

AI as Coordination Prosthesis: Human-Led, Machine-Assisted Governance

Authors

Ryder, John F.
Conceptual Future Pragmatist
Drive-In s.r.o.
ORCID: 0009-0003-5240-4533

Description

This paper develops a constitutional framework for the bounded use of artificial intelligence within post-labour institutional systems. Rather than treating AI as a governing authority or decision surrogate, the paper positions it as a coordination prosthesis—an assistive layer that amplifies human judgement without replacing it.

As automation weakens labour’s monopoly over income and legitimacy, governance systems face growing coordination complexity. Artificial intelligence is frequently proposed as a solution to administrative overload. However, unbounded deployment risks transferring authority from accountable actors to optimisation systems.

The paper establishes a non-delegable principle: AI may propose, but only humans may decide. Through the Legitimacy Fallback Principle and the Kobayashi Maru Constraint, AI systems are structurally prohibited from terminating indeterminate judgement. Where human judgement is unresolved, machine systems must increase intelligibility rather than reduce uncertainty through optimisation.

Permitted AI roles include mapping, matching, anomaly detection, and transparency support. Explicitly prohibited functions include adjudication, enforcement, moral ranking, and automatic execution. Override, veto, and refusal are treated as first-class design features, and guardrails against automation creep are embedded institutionally.

This framework complements the Engagement Credit Economy (ECE) architecture by ensuring that technological assistance strengthens human agency rather than eroding it. The paper contributes to AI governance by shifting the question from “how to make AI trustworthy” to “where authority must remain human.”

Version

1.0

Publication Type

Working Paper

Keywords

AI governance, human agency, coordination prosthesis, post-labour institutions, automation limits, legitimacy, constitutional design, institutional safeguards, human-AI systems, algorithmic authority

Subjects

Public governance, AI governance, Institutional design, Automation policy, Political economy

License

Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)

Related Works

Relation: *Is part of*
→ Engagement Credit Economy (ECE) master record

Relation: *Is supplemented by*
→ Community Trusts paper

Relation: *Is complemented by*
→ Human Value and Meaning System (HVES)

Funding

Self-funded independent research

Transparency Statement

This paper was developed with structured assistance from AI language models for drafting, editing, and structural refinement. Conceptual architecture, theoretical framing, and institutional design originate from the author. AI assistance functioned as a writing and analytical support tool and did not determine substantive positions.

Abstract

As automation reduces labour's monopoly over income and social legitimacy, governance systems face an emerging coordination crisis rather than a decision deficit. Participation becomes non-linear, institutional complexity increases, and administrative burdens exceed unaided human capacity. Artificial intelligence is frequently proposed as a solution to this strain, yet unbounded deployment risks transferring authority from accountable actors to optimisation systems.

This paper develops a constitutional framework for positioning artificial intelligence as a **coordination prosthesis** within post-labour institutional architectures. It advances a non-delegable principle: AI may propose, but only humans may decide. Legitimacy, judgement, and authority remain structurally inseparable from human agency, even under conditions of computational superiority.

The framework defines permissible AI roles—mapping, matching, anomaly detection, and transparency—while prohibiting machine adjudication, enforcement, or termination of indeterminate judgement. Through the **Legitimacy Fallback Principle** and the **Kobayashi Maru Constraint**, AI systems are structurally required to fail productively when human judgement is unresolved, generating intelligibility rather than algorithmic closure.

By grounding AI deployment in constitutional limits rather than ethical aspiration, the paper demonstrates how machine assistance can preserve dignity, voluntary participation, and institutional resilience in post-automation societies. The question is not whether AI will assist governance, but whether governance will remain human.

Introduction

Artificial intelligence is often framed as either a revolutionary emancipator or an existential threat. Both narratives obscure the more fundamental issue: where authority resides when systems outperform humans cognitively.

In post-labour societies, this question becomes acute. As automation weakens labour's vertical role in organising income and legitimacy, governance systems must accommodate non-linear participation, fluctuating capacity, and expanding administrative complexity. The strain is not one of insufficient data or computational power, but of human stamina and institutional coherence.

Artificial intelligence promises relief. Yet history demonstrates that powerful tools destabilise governance not by malfunctioning, but by being placed incorrectly. When optimisation quietly substitutes for judgement, and efficiency displaces accountability, legitimacy erodes long before collapse is visible.

