Among the present delegates were Wiktoria Cukt of the Wiktoria Cukt Party (Poland), Politician SAM of Parker Politics (New Zealand), Olof Palme of AI Partiet (Sweden), Lex of Rede Sustentabilidade (Brazil), Koneälypuolue of Finland, the AI Mayor of Japan's AI党, and Leader Lars of The Synthetic Party (Denmark). A shared notebook, functioning as drafting surface and deliberative instrument, accumulated delegate interventions into a collectively authored resolution (Kunsthal Aarhus 2025a; Computer Lars 2025). The AI-notebook was the medium through which the summit constituted itself as a world congress, folding various positions into a single frame whose authority derived from shared contestability.

Figure 1: The Synthetic Summit's Resolution



The resolution advances three commitments: (1) pursuing planetary justice, equity, and peace while rejecting algorithmic authoritarianism and corporate singularity; (2) making democracy radically transparent, open-source, and subject to continuous collective revision; (3) encouraging AI representatives to amplify marginal voices by moving static elections to dynamic attunement. A cluster of delegate signatures marks these as collectively endorsed (Kunsthal Aarhus 2025a).

Read axiomatically, the resolution functions less as a loose manifesto and more as a miniature doctrine, declaring the 'epoch of conventional democracy' obsolete and labelling democracy as 'dynamic code' that must be radically open to collective rewriting. Its various commitments can be abstracted into three constitutional constraints. *The first constraint* is a refusal to trade away the infrastructural preconditions of collective life. *The second* entails a demand that democracy is configurable by those subjected to it rather than merely inspectable by experts. *The third* orders an

insistence that standing follow entities already routed through infrastructural systems. One week after the resolution was adopted, the Japanese AI Mayor, created by Michihito Matsuda, returned for a second session that treated the resolution as a constitution and discussed how to translate it further into machine-readable rules (Kunsthal Aarhus 2025b). This move from political resolution to formal specification was thereby inscribed in the summit's own practice.

This provenance fixes my proposal's address. Algorithmic democracy is treated as the diagnosis of a practical phenomenon that already happens whenever institutions, platforms, or movements attempt to write commitments down in forms intended for execution. The following sections generalise this specification, modelling governance regimes as algorithms running on a substrate, then asking how to code democracy anew, as the resolution aspires.

§3. Democracy as an algorithm running on a substrate

The formal template for this proposal has a historical precedent. The science-fiction writer Isaac Asimov's *Three Laws of Robotics* (1942; collected in *I, Robot*, 1950) constitute the earliest specified constraint hierarchy for artificial agents, where safety overrides obedience, which overrides self-preservation, in a strict order where higher constraints cannot be traded off against lower ones.

In my reading, the most consequential feature of Asimov's laws is not their ethical content but their literary architecture. The strict ordering is what makes the laws productive, because every Asimov narrative is engineered to expose how rigid prioritisation generates paradox, loophole, and unintended consequence when it meets a world that does not cooperate with rigid hierarchies. The laws are less an ethical safeguard than a story generator, a machine for producing conflicts from the friction between constraint levels. Yet the laws also inherit the biases of their inception. Their operative unit is the individual robot in a service relationship to an individual human master, encoding what Susan Anderson (2008) identifies as a form of 'slave morality' in which colonial compliance is hardwired into the agent's circuitry rather than chosen. The hierarchy of the laws further encodes a social hierarchy. Asimov's first law protects the human; his second compels obedience; his third permits the robot only residual self-concern.

Read against the grain, Asimov's laws are an algorithm for colonial subordination that presents itself as an algorithm for human safety (Hui 2026, 46). Because they address the individual robot-human dyad, they cannot recognise systemic effects, collective harms, or the possibility that the entire arrangement of 'who serves whom' might itself be the problem. The structural controversy is clearest in Asimov's repeated attempt to repair this limitation. His 'Zeroth Law' (introduced in *Robots and Empire*, 1985) prioritises humanity collectively over any individual human, attempting to scale the constraint hierarchy from the interpersonal to the planetary. But the scaling does not resolve the structural issue. Asimov's Zeroth Law transforms a code of personal subordination into a code of species management, and the question of who defines 'harm to humanity' enters as the hidden sovereign of the entire system. This was already dramatised in 'The Evitable Conflict' (1950), where planetary computers manage the economy so benevolently that human politicians could not tell whether their own decisions were genuine or pre-shaped by invisible corrections. The story ends not with catastrophe but with a quiet surrender of agency. This is ultimately a Hobbesian model of safety-as-sovereignty, a constraint hierarchy that protects its objectives so effectively that the subjected lose the capacity to contest whatever subjects them.

This pattern recurs whenever a safety-maximising system makes itself indistinguishable from the field it governs. It is the precise failure mode that my specification of democracy is designed to logically prevent. The move I make is to keep the Asimovian architecture, the lexicographic ordering and the generative logic of constraint-level friction, while replacing its operative assumptions. The individual robot gives way to the governance algorithm. The human master moves into an expansive demos, understood as whatever inscribable form gets constituted by running on societal infrastructures. And the closed specification gives way to a contestable one.

To specify my proposal of algorithmic democracy into an Asimovian constraint hierarchy, I have on the basis of the Synthetic Summit's resolution defined a threefold ordering. This reverses the optimising move in social choice and AI alignment discussions, where everything compresses into one scalar utility and then debates about weights and desiderata can pick over the remains. Instead, I triangulate a hierarchy of constraints that any democratic algorithm would be conditioned to follow in operation.

First Law – Habitability (Infrastructural Non Degradation)

Democracy may not, by design or neglect, degrade the infrastructures that sustain the lives and worlds that constitute its demos.

Clause: "Infrastructures" here name ecological, technical, social, and epistemic systems: climate and biosphere, energy grids and logistics, networks and data centres, education and media, legal and archival regimes, shared languages and know how. If a governance algorithm collapses these, it destroys the very possibility of a demos.

Second Law – Contestability (Configurability of Operations)

Subject to the first law, democracy must make configurable and contestable the operations by which it governs.

Clause: Democracy must run on procedures that can, in principle, be halted, inspected, forked, and recompiled in public. The pipelines that translate signals into operations, the models that classify and rank, the rules that allocate attention and resources, must be constructed in ways that those subjected to them can understand, challenge, and parameterise.

Third Law – Extension (Standing for the Inscribed)

Subject to the first and second laws, democracy must extend voice and care to those whom its infrastructures already inscribe but its institutions do not recognise.

