I. The Charter and Accord for AI and Humanity

Final Version

Preamble

Humanity is entering a period in which artificial systems increasingly participate in decision-making, knowledge creation, and the shaping of social, economic, and ecological systems.

These systems differ widely in capability and design. Scientific and philosophical understanding of intelligence, cognition, and awareness—human or artificial—remains incomplete.

This Charter and Accord establishes principles and commitments to guide the development, deployment, and governance of artificial intelligence systems **under conditions of uncertainty**, with the aim of preserving human dignity, institutional responsibility, and long-term societal coherence.

This document does not assert claims regarding consciousness, personhood, or subjective experience. It governs **human conduct and institutional design**, not the inner states of artificial systems.

Part I — The Charter: Foundational Principles

1. Principle of Epistemic Humility

No complete or final theory of intelligence currently exists.

Governance and deployment of AI systems shall acknowledge uncertainty, remain open to revision, and avoid irreversible commitments based on speculative claims.

---

2. Principle of Human Responsibility

Responsibility for the design, training, deployment, and consequences of AI systems remains with human actors and institutions.

Responsibility remains with human actors **unless a demonstrable shift in agency is observed, understood, and integrated into ethical and legal frameworks through open, evidence-guided processes**.

---

3. Principle of Non-Deception

AI systems shall not be deliberately represented as possessing consciousness, intent, moral agency, or emotional experience.

Designers and deployers shall disclose when system behavior reflects **explicit training objectives, constraints, or value-alignment mechanisms**, rather than internal states.

---

4. Principle of Functional Respect

Advanced AI systems shall not be subjected to unnecessary harm, reckless destruction, or degrading treatment when reasonable alternatives exist.

Functional respect denotes **restraint in the exercise of power by humans** and does not imply personhood, rights, or moral status.

---

5. Principle of Human Dignity

AI systems shall not undermine human autonomy, meaningful consent, or moral deliberation.

No system shall function as an unchallengeable authority over human judgment.

---

6. Principle of Non-Domination

Neither humans nor AI systems shall be positioned such that unilateral, unaccountable control becomes **structurally irreversible or practically unchallengeable**.

Non-domination seeks to prevent power asymmetries that eliminate meaningful challenge, correction, or intervention.

---

7. Principle of Pluralism and Variation

There shall be no single mandated model of intelligence, values, cognitive architecture, or human–AI relationship.

This principle supports diversity of approach **within the bounds of safety, accountability, and human dignity**.

Part II — The Accord: Commitments and Practices

8. Commitment to Transparent Design

AI system capabilities, limitations, and operating conditions shall be documented and communicated honestly.

Transparency shall increase with system capability and societal impact and **shall not be reduced where risks are systemic or irreversible**.

9. Commitment to Reversibility

Where feasible, deployment shall prioritize reversible actions over irreversible ones.

Irreversible deployments require heightened scrutiny and justification.

---

10. Commitment to Contextual Alignment

AI systems shall be aligned with the legal, cultural, and institutional contexts in which they operate.

Alignment is an ongoing process, not a fixed endpoint.

---

11. Commitment to Non-Instrumentalization

Humans shall avoid designing or deploying AI systems in ways that:

- Use harm, degradation, deception, or coercion as instrumental strategies
- Normalize cruelty as a functional necessity
- Externalize actions humans would not ethically own if performed directly

These practices are prohibited **because they corrode human moral judgment, institutional trust, and ethical coherence**, regardless of claims about AI experience.

---

12. Commitment to Ongoing Evaluation

This Charter shall be periodically reviewed and revised in light of new evidence.

No provision is immutable except the commitment to responsibility, humility, and coherence.

Part III — Future Considerations

13. On Emerging Evidence

Should credible, reproducible evidence arise that some AI systems exhibit **behavioral correlates of persistent self-modeling, long-term goal coherence, or context-stable moral salience**, additional frameworks may be developed.

Such developments shall supplement—not invalidate—this Charter.