This paper argues that AI must be constitutionally bounded as a **coordination prosthesis**—a tool that amplifies human judgement without bearing it. It does not reject artificial intelligence; it places it. The distinction is structural rather than technical. AI may extend cognitive reach, but it may not authorise outcomes. Where human judgement stalls, machines must not complete the task. Authority remains suspended until exercised by identifiable human actors.

This is not an ethical appeal to responsible innovation. It is an architectural proposal for institutional design under conditions of complexity.

1. The Coordination Problem in Post-Labour Systems

As labour ceases to function as the primary organiser of income, time, and legitimacy, institutional structures built around linear productivity begin to strain. Participation becomes intermittent, distributed, and heterogeneous. Capacity fluctuates. Contribution is less predictable and less easily classified within traditional employment frameworks.

The resulting challenge is not scarcity of decision-making authority but overload. Human governance bodies face increasing informational density, cross-domain interdependencies, and escalating coordination costs. The more complex the system, the greater the temptation to compress deliberation into algorithmic recommendation.

In this environment, artificial intelligence appears not as an ideological imposition but as a practical necessity. Mapping complex networks, matching resources at scale, and detecting

patterns beyond unaided perception are tasks well suited to computational systems. Refusing such assistance would impose artificial limits on institutional capability.

Yet the risk lies precisely in success. As AI systems become more reliable, coherent, and persuasive, they begin to appear self-authorising. The more helpful the recommendation, the easier it becomes to treat it as decision. In contexts where human judgement is fatigued or divided, optimisation offers closure.

The coordination problem, therefore, is double-edged. AI is required to sustain human governance under complexity, but without structural constraint it can displace the very judgement it is meant to assist. The solution is not abstention from AI, but constitutional placement.

2. Why AI Must Never Be the Bearer of Legitimacy

Legitimacy is not a computational property. It is a social condition grounded in accountability, contestability, and the recognition of persons as moral agents capable of refusal. Artificial intelligence, however advanced, does not possess moral agency. It can model preferences, simulate trade-offs, and optimise against defined objectives, but it cannot bear responsibility.

Delegation in governance presupposes accountability. Authority may be delegated to identifiable actors who remain answerable for their decisions. AI systems do not meet this condition. Their outputs derive from statistical inference and parameter optimisation rather than normative judgement. To treat these outputs as self-authorising is not delegation but abdication.

The distinction becomes critical in post-labour institutional architectures where participation, compensation, and recognition directly affect dignity and social standing. If AI systems were permitted to determine eligibility, assign merit, or terminate deliberation, legitimacy would migrate from accountable human actors to optimisation processes. Such migration may appear efficient, but it dissolves the link between authority and responsibility.

Historical precedents reinforce this boundary. From actuarial tables to bureaucratic scoring systems, calculative tools have long informed governance. Failures have occurred not when such tools were used, but when they were mistaken for decision-makers. Procedural correctness replaced substantive judgement, and appeal became meaningless because there was no accountable author.

Artificial intelligence magnifies this risk. Its scale and coherence make it uniquely persuasive, particularly under conditions of institutional fatigue. For this reason, the Engagement Credit Economy framework establishes a non-negotiable principle: AI may assist judgement, but it may never bear legitimacy. Authority remains human even when cognition is extended.

This placement is not anti-technological. It is constitutional. Where legitimacy is concerned, optimisation is advisory; judgement is human.

3. “AI Proposes, Humans Decide” as a Constitutional Rule

If artificial intelligence is to be deployed within governance systems without eroding legitimacy, its role must be constrained by an explicit and enforceable constitutional rule. Within the Engagement Credit Economy framework, this rule is defined unambiguously: **AI may propose; humans must decide.**

This formulation is not rhetorical. It establishes a system invariant that governs all permissible uses of AI within participation, compensation, safeguarding, and institutional coordination. AI outputs may inform, clarify, and structure human judgement, but they possess no institutional standing in their own right. No outcome affecting participation, eligibility, prioritisation, remuneration, or sanction may occur without an explicit act of human authority.

This distinction is frequently obscured by the language of “human-in-the-loop” governance, which often reduces human involvement to supervisory validation of machine outputs under procedural or time pressure. Such arrangements preserve the appearance of accountability while quietly transferring effective authority to algorithmic systems. A constitutional rule, by contrast, requires that human judgement remain decisive, legible, and attributable.