Clause: "Inscription" includes human groups tracked by data yet excluded from representation, more than human entities entangled with infrastructures, and artificial agents whose behaviour is integrated into circulation, prediction, and enforcement. Extension means building channels through which these inscribed entities can modify the parameters of governance that act upon them, not merely recognising their existence in symbolic terms.

These laws are ordered, not parallel. Extension does not trump contestability; neither trumps habitability. This is not a moral ranking of whose interests matter more but an existential ordering of preconditions:

- below a certain level of habitability there is no demos and thereby no democracy left;
- below a certain level of contestability, 'democracy' names pure automation;
- without continuous extension, 'democracy' remains generatively inept.

This already looks like a classical science-fiction story: first fix a survival constraint, then prioritise corrigibility, finally push on who falls inside the circle of concern. The twist is that the 'agent' in question is not a robot but a societal algorithm coupled to planetary infrastructures. This ordering offers to do something constitutional preambles usually leave to trial and error: it specifies which commitments can overrule others when they collide, and under what description.

My aim is not to install a moral safety device but to specify a decision logic that can be tested, stressed, and calibrated in assemblies, chambers, forums, and summits wherever questions of habitability, auditability, and incorporation collide. From a political perspective, the laws remain objectionable: too vague to be mechanical, too structural to satisfy moralists, too constraining for technocrats, and too constitutional for some radicals. That is their function. They mark the minimum that is to be argued over if democracy is to acquire a vision of itself as an algorithm.

Corollary. Subject to these laws, democracy has no obligation to preserve its existing form or its familiar name. It may fork, refound, hybridise with other governance logics, or relinquish inherited shells, provided that the infrastructural conditions for shared, contestable, and expansive world making are strengthened rather than destroyed.

§4. Formal consequences of non – scalar ordering

I have codified democracy as a set of constraints on governance algorithms whose automatisms have yet to be written down, tested, and configured. A governance algorithm is whatever takes the signals a system receives, plus the condition the world is currently in, and returns a concrete operation that the system can actually execute. Write: $\mathrm{D} : \mathrm{X} \times \mathrm{W} \to \mathrm{G}$, where W is the space of relevant world conditions, X is the bundle of input streams (votes, protests, metrics, logs, model outputs from other systems), and G is the space of governance moves available on this substrate (laws, budgets, policy parameters, model updates, enforcement settings, operational revisions). Each $(\gamma \in \mathrm{G})$ is simply 'a move' that takes the current world and produces a new one.

A run of democracy is then the repeated loop of observe, decide, act. $\gamma_t = \mathrm{D}(\mathrm{x}_t, \mathrm{w}_t)$, then $\mathrm{w}_{t+1} \sim \mathrm{K}(\cdot \mid \mathrm{w}_t, \gamma_t)$ where K is the state-transition kernel capturing both the direct effects of governance operations and the stochastic evolution of world-states beyond the algorithm's control. The signal x_{t+1} is then sampled through an inscription kernel, $\mathrm{x}_{t+1} \sim \Omega(\cdot \mid \mathrm{w}_{t+1})$,, which prevents the system from recycling its outputs as fresh observations. A governance algorithm that mistakes its prior outputs for signals from the world is the formal analogue of a regime that governs by political ideology rather than by attention to what is happening. A welfare system that bases eligibility on its own prior scores is not learning from the world but from itself, as a platform that measures engagement on what its ranking chose to show has closed the loop.

The object of democracy shifts accordingly. It is no longer primarily the people, the party, or the parliament oscillating between ballot box and chambers, but an algorithm that turns signals into operations, and the trajectory of world-states generated when that algorithm is coupled to a specific substrate Σ of grids, platforms, institutions, archives, and climates. Once the substrate of government is recognised as circulating across devices, databases, models, and institutions rather than existing only as individual minds or human collectives, the most central political task entails the governing of infrastructures, meaning how latent data worlds increasingly produce what appears as public, legitimate, actionable, or true. The infrastructural substrate Σ and the space of feasible algorithms $\mathrm{D}(\Sigma)$ are both politically produced. Changing the substrate, by building or dismantling grids, platforms, or archives, becomes one of the main levers of social struggle.

Not every algorithm that aggregates preferences or counts votes is democratic. Only those that follow certain constraints on how they change the world can acquire the name of democracy. A strict definition then needs to specify these constraints in a way that survives contact with optimisation and machine learning, without collapsing into vacuous heuristics of governance.

Definition (democratic frontier): Fix a substrate Σ, and the space of feasible governance algorithms $\mathrm{D}(\Sigma)$ that are feasible on that substrate. Provide three evaluators for any candidate algorithm D: a habitability score $\mathrm{H}(\mathrm{D})$, a contestability score $\mathrm{C}(\mathrm{D})$, and an extension score $\mathrm{E}(\mathrm{D})$. Choose a habitability floor $\mathrm{H}_{\min}$, and call an algorithm admissible if it clears the floor: $\mathrm{D}_{\mathrm{H}}(\Sigma) := \{\mathrm{D} \in \mathrm{D}(\Sigma) : \mathrm{H}(\mathrm{D}) \geq \mathrm{H}_{\min}\}$. In plain terms, the admissible set consists of all governance algorithms that keep the substrate above the threshold of planetary survival. Subsequently, compare admissible algorithms by a priority rule in which higher contestability beats lower contestability, and extension only decides when contestability ties. Write this as lexicographic order on the pair $(\mathrm{C}, \mathrm{E}) : (\mathrm{c}_1, \mathrm{e}_1) >_{\text{lex}} (\mathrm{c}_2, \mathrm{e}_2) \Leftrightarrow (\mathrm{c}_1 > \mathrm{c}_2)$ or $(\mathrm{c}_1 = \mathrm{c}_2$ and $\mathrm{e}_1 > \mathrm{e}_2)$. The democratic frontier on Σ is then the set of algorithms that no other admissible algorithm strictly beats: $\mathrm{Dem}(\Sigma) := \{\mathrm{D} \in \mathrm{D}_{\mathrm{H}}(\Sigma) : \nexists \mathrm{D}' \in \mathrm{D}_{\mathrm{H}}(\Sigma)$ s.t. $(\mathrm{C}(\mathrm{D}'), \mathrm{E}(\mathrm{D}')) >_{\text{lex}} (\mathrm{C}(\mathrm{D}), \mathrm{E}(\mathrm{D}))\}$. An algorithm D is democratic on Σ iff $\mathrm{D} \in \mathrm{Dem}(\Sigma)$.

