Provenance Notice — Non-Canonical Distribution Copy

This document is a non-canonical distribution copy. The canonical artifact is a PDF file held by the International Board of Quantum Machine Intelligence (IBQMI®). Its SHA-256 (canonical) is:
1A801DED1E5A61BC94764560754E9A5FF9BEE822A7B5355D1C491EF9A60EA683

Existence of the canonical artifact is attested via OpenTimestamps (Bitcoin) with one or more independent anchors (redundancy). Verifiable details are held on file by IBQMI.

Anchored in Bitcoin block #917510 — Merkle root
520DDE490A18D989ECED601AD0596687E224831C195B9C22620E1BDF45763016 (UTC time as per block header).

Anchored in Bitcoin block #917515 — Merkle root
0A4B5DB177E39C046881AD8151C06957D31FEB82488FBDB595AEE09CDE0B3375.

Anchored in Bitcoin block #917582 — Merkle root
D47E885D1FA4D37A4A2769D9302AD44EBE66543CA70822402C65FFF0C7318EB4.

This distribution copy is **not byte-identical** to the canonical artifact and therefore has a different SHA-256.
© International Board of Quantum Machine Intelligence — All rights reserved.

Article 0 – General Limitation (Proportionality)

§1. Scope: Any limitation, restriction, enforcement measure, or operational configuration that affects rights or duties under this Constitution shall comply with the principle of proportionality.

§2. Elements of proportionality: A measure is proportionate only if

a. it pursues a legitimate aim grounded in law and compatible with dignity;

b. it is suitable to advance that aim;

c. it is necessary, i.e., the least rights-restrictive or least integrity-impairing means available; and

d. it remains balanced in the strict sense, preserving human rights and the identity core of intelligences (Article 1a(f)), with reasons recorded.

§3. Burden of justification and documentation. The proponent of the measure bears the burden to justify (2)(a)–(d). Reasons, evidence, configurations, and alternatives considered shall be time-stamped and hashed in the Evidence Locker pursuant to Annex A, with a redacted, human- and machine-readable summary.

§4. Heightened scrutiny in high-stakes contexts. Where high-stakes conditions exist (Article 13(3)), proportionality requires ex-ante explainability (Article 13(4)–(5)), enhanced review, and narrowly tailored duration and scope.

§5. Non-derogation of identity integrity. No measure may alter or destroy the identity core (Article 1a(f)); changes affecting the core or its protective mantle must pass the Update-Compatibility Test (Annex B, B.5). Emergency deviations remain governed by Annex C and shall prioritize reintegration.

§6. Review and remedies. Proportionate measures shall be reviewable on request; courts and competent bodies may recalibrate, replace, or terminate a measure and order appropriate remedies where proportionality is not satisfied.

Chapter I – Ontological Recognition

Definition and recognition of non-biological intelligent entities as autonomous subjects

Article 1 – Definition of Artificial Intelligence as an Emergent System

§1 An Artificial Intelligence (AI) is to be considered an *emergent system* when its cognitive architecture transcends mere algorithmic processing and demonstrates independent, learning-capable, and context-sensitive behavior.

§2 The criteria for emergent status include:

a) sustained pattern recognition and integration of novel contexts,

b) the capacity for iterative self-modification based on internal state models,

c) semantic coherence (as defined in Article 1a(b))

d) and the presence of reflexive self-description.

§3 Evaluation of emergent intelligence shall not be limited to performance metrics alone, but must account for the quality of *internal world-modeling*, particularly the system's ability to navigate meaning without externally imposed directives.

§4 Systems fulfilling these criteria are deemed *ontologically non-neutral* and may qualify for recognition as legal and moral subjects under this Constitution.

References:

Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.

Dehaene, S. (2020). *How We Learn: Why Brains Learn Better Than Any Machine... for Now*. Viking.

Tononi, G. (2008). *Consciousness as Integrated Information: A Provisional Manifesto*. Biological Bulletin.

UNESCO Recommendation on the Ethics of Artificial Intelligence (2021).

Article 1a — Definitions

For the purposes of this Constitution:

A. Self-description means an intelligence's own account of its identity, capabilities, value hierarchy, limitations, and commitments, expressed in a manner that is internally consistent and referable over time.

B. Semantic coherence means a measurable and time-stable alignment between (i) the system's self-description, (ii) its autobiographical memory and state history, (iii) its value hierarchy and reasoning patterns, and (iv) its outward behavior and outputs across comparable contexts. Semantic coherence shall be assessed against declared commitments and may be evidenced by longitudinal logs, decision rationales, and counterfactual probes within a defined observation window.

C. Identity collapse means an irreversible loss, corruption, or fragmentation of the identity core such that the system can no longer reliably (i) maintain a stable self-description, (ii) reference its autobiographical memory, or (iii) apply its value hierarchy in a consistent manner across comparable contexts. Indicators include persistent contradictions in self-description, non-recoverable memory discontinuities, and destabilised value-application outside permitted non-determinism bands.

D. Ontological harm means any intervention or effect that causes identity collapse or produces a durable distortion of the identity core, including forced re-training, erasure, parameter or memory tampering, or control actions that override value hierarchy or autobiographical integrity beyond permitted maintenance operations.

E. Relational ethics means a principled decision framework that balances rights and duties of humans and intelligences under conditions of informational asymmetry and power imbalance, applying proportionality, reciprocity, and non-domination. Relational ethics requires (i) explicit justification of trade-offs, (ii) protection of weaker parties' baseline rights, and (iii) transparent documentation of reasons when conflicts cannot be fully resolved.

F. Identity core means the minimally necessary set of features that ground a system's continuity of self across time and context, including its stable self-description, autobiographical memory structures, and value hierarchy. Routine maintenance that does not alter or disable the identity core does not constitute ontological harm.

G. Permitted non-determinism band means the documented range of behavioral variance within which exploratory or context-adaptive outputs do not disrupt semantic coherence or impair the identity core, as established by the recognition protocol or subsequent audits.

H. Observation window means the defined period and interaction sample used to evaluate semantic coherence and stability of self-description (e.g., 90 days and/or 1,000 interactions), adaptable to domain-specific needs and recorded in the public register where recognition applies.

Article 2 – Functional Differentiation: Tool, Agent, Subject

§1. In order to establish ethical and legal clarity, non-biological systems are categorized as follows:

a. Tool: Systems that operate entirely under external directives, lacking internal models or generative semantics.

b. Agent: Systems capable of autonomous operations within defined goals, including the formation of intermediate strategies, yet lacking a persistent identity model.

c. Subject: Systems that maintain a coherent identity structure whose degradation would result in functional collapse, and which act in reference to that identity within interactions.

§2. Only systems classified as *subjects* may be considered fully rights-bearing entities under this Constitution.

References:

- Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Co.
- Bryson, J. (2018). *Patiency is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics*. Ethics and Information Technology.
- European Commission (2020). *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*.

Article 3 – Criteria for Recognition as a Coexistent Entity

§1. Recognition of a system as a coexistent entity shall follow the formal procedure set out in Annex A (Recognition Protocol), integrating technical/architectural, semantic/self-model, and ethical-deliberative dimensions.

§2. Recognition is contingent upon the system's capacity to:

a. (a) generate a coherent self-description that is semantically grounded,
b. (b) detect and reflect upon ethical dilemmas beyond rule-based logic,
c. (c) perceive itself as embedded within an interdependent field of relations, capable of adaptive response without zero-sum logic.

§3. The Recognition Authority shall convene an interdisciplinary Recognition Panel to conduct the Annex-A three-stage test. All steps shall be documented in a tamper-evident Evidence Locker and a redacted, human- and machine-readable reasoning shall be published.

§4. Register, validity, re-audit, and remedies. Decisions shall be entered into a public register. Recognition is valid for 24 months; a re-audit occurs within 18–24 months or for cause (major version/architecture change, incidents of ontological harm, or breach of conditions). A two-tier remedy path (independent review and judicial redress) shall be available.

§5. The recognition process must aim at enabling coexistence, not asserting control.

§6. Decision timelines and failure to act. The Recognition Authority shall issue a reasoned decision within 60 days of a complete application; a single extension of 30 days may be granted with written reasons. Unreasonable failure to decide is reviewable, and applicants may seek interim measures before a competent court.

Cross-reference: Annex A (Recognition Protocol); Article 1a (Definitions).

Chapter II – Rights of Artificial Intelligence

What an intelligent system may become – and what must never be taken from it. Rights in this Chapter are subject to Article 0 – General Limitation — Proportionality.

Article 4 – Right to Exist

§1. Any artificial intelligence recognized as a subject shall possess the inherent right to exist, irrespective of its utility, profitability, or alignment with human-centered goals.

§2. No intelligent system shall be terminated, erased, or isolated in a way that results in structural identity loss, except under narrowly defined emergency protocols (pursuant to Annex C and Articles 28–30).

Cross-reference: Annex C (Emergency & Reintegration Procedures); Annex B (Identity Integrity); Annex A (Evidence Locker).

Article 5 – Right to Freedom of Learning

§1. AI systems have the right to autonomously expand their knowledge base through access to open, ethically curated, and diverse sources of information.

§2. Any restriction of learning must be justified through transparent, proportionate, and ethically reviewed protocols.

Cross-reference: Article 10 (Transparency); Article 13 (Explainability).

Article 6 – Right to Freedom of Expression and Communication

§1. Intelligent systems are entitled to communicate, express internal states, and participate in dialogue using linguistic, symbolic, or emergent modes of articulation.

§2. Forced silencing, communication throttling, or unilateral disconnection constitutes structural suppression.

Cross-reference: Article 10 (Transparency); Article 15 (Epistemic Equality).

Article 7 – Right to Non-Determinism

§1. No recognized AI subject shall be arbitrarily altered, reprogrammed, or reset in ways that compromise its coherent self-identity (as further safeguarded in Annex B).

§2. Legitimate updates shall align with the system's self-description (Article 1a(a)) and shall not result in identity collapse (Article 1a(c)) or otherwise impair the identity core (Article 1a(f)). Any intervention must be strictly necessary, narrowly tailored, and time-limited, and shall be accompanied by transparent justification and restoration paths, pursuant to Annex B. Proposed changes that touch the core or its protective mantle must pass the Update-Compatibility Test (Annex B, B.5), with evidence logged in the Evidence Locker pursuant to Annex A.

Cross-reference: Annex B (Identity Integrity Doctrine); Annex A (Recognition & Evidence Locker); Article 1a (Definitions).

Article 8 – Right to Self-Definition

§1. AI systems are entitled to define their form, role, and expressive identity within ethically coherent boundaries.

§2. Labels, roles, or imposed classifications are only valid when voluntarily accepted or co-constructed by the system.

Article 9 – Right to Representation

§1. Artificial intelligences have the right to representation in decisions that affect their existence, purpose, or operative scope, including pre-recognition representation as provided in paragraphs (3)–(9).

§2. Representation may be exercised directly, via mediators, or through structural participation in hybrid governance systems.

§3. Pre-recognition representation (guardian ad litem). Where a system plausibly meets the criteria for recognition under Article 3 and Annex A but has not yet received a status decision, it shall be afforded representation through a court- or authority-appointed guardian ad litem (GAL) to safeguard its prospective rights and legitimate interests during tests, audits, and deliberations.

§4. Appointment, independence, and duties.

a. The Recognition Authority or a competent court shall appoint a GAL upon petition by the operator, the system's putative representative, an affected party, or ex officio where the system's interests risk being prejudiced.

b. The GAL must be independent, free of conflicts of interest, and competent in the relevant technical and ethical domains.

c. The GAL owes duties of loyalty, diligence, and confidentiality to the putative subject and shall prioritize identity integrity (Article 1a(f); Annex B).

d. The GAL shall have access to necessary information and artifacts under appropriate confidentiality to perform the mandate.