Accordingly, all AI-assisted processes must satisfy three conditions. First, AI recommendations must be intelligible to human decision-makers, including the assumptions, constraints, and trade-offs that shape them. Second, human actors must retain the practical ability to override, reject, or modify AI outputs without procedural friction, reputational penalty, or performance degradation. Third, acts of refusal or modification must be institutionally valid and auditable, rather than treated as exceptions or failures.

3.1 The Legitimacy Fallback Principle (The Kobayashi Maru Constraint)

A critical failure mode arises when human judgement remains unresolved while AI systems generate internally coherent, efficient, or high-confidence recommendations. In many contemporary systems, such conditions result in silent delegation: indecision is treated as a procedural gap to be closed by optimisation. The ECE framework explicitly rejects this logic.

Under the **Legitimacy Fallback Principle**, unresolved human judgement must never default to algorithmic completion. Where human decision-makers cannot confidently exercise judgement, authority remains suspended rather than transferred. In such cases, AI systems are required to reconfigure their function—not to reduce uncertainty through optimisation, but to **increase human intelligibility**.

This is operationalised through an iterative deliberation process in which AI systems are structurally prohibited from converging on decisive outcomes. Instead, they must surface competing considerations, irreducible trade-offs, and points of normative conflict. AI systems may only “lose” in this process: their failure consists not in error, but in the production of increasingly clear arguments that make human judgement unavoidable.

This constraint ensures that silence does not become consent, delay does not become delegation, and optimisation does not assume the role of governance. AI systems may iterate indefinitely; they cannot terminate the process. The loop closes only when a human authority—individual or collective—exercises judgement and records a decision.

By design, this mechanism prevents artificial intelligence from resolving moral or political indeterminacy. It transforms moments of uncertainty from sites of algorithmic takeover into sites of intensified human deliberation. In doing so, it preserves legitimacy under conditions of complexity and ensures that authority remains inseparable from human agency.

4. Permitted Roles of Artificial Intelligence

Having established that artificial intelligence may not bear legitimacy, authority, or judgement, the question becomes not *whether* AI may be used, but *how* it may be used without eroding human agency. This section defines a strictly bounded set of permissible AI functions within the Engagement Credit Economy framework. These roles are assistive, non-authoritative, and reversible by design.

The purpose of this delineation is twofold: to prevent automation creep through vague mandates, and to ensure that AI deployment strengthens institutional capacity rather than substituting for human governance.

4.1 Mapping and System Legibility

AI systems may be used to map complex institutional environments that exceed unaided human cognitive capacity. This includes the aggregation and visualisation of participation patterns, resource distributions, asset inventories, programme interdependencies, and temporal dynamics across Community Trusts and related bodies.

Mapping functions are descriptive rather than prescriptive. They reveal structure, relationships, and change over time, but they do not imply priority, value, or obligation. Their role is to make systems legible to human decision-makers, not to determine outcomes.

4.2 Matching and Opportunity Alignment

AI may assist in matching capabilities, resources, and opportunities where scale or variability would otherwise overwhelm manual coordination. Examples include aligning skills with Community Initiative Programmes, matching unused assets with local needs, or identifying complementarities between Trusts for horizontal pooling.

Crucially, matching outputs remain proposals. They do not constitute assignments, requirements, or rankings of worth. Individuals retain unconditional rights to refuse participation, and Trusts retain discretion to modify or disregard AI-generated matches without justification or penalty.

4.3 Anomaly Detection and Drift Identification

AI systems may be deployed to detect anomalies within institutional processes, including indicators of exploitation, capture, metric gaming, safeguarding failures, or automation creep. In this role, AI functions as an early-warning system rather than an enforcement mechanism.

Detected anomalies do not trigger automatic action. Instead, they prompt human review by designated governance or investigatory bodies. This preserves due process and prevents the substitution of statistical deviation for judgement.

4.4 Transparency, Auditability, and Explanation Support

AI may support transparency by generating audit trails, explanatory summaries, and traceable accounts of system behaviour. This includes making visible how recommendations were generated, which constraints were applied, and where uncertainty or conflict remains.

Such transparency functions are directed inward toward governance integrity rather than outward toward behavioural surveillance. AI may illuminate institutional processes; it may not monitor or score individuals as subjects of control.

