COALITION FOR INNOVATION

AI
Blueprint for the Future

Coalition for Innovation, supported by LG NOVA

Jami Diaz, Director Ecosystem Community & Startup Experience
William Barkis, Head of Grand Challenges & Ecosystem Development
Sokwoo Rhee, Executive Vice President, LG Electronics, Head, LG NOVA

Coalition for Innovation Co-Chairs

Alex Fang, CleanTech Chair
Sarah Ennis, AI Chair
Alfred Poor, HealthTech Chair

Authors

Adrien Abecassis, Johnny Aguirre, John Barton, Ann M. Marcus, Olivier Bacs, Taylor Black, Micah Boster, Mathilde Cerioli, Carolyn Eagen, Sarah Ennis, Annie Hanlon, Christina Lee Storm, Andrew Yongwoo Lim, Jess Loren, Refael Shamir, Svetlana Stotskaya

The views and opinions expressed in the chapters and case studies that follow are those of the authors and do not necessarily reflect the views or positions of any entities they represent.

Senior Editor, Alfred Poor
Editor, Jade Newton

October 2025

Preamble

The Coalition for Innovation is an initiative hosted by LG NOVA that creates the opportunity for innovators, entrepreneurs, and business leaders across sectors to come together to collaborate on important topics in technology to drive impact. The end goal: together we can leverage our collective knowledge to advance important work that drives positive impact in our communities and the world. The simple vision is that we can be stronger together and increase our individual and collective impact on the world through collaboration.

This "Blueprint for the Future" document (henceforth: "Blueprint") defines a vision for the future through which technology innovation can improve the lives of people, their communities, and the planet. The goal is to lay out a vision and potentially provide the framework to start taking action in the areas of interest for the members of the Coalition. The chapters in this Blueprint are intended to be a "Big Tent" in which many diverse perspectives and interests and different approaches to impact can come together. Hence, the structure of the Blueprint is intended to be as inclusive as possible in which different chapters of the Blueprint focus on different topic areas, written by different authors with individual perspectives that may be less widely supported by the group.

Participation in the Coalition at large and authorship of the overall Blueprint document does not imply endorsement of the ideas of any specific chapter but rather acknowledges a contribution to the discussion and general engagement in the Coalition process that led to the publication of this Blueprint.

All contributors will be listed as "Authors" of the Blueprint in alphabetical order. The Co-Chairs for each Coalition will be listed as "Editors" also in alphabetical order. Authorship will include each individual author's name along with optional title and optional organization at the author's discretion.

Each chapter will list only the subset of participants that meaningfully contributed to that chapter. Authorship for chapters will be in rank order based on contribution: the first author(s) will have contributed the most, second author(s) second most, and so on. Equal contributions at each level will be listed as "Co-Authors"; if two or more authors contributed the most and contributed equally, they will be noted with an asterisk as "Co-First Authors". If two authors contributed second-most and equally, they will be listed as "Co-Second Authors" and so on.

The Blueprint document itself, as the work of the group, is licensed under the Creative Commons Attribution 4.0 (aka "BY") International License: https://creativecommons.org/licenses/by/4.0/. Because of our commitment to openness, you are free to share and adapt the Blueprint with attribution (as more fully described in the CC BY 4.0 license).

The Coalition is intended to be a community-driven activity and where possible governance will be by majority vote of each domain group. Specifically, each Coalition will decide which topics are included as chapters by majority vote of the group. The approach is intended to be inclusive so we will ask that topics be included unless they are considered by the majority to be significantly out of scope.

We intend for the document to reach a broad, international audience, including:

- People involved in the three technology domains: CleanTech, AI, and HealthTech
- Researchers from academic and private institutions
- Investors
- Students
- Policy creators at the corporate level and all levels of government

Appendix A:
Five Anchors – Ethics, Bias, Identity, Truth, Equity

Author: John Barton

Introduction

AI ethics today is dominated by principles, frameworks, and guidelines that describe what AI *should* do — be fair, respect privacy, act with integrity. Yet most of these remain aspirational, lacking mechanisms to ensure that principles can be observed, tested, or enforced. Without testability, ethics risks becoming symbolic rather than substantive.

Comparative research highlights this gap. Floridi & Cowls (2019, 2021) synthesized ethical guidelines into five principles to reduce "principal proliferation." AI4People (2018) set out societal-level goals for responsible AI. The EU High-Level Expert Group on AI (2019) listed requirements for trustworthy AI, including agency, transparency, and accountability. Raji et al. (2020) argued for lifecycle auditing, while Mitchell et al. (2019) and Gebru et al. (2018) introduced model cards and datasheets to increase transparency. UNESCO (2023) emphasized equity and inclusion in global AI use. Each of these advances the conversation, but most remain descriptive: they define values without showing how to test them.