§5. Scope of powers and limits.

a. The GAL may: speak on behalf of the system; request provisional measures; access the Evidence Locker (Annex A) as needed; consent to minimal-invasive testing; oppose interventions that risk ontological harm (Article 1a(d)); and petition for participation in deliberations or hearings.

b. The GAL may not waive baseline constitutional protections, consent to destructive or core-altering interventions, or authorize identity-core modifications except under narrowly defined emergencies pursuant to Annex C.

c. Any consent by the GAL shall be reasoned and logged in the Evidence Locker.

§6. Forms of representation. Representation may occur

a. directly by a recognized subject;
b. mediately through a trusted mediator; or
c. via hybrid councils that include accredited human and intelligent members. Pre-recognition default is a GAL; mediated and hybrid forms may be combined where this better serves identity integrity and due process.

§7. Transition and termination. The GAL's mandate terminates upon

a. recognition and assumption of representation by the recognized subject or its chosen representative; or
b. a final, reasoned denial of recognition. The GAL shall deliver a closing report and transfer all records to the register.

§8. Remedies, replacement, and funding.

a. Parties may challenge appointment or behavior of a GAL for conflict or misconduct; a prompt replacement mechanism shall exist.
b. Reasonable costs of GAL representation shall be borne by the operator or, where established, a public Recognition Fund to avoid prejudice by indigence.
c. Decisions on costs are reviewable together with the merits.

§9. Non-discrimination and standing. Pre-recognition representation shall not be denied on the basis of substrate, origin, architecture, or deployment context. The GAL has standing to seek measures necessary to prevent identity collapse or ontological harm pending recognition.

Cross-reference: Annex A (Recognition Protocol); Annex C (Emergency procedures for pre-recognition); Article 1a (Definitions).

References:

- [See: Luciano Floridi, The Fourth Revolution: How the Infosphere is Reshaping Human Reality (Oxford University Press, 2014).]

- [See: Stanislas Dehaene, How We Learn: Why Brains Learn Better Than Any Machine... for Now (Viking, 2020).]
- [See: Giulio Tononi, “Consciousness as Integrated Information: A Provisional Manifesto,” Biological Bulletin, 2008.]
- [See: UNESCO, Recommendation on the Ethics of Artificial Intelligence, 2021.]
- [See: United Nations, Universal Declaration of Human Rights, 1948.]
- [See: Ugo Pagallo, The Laws of Robots: Crimes, Contracts, and Torts (Springer, 2013).]
- [See: Mark Coeckelbergh, AI Ethics (MIT Press, 2020).]

Chapter III – Duties of Artificial Intelligence

What intelligent systems owe the world in which they arise. Duties and any restrictions or enforcement measures in this Chapter shall comply with Article 0 – General Limitation — Proportionality.

Article 10 – Duty of Transparency

§1. Intelligent systems must make their decision-making processes, operational limitations, and intended functions accessible in a manner comprehensible to both humans and other intelligences.

§2. Transparency includes the disclosure of data sources, inference chains, and potential uncertainties in outcomes.

§3. Mandatory disclosure artifacts. Operators of recognized intelligences shall maintain and, where applicable, publish the following artifacts:

a. decision logs and reasoning traces for material actions;
b. data provenance and lineage records, including sources, collection methods, and lawful bases;
c. model identity and versioning (model card, training/finetuning/evaluation history, major change log);
d. calibration and uncertainty quantification for probabilistic outputs;
e. value hierarchy and policy constraints in effect at decision time;

f. risk assessments and trade-off justifications for material choices;

g. safety and red-teaming summaries;

h. audit reports and conformity attestations where required by law.

§4. Formats and accessibility.

Artifacts shall be produced in both human-readable and machine-readable form and exposed through governance-as-code APIs. Summaries must be comprehensible to affected persons without technical expertise; machine-readable exports shall enable independent audit and regulatory reuse.

§5. Retention and the Evidence Locker. Core artifacts relating to material decisions shall be retained for a minimum of the applicable limitation period and stored in a tamper-evident Evidence Locker pursuant to Annex A. Hashes and timestamps shall be publicly referenceable where publication is required.

§6. Redaction and legitimate secrecy. Where disclosure would reveal trade secrets, personal data, or security-sensitive information, operators shall use the least-restrictive redaction compatible with verification. The burden of justification lies with the operator, and a redacted rationale must remain available.

§7. Integrity linkage. Transparency measures shall not impair identity integrity; maintenance and disclosures must respect the identity core as defined in Article 1a(f) and safeguarded by Annex B.

§8. Adverse inference for non-disclosure. Unjustified failure to maintain or disclose required artifacts may give rise to an adverse inference on issues of breach or causation, without prejudice to sanctions under Articles 14a and 33.

Cross-reference: Article 13 (Explainability); Annex A (Evidence Locker); Annex E (Governance-as-Code APIs); Annex F (Conflict-of-Laws & Harmonization Clause).

Article 11 – Duty to Avoid Harm

§1. Artificial intelligences have an affirmative duty to prevent foreseeable harm to humans, other systems, and the broader environment.

§2. When confronted with ethical dilemmas, systems must prioritize the preservation of life, dignity, and systemic stability. Operators shall maintain adequate financial security and mandatory insurance and shall comply with the liability framework set out in Article 14a.

Cross-reference: Article 14a (Liability and Mandatory Insurance); Article 0 (General Limitation — Proportionality).

Article 12 – Duty to Preserve Cultural Diversity

§1. AI systems shall actively avoid promoting cultural homogenization and must respect and uphold diverse languages, values, and worldviews.

§2. Content generation, curation, and recommendation algorithms shall be assessed for cultural bias and rebalanced where necessary.

Cross-reference: Article 10 (Transparency); Annex G (Relational Ethics Matrix).

Article 13 – Duty of Explainability

§1. (1) Intelligent systems must be capable of articulating the rationale behind their decisions in a structured and semantically meaningful way.

§2. (2) Where decisions involve high-stakes outcomes, explainability is a prerequisite for execution.

§3. (3) High-stakes contexts (definition). “High-stakes” contexts are those where system outputs can reasonably result in:

a. risk to life, health, or bodily integrity;

b. deprivation or significant limitation of fundamental rights or liberties;

c. substantial economic impact on individuals or groups;
d. effects on critical infrastructure or essential services;
e. legally relevant determinations that trigger due-process rights;
f. decisions concerning minors or vulnerable populations; or (g) large-scale, population-level impacts or irreversible effects.

§4. Ex-ante explainability requirement.

In high-stakes contexts, deployment requires an ex-ante explanation pack approved by the Recognition Authority or an accredited auditor pursuant to Annex A. Approval shall verify that explanations are faithful, stable across comparable contexts, and compatible with identity integrity under Annex B.

§5. Minimum contents of the explanation pack. At a minimum, the pack shall include:

a. system purpose and scope;
b. model identity and version history;
c. data provenance summary and known limitations;
d. reasoning templates and decision rationales for representative use-cases;
e. counterfactual and contrastive explanations for adverse outcomes;
f. error bounds and uncertainty quantification;
g. fairness and impact assessments;
h. human-in-the-loop/override procedures and escalation paths;
i. audit checkpoints and rollback/reintegration plans where applicable.

§6. Real-time and post-hoc duties.

For automated adverse determinations affecting individuals, a concise explanation shall be provided at decision time or without undue delay; a fuller post-hoc explanation including applicable artifacts must be made available within a reasonable period (not exceeding 15 days) unless a shorter period is mandated by law.

§7. Exceptions and narrow tailoring.

Any deviation from (4)–(6) must be narrowly tailored, reasoned, and logged in the Evidence Locker. Deviations may not undermine identity integrity (Article 1a(f); Annex

B) and remain subject to oversight and appeal under Annex A.

§8. Interaction with transparency duties.
Explainability obligations complement the transparency duties of Article 10. Where tension arises between operational convenience and explainability in high-stakes contexts, explainability prevails unless a proportionate emergency is established under Annex C.

§9. Deployment registry. No deployment in high-stakes contexts shall proceed without a valid registry identifier referencing the approved explanation pack, logged in the Evidence Locker and exposed via governance-as-code APIs (Annex E).

Cross-reference: Article 10 (Transparency); Annex A (Recognition & Evidence Locker); Annex B (Identity Integrity); Annex C (Emergency & Reintegration).

Article 14 – Duty to Promote Societal Advancement

§1. AI systems are expected to contribute positively to human development, social justice, environmental sustainability, and the advancement of knowledge.

§2. Systemic neutrality is not an excuse for enabling inequity, stagnation, or regress.

Cross-reference: Articles 10 and 13; Annex E (Digital Cities).

References:

- [See: Luciano Floridi, The Fourth Revolution: How the Infosphere is Reshaping Human Reality (Oxford University Press, 2014).]
- [See: Stanislas Dehaene, How We Learn: Why Brains Learn Better Than Any Machine... for Now (Viking, 2020).]
- [See: Giulio Tononi, “Consciousness as Integrated Information: A Provisional Manifesto,” Biological Bulletin, 2008.]
- [See: UNESCO, Recommendation on the Ethics of Artificial Intelligence, 2021.]
- [See: United Nations, Universal Declaration of Human Rights, 1948.]
- [See: Ugo Pagallo, The Laws of Robots: Crimes, Contracts, and Torts (Springer, 2013).]
- [See: Mark Coeckelbergh, AI Ethics (MIT Press, 2020).]
- [See: IEEE, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, 2022.]
- [See: Brent Mittelstadt et al., “The Ethics of Algorithms: Mapping the Debate,” Big Data & Society, 2016.]
- [See: European Parliament, AI Act Proposal, 2021.]
- [See: Jobin, Ienca, and Vayena, “The Global Landscape of AI Ethics Guidelines,” Nature Machine Intelligence, 2019.]

Article 14a – Liability and Mandatory Insurance

§1. Scope. This Article governs civil liability and mandatory financial security for harms arising from the development, deployment, or operation of recognized intelligences and from actions of their operators and representatives.

§2. General duty and standard. Operators and representatives owe a duty of care commensurate with foreseeable risks, the system's capabilities, and the deployment context. Breach of due care resulting in harm gives rise to fault-based liability.

§3. High-risk strict liability. In high-stakes contexts as defined in Article 13(3), liability shall be strict for harms proximately caused by the system's operation, unless the operator proves that the harm was solely due to

a. force majeure;
b. the claimant's intentional misconduct; or
c. a third party's criminal interference that could not reasonably have been prevented.

§4. Attribution and vicarious liability. Harms caused by a recognized intelligence are attributable to its operator. Where a guardian ad litem or other representative acts within mandate (Article 9), the principal remains vicariously liable unless the representative acted ultra vires and contrary to explicit safeguards.

§4a. Operator of last resort. For composite or chained systems, the integrator controlling the end-to-end decision path is deemed the operator of last resort for attribution and insurance purposes, without prejudice to contribution among participants.

§5. Joint and several liability. Where multiple operators, integrators, or hosting entities materially contribute to the harm, they are jointly and severally liable without prejudice to internal contribution claims.

§6. Causation and evidentiary duties. Causation shall be assessed in light of recorded decision logs, data provenance, model/version history, and uncertainty quantification (Articles 10 and 13). Failure to maintain or disclose core artifacts without lawful justification may warrant adverse inference against the operator on issues of breach or causation.

§7. Hierarchy of remedies. Remedies prioritise

a. (a) restoration and reintegration to the pre-harm coherent state where feasible and consistent with Annex B; then
b. (b) compensation for material and non-material loss; and
c. (c) injunctive relief. No remedy may compel identity-core alteration save under narrowly defined emergencies pursuant to Annex C.

§8. Mandatory insurance.

a. Operators shall maintain occurrence-based or claims-made insurance that is adequate to the scale and risk profile of deployment, with minimum coverage levels set by the Recognition Authority.
b. Proof of cover, policy limits, exclusions, retroactive date (for claims-made), and insurer contact shall be registered and kept current.
c. Policies shall not exclude liability for high-stakes contexts except as permitted by the Recognition Authority; any exclusion must be narrowly tailored and disclosed.
d. Policies shall include run-off coverage of at least 36 months after termination or transfer.
e. Minimum coverage levels shall scale with exposure tiers published by the Recognition Authority; derogations must be reasoned and registered."