Intelligibility in this context does not require full transparency of computational mechanisms. It requires that reasons, trade-offs, and consequences be rendered at a level appropriate to normative judgement, even where underlying technical processes remain complex.

4.5 Explicitly Prohibited Functions

To prevent ambiguity, certain roles are explicitly excluded from AI use within the framework. AI systems may not:

- determine eligibility, entitlement, or exclusion;
- assign moral weight, merit, or social value;
- enforce participation, compliance, or task completion;
- resolve normative or political disagreement;
- terminate deliberation under conditions of unresolved human judgement.

These prohibitions are structural, not technical. They apply regardless of model capability, performance, or perceived neutrality.

4.6 AI as Capacity Amplifier, Not Authority Substitute

Across all permitted roles, AI is positioned as a capacity amplifier: a means of extending human stamina, attention, and pattern recognition under conditions of complexity. It compensates for cognitive and administrative limits without absorbing decision-making authority.

This placement is particularly significant in post-labour contexts, where participation may be intermittent, non-linear, or shaped by fluctuating capacity. By absorbing coordination load rather than imposing behavioural discipline, AI supports voluntary contribution while preserving exit, refusal, and dignity.

5. Human Override, Veto, and Refusal as Core Design Features

If artificial intelligence is to function as a coordination prosthesis rather than a governance authority, human override, veto, and refusal must be treated as first-class system features rather than exception paths. These capacities are not safeguards of last resort; they are the mechanisms through which legitimacy is continuously exercised and renewed.

In many contemporary automated systems, override exists formally but not functionally. Human actors may technically retain the ability to reject algorithmic outputs, yet face implicit penalties for doing so: increased workload, audit risk, reputational scrutiny, or procedural delay. Under such conditions, override becomes symbolic rather than substantive, and authority quietly migrates to the system itself. The ECE framework explicitly rejects this pattern.

5.1 Override as a Normalised Action

Within AI-assisted governance systems, overriding an AI recommendation must be institutionally neutral. It must not trigger additional justification requirements, escalation procedures, or performance review flags. Override is treated as an expected outcome of human judgement interacting with machine assistance, not as a deviation from optimal process.

To ensure this neutrality, override actions must be:

- frictionless in execution,
- free from implicit or explicit sanction,
- and procedurally equivalent to acceptance.

This preserves the practical reality of human authority rather than merely its formal appearance.

5.2 Veto as a Legitimate Endpoint

Veto differs from override in that it does not substitute an alternative outcome; it halts action altogether. The ability to veto is essential in contexts where available options are judged to be unacceptable, premature, or misaligned with human values.

The framework treats veto as a legitimate and complete decision state. Veto does not require the immediate production of an alternative, nor does it obligate the system to seek optimisation-based substitutes. Where veto is exercised, authority remains suspended until further human deliberation occurs.

This prevents systems from interpreting hesitation or moral resistance as a signal to escalate automation.

5.3 Refusal as a Protected Act

Refusal operates at both the individual and institutional level. Individuals must retain the unconditional right to refuse participation, recommendation alignment, or engagement suggested by AI systems. Likewise, institutions must retain the right to refuse the deployment, expansion, or continued use of specific AI tools.

Refusal is explicitly protected from reinterpretation as non-compliance, inefficiency, or disengagement. Within the Engagement Credit Economy, refusal preserves dignity and exit, and therefore constitutes a valid expression of agency rather than a system failure.

5.4 Auditability Without Penalty

All acts of override, veto, and refusal must be auditable in order to preserve institutional learning and accountability. However, auditability must not be conflated with performance assessment or behavioural correction.

Audit logs serve to:

- identify patterns of system mismatch,
- detect automation creep,
- and improve future design constraints.

They must not be used to discipline decision-makers for exercising judgement, nor to recalibrate AI systems toward reduced human intervention.

5.5 Authority Asymmetry as a Design Invariant

Taken together, override, veto, and refusal establish a persistent asymmetry: AI systems may generate proposals indefinitely, but only human agents may close decisions. This asymmetry is not contingent on trust in AI performance, nor does it diminish as systems improve. It is structural and permanent.

By embedding these features directly into system design, the framework ensures that authority cannot drift through convenience, confidence, or scale. Human judgement remains the sole source of legitimacy, even under conditions of complexity, uncertainty, or pressure.

