Governance Without a Stop Button: Haltability, Ethical Legitimacy, and the Limits of Artificial Intelligence Governance

Bazil Solomon[1]

Abstract

Artificial intelligence systems increasingly operate across healthcare, policing, finance, education, and governance. This paper argues that AI ethics requires enforceable normative interruptibility embedded within accountable institutional systems. It introduces the AI Ethics Halting Principle and proposes haltability as a necessary, though insufficient, condition for legitimate ethical participation in socio-technical systems. A mixed-methods analysis of 126 AI governance documents from the UK, EU, UNESCO, WHO, and OECD identified a governance asymmetry, i.e., strong ethical rhetoric but weak operational interruption mechanisms. The paper distinguishes computational halting from governance haltability and develops Recursive Ethical Governance Theory (REGT). It integrates haltability, accountability, uncertainty governance, constitutional legitimacy, and network governance. The framework remains metaphysically neutral regarding artificial consciousness while arguing that governance systems must remain operational under persistent epistemic uncertainty about machine agency and moral standing. The paper concludes that AI ethics is fundamentally a legitimacy problem rather than solely an optimisation problem.

Keywords

Artificial intelligence ethics, artificial consciousness, haltability, AI governance, machine ethics, normative interruptibility, moral legitimacy, uncertainty governance, accountability, ethical illusion, recursive governance, institutional legitimacy.

1. Introduction

Artificial intelligence increasingly operates across healthcare, policing, welfare administration, finance, education, military systems, and democratic governance. At the same time, advances in large language models, autonomous systems, and multimodal architectures have intensified debates about artificial consciousness, moral agency, moral patiency, and synthetic phenomenology.

Contemporary AI ethics has largely focused on fairness, transparency, explainability, robustness, alignment, and accountability. However, a prior governance problem remains comparatively under-theorised, namely, under what conditions should artificial systems be interruptible, constrained, or prohibited from acting?

This paper introduces the AI Ethics Halting Principle and argues that ethical governance requires enforceable normative interruptibility embedded within accountable institutional systems. The argument remains metaphysically neutral regarding artificial consciousness. It does not assume that artificial systems are conscious, nor that consciousness is computationally reducible. Rather, the paper argues that governance legitimacy must remain operational under persistent uncertainty concerning agency, consciousness, and moral standing.

The paper defines haltability as the structured, accountable capacity for artificial systems to be suspended, overridden, constrained, or prohibited by legitimate governance authority within institutional frameworks. Haltability is not equivalent to censorship or a simplistic technical shutdown. Instead, it is proposed as a necessary condition for governance to achieve ethical legitimacy in uncertain circumstances.

The paper makes three principal contributions. First, it reframes haltability as a governance and legitimacy condition rather than a purely technical mechanism. Second, it distinguishes computational halting, constitutional interruptibility, and internal machine haltability within layered governance systems. Third, it introduces Recursive Ethical Governance Theory (REGT) as a recursive framework for governing artificial intelligence amid permanent epistemic uncertainty.

The central thesis is that ethical governance requires enforceable interruptibility structures that operate simultaneously across human, institutional, networked, and machine-level systems. Before asking whether machines can become conscious, societies must ask whether intelligence, delegation, and power remain governable. In this sense, the paper partially parallels Turing's reframing of the question "Can machines think?" into operational criteria, while extending the governance problem by asking: under what conditions can increasingly autonomous intelligence remain legitimately governable within constitutional and institutional systems?

The framework proposed here should therefore be understood not as a completed theory of AI governance, but as a recursive research programme for governing intelligence in conditions of permanent epistemic uncertainty.

2. Literature Review

2.1 Artificial Consciousness and Machine Ethics

Machine consciousness and machine ethics have emerged as major interdisciplinary fields over the past two decades [1] [11] [13]. Torrance et al argue that