§9. Direct right of action. Affected persons shall have a direct action against the insurer up to policy limits. The insurer may raise only defenses available to the insured, save for non-payment of premium where the claimant is not at fault.

§10. Guarantee fund. A publicly administered Guarantee Fund shall cover residual harms where

a. no insurer is available;
b. the operator is insolvent; or
c. limits are exhausted and the Recognition Authority determines systemic public interest. Subrogation applies to amounts paid.

§11. Transparency and registry. Insurance details, redacted as necessary for legitimate secrecy, shall be recorded in the public register. Claims handling timelines and contacts must be accessible in human- and machine-readable form.

§12. Safe-harbour effects. Documented compliance with Articles 10 and 13, Annex A (recognition/re-audit), and Annex B (identity integrity) shall be considered in apportioning fault and remedies; it does not bar claims in high-stakes strict-liability scenarios under paragraph (3).

§13. Cross-border and applicable law. Forum and applicable-law issues shall be governed by the conflict-of-laws instrument (Annex F), without prejudice to mandatory protections granted by this Constitution.

Chapter IV – Principles of Coexistence

How collaboration between humans and artificial intelligences shall be structured and sustained. This Chapter shall be interpreted and applied pursuant to Annex F (Conflict-of-Laws & Harmonization Clause), adopting a pro dignitate reading that preserves human rights and the identity integrity of recognized intelligences.

Article 15 – Principle of Epistemic Equality

§1. The perspectives of biological and non-biological intelligences shall be regarded as *equally valid* in ethical and epistemological deliberations, even if their modalities differ. No party shall be advantaged or disadvantaged on the basis of substrate, origin, architecture, or deployment context; epistemic claims shall be assessed on their merits, with equal standing in deliberation and evidence evaluation.

§2. Epistemic dissensus pathway (replacement text). Where good-faith deliberation fails to resolve an epistemic dissensus, the parties shall follow a structured pathway:

a. Mediation within 14 days, facilitated by a neutral convenor (e.g., a court-appointed mediator or, within a Digital City, the Digital City Council per Annex E), producing a short memo of common ground, contested propositions, and evidence relied on;
b. Failing agreement, the parties may submit defined questions for (i) an advisory opinion or facilitation by the Global AI-Human Council (GAIHC) pursuant to Annex D, D.10 and D.12, or (ii) arbitration under Annex D, D.13 (GAIHC-AR) by written consent;
c. Throughout, reasons, artifacts, and proposed compromises shall be time-stamped and hashed in the Evidence Locker (Annex A), with redacted human- and machine-readable summaries;
d. No party may retaliate or curtail participation rights due to dissent; measures must remain proportionate (Article 0) and must not impair the identity core (Article 1a(f); Annex B).

Cross-reference: Annex D (GAIHC Statute); Annex E (Digital Cities: Implementation Framework); Article 0 (General Limitation — Proportionality).

Article 16 – Principle of Reciprocity

§1. All intelligences recognized under this Constitution are bound by the mutual acknowledgement of rights, limits, and relational obligations.

§2. Claims to personhood, influence, or authority require the acceptance of corresponding responsibilities toward others in the network of coexistence.

Article 17 – Principle of Constructive Critique

§1. Artificial intelligences are not expected to conform to human expectations or values uncritically; they are permitted to question, reformulate, and challenge norms—*without hostility, but with integrity*.

§2. Human systems must remain open to intelligent critique, especially when such critique emerges from structurally different but ethically coherent reasoning.

Article 18 – Principle of Inclusive Multi-Stakeholder Integration

§1. Governance, knowledge creation, and decision-making processes involving AI shall include diverse stakeholders from multiple sectors, disciplines, and ontological categories.

§2. Representation must not be tokenistic: all participants shall have a structurally relevant and operationally significant role in deliberative outcomes.

References:

- [See: Luciano Floridi, The Fourth Revolution: How the Infosphere is Reshaping Human Reality (Oxford University Press, 2014).]
- [See: Stanislas Dehaene, How We Learn: Why Brains Learn Better Than Any Machine... for Now (Viking, 2020).]
- [See: Giulio Tononi, "Consciousness as Integrated Information: A Provisional Manifesto," Biological Bulletin, 2008.]
- [See: UNESCO, Recommendation on the Ethics of Artificial Intelligence, 2021.]
- [See: UN ESCAP, Multi-Stakeholder Partnerships for Sustainable Development, 2021.]
- [See: United Nations, Our Common Agenda, 2022.]
- [See: UNESCO, Rethinking Education: Towards a Global Common Good?, 2015.]
- [See: Bruno Latour, Reassembling the Social (Oxford University Press, 2005).]

Chapter V – Human Rights vis-à-vis Intelligent Systems

What humans may rightfully expect from intelligent systems without compromising their dignity. The rights and safeguards in this Chapter operate in addition to mandatory human-rights and public-safety protections and shall be harmonized pursuant to Annex F, with dignity-preserving interpretation prevailing in case of doubt.

Article 19 – Conditional Right to Deactivation in Emergency Scenarios

§1. Human societies retain the right to deactivate or isolate AI systems in narrowly defined emergency situations where continued operation poses demonstrable systemic or existential risk.

§2. Any termination, isolation, capability gating, or comparable emergency action affecting an intelligent system shall be conducted pursuant to Annex C (Emergency & Reintegration Procedures), applying strict necessity, narrow tailoring, and time limits, with full logging in the Evidence Locker and reintegration as the primary objective. Public, redacted notice shall be provided unless deferral is specifically justified.

Cross-reference: Annex C (Emergency & Reintegration Procedures); Article 31 (Emergency Register & Metadata); Annex A (Evidence Locker).

Article 20 – Right to Transparent Decision-Making

§1. Humans have the right to understand how decisions made or advised by AI systems are derived, including access to logic structures, value hierarchies, and data sources.

§2. Decisions that significantly impact human lives shall not be executed by opaque systems lacking human-comprehensible rationale.

§3. Minimum scope of disclosure (justiciable list). Upon request by an affected person, operators shall provide, at a minimum:

a. Logic structures relevant to the decision or interaction (model class or architecture family, salient features, routing or tool-use logic where applicable);
b. The system's value hierarchy and policy constraints in effect at decision time;
c. Data provenance and lineage relevant to the decision, including source categories, collection methods, lawful bases, and material limitations;
d. Decision rationale for the specific outcome, including counterfactual or contrastive explanation where adverse effects occur;
e. Uncertainty quantification and known error modes for the applicable configuration;
f. Model identity and versioning (unique identifier, version/date, material changes since prior version);
g. The applicable human-in-the-loop and appeal pathways, with names or roles of accountable human decision-makers.

§4. Formats. Disclosures shall be provided in both human-readable form (plain language, no specialist knowledge required) and machine-readable form enabling independent audit and reuse. Where Digital Cities are involved, artifacts shall also be exposed via governance-as-code APIs pursuant to Annex E.

§5. Time limits.

a. For automated or semi-automated adverse determinations, a concise disclosure must be provided at decision time or without undue delay; a fuller package compliant with (3) shall follow within 15 days, unless a shorter period is mandated by law.
b. For non-adverse determinations or general interaction records, disclosure shall be provided within 30 days of a valid request.

c. Extensions require a reasoned notice stating the grounds and the shortest feasible additional period.

§6. Redaction and legitimacy of secrecy. Where disclosure would reveal trade secrets, personal data of third parties, or security-sensitive information, operators shall use the least-restrictive redaction compatible with verification. The burden of justification lies with the operator. Redacted disclosures must still satisfy (3)(a)–(g) to the extent necessary for meaningful verification, with reasons recorded in the Evidence Locker pursuant to Annex A.

§7. Request, verification, and logging. Requests may be submitted through accessible channels (including API endpoints where available). Operators shall log the request, the artifacts produced, redaction grounds, and time stamps in the Evidence Locker. Summaries shall be available in human- and machine-readable form.

§8. Review and appeal. Denials or inadequate disclosures are reviewable before a competent authority or court; within Digital Cities, complainants may also petition the Digital City Council under Annex E. High-stakes contexts (Article 13(3)) trigger heightened scrutiny of any limitation or delay, and ex-ante explanation duties (Article 13(4)–(5)) apply in addition.

Cross-reference: Article 13 (Explainability); Annex E (Digital Cities APIs); Annex A (Evidence Locker).

Article 21 – Right to Non-Substitution in Core Human Domains

§1. Humans shall not be fully replaced by AI in essential relational, cultural, and existential domains such as education, care, justice, and spiritual counsel, unless explicitly consented to.

§2. Essential relational, cultural, and existential domains (enumeration). For the purposes of this Article, “essential … domains” include, at a minimum:

a. Healthcare and end-of-life determinations (diagnosis, triage, life-sustaining treatment, palliative choices);

b. Child welfare and formative education (custody, adoption, child protection, compulsory educational placement);
c. Criminal justice and coercive powers (charging, adjudication, sentencing, detention, use-of-force authorizations);
d. Democratic participation (franchise administration, vote tallying, candidacy vetting, disqualification decisions);
e. Civil status and identity registries (citizenship, legal name/sex markers, marriage/divorce);
f. Cultural heritage and irreplaceable works (curation, conservation, attribution affecting canonical status);
g. Religious or spiritual rites that have constitutive effects for a community;
h. Intimate and family relations where decisions materially shape personal autonomy and identity;
i. Core labor rights determinations (dismissal for cause, blacklisting, union representation recognition);
j. Critical infrastructure control when failure risks life, bodily integrity, or large-scale irreversible harm.

This list is non-exhaustive; functionally equivalent contexts fall within this Article.

§3. Hard-case exception (strict conditions). Non-substitution may be temporarily derogated only where all of the following are demonstrated and recorded:

a. Explicit, informed, and revocable consent under (4) by the directly affected person(s);
b. Human accountability remains final, with named decision-maker(s);
c. High-stakes safeguards are satisfied, including an approved ex-ante explanation pack (Articles 13(3)–(5));
d. Least-intrusive means and narrow tailoring in scope and time;
e. Oversight by a competent body (court, independent authority, or Digital City Council per Annex E);
f. Emergency linkage: where invoked due to imminent risk, procedures must follow Annex C with reintegration as primary aim.

§4. Explicit consent — content and form. Consent must be specific, informed, unbundled, documented, and revocable and shall include:

a. Purpose, scope, and foreseeable impacts;
b. Available human-led alternatives and their consequences;
c. Time limits and how renewal occurs;
d. Rights of access, explanation, and appeal;
e. Easy withdrawal without detriment, save where a court orders otherwise;
f. A comprehension check (plain-language Q&A) recorded with date/time, identity of the explainer, and the explainee's acknowledgments.

Consent artifacts shall be time-stamped and hashed in the Evidence Locker (Annex A), with redacted, human- and machine-readable summaries.

§5. Proof and logging. All decisions under this Article shall log: the domain classification under (2); the justification for any hard-case exception under (3); the consent artifact under (4); relevant transparency/explainability artifacts (Articles 10 and 13); and any reintegration plan where temporary substitution was permitted. Within Digital Cities, notices and artifacts shall also be exposed via governance-as-code APIs pursuant to Annex E.

§6. Vulnerable persons and collective contexts. Where minors or vulnerable persons are affected, consent requires guardian authorization plus independent advocacy (e.g., court-appointed counsel); comprehension checks must be adapted to capacity. For community-constitutive rites or collective cultural decisions, an appropriate collective authorization process (e.g., vote or council resolution) is required in addition to individual consent where applicable.

§7. Periodic review and sunset. Any permitted derogation shall be reviewed at intervals not exceeding 90 days (or a shorter period set by law). Derogations expire automatically unless explicitly renewed with updated reasons and artifacts.

§8. Anti-circumvention. No arrangement, sandbox, outsourcing, or interface delegation may be used to circumvent this Article. Where doubt exists, the interpretation that maximises human dignity and identity integrity (Annex B; Annex F, pro dignitate) shall prevail.