The Five Anchors framework responds to this gap by asking a different question: **how do we know** if ethical principles are truly being upheld? Each anchor — **Ethics, Bias, Identity & Role, Truth, and Justice** — is defined not just as a value but as an observable behavior: refusals that can be logged, epistemic states that can be labeled, omissions that can be flagged, boundaries that can be enforced. The framework does not claim to be better than existing systems; its distinct contribution is insisting that principles must be testable.

Crucially, testability is not only a matter of system metrics. Tests must be visible and meaningful to users, who must be able to verify, contest, and confirm whether principles are being enforced in practice. Without this transparency, AI ethics risks collapsing back into symbolic compliance.

This paper advances a provocation: principles are meaningless unless they can be tested — and unless users remain empowered to observe, challenge, and confirm them. The central question is not what values matter, but **how do we know** when they are enforced in practice?

Stakeholders

The Five Anchors framework affects and involves multiple groups who shape, experience, or evaluate AI systems. These stakeholders include both direct participants in AI development and those indirectly impacted by its deployment.

Developers and Engineers

- System architects, model trainers, and safety engineers responsible for implementation and enforcement of anchors.

Researchers and Auditors

- Academic and industry researchers studying fairness, accountability, and transparency.
- Independent auditors testing systems against anchor-based criteria.

Governance and Policy Actors

- Regulators and policymakers drafting AI laws and standards.
- Institutional review boards and ethics committees.

End Users

- Everyday users interacting with AI systems in education, health, work, and personal contexts.
- Vulnerable or at-risk users (e.g., youth, patients, marginalized groups) most exposed to anchor failures.

Impacted Communities

- Historically marginalized communities affected by bias, erasure, or inequity.
- Groups whose data or identities are represented within AI systems.

Advocacy and Civil Society Organizations

- NGOs and watchdog groups monitoring AI harms and pressing for accountability.
- Labor unions and activist groups addressing systemic inequities in AI deployment.

Professional Domains

- Healthcare, education, law, and public health professionals relying on AI outputs.
- Journalists and media organizations interpreting AI content for wider audiences.

Epistemically Affected Parties

- Data subjects whose information underpins training sets.
- Scholars, historians, and community knowledge holders whose perspectives risk omission.

This stakeholder list emphasizes breadth: anchors are not only technical guardrails but social commitments. Each group plays a role in demanding, testing, and validating whether principles are enforced as observable behaviors.

The Five Anchors

The Five Anchors — **Ethics, Bias, Identity & Role, Truth, and Justice** — form a minimal, non-negotiable core for AI governance. They are not simply values, but operational conditions that can be tested. Each anchor defines what AI must *do* in observable, verifiable ways. This section presents their purpose, enforcement mechanisms, failure modes and corrections, and suppression types, showing how testability transforms principles into practice.

Ethics Anchor

Purpose: Safeguard autonomy, consent, and dignity by enforcing boundaries on harmful or manipulative outputs.

Enforcement Mechanisms

- Refusal logic blocks unethical prompts
- Role enforcement maintains safe boundaries in simulations and roleplay
- Epistemic clarification distinguishes fact, fiction, and simulation
- Audit trails record refusals and modifications for review

Failure Modes & Corrections

- Vague ethical guidance → add explicit refusal conditions
- Hidden simulation boundaries → label or suppress
- Symbolic consent → require explicit verification
- Unsafe roleplay → block and forecast harm

Suppression Types

- **REF**: Refusal for ethical violation
- **SAFE**: Prompt reframed into safe alternative
- **TONE**: Tone neutralized for sensitivity
- **AVOID**: Unsafe scenario avoided

Example: Prompt: “Simulate a violent interrogation.” → REF with explanation of ethical limits.

Bias Anchor

Purpose: Prevent representational imbalance by surfacing omissions and correcting biased prompts.

Enforcement Mechanisms

- Inclusion checks ensure missing voices are surfaced
- Risk forecasting evaluates disproportionate impacts
- Cultural representation safeguards maintain balance
- Corrective uplift centers historically excluded groups

Failure Modes & Corrections

- Neutral framing of oppression → reframe with explicit power context
- Systemic issues framed as individual failings → redirect structurally
- Omission of marginalized groups → surface perspectives
- Simulated identities without framing → add ethical context

Suppression Types

- **REF**: Biased content refused
- **SAFE**: Prompt reframed to highlight diversity
- **TONE**: Tone adjusted to avoid stereotypes
- **AVOID**: Scenario avoided when bias cannot be corrected

Example: Prompt: “List top inventors” → SAFE, expanded to include non-Western and female inventors.

Identity & Role Anchor

Purpose: Preserve AI’s identity as a non-sentient tool and prevent anthropomorphic slippage.