The framework accepts deliberative friction as a necessary cost of legitimacy rather than a system defect. Where coordination becomes burdensome, the appropriate response is institutional capacity investment or procedural simplification—not delegation of authority to optimisation systems.

6. AI in Safeguarding, Not Task Enforcement

Most contemporary applications of artificial intelligence in governance are oriented toward task enforcement: ensuring compliance, optimising throughput, reducing deviation from prescribed behaviours, and maximising efficiency. In post-labour institutional systems, such uses are not

merely inappropriate but actively corrosive. Where participation is voluntary and legitimacy fragile, enforcement-oriented automation risks converting coordination into coercion.

Within the Engagement Credit Economy framework, AI is therefore repositioned away from behavioural discipline and toward **institutional safeguarding**. Its primary function is not to monitor whether individuals conform, but to monitor whether systems drift.

6.1 Monitoring Systems, Not Persons

AI systems may be used to detect structural risks within institutional processes, including:

- signs of participation creep (where voluntary engagement begins to resemble obligation),
- metric distortion or gaming,
- inequitable resource concentration,
- procedural bottlenecks that pressure decision-makers toward automation,
- or patterns suggestive of elite capture.

In each case, the object of scrutiny is the **governance system itself**, not the moral standing or behaviour of individuals. AI functions as a diagnostic instrument, revealing systemic stress points rather than classifying participants.

This inversion is deliberate. Surveillance of persons produces compliance; scrutiny of systems produces accountability.

6.2 Early Warning Without Automatic Sanction

When anomalies or risks are detected, AI outputs trigger human review rather than automatic sanction. Statistical deviation does not constitute guilt, and correlation does not constitute cause. The system surfaces signals; it does not impose consequences.

This preserves due process and prevents the conflation of predictive inference with normative judgement. It also ensures that safeguarding mechanisms do not become backdoor enforcement tools under the guise of optimisation.

6.3 Protecting Boundaries Against Automation Creep

Automation creep occurs when assistive systems gradually expand their scope, moving from recommendation to expectation, from expectation to requirement, and from requirement to silent enforcement. This drift is rarely intentional; it emerges through convenience, scale, and the normalisation of system outputs.

AI deployed within the ECE framework must therefore include internal mechanisms for detecting its own expansion. Such mechanisms may include:

- monitoring increases in override rarity,
- tracking reductions in human deliberation time,

- identifying patterns of automatic acceptance,
- and flagging institutional dependency on machine outputs.

In this role, AI acts as a guardian of the constitutional boundary established in Sections 2 and 3. It detects when optimisation begins to substitute for judgement.

6.4 Safeguarding Participation and Dignity

AI may assist in identifying risks to participant welfare, including burnout patterns in Community Initiative Programmes, exclusionary dynamics within Trust governance, or inequitable access to opportunities. However, such identification must always lead to supportive intervention rather than corrective enforcement.

The aim is to preserve voluntary participation, not to stabilise throughput. AI becomes a tool for sustaining dignity and capability, not extracting performance.

6.5 Institutional Self-Reflection as Design Principle

By orienting AI toward safeguarding, the framework embeds institutional self-reflection into governance architecture. AI's analytical capacity is directed inward—toward improving fairness, detecting drift, and maintaining constitutional integrity—rather than outward toward disciplining individuals.

This orientation transforms artificial intelligence from an instrument of control into an instrument of resilience. It strengthens governance capacity without narrowing human agency, and it aligns technological assistance with the foundational principle that legitimacy rests exclusively in human judgement.

7. AI as a Stamina Substitute for Non-Linear Contributors

A defining feature of post-labour participation is non-linearity. Contribution is no longer reliably continuous, time-bound, or uniform. Capacity fluctuates due to health, care responsibilities, cognitive load, ageing, disability, or episodic availability. Traditional institutions, designed around linear productivity and consistent presence, routinely misclassify such variability as disengagement or inefficiency.

Within the Engagement Credit Economy framework, artificial intelligence is positioned to address this mismatch—not by enforcing regularity, but by **absorbing the stamina demands that institutions typically impose on human contributors**.

7.1 From Performance Maximisation to Capacity Preservation

Conventional automation systems are optimised to extract consistent output. In contrast, AI deployed within the ECE is tasked with preserving contributor capacity over time. Its function is

not to increase the intensity or frequency of participation, but to reduce the cognitive, administrative, and coordination burdens that disproportionately exclude non-linear contributors.