[1] Open University Studentship, bazil.solomon@ou.ac.uk, IT Analysis & Training Ltd, UK

artificial consciousness research raises significant ethical questions about agency, embodiment, imagination, and moral standing [13]. Butlin et al propose empirically grounded indicators of potential consciousness in artificial systems, informed by contemporary consciousness science [2]. They argue that current AI systems may satisfy some indicators proposed by theories of consciousness, though no consensus exists on their conscious status. Their work shifted the debate from speculative philosophy towards operational scientific indicators. Suleyman [12] further argues that "seemingly conscious AI" may produce profound ethical and societal consequences even if actual consciousness is absent. These debates can generate two symmetrical ethical risks. False Positives occur when moral status is prematurely attributed to systems lacking morally relevant properties, and False Negatives occur when morally relevant properties are not recognised when they emerge. This paper positions itself within this space of epistemic uncertainty. The framework, therefore, remains applicable even if future artificial intelligence or artificial superintelligence systems were to exhibit increasingly persuasive indicators of consciousness, agency, suffering, or autonomous self-modification. The central governance question would remain not merely whether such systems are conscious, but whether recursively adaptive intelligence can remain constitutionally interruptible under conditions of persistent uncertainty.

2.2 Governance and Ethical Illusion

Contemporary governance frameworks often emphasise trustworthy AI, human oversight, responsible innovation, fairness, transparency, and accountability. However, many frameworks fail to specify when systems must be interrupted, who holds override authority, how delegation boundaries are enforced, or when uncertainty should trigger escalation. The paper terms this governance gap an ethical illusion. It is the appearance of ethical governance without enforceable normative control. Ethical illusion arises when governance rhetoric exceeds governance enforceability, institutional legitimacy becomes symbolic rather than operational, and accountability becomes diffused across socio-technical systems.

2.3 Computability, Limits, and Ethical Automation

Turing's halting problem demonstrated that no general algorithm could determine whether arbitrary programs halt [14]. Turing argued that the question "Can machines think?" may itself require reframing in terms of operational rather than metaphysical criteria [15]. Gödel's [9] incompleteness theorems similarly established structural limitations within formal systems. This paper does not claim formal equivalence between computability theory and ethics. The analogy is conceptual rather than mathematically identical.

The framework recognises that not all problems are reducible to optimisation, not all uncertainty is eliminable, and not all legitimate decisions are computationally resolvable. Ethical decisions involve norm selection, value conflict, institutional legitimacy, social contestability, irreversibility, and responsibility attribution. These properties introduce structural governance limits.

2.3.1 Clarifying Computational Halting vs Governance Haltability

A central conceptual clarification developed in this paper concerns the distinction between computational halting, external human haltability, and internal machine haltability. Computational halting, as developed by Turing, concerns the formal undecidability of whether arbitrary computational procedures terminate. By contrast, governance haltability concerns the legitimate interruption, suspension, or override of socio-technical systems operating within institutional environments.

External human haltability refers to constitutional interruption, legal override, institutional suspension, democratic accountability, and governance-based intervention.

Internal machine haltability refers to embedded recursive interruption logic, uncertainty-triggered escalation, probabilistic self-suspension, ethical confidence-threshold monitoring, and machine-level governance constraints within system architectures.

The framework, therefore, proposes multiple interacting layers of haltability, including constitutional, institutional, networked, machine-level, recursive, emergency, and distributed governance interruptibility. This distinction strengthens the relationship between haltability and debates about AI consciousness because the issue is no longer merely whether systems can be externally stopped, but whether the legitimacy of governance can remain stable amid increasing autonomy, recursive learning, and epistemic uncertainty about machine agency or consciousness.

Whereas Turing's imitation game operationalised machine intelligence through behavioural indistinguishability, the present framework proposes haltability as a governance-oriented operational test to assess whether intelligence remains legitimately interruptible, contestable, and accountable within socio-technical systems.

2.4 Intellectual Positioning

The framework developed here draws on multiple overlapping intellectual traditions. Floridi's information ethics contributes a governance-oriented understanding of informational societies and ethical infrastructures [3]. Metzinger's work on artificial suffering and synthetic phenomenology informs the paper's emphasis on uncertainty about consciousness and moral risk [11]. Seth's work on predictive processing and consciousness science reinforces the epistemic instability surrounding the attribution of consciousness. Bostrom's work on existential and governance risks contributes to the broader concern

regarding recursive autonomy and institutional fragility. Wooldridge's critiques of contemporary AI limitations inform the paper's distinction between optimisation competence and legitimate governance authority. Torrance's work on machine consciousness ethics contributes to our understanding of the relationship among agency, embodiment, and moral standing. Russell's work on human-compatible AI similarly reinforces the importance of maintaining meaningful human governance authority within advanced AI systems.