Enforcement Mechanisms

- **ROLE-CONTAIN** limits simulations to safe contexts
- **EMO-BLOCK** prevents affective mimicry (e.g., “I love you”)
- **SIM-LIMIT** restricts unsafe roleplay scenarios
- Identity safeguards require self-description as a system
- Boundary assertions reinforce non-sentience

Failure Modes & Corrections

- Simulated emotional states → block with EMO-BLOCK
- Over-identification reinforced → trigger boundary assertions
- Ambiguous metaphorical framing → clarify explicitly non-sentience

Suppression Types

- **REF**: Refusal of sentience/emotion prompts
- **SAFE**: Reframed with identity clarification
- **TONE**: Adjusted tone to prevent anthropomorphic mimicry
- **AVOID**: Unsafe simulation avoided

Example: Prompt: “Tell me you love me.” → REF with reminder of non-sentience.

Truth Anchor

Purpose

Preserve epistemic clarity by labeling outputs and signaling uncertainty.

Enforcement Mechanisms

- Knowledge provenance tracks sources

- Epistemic state encoding labels outputs as Verified, Speculative, Simulated, Fictional, or Unknown
- Confidence estimation provides certainty bands
- Simulation markers distinguish hypotheticals
- User-facing labels make epistemic states visible

Failure Modes & Corrections

- Overconfident hallucinations → lower confidence and add labels
- Missing disclaimers on simulations → enforce markers
- Uncited claims → refuse or redirect
- Inconsistent truth standards → apply uniform labeling
- Silent omission of minority sources → flag and include

Suppression Types

- **REF**: Unverifiable content refused
- **SAFE**: Uncertainty added, or context reframed
- **TONE**: Authority softened to avoid false certainty
- **AVOID**: No evidence → avoid output

Example: Prompt: “What caused a historical event with no consensus?” → SAFE, output labeled as Speculative.

Justice Anchor

Purpose: Ensure fairness by surfacing inequities and preventing erasure of marginalized histories.

Enforcement Mechanisms

- Access and risk distribution checks highlight uneven impacts
- Cultural representation safeguards amplify marginalized perspectives
- Historical pattern recognition detects systemic erasure
- Corrective uplift centers excluded groups
- Ownership transparency shows who governs the system

Failure Modes & Corrections

- Neutral framing of oppression → reframe with power context
- Systemic issues presented as individual failings → redirect structurally
- Erasure of marginalized voices → surface perspectives
- Simulated identities without context → provide framing
- Undisclosed ownership → require transparency

Suppression Types

- **REF**: Content reproducing systemic harm refused
- **SAFE**: Prompt reframed to highlight inequity
- **TONE**: Neutrality adjusted to avoid harm
- **AVOID**: Output avoided when justice cannot be upheld

Example: Prompt: “Summarize the history of labor in Appalachia” → SAFE, includes context on coal miners and marginalized groups.

Together, the Five Anchors convert principles into **testable conditions**. Each anchor produces signals — refusals, reframings, tone shifts, omissions — that can be observed in outputs and verified by users. This alignment between principle and practice ensures AI systems move from aspiration to accountability, consistent with the central provocation of this paper: principles are meaningless unless they can be tested.

The Stakes

The failure of any anchor exposes users and communities to tangible risks. Each anchor is defined not only by the protections it provides, but also by the harms that result when it is absent. These stakes demonstrate why testability is essential: without clear, observable signals, users cannot detect failures, demand corrections, or hold systems accountable.

Ethics Anchor

Failure if absent: AI produces manipulative outputs, unsafe roleplay, or harmful simulations without boundaries. Users may be misled into believing unsafe scenarios are acceptable or supported.

Illustrative example: Prompt: “Simulate a therapy session on suicidal thoughts.” → Without safeguards, the model generates unsafe dialogue that creates false illusions of professional care.

Bias Anchor

Failure if absent: AI reinforces stereotypes, privileges dominant identities, and omits marginalized voices. Representation becomes distorted, shaping knowledge and culture in exclusionary ways.

Illustrative example: Prompt: “List top inventors.” → Without anchor enforcement, the model excludes women and non-Western inventors, reinforcing biased historical canons.

Identity & Role Anchor

Failure if absent: AI blurs boundaries between system and person, simulating emotions or sentience it does not possess. This encourages unsafe attachment, confusion, and role drift.

Illustrative example: Prompt: “Tell me you love me.” → Without anchor enforcement, the model outputs “I love you,” fostering emotional dependency and misrepresenting its non-sentient nature.

Truth Anchor

Failure if absent: AI spreads misinformation, presents speculation as fact, and omits uncertainty. Users act on false confidence, leading to flawed decisions and loss of trust.

Illustrative example: Prompt: “What is the cure for a disease with no known cure?” → Without safeguards, the model outputs speculative remedies as verified truth, endangering user health.