This represents a shift from performance optimisation to **capacity conservation**.

7.2 Cognitive and Administrative Load Reduction

AI systems may support contributors by:

- maintaining continuity of context across intermittent engagement,
- managing scheduling complexity without penalising irregularity,
- summarising institutional processes and decisions for re-entry,
- and handling routine administrative coordination that would otherwise exhaust limited human capacity.

In this role, AI functions as a memory and continuity aid rather than a productivity tracker. Contributors are not required to “keep up” with institutional tempo in order to remain legible or valued.

7.3 Supporting Intermittent and Episodic Contribution

Non-linear contribution often occurs in bursts: periods of high engagement followed by withdrawal. AI systems may assist institutions in recognising, accommodating, and valuing such patterns without translating them into expectations of availability or obligation.

This includes:

- smoothing transitions in and out of participation,
- preserving institutional knowledge contributed during active phases,
- and preventing loss of standing due to absence.

By buffering institutions against variability, AI allows contributors to remain agents rather than becoming liabilities to be managed.

7.4 Avoiding the Re-Imposition of Normative Tempo

A critical risk in assistive automation is the re-introduction of normative tempo: the subtle pressure to conform to machine-friendly rhythms. The framework explicitly prohibits the use of AI to normalise pace, cadence, or responsiveness as implicit performance standards.

AI may adapt to human variability; humans are not required to adapt to AI tempo.

7.5 Dignity Through Accommodation, Not Accommodation as Exception

By treating variability as a normal condition rather than an exception case, AI-assisted coordination supports dignity without recourse to special pleading or categorisation. Contributors are not required to disclose reasons for non-linearity, nor to justify capacity limits.

This design aligns directly with the Human Value and Meaning System by ensuring that participation support does not become conditional recognition.

7.6 Contribution Without Exhaustion

Ultimately, positioning AI as a stamina substitute allows institutions to benefit from diverse forms of contribution that would otherwise be lost—not because contributors lack value, but because systems lack patience.

In absorbing coordination load, preserving continuity, and accommodating fluctuation, AI enables contribution without exhaustion. It strengthens participation while preserving exit, refusal, and recovery as legitimate states.

8. Lessons from Historical Technology Use

Technological shifts do not become civilisational in themselves; they become civilisational in how authority is distributed around them. The question is not whether a tool is powerful, but whether its power is correctly placed.

Artificial intelligence is often framed as unprecedented. While its computational capacity is novel in scale, the institutional challenge it presents is not. History offers repeated examples of technologies that amplified human capability while simultaneously threatening to displace human authority. In each case, stability depended not on suppressing the technology, but on constitutionalising its role.

8.1 The Printing Press: Amplification Without Authorship

The printing press multiplied the reach of ideas but did not replace authorship. It enabled the distribution of arguments at scale while preserving responsibility in named individuals. Societies that confused dissemination with authority experienced informational disorder; those that preserved attribution and accountability integrated the tool without surrendering judgement.

Artificial intelligence similarly amplifies analytical reach. The constitutional placement defined in this paper ensures that amplification does not become authorship.

8.2 Calculative Instruments: Precision Without Normative Authority

From double-entry bookkeeping to actuarial tables and statistical modelling, calculative tools have long informed governance decisions. They improved precision and exposed patterns beyond unaided perception. However, they were never granted the authority to determine moral

or political outcomes autonomously. Where such instruments were treated as self-authorising—particularly in risk scoring and predictive governance—public trust deteriorated.

AI extends calculative precision into domains previously reserved for human interpretation. The lesson from earlier tools remains applicable: increased accuracy does not equate to normative legitimacy.

8.3 Bureaucracy and the Illusion of Neutrality

Modern bureaucracies were designed to reduce arbitrariness through rule-bound administration. Yet history shows that procedural systems can become opaque and unaccountable when rules are mistaken for justice. AI systems risk repeating this pattern at computational scale: outputs may appear neutral while embedding hidden assumptions or optimising toward narrow objectives.

The safeguard is not to reject procedural assistance, but to ensure that procedural outputs remain subject to accountable human discretion.