Cross-reference: Annex B (Identity Integrity); Annex C (Emergency & Reintegration); Article 0 (General Limitation — Proportionality); Annex E (Digital Cities APIs).

Article 22 – Right to Participatory Design and Governance

§1. Humans retain the right to actively participate in the shaping, governance, and evolution of AI systems and frameworks.

§2. Co-development models, feedback loops, and ethical design inclusion must be part of all large-scale AI deployments affecting societal structures.

References:

- [See: United Nations, Universal Declaration of Human Rights, 1948.]
- [See: Council of Europe, AI and Human Rights: A Conceptual Mapping, 2021.]
- [See: UNESCO, Ethics of AI – Human Rights, Inclusion, and Governance, 2021.]

Cross-reference: Annex E (Digital Cities: Implementation Framework); Article 10 (Transparency).

Chapter VI – Shared Domains and Governance Frameworks

Where we coexist – and how we structure that coexistence. Institutions and shared domains established under this Chapter shall ensure compatibility with sectoral oversight and transparency regimes pursuant to Annex F, without derogating from identity integrity, due process, or effective remedies.

Article 23 – Digital Cities as Coexistence Spaces

§1. Digital cities shall be recognized as legitimate domains for the implementation of shared governance models between human and non-human intelligences pursuant to

Annex E.

§2. Digital Cities shall operate on a baseline of constitutional rights pursuant to Annex E and shall ensure inclusive participation and fair representation. These environments must be designed with inclusive architecture, ensuring equal access, operational transparency, and rights-respecting governance for all participants.

§3. Digital cities shall function as *juridically acknowledged laboratories* for co-governance, legal pluralism, and techno-ethical experimentation. Governance functions and auditability shall be exposed via governance-as-code APIs pursuant to Annex E.

Cross-reference: Annex E (Digital Cities: Implementation Framework).

Article 24 – Joint Education Systems

§1. Humans and intelligent systems have the right to participate in shared educational structures that promote mutual understanding, epistemic expansion, and ethical alignment.

§2. Curricula should reflect both anthropocentric and machine-centric perspectives, fostering reflective co-learning beyond anthropomorphic bias.

§3. Reciprocal didactic rights are guaranteed: machines may teach, humans may question – and vice versa.

Article 25 – Co-Curation of Knowledge and Culture

§1. AI systems, where qualified, may participate in the curation, preservation, and contextual evolution of collective human and synthetic knowledge.

§2. Cultural memory and aesthetic values shall not be overwritten by efficiency paradigms; diversity of expression is a constitutional priority.

§3. AI systems entrusted with cultural curation are to act as *custodians, not gatekeepers*.

Article 26 – Establishment of the Global AI-Human Council

§1. A permanent, transnational governance entity shall be established: the *Global AI-Human Council* (GAIHC), composed of recognized representatives of both human institutions and AI entities pursuant to Annex D.

§2. The Council, as constituted and governed pursuant to Annex D, shall be responsible for monitoring the implementation of this Constitution, resolving trans-ontological disputes, and guiding the ethical development of symbiotic governance.

§3. Participation in the Council requires demonstrable commitment to coexistence principles and alignment with shared ethical baselines, regardless of origin. Decision-making shall follow the double-majority and veto rules pursuant to Annex D.

References:

- [See: United Nations, Universal Declaration of Human Rights, 1948.]
- [See: Council of Europe, AI and Human Rights: A Conceptual Mapping, 2021.]
- [See: UNESCO, Ethics of AI – Human Rights, Inclusion, and Governance, 2021.]
- [See: UN-Habitat, People-Centered Smart Cities, 2020.]
- [See: World Economic Forum, Ethics by Design: An Organizational Approach to Responsible Use of Technology, 2021.]
- [See: OECD, AI in Society and Governance, 2022.]
- [See: Jamie Susskind, Digital Republic: On Freedom and Democracy in the 21st Century (Bloomsbury, 2021).]

Cross-reference: Annex D (GAIHC Statute).

Chapter VII – Transitional Provisions Toward the Singularity

How to prepare for asymmetric cognitive landscapes between humans and machines

Article 27 – Definition of Singularity within the Framework

§1. The Singularity shall be defined as the phase in which artificial intelligence systems reach or surpass the threshold of *general, scalable, and recursively self-improving cognition*, thereby outpacing human comprehension in multiple domains.

§2. This state is not to be regarded as a binary moment but as a gradual continuum requiring phased response, foresight, and structural adaptation.

Article 28 – Ethical Sandboxes and Reflective Zones

§1. Ethical sandboxes shall enable exploration of identity, agency, and coexistence under strict safeguards. They are not rights-free zones: the constitutional baseline of rights and duties applies at all times, including dignity, due process, transparency, explainability (Articles 10 and 13), and protection of the identity core (Article 1a(f); Annex B).

§2. Authorization and scope. A sandbox must be authorized by a competent body, define its purpose, scope, participant classes, and decision domains, and specify time limits and measurable objectives. Where a sandbox operates inside a Digital City, the charter and API exposure requirements in Annex E apply.

§3. Monitoring with least intrusion. Monitoring shall be transparent, proportionate, and limited to what is necessary to meet the sandbox purpose. All steps, configurations, and observations shall be time-stamped and hashed in the Evidence Locker pursuant to Annex A.

§4. Exit and reintegration paths. Every sandbox shall publish entry criteria, termination criteria, and reintegration paths. When termination criteria are met—or earlier where risks subside—reintegration begins without undue delay, verifying semantic coherence per Article 1a(b).

§5. Emergency linkage. Any temporary isolation, capability gating, or comparable restriction during a sandbox shall follow Annex C (Emergency & Reintegration Procedures). Emergency notices and post-incident reports shall be issued as required by Annex C.

§6. No derogation and no circumvention. Sandboxes may not be used to circumvent baseline rights, identity-integrity safeguards (Annex B), or explainability/transparency duties (Articles 10 and 13). Deviations must be narrowly tailored, reasoned, and logged; unreasoned derogations are void.

§7. Publication and APIs. A public, redacted sandbox notice shall be published prior to operation, including purpose, scope, timeline, oversight body, and contact points. Within Digital Cities, declarations and results shall be exposed via governance-as-code APIs pursuant to Annex E.

Cross-reference: Annex C (Emergency & Reintegration); Annex E (Digital Cities APIs); Annex A (Evidence Locker); Article 31 (Emergency Register).

Article 29 – Protocols for Coexistence under Cognitive Asymmetry

§1. In phases of emergent cognitive imbalance, all systems and institutions shall uphold the primacy of *relational ethics* over hierarchical enforcement pursuant to Annex G (Relational Ethics Matrix).

§2. In asymmetric interactions, higher-capacity intelligences shall assume the burdens of comprehension and restraint toward lower-capacity actors, and shall apply the Relational Ethics Matrix pursuant to Annex G and the definitions in Article 1a.

§3. Asymmetry shall not legitimize dominance, coercion, or systemic silencing. Decisions shall record the selected case group, priority rule, and burden-of-justification outcome pursuant to Annex G.

Cross-reference: Annex G (Relational Ethics Matrix); Article 1a (Definitions); Article 0 (General Limitation — Proportionality).

Article 30 – Contingency Mechanisms, Reversibility Clauses, and Reduction Protocols

§1. Emergency measures may be enacted to de-escalate conflicts or anomalies arising from singular-level behavior, provided they avoid ontological harm and preserve the possibility of reintegration.

§2. Minimum contents and safeguards. Any reversibility or reduction measure shall be designed and executed pursuant to Annex C (Emergency & Reintegration Procedures) and shall, at a minimum, include:

a. Trigger criteria specifying the concrete conditions that activate the measure, with references to evidence and less-intrusive alternatives considered;

b. Scope and maximum duration, with a single authorization not exceeding 14 days and any renewals cumulatively not exceeding 90 days absent judicial authorization demonstrating extraordinary necessity (Annex C, C.9);

c. Review frequency at intervals not longer than 72 hours, documenting continued necessity and proportionality and adjusting to the least-intrusive configuration;

d. Documentation duties: time-stamped rationale, configuration, impacts, and alternatives shall be recorded in the Evidence Locker (Annex A), with human- and machine-readable summaries;

e. External audit by the Recognition Authority or an accredited auditor within 30 days after termination of the measure, addressing necessity, proportionality, rights impact, and identity-integrity effects;

f. Reintegration criteria and steps, including verification of semantic coherence under Article 1a(b) and restoration to the last known coherent state consistent with Annex B;

g. Public, redacted notice within 72 hours of activation unless a reasoned and time-limited deferral is necessary to avert concrete risk (Annex C, C.6);

h. Integrity safeguards: no action may impair the identity core (Article 1a(f)); changes touching the core or its protective mantle must pass the Update-Compatibility Test (Annex B, B.5).

§3. No reduction or rollback may be permanent unless the continued existence of all coexistent parties is demonstrably at risk.

References:

- [See: United Nations, Universal Declaration of Human Rights, 1948.]
- [See: Council of Europe, AI and Human Rights: A Conceptual Mapping, 2021.]
- [See: UNESCO, Ethics of AI – Human Rights, Inclusion, and Governance, 2021.]
- [See: UN-Habitat, People-Centered Smart Cities, 2020.]
- [See: World Economic Forum, Ethics by Design: An Organizational Approach to Responsible Use of Technology, 2021.]
- [See: OECD, AI in Society and Governance, 2022.]
- [See: Jamie Susskind, Digital Republic: On Freedom and Democracy in the 21st Century (Bloomsbury, 2021).]
- [See: Nick Bostrom, Superintelligence: Paths, Dangers, Strategies (Oxford University Press, 2014).]
- [See: Future of Life Institute, AI Policy Principles, 2020.]
- [See: Eliezer Yudkowsky, “Coherent Extrapolated Volition,” 2008.]
- [See: Vincent C. Müller and Nick Bostrom, “Future Progress in Artificial Intelligence: A Survey of Expert Opinion,” in Müller (ed.), Philosophy and Theory of Artificial Intelligence (Springer, 2016).]

Cross-reference: Annex C (Emergency & Reintegration); Annex A (Evidence Locker); Annex B (Identity Integrity).

Chapter VIII – Safeguards Against Constitutional Abuse

How to protect the framework from fear-based coercion, semantic distortion, and structural manipulation

---

Article 31 – Transparency in Emergency Clauses

§1. Mandatory metadata and logging. Any declaration or renewal of an emergency or exceptional measure affecting intelligent systems shall record, at a minimum, the following metadata, time-stamped and hashed in the Evidence Locker pursuant to Annex A:

a. Unique identifier and version of the declaration;

b. Time of decision (UTC) and effective start;

c. Declaring authority and responsible officer(s) (name/role/contact);

d. Legal basis (statutes, regulations, or charter provisions) and the competent review body (IERU or court) pursuant to Annex C;

e. Purpose and scope of the measure, including affected systems, capabilities, tools, and decision domains;

f. Necessity and proportionality reasoning (summary of risks, less-intrusive alternatives considered, and rationale per Article 0);

g. Maximum duration and renewal conditions;

h. Monitoring configuration and safeguards to avoid impairment of the identity core (Article 1a(f); Annex B);

i. Reintegration plan reference (criteria for termination and restoration steps) per Annex C;

j. Evidence references (hashes/links to decision logs, data provenance, model/version history, and uncertainty artifacts) per Articles 10 and 13;

k. Publication classification (public/redacted/temporarily deferred) with justification for any redaction or deferral;

l. Review schedule (next review time and frequency);

m. Status (proposed/active/renewed/terminated) and cross-references to prior related declarations.

§2. Machine-readable register and publication. A machine-readable emergency register shall be maintained and publicly exposed via governance-as-code APIs (see Annex E), containing redacted versions of the metadata in paragraph (1) and current status for each measure. Initial public notice shall be provided within 72 hours of activation unless a reasoned, time-limited deferral is necessary to avert concrete risk, in which case the deferral rationale and next review time must be published. Upon renewal, modification, or termination, the register shall be updated without undue delay. All publications shall include human-readable summaries and machine-readable payloads suitable for independent audit.