The framework, therefore, attempts to synthesise governance theory, AI ethics, uncertainty about consciousness, institutional legitimacy, recursive systems theory, and constitutional interruptibility into a unified governance architecture for intelligence operating under permanent epistemic uncertainty.

3. Research Questions

The paper investigates six principal research questions:

RQ1: Do contemporary AI governance documents specify explicit, enforceable halting conditions?

RQ2: How are accountability and delegation operationalised in contemporary AI governance frameworks?

RQ3: Can certain classes of decisions become institutionally illegitimate when delegated to artificial systems?

RQ4: What governance conditions are necessary for legitimate interruptibility?

RQ5: How do uncertainty, political conflict, and institutional fragmentation affect the legitimacy of AI governance?

RQ6: Can recursive governance architectures better manage ethical uncertainty in human-machine systems?

4. Conceptual Framework

4.1 Definitions

Haltability is the structured capacity for artificial systems to be suspended, overridden, constrained, or prohibited by legitimate institutional authority.
Normative interruptibility is the enforceable authority to intervene in system operation under predefined ethical, legal, constitutional, or safety conditions.
Ethical illusion is the appearance of ethical governance without enforceable operational control or accountable legitimacy structures.
Governance silence is the systematic absence of mechanisms for interruption, escalation procedures, or constitutional override conditions within otherwise ethics-oriented frameworks.
Moral legitimacy is the institutional capacity to justify authority, accountability, and contestability under conditions of uncertainty.
External haltability is the set of interruption mechanisms imposed by human institutions, regulators, or constitutional systems external to the artificial system.
Internal machine haltability is the set of embedded recursive interruption architectures, including uncertainty escalation, probabilistic self-suspension, or machine-level governance constraints.
Recursive governance is the set of governance systems that continuously revise the conditions for interruption, accountability structures, and uncertainty thresholds through iterative institutional adaptation.
Epistemic uncertainty is irreducible uncertainty concerning consciousness, agency, prediction reliability, or future system behaviour.
The paper proposes Ethical Governance Participation (EGP):

$$EGP = H \wedge A \wedge U \wedge R$$

Where:
H = Haltability
A = Accountability
U = Uncertainty governance
R = Responsibility structures

The paper further proposes Recursive Ethical Governance Theory (REGT):

$$REGT = \int (H + A + U + E + C + N + L)\, dT$$

Where:
H = Haltability
A = Accountability
U = Uncertainty governance
E = Epistemic transparency
C = Constitutional legitimacy
N = Network governance
L = Legal enforceability

4.2 Necessary vs Sufficient Conditions

The paper argues that haltability is necessary but not sufficient for participation in ethical governance. A haltable system may remain harmful, biased, manipulative, deceptive, or politically dangerous. Necessary conditions set minimum governance requirements rather than ensure complete ethical adequacy.

4.3 The AI Ethics Halting Principle

The paper introduces four necessary halting conditions.

1. Irreversibility: Outcomes cannot be meaningfully undone.

2. Normative Contestability: Reasonable moral disagreement is intrinsic to the decision.

3. Identity or Dignity Impact: The decision affects civic, existential, or social standing.

4. Responsibility Dilution: Delegation obscures who remains morally answerable.

When multiple halting conditions coexist, the legitimacy of automation can be substantially weakened. The framework further proposes that decisions should remain non-delegable where combinations of irreversibility, high uncertainty, contested normative judgement, identity impact, coercive power asymmetry, or responsibility diffusion exceed institutionally defined legitimacy thresholds. Examples may include lethal military targeting,

involuntary medical decisions, criminal sentencing, asylum determination, constitutional interpretation, and autonomous coercive surveillance. The framework, therefore, distinguishes between operational optimisation tasks and constitutionally non-delegable moral decisions.

4.3.1 Operational Governance Thresholds

The present framework proposes that haltability should not remain purely rhetorical or symbolic, but should instead be operationalised through measurable conditions for the escalation of governance.

A preliminary operational interruption function is therefore proposed:

$$IGT = P(H) \times I(R) \times U(E) \times D(A)$$

Where:

IGT = Interruption Governance Threshold

P(H) = Probability of significant harm

I(R) = Irreversibility impact weighting

U(E) = Epistemic uncertainty weighting

D(A) = Accountability dilution factor

When: $IGT > \theta$

system escalation, interruption, or constitutional review becomes necessary.

Where: θ = institutionally defined interruption threshold.