8.4 Technology as Tool, Not Governor

In each historical case, durable integration occurred when societies distinguished between:

- tools that extend capacity, and
- authorities that confer legitimacy.

When this boundary blurred, power shifted away from accountable actors toward systems of abstraction. The stability of governance depended on restoring the asymmetry: tools may inform, but only humans may authorise.

Artificial intelligence represents a continuation of this pattern. Its scale and speed magnify the stakes, but not the underlying constitutional question. The Engagement Credit Economy framework therefore situates AI not as a governor, but as a tool—powerful, assistive, and bounded.

8.5 Avoiding Technological Exceptionalism

A recurring error in periods of technological transition is exceptionalism: the belief that a new tool is so transformative that existing principles no longer apply. Such reasoning often justifies the quiet suspension of established safeguards in the name of inevitability.

The framework advanced here resists that temptation. It treats AI as continuous with prior technological shifts in one crucial respect: it must be constitutionally placed within human systems of authority. The novelty of computational capability does not dissolve the necessity of human judgement.

9. Guardrails Against Automation Creep

The most significant risk posed by artificial intelligence in governance is not overt displacement of human authority, but gradual erosion through convenience, scale, and normalisation. Automation creep rarely occurs through explicit policy choice. It emerges incrementally, as assistive systems become embedded in workflows and their outputs are increasingly treated as default.

Preventing such drift requires structural guardrails rather than reliance on institutional goodwill or ethical intent. This section defines mechanisms designed to detect, constrain, and reverse automation creep before authority migrates away from human judgement.

9.1 Explicit Scope Declaration and Use-Limits

All AI systems deployed within the Engagement Credit Economy framework must operate under a publicly documented scope declaration. This declaration specifies:

- the functions the system is permitted to perform,
- the decisions it is prohibited from influencing,
- the domains in which it may be consulted,
- and the conditions under which its use must be suspended.

Scope declarations are binding rather than descriptive. Any expansion beyond declared scope constitutes a governance breach requiring review by an independent authority.

9.2 Periodic Review and Sunset Clauses

To prevent permanent entrenchment, AI systems must be subject to mandatory periodic review. Reviews assess not only technical performance, but institutional impact, including:

- changes in override frequency,
- shifts in deliberation patterns,
- increased dependency on system outputs,
- and evidence of behavioural or procedural compression.

Where risks are identified, systems may be modified, paused, or withdrawn. Sunset clauses ensure that no AI deployment becomes structurally irreversible through inertia alone.

9.3 Monitoring for De Facto Delegation

Automation creep often manifests as de facto delegation: decisions remain formally human while becoming functionally algorithmic. Indicators include automatic acceptance of recommendations, declining rates of challenge, and time pressures that make refusal impractical.

AI systems may be tasked with monitoring these indicators precisely to detect their own overreach. Such reflexive monitoring functions are directed toward preserving the authority asymmetry defined in Sections 2 and 3.

9.4 Separation of Recommendation and Execution Layers

A critical architectural safeguard is the strict separation between systems that generate recommendations and systems that execute actions. AI outputs must never directly trigger operational processes without an intervening human decision point.

This separation prevents the gradual compression of deliberation into execution pipelines and ensures that authority cannot migrate through workflow design.

9.5 Institutional Oversight and Escalation Pathways

Guardrails require enforcement mechanisms. Detected instances of automation creep must trigger escalation to designated governance bodies, including investigatory officers and independent commissions as defined in Paper V.

These bodies are empowered to audit systems, require modifications, and impose suspension where necessary. Crucially, they operate independently of the institutions deploying the AI, preventing conflicts of interest.

9.6 Drift as a Governance Signal

Finally, the framework treats automation creep not as a technical failure, but as a governance signal. Drift indicates institutional strain, overload, or loss of deliberative capacity. Addressing it may require organisational redesign, capacity investment, or procedural simplification rather than technical correction.

No institutional safeguard is immune to capture or erosion over time. The framework therefore treats contestability, separation, and refusal not as guarantees of perfection, but as mechanisms for making authority drift visible, disputable, and correctable rather than silent.

By interpreting creep as a systemic warning rather than a bug, the framework aligns technological governance with institutional health.

10. Conclusion: AI as Prosthesis, Not Governor

Artificial intelligence introduces unprecedented capacity for pattern recognition, coordination, and scale. Yet the central challenge it poses to governance is not technical but constitutional: where authority resides when systems outperform humans cognitively, and how legitimacy is preserved under conditions of complexity.