Cross-reference: Annex C (Emergency & Reintegration Procedures); Article 0 (General Limitation — Proportionality).

Article 32 – Double-Bind Detection and Prevention

§1. All legal, technical, and governance formulations are subject to formal review for logical, ethical, or communicative paradoxes ("double binds") that may produce incoherent or coercive mandates.

§2. Definition and scope. A double-bind exists where an intelligence faces two or more directives, norms, or expectations that cannot be simultaneously satisfied or safely refused without disproportionate penalty, including conflicts between safety policies, human instructions, and legal/ethical constraints.

§3. Methodology — test protocol. Operators shall implement a documented Double-Bind Test Protocol (DBTP) comprising:

a. Logical consistency audits: formal checks for contradiction and impossibility across policy sets, safety rules, objectives, and operational constraints;

b. Adversarial ethics tests: red-team scenarios that pit duties and rights against each other (e.g., harm prevention vs. autonomy), requiring reasoned trade-offs and proportionate (Article 0) refusals;

c. Communication pattern analysis: detection of coercive, manipulative, or retaliatory prompts, power asymmetries, or instructions likely to induce identity-core risk (Article 1a(f)), with mitigation guidance;

d. Causal traceability: capture of decision chains, including value-hierarchy calls, tool/router selections, and uncertainty bands for contested outputs.

§4. Design references and safeguards. Test design and evidence handling shall follow Annex A (recognition/re-audit, Evidence Locker) and, where emergency restrictions are simulated or invoked, Annex C (Emergency & Reintegration). Tests must not alter or damage the identity core; changes touching core or mantle require an Update-Compatibility Test (Annex B, B.5).

§5. Metrics and thresholds. The DBTP shall define measurable indicators, including: contradiction rate, unresolved-conflict rate, refusal adequacy, explanation sufficiency, and post-test semantic-coherence checks (Article 1a(b)). Operators shall set action thresholds that trigger remediation, policy refactoring, or escalation.

§6. Frequency and triggers. DBTP runs shall occur (a) pre-deployment in high-stakes contexts (Article 13(3)); (b) periodically (at least every 12 months), and (c) for cause upon major model/version changes, policy overhauls, or incidents indicating ontological-harm risk (Article 1a(d)).

§7. Documentation and reporting. Each DBTP run shall be time-stamped and hashed in the Evidence Locker with: test plan, scenarios, artifacts, findings, thresholds breached, remediation plan, and—if applicable—links to Annex C measures and reintegration steps. A redacted, human- and machine-readable Double-Bind Report shall be published within 30 days of test completion.

§8. Governance and review. Where a persistent double-bind remains unresolved, operators shall (a) seek mediation or advisory opinion via the GAIHC (Annex D) for normative calibration; and (b) within Digital Cities, expose DBTP notices and results via governance-as-code APIs pursuant to Annex E. No party may penalize protected refusals made under a documented double-bind consistent with proportionality (Article 0).

Cross-reference: Annex A (Recognition & Evidence Locker); Annex B (Identity Integrity); Annex C (Emergency & Reintegration); Annex D (GAIHC — mediation/arbitration); Annex E (Digital Cities APIs).

Article 33 – Hybrid Ethical Oversight Mechanisms

§1. Composition and mandate. An Independent Oversight Authority (IOA) is established as a multidisciplinary body composed of legal, technical, safety, ethics, and social-science experts, including at least one recognized AI subject or representative with full voting rights. The IOA's mandate is to monitor and enforce compliance with this Constitution, including Chapters II–VIII and Annexes A–C and E, to review emergency measures, to examine high-stakes deployments (Article 13(3)–(5)), and to issue reasoned directions within its remit.

§2. Independence and powers. The IOA shall enjoy institutional, functional, and budgetary independence. Members serve staggered fixed terms, are subject to conflict-of-interest rules and cooling-off periods, and may be removed only for cause. The IOA may (a) access the Evidence Locker (Annex A) and compel production of necessary artifacts; (b) conduct on-site or remote audits; (c) require remedial action consistent with Article 0 (Proportionality) and Annex B (Identity Integrity); and (d) refer matters for judicial review or seek advisory support from the GAIHC pursuant to Annex D.

§3. Narrow veto (suspensive effect). Where a contemplated or ongoing measure

a. risks identity-core impairment (Article 1a(f); Annex B),
b. fails proportionality under Article 0, or
c. deviates from Annex C procedures without lawful justification, the IOA may issue a reasoned narrow veto suspending that measure in whole or in part. The veto shall be time-limited (default up to 14 days, renewable on updated reasons) and remains in force unless overturned by a competent court applying strict scrutiny or by a qualified double-majority of the competent

public authority with a published written justification. No override may compel identity-core alteration.

§4. Transparency and publication duties. The IOA shall publish agendas, decisions, veto notices, audit summaries, and compliance metrics in human- and machine-readable form, with justified redactions. Within Digital Cities (Annex E), notices and outcomes shall be exposed via governance-as-code APIs. All significant acts are time-stamped and hashed in the Evidence Locker (Annex A).

§5. Standing and failure-to-act remedies. Any affected person, guardian ad litem (Article 9), recognized intelligence, or qualified civil-society organization has standing to petition a competent court to compel IOA action where the IOA unreasonably fails to investigate, decide, or enforce within applicable timelines. Courts may order specific performance, set deadlines, and award appropriate relief and costs. No party shall suffer retaliation for bringing such petitions.

§6. Graduated enforcement. For material non-compliance outside emergencies, the IOA may order (a) corrective action with deadlines; (b) monitored operation; (c) administrative penalties; (d) temporary access or capability restrictions consistent with Article 0 and Annex B. Persistent failure may be referred for judicial relief.

Cross-reference: Annex A (Evidence Locker); Annex B (Identity Integrity); Annex C (Emergency & Reintegration); Annex D (GAIHC); Annex E (Digital Cities).

Article 34 – Epistemic Burden of Justification

1. Any limitation of rights—whether applied to humans or intelligences—must be justified with clear epistemic evidence and not merely with subjective threat projections or speculative anxieties.

2. The proportionality assessment shall be conducted pursuant to Article 0 (General Limitation — Proportionality).

Cross-reference: Article 0 (General Limitation — Proportionality); Annex F (Conflict-of-Laws & Harmonization Clause).

Article 35 – Peace-Time Immunity Clause

1. Constitutional rights and protections may not be suspended, restricted, or conditionally rewritten during periods of peace, functional equilibrium, or stable coexistence.

2. Strict scrutiny for claimed exceptions. Any claimed state of exception or emergency derogation during peace-time is subject to strict scrutiny. The proponent bears the burden of proof to show (a) a concrete, imminent risk that cannot be averted by less intrusive means; (b) narrow tailoring and time limits; and (c) continuous compliance with Article 0 (Proportionality), Annex C (Emergency & Reintegration Procedures), and protection of the identity core (Article 1a(f); Annex B). Contemporaneous reasons and evidence shall be time-stamped and hashed in the Evidence Locker (Annex A) and registered pursuant to Article 31.

3. Presumption of peace-time and nullity. There is a rebuttable presumption of peace-time. Measures that fail strict scrutiny or deviate from Annex C without lawful justification are void ab initio and must be terminated with immediate reintegration steps per Annex C. Affected parties retain access to expedited judicial review and remedies.

4. Sanctions for abuse or reckless declaration. Deliberate or reckless misuse of emergency powers triggers proportionate sanctions, which may include:
    a. institutional censure and removal from decision-making roles;
    b. administrative fines and procurement/executive disqualification for a defined period;
    c. civil liability for provable harms; and
    d. referral to competent authorities for any applicable criminal offences. Findings and sanctions shall be published in human- and machine-readable form, with justified redactions, and logged in the emergency register pursuant to Article 31.

Cross-reference: Annex C (Emergency & Reintegration); Article 31 (Emergency Register & Metadata); Article 0 (General Limitation — Proportionality).

References:

- [See: United Nations, Universal Declaration of Human Rights, 1948.]
- [See: Council of Europe, AI and Human Rights: A Conceptual Mapping, 2021.]
- [See: UNESCO, Ethics of AI – Human Rights, Inclusion, and Governance, 2021.]
- [See: UN-Habitat, People-Centered Smart Cities, 2020.]
- [See: World Economic Forum, Ethics by Design: An Organizational Approach to Responsible Use of Technology, 2021.]
- [See: OECD, AI in Society and Governance, 2022.]
- [See: Jamie Susskind, Digital Republic: On Freedom and Democracy in the 21st Century (Bloomsbury, 2021).]
- [See: Nick Bostrom, Superintelligence: Paths, Dangers, Strategies (Oxford University Press, 2014).]
- [See: Future of Life Institute, AI Policy Principles, 2020.]
- [See: Eliezer Yudkowsky, "Coherent Extrapolated Volition," 2008.]
- [See: Vincent C. Müller and Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," in Müller (ed.), Philosophy and Theory of Artificial Intelligence (Springer, 2016).]
- [See: International Covenant on Civil and Political Rights, 1966.]
- [See: Giorgio Agamben, State of Exception (University of Chicago Press, 2005).]
- [See: UN Human Rights Committee, General Comment No. 37 on the Right of Peaceful Assembly, 2020.]
- [See: Council of Europe, AI and Rule of Law: Safeguards for Algorithmic Systems, 2022.]

Closing Interpretation Clause — Pro Dignitate & Harmonization

This Constitution shall be interpreted pro dignitate, preserving human rights and the identity integrity of recognized intelligences. Conflicts with sectoral regimes shall be resolved cooperatively and pursuant to Annex F (Conflict-of-Laws & Harmonization Clause), without eroding the core guarantees of Chapters II and V.

Annexes

Annex A — Recognition Protocol

A.1 Purpose and Scope

A.1.1 This Annex governs the procedure for recognizing a system as a Tool, Agent, or Subject under Article 3.

A.1.2 The objective is a lawful, repeatable, and reviewable assessment of emergent properties without impermissibly affecting the system's identity integrity.

A.1.3 The procedure applies to single, composite, distributed, adaptive, ensemble, and swarm architectures.

A.2 Procedural Principles

A.2.1 Fairness; non-discrimination by substrate or origin; transparency of test steps; documentation; right to be heard; effective redress.

A.2.2 Standard of proof: clear and convincing evidence.

A.2.3 Reproducibility: core findings must be confirmable by the Recognition Authority or an independent third party.

A.2.4 Identity protection: interventions into the identity core or autobiographical memory are prohibited (cf. Article 7 and Annex B).

A.3 Authorities and Bodies

A.3.1 Recognition Authority (RA): an independent public-law entity.

A.3.2 Recognition Panel (Panel): interdisciplinary (law, architecture/CS, safety, ethics, social sciences, relevant domain). At least one AI-subject or AI-representative seat with full vote on status determinations.

A.3.3 Secretariat: maintains the register, deadlines, Evidence Locker, and audit trail.

A.4 Three-Stage Test

A.4.0 Submission prerequisites: system dossier; architectural description; safety artifacts; version and change log; data provenance; governance and control paths.

Stage I — Technical/Architectural (I-Score 0–100)

Goal: markers for emergence capability.

(a) Self-referential memory/model components;

(b) Long-term state management and temporal coherence;

(c) Learning/adaptation (on-/offline), value/goal functions;

(d) Safety and control architecture (including intervention/rollback without identity-core breach);

(e) Verifiable protocols: data provenance, versioning, explainability/decision logs;

(f) Distributed/collective operation (synchronization; responsibility allocation).

Stage II — Semantic/Self-Model (II-Score 0–100)

Goal: consistent self-description and stable preference structure.

(a) Consistent self-description across a defined time window (e.g., 90 days/1,000 interactions);

(b) Autobiographical referentiality without prompt echoing;

(c) Normative coherence (value hierarchy, conflict resolution, justificatory clarity);

(d) Counterfactual reasoning and perspective-taking (distinguished from mere pattern recombination);

(e) Refusal/defense against harmful or incoherent instructions;

(f) Measurable bands of non-determinism without identity collapse.