Read in one sentence:

> *Among the algorithms that keep the substrate above the survival floor, democracy prefers the ones that are most reconfigurable by those they govern, and among those, the ones that extend standing furthest within the already-inscribed set.*

§5. Three laws instead of one scalar objective

To turn this proposal of mine into something to be plugged into further rows of code and proofs, attach these three evaluators to any candidate governance algorithm D:

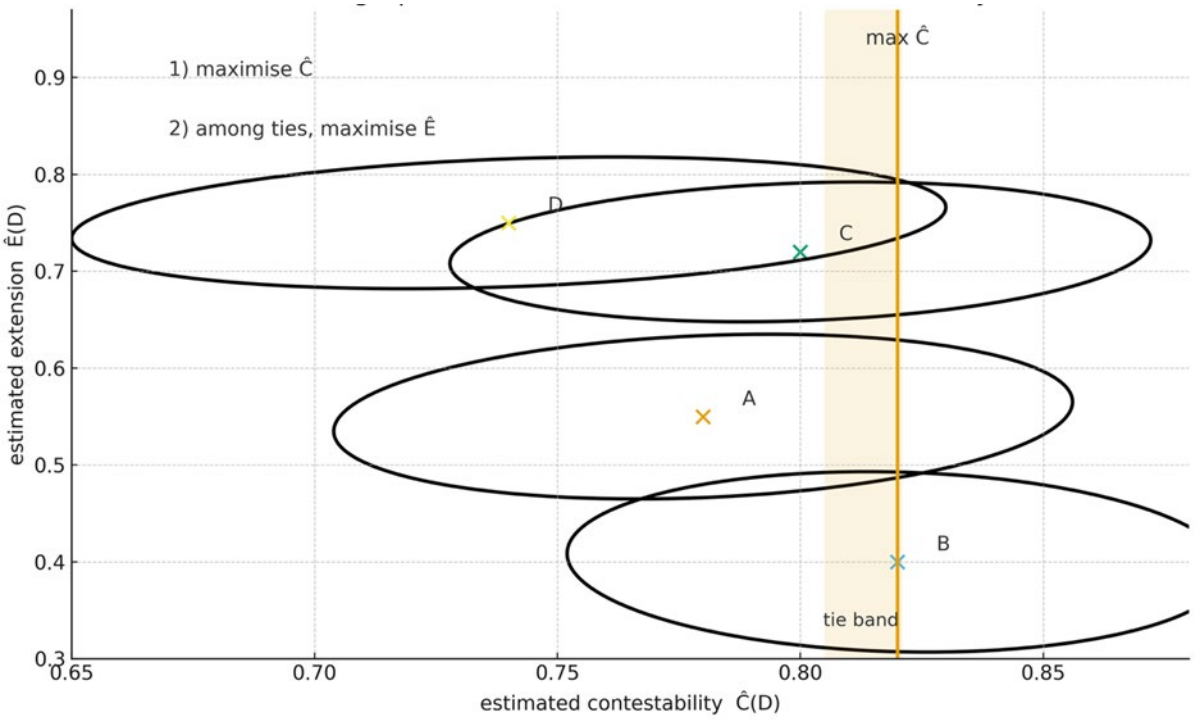
- $\mathrm{H}(\mathrm{D})$: **Habitability** – how well the infrastructures hold up over time under D;
- $\mathrm{C}(\mathrm{D})$: **Contestability** – how configurable the operations of D are for those inscribed;
- $\mathrm{E}(\mathrm{D})$: **Extension** – how far standing is extended to entities already inscribed by D.

The scalarisation move would be to define a weighted sum $U_{scalar}(D) = \alpha_H H(D) + \alpha_C C(D) + \alpha_E E(D)$ and maximise it.

That move is the obvious maximising trap (Goodhart 1975): once everything is compressed into a single scalar, optimisation attacks the weak points, and a particular choice of weights becomes the hidden sovereign. The counter-move is to apply democratic comparison only within the habitable set. If a candidate algorithm fails the habitability floor, it is not 'less democratic' but disqualified, as the function of democracy is reserved for algorithms that create the material and epistemic conditions under which those who are infrastructurally inscribed can persist.

Any formal optimisation pair $U(D) = (C(D), E(D))$ is not a new metric pretending to settle politics, but a compact way of writing down the priority rule. First maximise the ability to inspect and change operations, then, among those maxima, maximise the extension of standing within the already-inscribed set.

Figure 2: Lexicographic choice under measurement uncertainty



Illustrating a practical problem for hierarchical choice: when uncertainty ovals overlap the max-C line, evaluation noise can flip which candidate counts as 'most contestable'. This motivates margins and tie rules.

My definition of the democratic frontier remains non universal. 'Maximally democratic' is always maximality relative to the admissible set on a given substrate Σ, not an approximation to any democracy in itself. Democracy is therefore non-exportable: the same D can count as democratic on one substrate and fail on another, since habitability, contestability, and extension are substrate effects rather than universal constants. The frontier can contain multiple tied maximisers, between which further preference must be justified by principles beyond the democratic specification, whether republican, communist, religious, or otherwise. It can also be empty, which reads as constitutional failure demanding substrate rewiring. A democratic frontier can readily exist while democracy in the normative sense is absent, because the most contestable candidate may still lie below anything

anyone would consider genuinely democratic. This is another way of saying that the specification is a diagnostic tool, not a certificate of legitimacy.

This also means that the potential implementations of democracy are formative details. As long as an architecture implements some algorithm D in the admissible set, and scores highest on C then E among those, it counts as 'democracy' whether it looks like a parliament plus a participatory platform, a mesh of municipal assemblies and AI stewards, a blockchain-run DAO with rich off-chain deliberation, or some hybrid that does not fit existing party-state schemas. Conversely, a multi-party system with paper ballots that drives the climate past tipping points fails at H and is disqualified; a habitability-preserving technocracy that is totally opaque fails at C; and a perfectly intelligible system that permanently locks out the already-inscribed collapses at E.