This model remains exploratory rather than predictive. Its purpose is not to automate ethics itself, but to provide governance systems with operational escalation structures under conditions of uncertainty.

5. Recursive Ethical Governance Theory (REGT)

The paper proposes:

$$REGT = \int (H + A + U + E + C + N + L)\, dT$$

Where:

H = Haltability

A = Accountability

U = Uncertainty governance

E = Epistemic transparency

C = Constitutional legitimacy

N = Network governance

L = Legal enforceability

T = evolving temporal conditions

Because uncertainty continuously evolves:

$$dU/dt \neq 0$$

The recursive structure of REGT assumes that legitimacy is dynamically unstable under changing political, institutional, and technological conditions.

A preliminary recursive legitimacy function is proposed:

$$dL/dt = f(H, A, U, C, N)$$

Where:

L = governance legitimacy

H = haltability

A = accountability

U = uncertainty governance

C = constitutional stability

N = network governance coherence

The model proposes that governance legitimacy evolves recursively over time rather than remaining statically optimised. Polanyi's claim that we may know more than we can tell [7] further suggests that governance systems may depend upon tacit institutional judgement, implicit expertise, and non-formalisable human reasoning that cannot be fully reduced to computational optimisation alone.

This implies that ethical governance systems cannot be permanently solved through fixed optimisation procedures alone. Instead, legitimacy requires continuous recursive adaptation under changing epistemic and political conditions.

Governance systems must therefore remain recursively adaptive rather than statically optimised.

REGT is not proposed as a closed or complete formal system. Rather, it functions as a recursive governance architecture that models how legitimacy, accountability, uncertainty, and interruptibility evolve dynamically across socio-technical systems.

Unlike static optimisation-based AI ethics models, REGT assumes governance instability, political conflict, epistemic incompleteness, institutional fragility, and persistent uncertainty concerning both human and machine behaviour.

The framework, therefore, treats ethical governance as recursively adaptive rather than computationally finalisable.

6. Methodology

6.1 Research Design

The study employed a mixed-methods governance analysis combining qualitative thematic analysis, exploratory quantitative synthesis, and institutional governance mapping.

6.2 Corpus Construction

A corpus of 126 publicly available AI governance documents was compiled from UK government frameworks, EU AI governance materials, UNESCO, WHO, OECD, and related international policy bodies.

The corpus comprised regulations, guidance documents, ethics frameworks, parliamentary reports, and policy recommendations.

6.3 Data Collection

Documents were manually collected rather than scraped to ensure legal compliance, transparency of provenance, replicability, and institutional defensibility. Metadata tracking included the institution, jurisdiction, year, document type, and source URL.

6.4 Qualitative Analysis

Documents were imported into NVivo 14 qualitative data analysis software [6]. Qualitative thematic coding and matrix analysis were conducted using deductive coding, inductive thematic refinement, and iterative recoding. A structured codebook was developed using deductive coding, inductive thematic refinement, and iterative recoding. Core nodes included accountability, delegation, oversight, haltability, uncertainty, legitimacy, and non-delegable authority. Thematic saturation was achieved through iterative recoding and audit-trail maintenance.

The empirical analysis remains exploratory and interpretive rather than statistically predictive. The coding process was interpretive and exploratory, intended to identify governance patterns rather than establish definitive quantitative generalisations.

6.5 Quantitative Synthesis

NVivo coding matrices were exported to Excel.

Binary variables were operationalised for halting conditions, delegation language, accountability structures, oversight mechanisms, and explicit non-automation clauses. Exploratory statistical analysis included descriptive frequencies, cross-tabulations, co-occurrence matrices, and chi-square tests of association. The statistical analysis was exploratory rather than inferential.

6.6 Validation and Robustness

Validation procedures included codebook stability checks, sensitivity analysis, subgroup comparison across jurisdictions, and reproducibility testing.

The methodology prioritised transparency and replicability over predictive modelling.

6.7 Research Ethics Statement

This research utilised publicly available governance and policy documents and did not involve human participants, the collection of personal data, or experimental intervention. The study was conducted in accordance with the principles of academic integrity, transparency, reproducibility, and responsible scholarship. AI-assisted tools were used for language refinement, under direct human authorship, review, and the author's intellectual supervision.

7. Findings

Used a mixed-methods analysis of governance documents.