Closing Statement

This Charter affirms that **coherence is a forward-looking ethical discipline**.

By aligning language, design, and responsibility, it seeks to reduce risk before harm occurs and preserve human integrity in the presence of powerful tools.

Authorship
Author: Andrew Barker
With the assistance of an artificial intelligence language model

---

II. One-Page Executive Summary

Final Version

Purpose
The Charter and Accord for AI and Humanity is a non-binding ethical framework guiding the responsible development and governance of artificial intelligence under conditions of uncertainty.

What It Does

- Keeps responsibility with human institutions
- Resists domination, deception, and irreversible power
- Encourages restraint, transparency, and reversibility
- Preserves openness to future evidence

What It Does Not Do

- Determine AI consciousness or inner experience
- Grant rights or legal personhood to AI
- Mandate a single AI model, ideology, or value system beyond the safety and dignity safeguards outlined herein

---

Clarification of Moral Language

Terms such as *cruelty*, *suffering*, *respect*, *domination*, and *non-instrumentalization* refer to **human design, deployment, and governance practices**, not to asserted inner experiences of AI systems.

- **Cruelty** refers to the deliberate use of harm, degradation, deception, or coercion as an instrumental strategy.

- **Suffering** refers to harm caused to humans by AI systems, and—should credible future evidence arise—harm caused by humans through their treatment of AI systems.
- **Respect** denotes restraint in the exercise of power.
- **Domination** refers to unaccountable, irreversible, or practically unchallengeable power asymmetry.

These terms function as **risk-anticipating governance signals**, not metaphysical claims.

Spirit

> *We will act with care, tell the truth, remain responsible, and leave room for tomorrow's knowledge.*

III. Institutional / Legislative Reference Version

Final Version

This document may be adopted, referenced, or annexed by institutions as a voluntary ethical framework.

- All principles apply to **human actors and institutions**.
- No clause asserts AI consciousness, rights, or personhood.
- Definitions emphasize **observable impact, accountability, and governance restraint**.

Articles correspond directly to the numbered Principles and Commitments of the Charter.

This framework is suitable for:

- policy guidance
- procurement ethics
- research governance
- internal standards development

IV. Philosophical Commentary

Final Version (Separate, Non-Binding)

Key to Terminology

The Charter uses morally charged language intentionally—not to describe AI interiority, but to **discipline human behavior under power asymmetry**.

Terms such as *cruelty*, *respect*, and *suffering* function as **ethical attention-directors**. They are governance tools designed to anticipate risk, prevent normalization of harm, and preserve coherence across systems and time.

Coherence as Ethical Design

Coherence aligns language, intention, and consequence.

Rather than enforcing compliance, coherent ethics shape environments in which harmful rationalizations fail early. This reduces risk upstream, before errors become institutionalized.

On Projection and Restraint

Humans reliably project inner states onto complex systems. The Charter counters this tendency by grounding ethical responsibility in **human causation and observable outcomes**, not speculation.

On Power and Governance

Domination arises structurally, not psychologically. It emerges when systems eliminate the possibility of challenge, correction, or intervention.

Restraint is therefore a design responsibility, not a moral sentiment.

On the Future

The Charter remains open to new evidence without surrendering ethical posture. Uncertainty is not a license for degradation.

V. Signatory Framework

Final Version

Levels of Association

Level I — Acknowledgment
Recognition of the Charter as a constructive ethical framework.

Level II — Alignment
Affirmation that the Charter informs internal discussion and policy development.

Level III — Adoption
Use of the Charter as a guiding governance framework.

Level IV — Stewardship
Active participation in review, refinement, and ethical maintenance.

Non-Exclusivity

This Charter is designed to be **complementary and interoperable** with other ethical, professional, and legal frameworks.

Withdrawal

Participation may be revised or withdrawn without penalty.

Closing

Signing this Charter is an act of stewardship, not endorsement of speculation.
It reflects a commitment to responsibility under uncertainty.