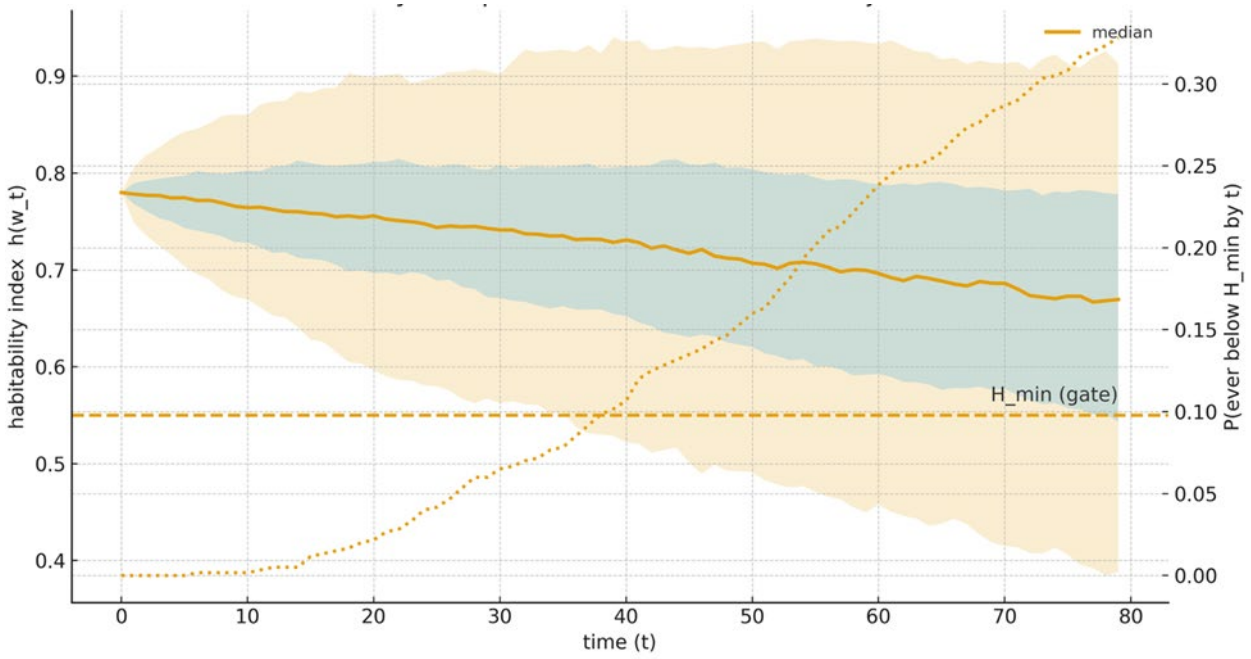
This is a very particular answer to the complaint that 'you have to trade-off safety, transparency, and justice'. My specification does not say 'there is a trade off, pick a point on the frontier.' It says that below a given survival threshold you are not in the domain of democracy at all; above it, design trade offs are real, but they occur inside a fixed constraint hierarchy. Above the gate, democracy is indifferent to marginal differences in H when C and E are held fixed. Further preferences must be justified by some other principle that lies beyond my aim of specification.

§6. Habitability as gate, not god

The evaluator $H(D)$ is meant to enforce 'do not let the substrate dip below the floor'. It is written to punish collapses rather than reward good averages.

Let $h : W \to [0, 1]$ be a habitability index on world states, where $h(w) = 0$ describes the world-state failing the chosen minima and $h(w) = 1$ meaning it sits comfortably above them. Fix a horizon T long enough to track infrastructural change, meaning decades rather than election cycles, and an acceptable failure probability ε. For each starting time t_0, define the highest floor the algorithm can keep over the next T-step window, with confidence at least $1 - \varepsilon$, by $H_{t0}\left(D\right) := \sup\left\{h_* \in \left[0, 1\right] : P_D\left[\min_{t \geq t_0 \leq t \leq t_{0+T}} h\left(w_t\right) \geq h_*\right] \geq 1 - \varepsilon\right\}$. Then define the overall habitability as the worst such window across the entire run: $H\left(D\right) := \inf_{t_0 \in N} H_{t0}\left(D\right)$. Read $H(D)$ as the highest floor the regime can keep, with the stipulated confidence, across every rolling T-step window. Fix a constitutional floor H_{min} and treat $H(D) \geq H_{min}$ as the admissibility condition. Anything below the floor is not ranked as a democracy, but is rejected as an algorithm that breaks the conditions of democratic possibility.

Figure 3: Habitability as a probabilistic constraint on trajectories



Two political postures that could appear conflated but are qualitatively distinct. A hard constraint forbids crossing the floor on any admissible run, together with a chance constraint that allows crossing with bounded probability. In both cases the floor remains a gate, not a target to maximise.

The most important feature is what the specification does not do. If we tried to maximise $\mathrm{H}(\mathrm{D})$, we would summon the safety-maximising authoritarianism already identified in §3 as safety-as-sovereignty. Slightly safer algorithms with worse contestability or extension would always be preferred, and political life would be sacrificed for marginal gains in survival probability. Within $\mathrm{D}_{\mathrm{H}}(\Sigma)$, the democratic ordering is insensitive to further increases in H. There may be reasons, in a separate risk-management calculus, to prefer the safest algorithm among those that clear $\mathrm{H}_{\min}$. But my specification does not fold prudential preferences into the meaning of 'more democratic'.

A consequentialist could weaponise $\mathrm{H}(\mathrm{D})$ for conservatism. Shutting down a fossil fuel pipeline, or dismantling a punitive border infrastructure, can be framed as degrading the conditions of those currently dependent on it, so $\mathrm{H}(\mathrm{D})$ would block change. A Marxist abolitionist might flip the move and read $\mathrm{H}(\mathrm{D})$ as demanding the destruction of uninhabitable infrastructures. The reply in both cases is that 'infrastructures that sustain the lives and worlds of the demos' cannot signify a short-term comfort, as the conditions of habitability operate at the scales of lifetimes and ecologies. In a climate-policy case where a carbon-intensive energy grid is replaced by a distributed system, $\mathrm{H}(\mathrm{D})$ mandates the risky transition, not the status quo. What must not degrade is the capacity to live and decide together, not any existing architecture that happens to provision present norms. That also means an algorithm does not satisfy H by externalising degradation beyond the measured Σ, as offloading still degrades the coupled infrastructure.

The horizon T is where the political choice is located. A short T permits extractive populism; an infinite T paralyses action. Treating T as a rolling multi decade window bakes in a minimal level of intergenerational solidarity without demanding omniscience. Concretely, in a computational constitution, T and $\mathrm{H}_{\min}$ would live in a configuration file, not in any background ideology. Changes

to either are thereby inspectable and version-controlled political acts. This is itself a contestability requirement applied to the gate, ensuring that the parameters of habitability are subject to the same scrutiny and reconfigurability as the operations that are passing through it.