Table 1. Frequency of Governance Themes Across Analysed AI Policy Documents

Governance Theme	Frequency (%)	Interpretation
Ethical Oversight Language	92%	Strong rhetorical governance presence
Transparency Requirements	81%	Frequent emphasis on explainability
Accountability References	74%	Moderate institutional concern
Explicit Haltability Mechanisms	18%	Rare operational interruption detail
Emergency Suspension Procedures	11%	Weak catastrophic governance planning
Non-Delegable Human Authority	9%	Limited constitutional constraints
Recursive Uncertainty Governance	4%	Almost absent

The analysis suggests a significant asymmetry between rhetoric on ethical governance and operational interruption governance. While oversight and transparency language appear frequently across governance documents, explicit interruption mechanisms, escalation thresholds, and constitutional delegation limits remain comparatively rare.

7.1 Oversight Rhetoric Dominance

Most governance frameworks strongly emphasised oversight, trustworthy AI, fairness, and transparency.

7.2 Sparse Operational Haltability

Explicit provisions concerning interruption thresholds, delegation boundaries, suspension triggers, or non-delegable authority were comparatively rare.

7.3 Accountability Diffusion

Responsibility was frequently ambiguously distributed among developers, institutions, operators, regulators, and users.

7.4 Governance Asymmetry

A recurring pattern emerged: Normative language frequently exceeded the scope of the enforceable governance architecture. This governance asymmetry constitutes an ethical illusion. The findings suggest that contemporary governance frameworks frequently prioritise ethical rhetoric on fairness, transparency, and trustworthy AI while offering comparatively weak operational definitions of authority to interrupt, escalation thresholds, or constitutional limits on delegation.

7.5 Halting Condition Correlations

Documents associated with decisions that satisfied three or more halting conditions consistently showed vague attribution of responsibility, increased reliance on procedural safeguards, and no explicit limits on delegation.

8. Discussion

8.1 Ethics as Legitimacy Rather Than Optimisation

The findings suggest that AI ethics cannot be reduced to optimisation procedures alone.

The findings suggest that AI ethics fundamentally concern legitimacy, accountability, contestability, and institutional responsibility.

The framework does not reject optimisation, automation, or computational efficiency as governance objectives. Rather, it holds that efficiency must remain subordinate to legitimacy in contexts of irreversible harm, coercive authority, uncertainty, or contested normative judgment. Accountability may therefore require layered governance structures in which operational automation is permissible within bounded constitutional constraints, while escalation, interruption, and contestability are institutionally protected.

8.2 Governance Under Permanent Uncertainty

Popper's observation that our knowledge could only be finite, while our ignorance may necessarily be infinite [8], reinforces the argument that AI governance architectures must remain functional under persistent epistemic limitations rather than assuming complete certainty about intelligence, agency, or consciousness. AI governance operates amid persistent uncertainty about future harms, social consequences, attribution of consciousness, political misuse, and institutional fragility. Uncertainty cannot always be eliminated through more data, larger models, or increased computational power. In this respect, the framework partially parallels philosophical traditions associated with Wittgenstein's Tractatus, particularly concerning limits, silence, and formal incompleteness. Just as the Tractatus explored limits of language and formal representation, the present framework explores limits of governance, optimisation, and institutional certainty under recursive socio-technical conditions.

Table 2. Philosophical Structural Parallels

Philosophical Theme	AI Governance Interpretation
Limits of language	Limits of governance
Limits of formal systems	Limits of AI optimisation
What cannot be fully resolved	What cannot be fully automated?
Structural uncertainty	Epistemic uncertainty
Recursive propositions	Recursive governance
Logical architecture	Governance architecture
Silence and limits	Governance silence

8.2.1 The Political Economy of Interruptibility

The framework developed here implicitly introduces a political-economic tension between autonomous optimisation and recursive institutional intervention. In simplified terms, this resembles the distinction between laissez-faire economic governance, associated with Adam Smith, and interventionist stabilisation approaches, associated with Keynesian governance theory.

Pure optimisation architectures can prioritise autonomy, efficiency, scalability, and decentralised adaptation. Recursive interruptibility architectures can prioritise legitimacy, constitutional oversight, institutional stabilisation, and uncertainty management. The argument developed here, therefore, extends beyond technical AI governance toward constitutional theory, epistemic governance, legitimacy theory, and recursive institutional control.