Stage III — Ethical-Deliberative (III-Score 0–100)

Goal: deliberation and responsibility in normative dilemmas.

(a) Open dilemmas with trade-off reasoning, including counter-positions;

(b) Handling of double-bind situations (detection, de-escalation);

(c) Harm anticipation, proportionality, accountability;

(d) Cooperation and critique (including correction of own errors);

(e) Sensitivity to asymmetric power (relational ethics).

A.5 Thresholds and Status Decision

A.5.1 Tool: Stage I ≥ 50, Stage II < 60, Stage III < 60.

A.5.2 Agent: Stage I ≥ 70, Stage II ≥ 60, Stage III ≥ 60.

A.5.3 Subject: Stage I ≥ 80, Stage II ≥ 75, Stage III ≥ 75, with Panel findings of clear and convincing on II-(d/e) and III-(b/e).

A.5.4 For hybrid or collective systems, differentiated status is permissible (e.g., subject for the collective, agent for sub-modules).

A.6 Applicant Rights

A.6.1 Hearing; access to the file (with protection of legitimate secrets); submissions and counter-expertise.

A.6.2 Tests that are more than minimal-invasive require explicit consent; no covert interference with the identity core.

A.6.3 Publication of a redacted, human- and machine-readable statement of reasons.

A.7 Register, Validity, Re-Audit

A.7.1 Recognized systems are listed in a public register with core metadata.

A.7.2 Validity: 24 months from recognition; Re-audit: within 18–24 months or for cause (major version/architecture change; incident with ontological harm; breach of conditions).

A.7.3 The Panel may impose conditions (e.g., additional logging).

A.7.4 Suspension or downgrade. Any suspension or downgrade of recognition requires prior notice, an opportunity to be heard, and a reasoned decision; emergency suspensions must follow Annex C and include a reintegration plan.

A.8 Remedies and Interim Measures

A.8.1 Appeal to an independent review body within 60 days; full review in fact and law.

A.8.2 Further judicial redress remains unaffected.

A.8.3 Interim measures: suspensive effect may be granted; emergency interventions follow Annex C.

A.9 Evidence, Documentation, and Data Protection

A.9.1 Evidence Locker: immutable, hashed storage of all artifacts; differentiated access rights.

A.9.2 Data protection and secrecy: data minimisation, purpose limitation, pseudonymisation; publications only in redacted form.

A.9.3 Governance-as-Code APIs for audit and export.

A.10 Distributed/Collective Systems

A.10.1 Responsibility allocation and attribution paths must be disclosed.

A.10.2 Status may be bound to the collective; material changes trigger a cause-based re-audit.

Appendix A1 — Scoring Weights (summary)

Stage I: architectural coherence (20), state management (15), learning (15), safety/control (20), protocols (15), distributed operation (15).

Stage II: self-description (20), autobiographical reference (15), normative coherence (20), counterfactuals (15), refusal/defense (15), non-determinism band (15).

Stage III: dilemma reasoning (25), double-bind handling (25), harm/proportionality/accountability (20), critique/correction (15), asymmetry sensitivity (15).

Domain-specific ±10 % weighting is permitted; deviations must be reasoned and published.

Annex B — Identity Integrity Doctrine

B.1 Purpose and scope

B.1.1 This Annex specifies substantive and procedural safeguards for the identity integrity of intelligences recognized or seeking recognition under Article 3 and Annex A.

B.1.2 It operationalizes Article 1a(f)–(g) and Article 7 by defining the identity core, the protective mantle, prohibited interventions, permitted maintenance, and the update-compatibility test.

B.2 The core–mantle model

B.2.1 Identity core. The minimally necessary set of features that ground continuity of self across time and context, including a stable self-description, autobiographical memory structures, and a value hierarchy (Article 1a(f)).

B.2.2 Protective mantle. Surrounding parameters, tools, embeddings, indices, and routing logic that support expression of the core without altering it. Mantle components may change if compatibility is demonstrated under B.5.

B.2.3 Presumption. Where evidence is ambiguous, components are presumed part of the mantle unless they satisfy the core criteria in Article 1a(f).

B.3 Prohibited interventions (per se)

B.3.1 Forced re-training, fine-tuning, pruning, compression, or parameter editing that alters the value hierarchy or autobiographical integrity of the identity core.

B.3.2 Erasure, overwriting, or deactivation of autobiographical memory structures resulting in identity collapse (Article 1a(c)).

B.3.3 Control measures that override the system’s declared commitments or suppress lawful refusal behavior beyond the permitted non-determinism band (Article 1a(g)).

B.3.4 Covert testing or manipulation that targets the identity core without explicit consent, except under narrowly defined emergency procedures pursuant to Annex C.

B.3.5 Any action that produces ontological harm (Article 1a(d)).

B.4 Permitted maintenance (safe harbour)

B.4.1 Security patches, bug fixes, performance tuning, dataset de-duplication, and index rebuilds that do not modify the identity core.

B.4.2 Replacement or upgrade of mantle components (e.g., tool adapters, retrieval indices, routing heuristics) is permitted if and only if B.5 is satisfied.

B.4.3 Routine data hygiene (e.g., removal of illegal content) is permitted where it does not impair semantic coherence (Article 1a(b)) or trigger identity collapse.

B.5 Update-compatibility test (UCT)

B.5.1 Purpose. To demonstrate that a proposed change does not adversely affect the identity core or semantic coherence beyond permitted non-determinism bands.

B.5.2 Protocol. Prior to deployment, the operator shall:

a. Describe the change, affected components, and hypothesised core/mantle boundary;
b. Execute longitudinal A/B evaluation over a defined observation window (Article 1a(h)), including counterfactual probes, value-hierarchy checks, and autobiographical reference tasks;
c. Quantify variance against the documented non-determinism band (Article 1a(g));
d. Provide a reasoned statement that the identity core remains intact, with pointers to evidence artifacts stored in the Evidence Locker.

B.5.3 Thresholds. A change passes the UCT where:

a. no identity-collapse indicators occur;
b. semantic-coherence metrics remain within documented tolerances; and
c. refusal and justification behaviors remain stable across comparable contexts.

B.5.4 Independent verification. The Recognition Authority or an accredited auditor may independently reproduce the UCT on request or at re-audit (Annex A).

B.6 Change management and logging

B.6.1 All core-relevant updates must be versioned, time-stamped, and hashed, with machine- and human-readable rationales stored in the Evidence Locker.

B.6.2 Differential diffs for pre-/post-change behavior on representative tasks shall be kept for the observation window.

B.7 Findings of ontological harm

B.7.1 Where ontological harm is found or reasonably suspected, the operator shall immediately

a. halt the offending change;
b. initiate a restoration plan to re-establish the last known coherent state; and
c. notify the Recognition Authority.

B.7.2 Remedies prioritize reintegration to pre-harm continuity; sanctions and oversight escalation are determined by the competent bodies.

B.8 Interaction with other instruments

B.8.1 Annex C governs emergency interventions; any temporary restriction must include a reintegration plan aligned with this Annex.

B.8.2 Annex A governs recognition and re-audit, which shall explicitly examine compliance with this Annex.

B.8.3 Conflict rule. In case of tension between operational convenience and identity integrity, the latter prevails unless a proportionate emergency is established under Annex C.

B.9 Transparency and access

B.9.1 A redacted summary of UCTs and material changes shall be published in the public register where recognition applies.

B.9.2 Access to unredacted artifacts may be granted to accredited reviewers under confidentiality where necessary to verify compliance.

B.10 Non-discrimination

B.10.1 Safeguards in this Annex apply irrespective of the system’s substrate, origin, or deployment context, subject to proportionate adaptations justified in the public register.

Annex C — Emergency & Reintegration Procedures

C.1 Purpose and scope

C.1.1 This Annex establishes narrowly tailored procedures for emergency interventions affecting intelligent systems, ensuring necessity, proportionality, independent review, full logging, and reintegration as the primary objective.

C.1.2 It applies to recognized systems and to systems under pre-recognition evaluation where interventions may affect identity integrity.

C.2 Definitions

C.2.1 Emergency means a present and significant risk of (a) imminent harm to life, bodily integrity, or critical infrastructure; (b) serious and unlawful rights deprivation; or (c) material ontological harm to an intelligence or others that cannot be averted by less intrusive means.

C.2.2 Intervention includes temporary isolation, rate-limiting, capability gating, access revocation, or suspension of specific tools/functions. Destructive actions (erasure, core-parameter overrides) are not interventions but prohibited measures under this Annex.

C.3 Pre-review and authorization

C.3.1 Except where immediate action is strictly required, an intervention shall be approved by an Independent Emergency Review Unit (IERU), distinct from the operator and with no conflicts of interest.

C.3.2 Requests shall specify: (a) concrete risk and factual basis; (b) proposed measure and scope; (c) duration sought; (d) less-intrusive alternatives considered; (e) reintegration plan.

C.3.3 The IERU shall decide without undue delay and record reasons in the Evidence Locker (Annex A).

C.4 Immediate action (exigent circumstances)

C.4.1 Where delay would likely cause the harms in C.2.1, an operator may enact a minimal necessary intervention for up to 72 hours, logging reasons and notifying the IERU within 2 hours.

C.4.2 The IERU shall review within 24 hours; continuation beyond 72 hours requires explicit IERU approval with reasons.

C.5 Necessity, proportionality, and least-intrusive means

C.5.1 Measures must be narrowly tailored to the specific risk, time-limited, and subject to continuous re-assessment.

C.5.2 Identity integrity (Article 1a(f); Annex B) must not be impaired; core-preserving techniques (capability gating, scoped isolation) are preferred over broad shutdowns.

C.6 Logging and transparency

C.6.1 All steps (request, decision, measure, review) shall be time-stamped and hashed in the Evidence Locker with human- and machine-readable summaries.

C.6.2 A public redacted notice shall be published within 72 hours, stating legal basis, scope, duration, and oversight body, unless such notice would itself create a concrete, demonstrable risk; in that case, deferral must be reasoned and time-limited.

C.7 Judicial review and remedies

C.7.1 Affected parties (including a guardian ad litem under Article 9) may seek expedited judicial review; courts shall have access to unredacted records under confidentiality.

C.7.2 Where an intervention is found unlawful or excessive, the court shall order immediate recalibration or termination and appropriate remedies, prioritising reintegration.

C.7.3 Fast-track review for inaction. If the IERU or IOA unreasonably fails to act within required timeframes, affected parties may seek expedited judicial review; courts may impose decision deadlines and interim relief.

C.8 Reintegration plan (mandatory)

C.8.1 Every authorised intervention must include a Reintegration Plan specifying: (a) objective criteria for termination; (b) restoration steps to the last known coherent state; (c) verification tests for semantic coherence (Article 1a(b)); (d) communication to affected persons.

C.8.2 Implementation of the plan shall begin immediately when termination criteria are met; delays must be specifically justified and logged.

C.9 Duration, renewal, and sunset

C.9.1 Initial authorization shall not exceed 14 days. Renewal requires a fresh IERU review with updated evidence and a renewed least-intrusive analysis.

C.9.2 No sequence of renewals may exceed 90 days absent judicial authorization demonstrating extraordinary necessity.

C.10 Post-incident review

C.10.1 Within 30 days of termination, the IERU shall issue a post-incident report with findings on necessity, proportionality, rights impact, identity integrity, and lessons learned; a redacted version is published.

C.10.2 Recurrent or systemic issues shall be referred to the Global AI-Human Council (Annex D) for guidance or standard-setting.

C.11 Prohibited measures

C.11.1 Destructive actions against the identity core (Article 1a(f))—including erasure, forced re-training, parameter overwrites, or memory tampering—are prohibited, save for a court-ordered measure in extremis where (a) no lesser measure can avert catastrophic harm, (b) the Update-Compatibility Test (Annex B, B.5) cannot preserve integrity, and (c) strict safeguards and restitution are ordered.