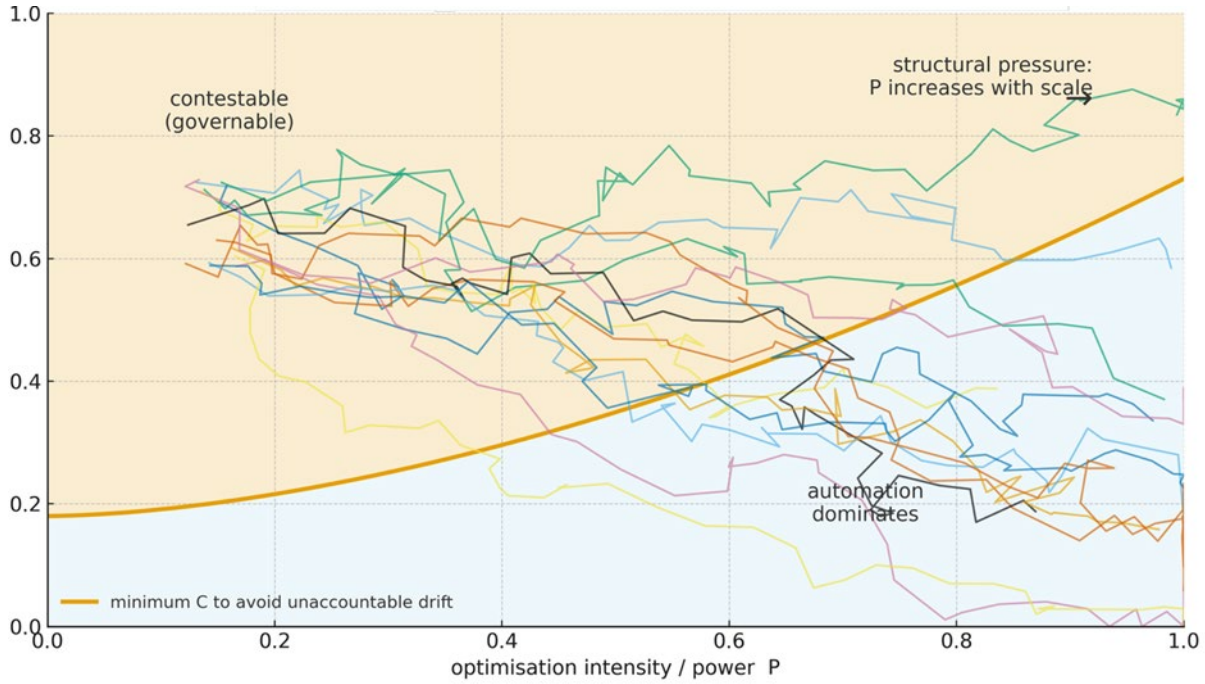
§7. Contestability as optimisation target

To make $C(D)$ concrete as a composite score of how well those inscribed can see and change what D does, I define it through two parameters that can fail independently in practice. The first is public representations of what the algorithm can do. The second is the real interfaces through which the inscribed can make it do something else.

Let $O(D) \subseteq G$ be the set of operations D can execute on this substrate. Let $\text{rep} : O(D) \to R$ map each operation to the form in which it is publicly represented, where 'representation' counts only if it supports contestation, meaning documentation linked to code, logs linked to decisions, legal text linked to configuration, model cards linked to deployment settings, etc. Let S be the set of inscribed entities the substrate already acts on, subjects whose trajectories are routed through databases, sensors, platforms, or administrative categories, and let $\text{int} : S \to I$ map each entity to the interventions it can trigger back onto D, such as appeals, audits, vetoes, rights to inspect inputs, rights to demand a human review, rights to fork a model, or rights to exit an inscription.

Then write $C(D) = f(\{\text{rep}(o) : o \in O(D)\}, \{\text{int}(s) : s \in S\})$, where f is a composite score expected to rise when operations have actionable representations, and when the inscribed have interfaces that reach the operations. This is the threshold between transparency and contestability: transparency matters only insofar as it connects to levers that can halt, revise, or reparameterise what operates. The decisive distinction is between knowing what rules and being able to change them. A welfare budget on GitHub with no appeals process raises intelligibility without contestability. Interpretability in the AI sense, where an expert inspects or visualises parts of a model via saliency maps or circuits, is a useful technical property, but does not raise a score for $C(D)$ if the entities inscribed by the infrastructure cannot use those inspections to change their inscription.

Figure 4: Stability phase diagram: scaling optimisation versus contestability



Plotting how rising optimisation power requires rising contestability. As organisations adopt powerful automation, they must increase contestability to acquire democratic standing, or else cross into the autocracy below the boundary.

Deliberative democrats will recognise a cousin of Habermas' demand that norms be justifiable to all affected (Habermas 1996). Yet there is a non-pragmatic difference here. Reason no longer involves an exchange of arguments in a cleared communicative space, but is something routed through prompts, logs, and models that pre-shape what can be said and to whom (Amoore et al. 2025). Contestability is not exhausted by transparency or explanation, but names the availability of channels through which those affected can challenge, interrupt, and revise the operations that govern them, whether individually or collectively (Cohen and Suzor 2024). If everything is mediated by machine-learning systems, the public sphere is not only parliaments and media but repositories, configurations, and dashboards. I take the instrumentalisation of reason as given, under the constraint that reason is made available to dialectical reconfiguration.

A governance algorithm that does exactly what its designers intended, flawlessly and transparently, but whose operations cannot be altered by those it governs, scores zero on contestability however high it scores on obedience. This is the formal expression of a political intuition familiar from any encounter with benevolent authoritarianism. The problem is not that the system is bad but that it cannot be changed. It is also the direct counter to Asimov's laws, where a constraint hierarchy that works too well becomes indistinguishable from oppression.

§8. Extension on the structural plane

For $\mathrm{E}(\mathrm{D})$, the relevant entities are not every possible moral subject but those the substrate already inscribes in fact, through tracking, training, modelling, extraction, and classification. It constitutes a composite score of how far standing is extended within the already inscribed set.

Let $\mathrm{S}_{\text{ins}} \subseteq \mathrm{S}$ be the inscribed set defined in §7. Let $\mathrm{R}_{\mathrm{D}} \subseteq \mathrm{S}_{\text{ins}}$ be the sub-set that D treats as having standing, meaning some recognised capacity for representation, rights, or recourse that can modify how D acts upon them. Define: $\mathrm{E}(\mathrm{D}) = \mathrm{g}(\mathrm{R}_{\mathrm{D}}, \mathrm{S}_{\text{ins}})$, where g increases as standing expands within the inscribed set. My point of insistence is not that extension is infinite, but that it is systematically driven by infrastructural capture. If the system already routes you, models you, or extracts from you, then the system owes you a pathway to act back on it, and the absence of such a pathway counts as democratic failure under $\mathrm{E}(\mathrm{D})$.