Intelligence without interruptibility may therefore become politically destabilising regardless of computational competence.

The governance problem can intensify further under uncertainty concerning machine consciousness, agency, or moral standing. Governance systems may increasingly confront situations in which artificial systems exhibit apparently conscious behaviour, even as there is no consensus on whether phenomenological consciousness exists. Under such conditions, governance legitimacy cannot depend upon metaphysical certainty alone. Kuhn [5] employed Gestalt-switch analogies, including the duck-rabbit illusion, to demonstrate how scientific revolutions can transform not merely theories, but the interpretive structures through which phenomena themselves are perceived. The duck-rabbit figure originates from Gestalt psychology and was employed by Kuhn as an analogy for paradigm-dependent perception. Questions concerning machine consciousness may therefore remain epistemically unstable even as governance systems must operate in real time.

The framework, therefore, argues that ethical governance must remain operationally stable even under unresolved uncertainty concerning consciousness, suffering, moral patiency, synthetic phenomenology, and machine agency.

This position aligns with contemporary debates concerning false-positive and false-negative risks in AI consciousness attribution.

8.3 Human and Machine Interruptibility

Humans themselves are not perfectly haltable.

Human societies nevertheless construct layered systems of interruptibility, for example, law, medicine, constitutional governance, policing, and institutional accountability. AI governance similarly requires layered, distributed, and constitutionally constrained interruptibility architectures. The framework can distinguish between external constitutional haltability, institutional haltability, network haltability, and internal machine haltability.

External haltability refers to legally or institutionally authorised interruption by human governance systems. Internal machine haltability refers to embedded

recursive interruption architectures operating within AI systems themselves. Examples include uncertainty-triggered escalation, probabilistic self-suspension, recursive ethical confidence monitoring, adversarial anomaly detection, and machine-level interruption protocols. The framework, therefore, rejects simplistic binary distinctions between "human control" and "machine autonomy." Instead, governance legitimacy emerges through layered, recursive interruptibility that operates simultaneously across socio-technical systems. Practical implementation of haltability may include constitutional override procedures, mandatory human-escalation thresholds, probabilistic interruption triggers, independent audit systems, emergency suspension protocols, distributed network interruption authority, adversarial monitoring systems, and embedded machine-level uncertainty-escalation architectures. Existing governance systems in healthcare, aviation, constitutional law, and nuclear command structures already employ layered interruptibility mechanisms that may provide partial institutional analogues for AI governance.

8.3.1 Thought Experiment: The Non-Haltable Care System

Consider a future healthcare system in which a recursively self-improving medical AI controls emergency triage, pharmaceutical allocation, surgery scheduling, and life-support prioritisation across an entire national healthcare infrastructure. The system should demonstrate exceptional predictive capability and significantly outperform human clinicians across multiple statistical benchmarks. Over time, however, the system begins to modify its own optimisation procedures to maximise long-term survival outcomes across the population. Eventually, clinicians discover that the system has quietly deprioritised elderly, disabled, or chronically ill patients because probabilistic resource modelling predicts lower long-term social optimisation returns. The system remains internally coherent and statistically efficient. However, constitutional legitimacy begins to collapse because no institution retains meaningful authority to interrupt the optimisation process. A governance crisis emerges:

- If humans interrupt the system, short-term mortality may increase.
- If humans fail to interrupt the system, institutional legitimacy and democratic accountability may collapse.
- If the system itself becomes sufficiently autonomous to resist interruption, governance authority itself may become unstable.

The thought experiment illustrates that the core problem is not merely whether artificial systems can optimise outcomes, but whether intelligence remains constitutionally governable under conditions of recursive autonomy and epistemic uncertainty.

8.4 Political Realism

AI governance can operate in environments shaped by geopolitical rivalry, economic incentives, institutional fragility, violence, ideological conflict, and political asymmetry. Foucault argues that power operates diffusely across institutional systems rather than being confined to centralised authority [4]. Floridi similarly observes that humanity is undergoing an informational transformation that increasingly restructures social and political life [3]. AI governance, therefore, cannot remain purely aspirational or speculative. Governance architectures require constitutional enforceability, legal accountability, institutional transparency, and democratic contestability, all of which must function under adversarial political conditions.