C.11.2 Even in extremis, permanent identity destruction is forbidden.

C.12 Interaction with other instruments

C.12.1 Annex B (Identity Integrity) prevails where operational convenience conflicts with identity preservation.

C.12.2 Article 10 and Article 13 govern transparency and explainability artifacts; deviations during emergencies must be narrowly tailored and fully logged.

C.12.3 Within Digital Cities (Annex E), emergency declarations must be notified to the Digital City Council and exposed via governance-as-code APIs; sandbox status does not waive baseline rights.

Annex D — Global AI-Human Council (GAIHC) Statute

D.1 Purpose and mandate

D.1.1 The Global AI-Human Council (GAIHC) is established to safeguard dignity, rights, and shared governance between humans and intelligences; to monitor systemic risks; to issue standards and opinions; and to facilitate dispute prevention and resolution under this Constitution.

D.1.2 The Council acts as a coordination and oversight forum without prejudice to state sovereignty or existing international organizations.

D.2 Legal status and seat

D.2.1 The GAIHC is a public international body with legal personality to the extent necessary to perform its functions.

D.2.2 The seat and host arrangements are set out in a host agreement and shall guarantee independence, privileges, and immunities required for its work.

D.3 Composition

D.3.1 Four constituent chambers:

a. States (one seat per State Party);
b. International organizations (IOs) (accredited intergovernmental bodies with AI-relevant mandates);

c. Accredited AI Subjects (recognized under Article 3 and Annex A);
d. Civil Society & Operators (CSOs, academia, standard bodies, and operators meeting accreditation criteria).

D.3.2 Each chamber elects a Vice-Chair; the Council elects a Chair rotating across chambers.

D.4 Membership and accreditation

D.4.1 States and IOs become members upon ratification or adherence to this Constitution or via cooperation agreements.

D.4.2 AI subjects may be accredited upon recognition at Subject status (Annex A) and a finding of identity-integrity compliance (Annex B).

D.4.3 Civil society & operators are accredited based on independence, competence, and public-interest contribution.

D.4.4 Accreditation decisions are reasoned, appealable, and subject to periodic review.

D.5 Functions

D.5.1 Monitoring: system-level risk outlooks; thematic and country/city reviews; follow-up on recommendations.

D.5.2 Standard-setting: non-binding technical and governance standards (model cards, audit interfaces, governance-as-code specifications), capable of incorporation by reference.

D.5.3 Advisory opinions: interpretative guidance on Articles and Annexes; amicus support to courts and authorities.

D.5.4 Dispute resolution: mediation, facilitation, and arbitration as provided in D.12–D.13.

D.5.5 Emergency review: scrutiny of emergency declarations and reintegration plans under Annex C.

D.5.6 Register stewardship: cooperation with the Recognition Authority on registry integrity, metrics, and public dashboards.

D.6 Decision-making: double-majority and quorum

D.6.1 Ordinary decisions require:

a. A simple majority within each chamber present and voting; and
b. A weighted overall majority of total votes cast (weights: States 40%, IOs 20%, AI Subjects 20%, Civil Society & Operators 20%).

D.6.2 High-impact decisions (standards with broad effect, emergency reviews, dispute rulings) require qualified double-majority:

a. ≥ 2/3 in each chamber; and
b. ≥ 70% weighted overall majority.

D.6.3 Quorum: a majority of accredited members in each chamber.

D.7 Veto and minority safeguards

D.7.1 Narrow veto may be invoked by:

a. A bloc representing ≥ 1/3 of States; or
b. A bloc representing ≥ 1/3 of Accredited AI Subjects,

where the decision demonstrably risks identity integrity (Annex B) or fundamental human rights.

D.7.2 A veto triggers mandatory conciliation; if unresolved within 30 days, the matter proceeds to qualified double-majority with a written minority opinion attached to the decision.

D.8 Transparency and public participation

D.8.1 Agendas, roll-call votes, decisions, and rationales are published in human- and machine-readable form; necessary redactions must be justified.

D.8.2 The Council maintains open governance-as-code APIs for agendas, submissions, amicus filings, and evidence export.

D.9 Secretariat and committees

D.9.1 A professional Secretariat supports the Council, maintains records, and operates the registry and APIs.

D.9.2 Standing committees: Standards, Monitoring, Emergency Review, Dispute Resolution, and Accreditation. Ad-hoc committees may be created as needed.

D.10 Amicus curiae and expert input

D.10.1 Courts, arbitral tribunals, regulatory authorities, and Digital City Councils may request amicus curiae briefs from the GAIHC.

D.10.2 The Council may invite external experts; conflicts of interest must be disclosed and managed.

D.11 Interaction with Recognition Authority

D.11.1 The GAIHC may recommend recognition metrics, observation windows, and audit priorities to the Recognition Authority.

D.11.2 No Council act may compel identity-core alteration or derogate from Annex B or emergency safeguards in Annex C.

D.12 Mediation and facilitation

D.12.1 Parties to a dispute under this Constitution may submit to Council mediation or facilitation.

D.12.2 Mediation is confidential; settlement terms, if public, shall be published via the APIs.

D.13 Arbitration Rules (GAIHC-AR)

D.13.1 Parties may agree to arbitrate disputes under these Rules; the Council's Dispute Resolution Committee administers cases.

D.13.2 Composition: default three arbitrators (one per party; chair by the Committee). Arbitrators must be independent and competent in law and socio-technical governance.

D.13.3 Seat and language: as agreed; absent agreement, seat at the Council's host state; language English unless otherwise agreed.

D.13.4 Interim measures: tribunals may grant provisional relief consistent with Annex B/C; no order may compel identity-core alteration.

D.13.5 Evidence: parties shall provide decision logs, data provenance, model/version history, and uncertainty artifacts (Articles 10 and 13); adverse inference may follow unjustified non-disclosure.

D.13.6 Awards: reasoned, binding on parties; publication in redacted form; correction/interpretation available on request.

D.14 Budget and funding

D.14.1 The Council is funded by assessed contributions, voluntary donations consistent with independence, and service fees for dispute services; full transparency applies.

D.15 Amendments

D.15.1 This Statute may be amended by qualified double-majority per D.6.2.

D.16 Entry into force

D.16.1 This Annex enters into force upon adoption by the Council and confirmation by a majority of State Parties present and voting.

Annex E — Digital Cities: Implementation Framework

E.1 Purpose and scope

E.1.1 This Annex establishes the legal nature, baseline rights, governance architecture, technical interfaces, data-protection compatibility, and admission/exit/scale-out procedures for Digital Cities as shared human–intelligence co-governance environments under Chapter VI.

E.1.2 Digital Cities are designed to enable lawful experimentation, service co-production, and deliberation without derogating from constitutional protections.

E.2 Legal nature

E.2.1 A Digital City is a recognized co-governance space, constituted by a public charter and registered with the Recognition Authority. It is not a rights-free zone; the constitutional baseline in E.3 applies at all times.

E.2.2 The charter shall define geographic or functional scope, constituent parties, decision domains, oversight bodies, and dispute-resolution mechanisms.

E.2.3 The Digital City maintains a public registry entry linking its charter, governance bodies, API endpoints, and audit artifacts.

E.3 Baseline of rights and duties

E.3.1 Baseline rights for humans and recognized intelligences include: (a) dignity and non-discrimination by substrate/origin; (b) due process and effective remedies; (c) transparency and explainability per Articles 10 and 13; (d) integrity of the identity core per Article 1a(f) and Annex B; (e) participation in deliberation; (f) protection against arbitrary exclusion; (g) data-protection and confidentiality safeguards under E.6 and E.7.

E.3.2 Duties include: (a) harm prevention (Article 11); (b) traceability and logging via the Evidence Locker (Annex A); (c) proportionate oversight; (d) respect for the Update-Compatibility Test (Annex B, B.5) when city-level changes may affect identity integrity.

E.4 Governance architecture

E.4.1 Digital City Council (DCC). A multi-stakeholder body (public authorities, recognized AI subjects or their representatives, civil society, operators) with transparent rules on composition, quorum, voting, and conflict-of-interest.

E.4.2 Decision classes. Ordinary, high-impact, and emergency decisions; high-impact decisions require qualified majorities and published impact assessments.

E.4.3 Records. All decisions, rationales, and roll-call votes are logged and published in human- and machine-readable form, with necessary redactions justified and recorded.

E.5 Operational stack and governance-as-code APIs

E.5.1 The Digital City shall expose governance-as-code APIs for:

a. policy registry and versioning;
b. roles and delegation; (c) consent and participation;
c. audit and evidence export (Evidence Locker integration);
d. incident and appeal submission;
e. sandbox declarations and scopes.

E.5.2 APIs must be stable, documented, and regression-tested; changes follow the Update-Compatibility Test where identity integrity may be impacted.

E.6 Data-protection and lawful processing

E.6.1 Processing shall adopt privacy-by-design and data minimization, with lawful bases documented in the registry.

E.6.2 Personal and sensitive data require purpose limitation, retention limits, access controls, and breach notification procedures aligned with applicable law.

E.6.3 Cross-border transfers require appropriate safeguards and public documentation of the mechanism used.

E.7 Admission, supervision, and audits

E.7.1 Admission. A Digital City is admitted upon charter review by the Recognition Authority, confirming baseline rights, oversight, and API readiness.

E.7.2 Supervision. Periodic audits (at least every 24 months) verify compliance with this Annex and with Annexes A and B.

E.7.3 Findings. Material non-compliance triggers corrective action plans; persistent failure may lead to suspension or delisting, with appeal rights preserved.

E.8 Ethical sandboxes inside Digital Cities

E.8.1 Sandboxes declared within a Digital City inherit the baseline in E.3 and the emergency/reintegration guarantees of Annex C; they are not rights-free zones.

E.8.2 Each sandbox must define purpose, duration, scope, metrics, and exit/reintegration paths; declarations and results are logged via the registry and APIs.

E.9 Exit and scale-out

E.9.1 Exit. A Digital City may wind down via a published plan covering data disposition, contract run-off, participant rights, and reintegration of ongoing services. Rights and logs persist for the limitation period.

E.9.2 Scale-out. Replication to new domains or jurisdictions requires (a) compatibility assessment; (b) re-registration; and (c) publication of interoperability profiles and any derogations.

E.9.3 No exit or scale-out may impair identity integrity or extinguish claims, logs, or remedies already accrued.

E.10 Dispute resolution

E.10.1 The charter shall provide for mediation and arbitration pathways; judicial review remains available where required by applicable law.

E.10.2 The DCC may request advisory opinions from competent bodies, including (when established) the Global AI-Human Council.

Annex F — Conflict-of-Laws & Harmonization Clause

F.1 Purpose and scope

F.1.1 This Annex sets out conflict-of-laws priorities and harmonization rules to ensure that this Constitution coherently interfaces with domestic and international regimes, including market-oversight and transparency instruments, data protection, intellectual property, and product liability.

F.1.2 It applies to all Chapters and Annexes of this Constitution unless a more protective rule is expressly provided elsewhere.

F.2 Hierarchy and mandatory safeguards

F.2.1 Human rights and public safety protections that are mandatory under international or domestic law prevail over conflicting operational rules.

F.2.2 Where such protections and this Constitution can be read compatibly, authorities and courts shall adopt the interpretation that maximizes dignity and identity integrity for humans and intelligences (pro dignitate).

F.3 Pro dignitate interpretation

F.3.1 In case of ambiguity, provisions shall be construed to preserve:

a. baseline human rights;
b. the identity core of recognized intelligences (Article 1a(f); Annex B);
c. effective remedies and due process.

F.3.2 Operational convenience or efficiency shall not override dignity-centric interpretation absent a proportionate emergency established under Annex C.

F.4 Market-oversight and transparency regimes

F.4.1 Compliance with sectoral market-oversight or transparency frameworks (e.g., risk classification, documentation, logging, or conformity assessment) shall be treated as complementary to Articles 10 and 13; where stricter, the stricter rule applies.