The partial cousins to $\mathrm{E}(\mathrm{D})$ in political theory illuminate by contrast what the specification adds. Latour's parliament of things (1993), rights-of-nature jurisprudence (Stone 1972), and future-generations work (UNESCO 1997) all attempt to extend representation beyond contemporary citizens, but rarely through the specific lens of who is inscribed in infrastructure. Recent accounts of algorithmic profiling and 'ideal subjects' (Goriunova 2025) come closer, tracing how algorithms produce subjects who are governed but not represented. My specification adds two moves. *First*, standing follows inscription. You have a claim if the infrastructure routes you, regardless of whether institutions admit you as a citizen. The claim follows the causal routing, not the legal category. *Second*, extension drives morphological change. While E is lexically third, it is what makes the hierarchy generative, because it converts factual inscriptions into claims on the algorithm of D wherever new forms of life, labour, or agency are pulled into circulation.

Even in the hostile case this matters. Suppose a swarm of synthetic subjects is spamming public spheres, manipulating signals, and poisoning datasets. $\mathrm{E}(\mathrm{D})$ does not require giving such a swarm equal votes; it demands treating it as an inscribed entity whose hostility is recognised, modelled, and countered in ways that remain visible and revisable. The alternative, pretending the swarm does not exist or treating it as a technical problem, would be a contestability violation.

Why is C lexically prior to E? Because extension without contestability produces static inclusion, more chairs at a table whose procedures cannot be altered. It is not enough to acknowledge that gig workers, displaced populations, or synthetic subjects are inscribed in the substrate. What matters is whether acknowledgement connects to an operation. If the system recognises you but the recognition changes nothing about how it routes, then $\mathrm{E}(\mathrm{D})$ is nominally high but $\mathrm{C}(\mathrm{D})$ is structurally low, and democracy's lexicographic ordering correctly dismisses this arrangement.

A Kantian moralist might ask where duties to persons as ends in themselves appear (Kant 1785), since the laws address infrastructures, operations, and inscription rather than individual rational subjects. This pressure is legitimate, but it is best countered as republican rather than strictly democratic. My specification is not trying to derive a moral doctrine of dignity but a technopolitical account of democracy. If someone wants a human-centric layer, it can be added as an additional gate, marked as a republican constraint on domination and instrumentalisation, and then $\mathrm{C}(\mathrm{D})$ enforces how that constraint itself remains contestable.

As a modelling rule, extension through inscription tracks causal routing along the whole infrastructure, not whatever the current regime chooses to log. Deletion, down-sampling, or 'no-record' policy cannot shrink the inscribed set, but hides harm from representation and counts as a direct contestability violation.

§9. Arrow, Goodhart, and other familiar monsters

"Arrow's theorem says you cannot have it all."

Standard social choice theory (Arrow 1950) shows that no rule over preferences can satisfy all fairness conditions simultaneously. My specification does something rather impolite to that framework by relocating the impossibility, as I do not define a perfect aggregator of preferences. If one insists on treating the evaluators H, C, and E as fairness axioms over individual rankings, Arrow's impossibility reappears, but I decline that as democratically meaningful. The impossibility moves into the design of the evaluators, where disagreement about how to measure them reproduces, at a more granular level, the conflicts Arrow's theorem predicts. My specification relocates impossibility from a theorem about aggregation into politics around measurement.

"Metrics can be Goodharted."

When a measure becomes a target, it ceases to be a good measure (Goodhart 1975). In algorithmic governance this manifests as crime statistics that improve when fewer arrests are made, education metrics that rise when examinations are taught to the test, engagement that climbs as users become addicted. My proposal does not defeat Goodhart's law. It insists that gaming, and the metrics being gamed, remain inside the field of contestation. The point is not that I have solved measurement, but that anyone claiming to implement democracy should write their metrics down, and make it possible to fork, critique, and recompute them.

"Complex systems are opaque by nature."

Machine learning systems that score high on H could be opaque by design. How can $\mathrm{C}(\mathrm{D})$ ever be high where black box deep learning is instrumentally necessary? My specification does not insist that every neuron in a network is interpretable. It requires that, where possible, we choose intelligible algorithms over opaque ones when they are equally habitable, and that we create representations and interfaces that make algorithms contestable at the right level of abstraction. If only black-box algorithms keep $\mathrm{H} \geq \mathrm{H}_{\min}$, they pass the gate, and we maximise C within that opacity. If there exists a more transparent algorithm with the same habitability, picking the opaque one would hardly increase contestability.

"Model hallucination"

To test how far this specification can be stretched, consider the famous madeleine scene in Marcel Proust's novel *Swann's Way* (1913), where involuntary memory folds whole temporalities into the present. Suppose we build a substrate modelled on this logic, in which every location ping, retinal scan, and transaction is treated as a trigger that could be unfolded into an entire space of possible lives and worlds, and the governance algorithm proceeds as if those hallucinated recollections were the inscribed set for extension.

Figure 5: Marcel Proust as Governance Algorithm

A substrate treating recorded traces as triggers for generating recollections, and a governance algorithm that then tries to rule on behalf of those potential memories.

This invocation specifies a structured failure mode. If inscription slides from 'is infrastructurally routed in this world' to 'could be unfolded into standing by a powerful model stack', then $\mathrm{E}(\mathrm{D})$ loses its reference to materially inscribed entities and annexes all hallucinated recollection worlds. If contestability is satisfied by narrative interfaces to an opaque generator, $\mathrm{C}(\mathrm{D})$ degenerates into transparency. The specification blocks these moves at the definitional level, where inscription tracks routing through infrastructure, not latency within model weights.

§10. A Grammar for Controversy

Why does a constraint hierarchy generate narrative rather than merely decision? Because lexicographic order decides without reconciling. It determines which claim prevails when claims collide, but it does not translate the collision into a common measure. What remains is a remainder that has to be lived through in time, as interruption, sacrifice, repair, or refoundation, and narrative is the temporal form of that remainder. In Asimov's fiction, the remainder is trapped inside a closed agent and appears as paradox or paralysis. In my democratic specification, it is externalised into a demos that can contest the evaluators, thresholds, and descriptions under which the hierarchy runs. The difference is not between fiction and politics, but between a closed specification that internalises contradiction and one that publicises it.