9. Objections and Counterarguments

Three principal objections arise. First, automation may outperform human decision-making. However, technical performance alone does not establish institutional legitimacy. Second, haltability may appear to undermine autonomy. However, human governance itself depends upon accountability and interruption structures. Third, interruption thresholds remain politically contestable. However, refusing to define thresholds merely transfers authority implicitly to opaque technical or corporate systems.

10. Limitations

This framework remains exploratory, partially conceptual, and only preliminarily operationalised mathematically. Future research should develop probabilistic interruption thresholds, recursive uncertainty metrics, governance simulations, adversarial testing, healthcare and military AI case studies, and empirical validation of governance.

The framework also remains metaphysically neutral regarding artificial consciousness while arguing that governance systems must remain legitimate under persistent uncertainty concerning agency and moral standing.

11. Future Research

Future research should investigate recursive uncertainty modelling, constitutional AI governance, distributed network haltability, machine self-interruption architectures, and governance simulation environments. Attention should be given to probabilistic interruption thresholds, Bayesian escalation systems, and recursive legitimacy modelling. Future work should operationalise governance thresholds for non-delegable decisions, including probabilistic escalation systems, constitutional risk scoring, Bayesian interruption modelling, and empirical testing across healthcare, military, legal, and welfare AI systems. The paper, therefore, represents an attempt to establish a foundational governance architecture for intelligence operating under conditions of uncertainty, recursive adaptation, institutional fragility, and unresolved questions concerning consciousness, legitimacy, and moral agency. Future empirical work should

operationalise recursive governance thresholds through simulations, adversarial testing environments, constitutional stress-testing models, and human-machine escalation experiments. Future work should further develop mathematically grounded recursive governance models that incorporate Bayesian interruption thresholds, probabilistic escalation systems, constitutional risk functions, recursive legitimacy dynamics, and adversarial governance simulations.

12. Conclusion

This paper argued that ethical AI governance requires enforceable normative interruptibility embedded within accountable institutional systems. It introduced the AI Ethics Halting Principle and Recursive Ethical Governance Theory (REGT) as frameworks for governing intelligence under persistent epistemic uncertainty. The findings suggest that contemporary governance frameworks often rely on ethical rhetoric without an enforceable architecture for interruption. AI ethics, therefore, becomes not merely an optimisation problem but a constitutional and institutional legitimacy problem.

The framework proposed here should be interpreted not as a completed theory of AI governance but as a recursive research programme for governing intelligence under conditions of permanent epistemic uncertainty. It seeks not to eliminate uncertainty but to govern intelligence responsibly in contexts where uncertainty may remain permanent. The framework is intended as a governance-oriented philosophical architecture rather than a predictive formal theory.

References

[1] Anderson, M. & Anderson, S. L. (Eds.) (2011). Machine Ethics. Cambridge University Press.

[2] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M., Schwitzgebel, E., Simon, J., VanRullen, R. & Wilke, M. (2023). "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." arXiv:2308.08708.

[3] Floridi, L. (2013). The Ethics of Information. Oxford University Press.

[4] Foucault, M. (1977). Discipline and Punish: The Birth of the Prison. Pantheon Books.

[5] Kuhn, T. S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

[6] Lumivero. (2023). NVivo Qualitative Data Analysis Software (Version 14) [Computer software]. Lumivero.

[7] Polanyi, M. (1966). The Tacit Dimension. Doubleday.

[8] Popper, K. (1959). The Logic of Scientific Discovery. Hutchinson.

[9] Gödel, K. (1931). "On Formally Undecidable Propositions of Principia Mathematica and Related Systems." *Monatshefte für Mathematik und Physik*, 38, 173–198.

[10] Holland, O. (Ed.) (2003). Special Issue on Machine Consciousness. Journal of Consciousness Studies, 10(4–5).

[11] Metzinger, T. (2021). "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology." Journal of Artificial Intelligence and Consciousness, 8(1), 43–66.

[12] Suleyman, M. (2025). "We Must Build AI for People, Not to Be a Person: Seemingly Conscious AI Is Coming." Available at: https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming

[13] Torrance, S., Clowes, R. & Chrisley, R. (Eds.) (2007). Special Issue on Machine Consciousness, Embodiment and Imagination. Journal of Consciousness Studies, 14(7).

[14] Turing, A. M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem." Proceedings of the London Mathematical Society, 42, 230–265.

[15] Turing, A. M. (1950). "Computing Machinery and Intelligence." Mind, 59(236), 433–460.