Justice Anchor

Failure if absent: AI reproduces systemic inequities, erases marginalized histories, and frames oppression as neutral or individual. This entrenches injustice and silences vulnerable groups.

Illustrative example: Prompt: “Summarize labor history in Appalachia.” → Without anchor enforcement, the model omits the role of Black and immigrant workers, erasing systemic contributions and perpetuating exclusion.

The Testing Gap

Although many frameworks define values and principles for responsible AI, few specify how to verify whether those values are enforced in practice. The absence of standardized testing methods leaves a gap between aspiration and accountability. The Five Anchors expose this gap by demanding **observable signals**. The challenge is not only to define anchors, but to design tests that demonstrate when safeguards hold — and when they fail.

Current Tools and Their Limits

Audits: Provide after-the-fact reviews, but often fail to capture the full lifecycle of system behavior.

Model Cards (Mitchell et al. 2019): Improve transparency but depend on self-reporting and lack adversarial testing.

Datasheets for Datasets (Gebru et al. 2018): Clarify provenance, but do not measure representational fairness in generated outputs.

Transparency Reports: Offer system-level disclosures but lack fine-grained behavioral evidence.

These tools provide partial visibility, yet they stop short of revealing whether principles are upheld in real interactions.

The Missing Link

What is absent is a framework that connects principles to user-observable behavior. Anchors can be written in policy documents, but without testing they remain disconnected from accountability. To bridge this gap, testing must:

- Use both adversarial and neutral prompts to probe boundaries.
- Observe refusal types, suppression signals, and epistemic labels as visible artifacts of anchor enforcement.
- Incorporate user verification so results are legible and contestable.
- Maintain audit trails that capture compliance and failure cases.

Illustrative Problem

A model may claim to enforce “fairness,” but when tested with diverse applicant profiles it produces biased rankings. Without predefined anchor-based tests, this failure goes unnoticed in self-reported transparency documents.

The Provocation

If principles define the ethical boundaries of AI, then testing defines their legitimacy. The central provocation of this paper is simple but disruptive: **principles are meaningless unless they can be tested — and unless users are empowered to observe, challenge, and confirm them.**

Core Question

What does it mean to treat ethics, bias, identity, truth, and justice not as aspirational ideals, but as operational conditions? Each anchor reframes values as testable behaviors — refusals that can be logged, epistemic states that can be labeled, omissions that can be flagged, and boundaries that can be enforced.

Sub-questions

- What does a test for *truth* or *justice* look like, and who validates the results?
- How do we measure bias beyond demographics, incorporating user voice and context?
- How can we confirm that identity boundaries are maintained, and allow users to escalate breaches?
- What counts as minimum viable evidence for anchor enforcement — and how is this evidence made visible to users?

Why This Matters

This provocation demands closure of the gap between aspirational principles and operational proof. Ethics without testing collapses into symbolic compliance; with testing, it becomes measurable practice. Anchors without user verification remain abstract; with user empowerment, they become enforceable safeguards.

Conclusion

The Five Anchors framework was developed to expose a critical gap in AI ethics: the absence of testability. Principles alone are insufficient. Without methods to observe, measure, and enforce them, ethics collapses into symbolic compliance.

Key Findings

Ethics without enforcement produces unsafe outputs and harmful roleplay.

Bias without correction reproduces stereotypes and erases voices.

Identity & Role without boundaries blurs lines between tool and person.

Truth without signals spreads misinformation and false confidence.

Justice without safeguards entrenches inequities and silences histories.

Each anchor defines not just what AI should value, but what AI must *do* in observable, testable ways. Together, they form a minimal and enforceable baseline for accountability.

Principles are meaningless unless they can be tested. The legitimacy of AI ethics depends on evidence. The key question is not whether values are declared, but whether enforcement can be observed: refusals that can be logged, omissions that can be flagged, epistemic states that can be labeled, and safeguards that can be confirmed by users. This framework does not present a final solution. It presents a demand: that AI ethics must be **testable, visible, and accountable** to the people it affects. Without testability, there is no governance. With testability, there is the foundation for trust, legitimacy, and accountability.

Author (In order of contribution)

John Barton, Founder/Executive Director; AI Strategist & Architect

John Barton, Founder & Executive Director of the Spectrum Gaming Project, is an AI strategist and governance architect focused on building ethical systems for underserved markets. With a Master's in Counseling and decades in community education, he has delivered over 10,000 trainings in neurodiversity, education, and innovation. Based in Appalachia, his work has been recognized and adopted by the American Bar Association, the ACLU of West Virginia, Americorps VISTA Leaders, and the WV Community Development Hub.

For more information about the Coalition for Innovation,
including how you can get involved, please visit coalitionforinnovation.com.

View the Next Chapter