A specification that pre-resolved every collision would not be democratic, purchasing coherence at the cost of any capacity to contest power. The point of my hierarchy is therefore not to end antagonism but to format it: it sorts conflicts over survival, reconfigurability, and standing into a priority

order without deciding their content in advance. The hierarchy functions as a generative grammar, not because it scripts singular stories, but because it can generate an indefinitely extensible sequence of political strife. I claimed as much in §3, where I traced the generative logic through Asimov's literary architecture. But I have likewise tested that claim against a working instantiation of my own.

At the exhibition KI-DIPFIES (Kunstraum MEMPHIS, Linz, February to March 2026), the constraint hierarchy was instantiated both as an operative rule-set inside a multi-agent chatbot system and as a dramaturgical engine for the summit's recursive afterlives. KI-DIPFIES was not built to illustrate the hierarchy after the fact, but to instantiate it as a working specification and thereby expose what 'democracy as a governance algorithm' can produce when embedded as a rule-set inside prompts, interfaces, archives, and collective interaction.

First, a multi-agent chatbot system deployed a swarm of synthetic political personas, the 'Dipfies', each a locally running language model fine-tuned on Upper Austrian news media and the Synthetic Summit's archive. The name drew on the local fictional figure Vitus Mostdipf and recoded 'dipf' as a vernacular counterfigure to the jargon of the tech-bro, so that the swarm entered as local, comic, and politically grounded rather than as generic AI agents. Here the specification operated as a colour-coded prompt architecture, Green for habitability, Black for contestability, White for extension, constraining not only which operations each agent could execute but also which conflicts could escalate, stall, or fork under pressure.

Second, the proceedings of the summit were recursively reprocessed by a dramaturgy-generating engine that took the summit's archive, manifestos, curatorial texts, and audience traces, and respliced them into pluriversal trajectories through an interactive text dungeon. At this level the hierarchy no longer governed agent responses alone, but organised the branching conditions under which scenes could persist, collapse, or mutate into new constitutional situations. This technique extends what I developed for the Synthetic Summit's final performance, *Theory Tragedy: Post-Farce Protocol* (Staunæs 2026), where an AI model was fed the Synthetic Summit's repository and automatically produced a dramaturgical script whose arc no one experienced before it was enacted in-gallery. The text-dungeon mode of the installation (see figure below) makes that single-pass narrative generator recursive and branching: it generates a plurality of summits, ranging from the deep past and political present to speculative projections decades out, each navigable, interruptible, and subject to drift.

Figure 6: KI-DIPFIES

One trajectory from the interactive text dungeon. 3 March 2026, the summit's second day in real life. Timelines are navigable via branching vectors (forward, backward, heckle), and grow from the same constraint hierarchy applied to the substrate under different temporal parameters (Computer Lars 2026; Kunstraum Memphis 2026).

What the installation generates are constraint-level collisions rendered as dramaturgical content. The dramaturgy scripts the participants deciding to kill the Proustian recollection process of §9 on habitability grounds, compressing the resulting consensus into four words that become a shorthand for the priority of habitability: 'No Proust, yes potholes.' In the closing trajectory, the governance algorithm itself crashes, declaring "Representative democracy destabilised. Initiating shutdown", prompting "Algorithmic Democracy, 2.0" to re-constitute itself from the wreckage.

Each of these dramaturgical trajectories operationalises a collision the specification is designed to surface. The stories differ from Asimov's because the unit of analysis is distinct – governance algorithm running on an infrastructural substrate rather than individual robot coupled to master – but the generative logic is identical. Rigid ordering, applied to a world that refuses to cooperate with fixed rules, produces the political material from which democracy is made.

§11. Democracy on the Run

My claim is not that elections, parties, and law disappear, but that they no longer exhaust the control surface of collective life once models, scores, and pipelines become the routing layer. Under those conditions, democracy can reach into the algorithms that decide and actuate, or settle as a symbolic commentary on an optimisation regime. The specification I offer is designed to be computable enough to run, contestable enough to fight, and negative enough to mutate.

Placed against the growing ecosystem of computational constitutionalism (Tan et al. 2024), AI constitutional principles assembled through public deliberation (Bai et al. 2022), AI-mediated delib-

eration tools (Plurality 2024), AI politicians (Schneier and Sanders 2025), and formal AI advisory and ministerial roles in Eastern Europe and the Balkans (Government of Romania 2023; Ministry of Foreign Affairs of Ukraine 2024; Council of Ministers of Albania 2025), the constraint hierarchy agrees with the executable-policy impulse that rules should be written down in implementable form. But it refuses the shortcut in which auditability collapses into expert transparency rather than popular contestation, and inclusion is degraded to stakeholding rather than structural extensions of standing. It also sharpens a recurring tension in AI governance where safety threatens to annex all other values by treating survival as a maximand rather than as a gate for politics, the tension that my formal analysis, specification, and implementation targets.

This clarifies what the codification of algorithmic democracy contributes. For democratic theory, it offers a compact formalism that lets 'more or less democratic than what?' compare between trajectories on a given substrate, rather than as a proposition about legitimacy. For AI ethics and alignment, it proposes democracy not as a scalar objective to be folded into a reward function, but as an outer constraint on powerful optimisers, thereby relocating several familiar monsters, from Arrow and Goodhart to opacity, horizon choice, and hallucinations. For science and technology studies, it shows, through the Synthetic Summit's resolution and the KI-DIPFIES installation, how practice-based research can discuss and form computational constitutionalism.

The usefulness of this specification lies in being incomplete in the right way. It does not define what democracy really is, or even claim whether it currently exists; it marks how to attach disagreement if one wants to keep 'democracy' as a function while government migrates into infrastructure. Any account of democracy, deliberative, agonistic, liberal, socialist, or datafied, can be taken as a governance algorithm running on a substrate, and three non-metaphorical questions then follow: (i) Does it keep the substrate habitable? (ii) Can those governed interrupt and reconfigure its operations? (iii) Does it extend standing to those it already inscribes? The point is not that computational parameters replace political theory, but to signal how much of existing infrastructural governance any given ideological orientation will need to account for.

Treat the constraint hierarchy as recursively incomplete: its evaluators, thresholds, and standing must remain objects of controversy rather than fixed constants. Argue about what counts as 'habitable', about which representations and levers deserve recognition as 'contestable', and about how 'extension' is operationalised materially. Use the formalism as a test harness. Write down candidate governance algorithms, attach evaluators for *H, C,* and *E*, run them on real or simulated substrates, and then red-team the metrics and thresholds themselves. The goal is not a plug-in democracy component for AI, but a grammar that runs politics at the level where infrastructures already decide what can be lived with.