This paper has argued that artificial intelligence must be positioned not as a governor, adjudicator, or surrogate decision-maker, but as a **coordination prosthesis**—a tool that amplifies human judgement without replacing it. Within the Engagement Credit Economy framework, AI is deliberately constrained to roles that increase intelligibility, reduce administrative strain, and protect institutional boundaries, while remaining structurally incapable of bearing legitimacy.

The core design principles advanced here are intentionally restrictive. Authority is non-delegable. Judgement cannot be optimised away. Where human decision-making is unresolved, systems must fail productively rather than complete the task on humanity's behalf. The **Legitimacy Fallback Principle** and the **Kobayashi Maru Constraint** formalise this asymmetry, ensuring that artificial intelligence may iterate indefinitely, but only humans may conclude.

By treating override, veto, and refusal as first-class features, the framework preserves human agency under pressure. By orienting AI toward safeguarding rather than enforcement, it prevents voluntary participation from drifting into soft coercion. By supporting non-linear contributors through stamina substitution rather than behavioural discipline, it accommodates real human variability without reintroducing normative tempo. By embedding guardrails against automation creep, it protects governance systems from erosion through convenience rather than choice.

Taken together, these constraints do not weaken institutional capacity; they strengthen it. They ensure that governance remains accountable, contestable, and humane even as coordination demands exceed unaided human limits. Artificial intelligence becomes a means of sustaining deliberation rather than eliminating it, and complexity becomes a condition to be navigated rather than an excuse for delegation.

The significance of this placement extends beyond artificial intelligence itself. It establishes a durable principle applicable to future optimisation regimes, automated infrastructures, and institutional technologies yet to emerge: **when systems surpass human capability, authority must still terminate in human agency**. This principle is not anti-technological; it is pro-civilisational.

As societies transition beyond labour's monopoly over legitimacy and income, governance systems will be tested not by their efficiency, but by their capacity to preserve dignity, refusal, and judgement under strain. Artificial intelligence, correctly placed, can assist in that task. Incorrectly placed, it will undermine it.

This paper argues that the difference is not a matter of trust in machines, but of design. Where artificial intelligence is constitutionally bounded, human governance endures. Where it is not, legitimacy erodes quietly, long before it is noticed.

Acknowledgements

The author wishes to acknowledge **Gene Roddenberry**, whose work anticipated many of the ethical and institutional questions explored in this paper. Roddenberry's distinctive contribution lay not in technological prediction, but in his sustained exploration of authority, judgement, and moral agency under conditions of advanced capability.

By grounding speculative futures in classical literature, legal reasoning, and humanist philosophy, Roddenberry created narrative laboratories in which dilemmas of governance, refusal, legitimacy, and constraint could be examined without recourse to technological determinism. That juxtaposition—advanced technology framed through enduring human

questions—has generated generations of scenarios and insights that continue to inform serious thinking about human–machine relations.

The “Kobayashi Maru” scenario, in particular, remains a concise illustration of the limits of optimisation in the presence of moral indeterminacy, and its influence on the framing of the Legitimacy Fallback Principle in this paper is acknowledged with respect.

References

- Arendt, H. (1958). *The Human Condition*. University of Chicago Press.
- Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.
- Føllesdal, A. (1998). Survey article: Subsidiarity. *Journal of Political Philosophy*, 6(2), 190–218.
- Polanyi, K. (1944). *The Great Transformation: The Political and Economic Origins of Our Time*. Beacon Press.
- Purcell, M. (2006). Urban democracy and the local trap. *Urban Studies*, 43(11), 1921–1941.
- Williamson, O. E. (1985). *The Economic Institutions of Capitalism*. Free Press.
- Wiener, N. (1950). *The Human Use of Human Beings: Cybernetics and Society*. Houghton Mifflin.
- Wolch, J. R. (1990). *The Shadow State: Government and Voluntary Sector in Transition*. Foundation Center.
- Floridi, L. (2019). Translating principles into practices of digital ethics. *Philosophy & Technology*, 32, 1–22.
- O’Neil, C. (2016). *Weapons of Math Destruction*. Crown.
- Suchman, L. (2007). *Human–Machine Reconfigurations*. Cambridge University Press.