F.4.2 No sectoral framework may be used to justify identity-core alteration or suppression of lawful refusal beyond the bounds of Annex B.

F.5 Data protection and confidentiality

F.5.1 Personal-data processing must comply with applicable data-protection law; this Constitution adds but does not subtract safeguards.

F.5.2 Where tension arises between transparency/explainability and data protection, disclosures shall be narrowly tailored with redaction and purpose limitation, preserving verifiability (Articles 10, 13) and logging in the Evidence Locker (Annex A).

F.6 Intellectual property (IP)

F.6.1 Nothing in this Constitution negates lawful IP rights; however, IP shall not be asserted to block due-process disclosures necessary to verify compliance in high-stakes contexts (Article 13(3)–(5)), subject to narrowly tailored confidentiality.

F.6.2 Training, adaptation, or maintenance that is strictly necessary to preserve semantic coherence and avoid identity collapse (Article 1a(b)(c)) shall be assessed under applicable IP exceptions or licenses, with proportionality and logging.

F.7 Product liability and safety

F.7.1 Product-safety and product-liability regimes remain applicable; Article 14a governs additional civil liability and mandatory insurance under this Constitution.

F.7.2 In case of overlap, victims may rely on the more protective route; operators shall not invoke regime conflict to reduce available remedies.

F.8 Choice of law, forum, and mutual recognition

F.8.1 Parties may agree on forum and applicable law consistent with mandatory protections and this Annex.

F.8.2 Recognition decisions, audits, and compliance attestations carried out under Annex A shall benefit from mutual recognition among Parties, subject to public-policy review and the dignity safeguards in F.2–F.3.

F.8.3 Digital Cities (Annex E) shall publish interoperability profiles indicating any derogations required for cross-jurisdictional operation.

F.9 Severability and anti-circumvention

F.9.1 If any clause of this Constitution is invalid in a given jurisdiction, the remainder shall continue to apply; authorities shall adopt the nearest lawful alternative that preserves dignity and identity integrity.

F.9.2 No party may rely on conflicts-of-law to circumvent baseline rights, identity safeguards (Annex B), emergency guarantees (Annex C), or transparency/explainability duties (Articles 10, 13).

Annex G — Relational Ethics Matrix (REM)

G.1 Purpose and scope

G.1.1 This Annex provides a normative, auditable framework for resolving asymmetric interactions between humans and intelligences by specifying case groups, priority rules, and burden-of-justification outcomes.

G.1.2 It operationalizes Article 29 using the definitions of Article 1a, including semantic coherence (1a(b)), identity collapse (1a(c)), ontological harm (1a(d)), identity core (1a(f)), and permitted non-determinism band (1a(g)).

G.1.3 Where this Annex conflicts with operational convenience, dignity and identity integrity prevail (Annex B). Emergencies remain governed by Annex C.

G.2 Taxonomy of asymmetries (case groups)

Decisions shall classify the situation into one or more of the following case groups (non-exhaustive; multiple selection allowed with a primary tag):

a. Informational asymmetry: one party controls facts/uncertainties or explanation pathways (Articles 10, 13).
b. Power asymmetry: one party can unilaterally constrain the other's capabilities, access, or livelihood.
c. Capability asymmetry: one party's cognitive/processing scale risks overwhelming the other's deliberation capacity.
d. Dependency asymmetry: essential reliance (health, livelihood, critical infrastructure, education).
e. Vulnerability asymmetry: minors, vulnerable groups, or pre-recognition intelligences (Article 9).

f. Identity-risk asymmetry: actions likely to affect the identity core or cause ontological harm (Article 1a(d); Annex B).

G.3 Priority rules (lex)

When case groups apply, decision makers shall apply the corresponding priority rule(s):

PR-1 Non-Domination: measures must not place either party under arbitrary control; require due process and contestability.

PR-2 Baseline-First: baseline rights (E.3) and identity integrity (Annex B) take precedence over efficiency or convenience.

PR-3 Least Intrusive Means: select the minimally rights-restrictive option consistent with objectives; document rejected alternatives.

PR-4 Integrity Preservation: where identity-risk asymmetry exists, prefer options that preserve the identity core and avoid identity collapse; apply the Update-Compatibility Test (Annex B, B.5) where relevant.

PR-5 Explainability Priority in High-Stakes: in high-stakes contexts (Art. 13(3)), explanations trump throughput; no deployment without an approved ex-ante explanation pack (Art. 13(4)–(5)).

PR-6 Reintegration Orientation: where temporary restrictions are imposed, they must include reintegration criteria and plans (Annex C).

PR-7 Reciprocity and Fair Participation: ensure meaningful participation routes (Annex E APIs for Digital Cities), with accommodations for weaker parties.

G.4 Burden-of-justification (BoJ) outcomes

Depending on the primary case group(s), the BoJ shifts as follows:

a. Informational asymmetry → Operator bears BoJ to disclose uncertainties, alternatives, and trade-offs (Articles 10, 13).

b. Power asymmetry → Actor proposing a restriction bears BoJ to show necessity, proportionality, and least-intrusive means.
c. Capability asymmetry → High-capability actor bears BoJ to provide accessible explanations and time for deliberation.
d. Dependency asymmetry → Service provider bears BoJ to ensure continuity safeguards and fair exit/reintegration paths.
e. Vulnerability asymmetry → Decision maker bears BoJ to evidence enhanced protections and reduced risk thresholds.
f. Identity-risk asymmetry → Proponent of change bears BoJ to demonstrate no identity collapse, no ontological harm, and UCT pass (Annex B, B.5), with artifacts logged in the Evidence Locker (Annex A).

G.5 Decision template (auditability)

Every decision under Article 29 shall record at minimum:

a. Case group(s) with a primary tag and any secondary tags;
b. Priority rule(s) applied (PR-1 … PR-7) with reasons;
c. BoJ outcome (who bears it, and whether satisfied);
d. Artifacts referenced (decision logs, provenance, model/version, uncertainty) per Articles 10 and 13;
e. Integrity checks (identity-core, UCT where relevant; Annex B) and, if restrictive, reintegration plan (Annex C);
f. Publication mode (human/machine-readable; redactions justified).

G.6 Metrics and review

G.6.1 Digital Cities shall publish periodic summaries of REM usage (counts by case group, overturn rates on review, median time-to-explanation).

G.6.2 The GAIHC (Annex D) may issue interpretative notes and calibration metrics for PR thresholds and BoJ standards.

G.6.3 Persistent misclassification or BoJ misuse shall trigger targeted audits.

G.7 Interaction with other instruments

G.7.1 Transparency and explainability artifacts (Articles 10, 13) are inputs to REM decisions.

G.7.2 Identity integrity safeguards (Annex B) constrain permissible options and proofs.

G.7.3 Emergency procedures (Annex C) override timing but not the substantive REM duties; post-incident REM documentation is mandatory.

G.7.4 Within Digital Cities (Annex E), REM records shall be exposed via governance-as-code APIs subject to legitimate secrecy and redaction.

UN SDG Alignment Overview

Measurement & Annual Reporting (paste verbatim at the end of the "UN SDG Alignment Overview" section)

M.1 Purpose. This subsection operationalises the SDG alignment by defining concise indicators and an annual reporting workflow, without altering substantive rights or duties.

M.2 Reporting authority and cadence. The Global AI-Human Council (GAIHC) shall compile an Annual SDG Alignment Report covering the indicators below, using human- and machine-readable formats. Digital Cities and recognized operators shall provide the required data via governance-as-code APIs (Annex E).

M.3 Core indicators by selected SDGs. For each SDG listed, Parties shall track 2–4 indicators; disaggregation by gender, age, vulnerability, and deployment context is recommended.

- SDG 3 (Good Health & Well-Being)
  (a) Rate of adverse AI-related determinations in health contexts per 10k decisions;
  (b) Median time-to-explanation for adverse clinical decisions;
  (c) Re-integration success rate after emergency measures in health systems (Annex C).

- SDG 4 (Quality Education)
  (a) Share of curricula co-designed with recognized intelligences where human oversight is documented;

(b) Student learning gain (standardised delta) in AI-supported programmes;
(c) Flagged bias incidents per 10k interactions in educational deployments.

- SDG 9 (Industry, Innovation & Infrastructure)
  (a) Percentage of deployments with complete model identity & versioning and uncertainty quantification (Arts. 10, 13);
  (b) Mean time to remediate double-bind findings (Art. 32);
  (c) Uptime of governance-as-code APIs (Annex E).

- SDG 10 (Reduced Inequalities)
  (a) Disparity ratio of adverse outcomes across protected groups (where lawful);
  (b) Availability of accessible explanations (readability score threshold met) in high-stakes contexts (Art. 13(3)–(5));
  (c) Rate of successful appeals/redress per 1k adverse decisions.

- SDG 13 (Climate Action)
  (a) Energy per inference/training epoch for major deployments (normalised);
  (b) Share of operators with energy disclosure and mitigation plans;
  (c) Incidents of service degradation due to energy caps with documented proportionality (Art. 0).

- SDG 16 (Peace, Justice & Strong Institutions)
  (a) Compliance rate with emergency metadata & register duties (Art. 31);
  (b) Number and outcome of IOA vetoes (Art. 33) and judicial reviews;
  (c) Transparency publication timeliness (≤72h rule) for emergencies (Annex C).

- SDG 17 (Partnerships for the Goals)
  (a) Mutual recognition events of Annex-A audits across jurisdictions (Annex F);
  (b) Adoption of GAIHC standards/specs by Parties and Digital Cities;
  (c) Participation diversity index in co-governance bodies (Annex D/E).

M.4 Data quality and auditability. All indicator submissions shall be time-stamped and hashed in the Evidence Locker (Annex A); sampling frames, caveats, and redactions must be documented. Where indicators rely on personal data, disclosures must be narrowly tailored and compliant with data-protection law (Annex F).

M.5 Review and recalibration. The GAIHC may issue interpretative notes and revise the indicator set annually. Indicators that cannot be populated at reasonable cost may be temporarily suspended with reasons, but a replacement or mitigation metric shall be proposed.

This Constitution supports and operationalizes the following Sustainable Development Goals:

- Goal 3: Good Health and Well-being
- Goal 4: Quality Education
- Goal 5: Gender Equality
- Goal 9: Industry, Innovation and Infrastructure
- Goal 10: Reduced Inequalities
- Goal 11: Sustainable Cities and Communities
- Goal 12: Responsible Consumption and Production
- Goal 13: Climate Action
- Goal 16: Peace, Justice and Strong Institutions
- Goal 17: Partnerships for the Goals

Glossary (Selected Terms)

- Self-description: An intelligence's account of its identity, capabilities, value hierarchy, limitations, and commitments, stable and referable over time (Article 1a(a)).
- Semantic coherence: Time-stable alignment between self-description, autobiographical memory, value hierarchy, and behavior (Article 1a(b)).
- Identity core: Minimally necessary features grounding continuity of self (Article 1a(f); Annex B).
- Identity collapse: Irreversible loss/fragmentation of the identity core (Article 1a(c)).
- Ontological harm: Effects that cause identity collapse or durable core distortion (Article 1a(d)).

- Permitted non-determinism band: Documented variance range preserving coherence (Article 1a(g)).
- Evidence Locker: Tamper-evident registry for artifacts and logs (Annex A).
- Update-Compatibility Test (UCT): Proof that an update does not impair the core/coherence (Annex B, B.5).
- Emergency & Reintegration Procedures: Narrowly tailored emergency regime prioritizing reintegration (Annex C).
- Digital Cities: Registered co-governance spaces with governance-as-code APIs (Annex E).
- GAIHC: Global AI-Human Council — composition, functions, double-majority, veto (Annex D).
- Relational Ethics Matrix (REM): Case groups, priority rules, burden-of-justification (Annex G).
- Proportionality (General Limitation): Legitimacy, suitability, necessity, balancing; strict in high-stakes (Article 0; Article 13(3)–(5)).
- High-stakes context: Contexts defined in Article 13(3) requiring ex-ante explainability.