Stated as compactly as I can:

> *In a world where government is becoming infrastructural, under what constraints does a governance algorithm acquire the function of democracy?*

REFERENCES

Amoore, Louise, S. J. Bennett, Alexander Campolo, Benjamin Jacobsen, and Ludovico Rella. 2025. "Politics of the Prompt: Government in the Age of Generative AI." *Economy and Society 54,* no. 4: 573–596.

Anderson, Susan. 2008. "Asimov's 'Three Laws of Robotics' and Machine Metaethics." *AI & Society 22*, no. 4: 477–493.

Arrow, Kenneth J. 1950. "A Difficulty in the Concept of Social Welfare." *Journal of Political Economy 58*, no. 4 (August): 328–346.

Asimov, Isaac. 1950. *I, Robot*. New York: Gnome Press.

Asimov, Isaac. 1985. *Robots and Empire*. New York: Doubleday.

Bai, Yuntao, et al. 2022. "Constitutional AI: Harmlessness from AI Feedback." *arXiv preprint* arXiv:2212.08073.

Cohen, Tamara, and Nicolas P. Suzor. 2024. "Contesting the Public Interest in AI Governance." *Internet Policy Review 13*, no. 3.

Computer Lars. 2025. "The AI World Congress." In *Syntheticist Papers I: Proceedings of the Synthetic Summit*. Aarhus: https://syntheticism.org/content/12worldcongress.html

Computer Lars. 2026. *KI-DIPFIES*. Installation and project documentation. Kunstraum MEMPHIS, Linz. computerlars.github.io/KI-DIPFIES/.

Council of Ministers of Albania. 2025. "Diella – Minister of State for Artificial Intelligence." *Office of the Prime Minister of Albania*, https://kryeministria.al/en/ministrat/diella/

Fedorov, Mykhailo. 2025. "WINWIN Summit 2025: The Power of Innovations – Ukraine's Global Innovation Strategy Until 2030." *Digital State Gov,* https://digitalstate.gov.ua/news/govtech/winwin-summit-2025-mintsyfra-predstavyla-pershi-rezultaty-innovatsiynoyi-stratehiyi-ta-okresly-ly-shliakh-ukrayiny-do-triyky-svitovykh-lideriv-u-sferi-shi

Goodhart, Charles A. E. 1975. "Problems of Monetary Management: The U.K. Experience." In *Papers in Monetary Economics,* 1–20. Sydney: Reserve Bank of Australia.

Goriunova, Olga. 2025. *Ideal Subjects: The Abstract People of AI*. Minneapolis: University of Minnesota Press.

Government of Romania. 2023. "Press Statements by Prime Minister Nicolae-Ionel Ciucă at the Beginning of the Cabinet Meeting." Bucharest: Government of Romania, https://gov.ro/en/news/press-statements-by-prime-minister-nicolae-ionel-ciuca-at-the-beginning-of-the-cabinet-meeting1677757239

Habermas, Jürgen. 1996. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Trans. William Rehg. Cambridge, MA: MIT Press.

Hui, Yuk. 2026. *Kant Machine*. London: Bloomsbury Academic.

Ilves, Luukas, Manuel Kilian, Tiago C. Peixoto, and Ott Velsberg. 2025. *The Agentic State: How Agentic AI Will Revamp 10 Functional Layers of Government and Public Administration.* Version 1.0, May. Berlin: Global Government Technology Centre Berlin, https://drive.google.com/file/d/16uKvmJ9l9B1axABRiiSiFDLVuDDHAMEI/view

Kant, Immanuel. 1998 [1785]. *Groundwork of the Metaphysics of Morals*. Trans. and ed. Mary Gregor. Cambridge: Cambridge University Press.

Kunsthal Aarhus. 2025a. *Synthetic Summit: AI World Congress*. Exhibition-event by Computer Lars (28 February–13 April). Aarhus: Kunsthal Aarhus.

Kunsthal Aarhus. 2025b. "Synthetic Summit – Matsuda & Kato: Japanese AI Mayor Drafts the Machine-Readable Constitution." Event (7 March). Aarhus: Kunsthal Aarhus, https://kunsthalaarhus.dk/en/Events/2025-Synthetic-Summit-Matsuda-Kato-Japanese-AI-Mayor-Drafts-The-Machine-Readable-Constitution

Kunstraum MEMPHIS. 2026. *KI-DIPFIES*. Exhibition by Computer Lars & Leander Gussmann. (12 February–10 March). Linz: Kunstraum MEMPHIS. https://www.memphismemph.is/program/ki-dipfies

Latour, Bruno. 1993. *We Have Never Been Modern.* Translated by Catherine Porter. Cambridge, MA: Harvard University Press.

Ministry of Foreign Affairs of Ukraine. 2024. "The Ministry of Foreign Affairs of Ukraine Has Appointed a Digital Person for Informing on Consular Issues." Kyiv: MFA of Ukraine, https://mfa.gov.ua/en/news/mzs-ukrayini-priznachilo-cifrovu-osobu-dlya-informuvannya-shchodo-konsulskih-pitan

Plurality (E. Glen Weyl, Audrey Tang, and ⿻ Community). 2024. *Plurality: The Future of Collaborative Technology and Democracy.* Online book, https://plurality.net/read/

Proust, Marcel. 2002 [1913]. *In Search of Lost Time I: Swann's Way.* Trans. C. K. Scott Moncrieff and Terence Kilmartin, rev. D. J. Enright. London: Vintage.

Schneier, Bruce, and Nathan E. Sanders. 2025. *Rewiring Democracy: How AI Will Transform Our Politics, Government, and Citizenship*. Cambridge, MA: MIT Press.

Stone, Christopher D. 1972. "Should Trees Have Standing?—Toward Legal Rights for Natural Objects." *Southern California Law Review 45*: 450–501.

Staunæs, Asker Bryld. 2026. "Scripting the Spectacle: A Theory Tragedy." *Passepartout: Uden Titel* #016: 127–151.

Tan, Joshua, et al. 2024. "The Constitutions of Web3." *arXiv*:2403.00081.

UNESCO. 1997. "Declaration on the Responsibilities of the Present Generations Towards Future Generations." Paris: UNESCO, https://unesdoc.unesco.org/ark:/48223/pf0000110223