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

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

Chapter 10: Making AI Safe: An Organizational Perspective

Author: Ann M. Marcus

What Does "AI Safety" Mean?

When we talk about “Safe AI”, what do we mean? Suppose the AI application your organization has developed and deployed suddenly:

- Provided erroneous or dangerous advice in a situation.
- **Delivered only certain content due to restrictions by a particular organization, possibly for its benefit.**
- Became unreliable due to power or communications failures.

The National Institute of Standards and Technology (NIST) identifies seven characteristics of trustworthy or safe AI:

1. **Valid & Reliable**: Performs as intended even under unexpected conditions.
2. **Safe**: Minimizes physical, emotional, economic, and environmental harm.
3. **Secure & Resilient**: Withstands attacks, accidents, or misuse.
4. **Explainable & Interpretable**: Operates intuitively so that users and stakeholders can understand how it works.
5. **Privacy-Enhanced**: Respects and protects personally identifiable information (PII).
6. **Fair** (Bias Managed): Avoids discriminatory or unjust outcomes.
7. **Accountable & Transparent**: Follows a clear chain of responsibility.

Real-World Harms: Why This Matters

Adverse AI outcomes can take many forms and impact people, organizations, and processes.

Without managing your organization's AI processes, the company's productivity and reputation could suffer significantly.

Below we've drawn from a number of knowledgeable sources to identify some key areas of AI vulnerability.

What To Watch For	Why it Matters & Recent Examples	Primary Safeguards & Where to Find Them
Jailbreak & prompt-injection loopholes	A May 2025 Ben-Gurion University team demonstrated a single "universal" jailbreak that bypassed guardrails in five leading chatbots, letting them give step-by-step hacking, bomb-making, and hate-speech instructions.	Layered input & output filters (regex, semantic classifiers) "Chain-of-thought" suppression or sandbox-inference for sensitive queries Continuous red-teaming with external researchers (now mandatory in EO 14110 & Seoul "Frontier AI" pledge) ResearchGateGOV.UK
Deepfakes & influence operations	In Jan 2024, New Hampshire voters received AI-generated robocalls mimicking President Biden urging them not to vote — an incident that led to FCC fines and criminal charges – showing how cheaply and readily disinformation can scale.	Cryptographic provenance & watermarking (C2PA / Content Credentials) Platform-side authenticity labelling; FCC & EU rules on AI robocalls and deepfake ads Public-sector media checksums for all official releases The VergeC2PA
Bias / discrimination in high-risk sectors	The EU AI Act (final text 2024) classifies employment, credit, health care and policing tools as "high-risk," obliging providers to run bias tests, log incidents and keep a human-oversight chain because statistically significant	Pre-deployment disparity testing + yearly audits (EU AI Act "high-risk" stack) European Parliament

What To Watch For	Why it Matters & Recent Examples	Primary Safeguards & Where to Find Them
	disparities are still appearing in production models.	Diverse test suites (OOD, intersectional), veto thresholds in procurement SLAs ISO / IEC 42001 clause 8.2: risk-impact assessment & human-oversight controls ISO
Adversarial & data-poisoning attacks	A Nature Medicine paper showed that "poisoning" only 0.01% of a popular medical dataset could make a healthcare LLM consistently output dangerous misinformation showing how fragile training pipelines remain.	Data lineage + signed ML-BOMs; immutable storage for "gold" datasets Automated anomaly filters & loss-spike monitors during training and inference OWASP LLM04 hardening guide for open-source models genai.owasp.org
Interpretability & "black-box" failure modes	Reportedly "mechanistic interpretability" techniques in frontier labs scale slower than the models, leaving developers blind to rare but potentially catastrophic behaviors before deployment.	Mechanistic-interpretability dashboards (circuits, attribution maps) "Test-time tool" isolation: no tool-calling without explicit policy approval Responsible Scaling Policy (Anthropic) ties model size to proof-of-understanding levels. Anthropic
Privacy leakage & data governance	U.S. Executive Order 14110 (Oct 2023) requires red-team reports for privacy leaks after researchers showed membership- inference attacks can recover personal data or copyrighted text from LLMs.	Differential-privacy fine-tuning or synthetic-data augmentation Red-team drills required by U.S. Executive Order 14110 &

What To Watch For	Why it Matters & Recent Examples	Primary Safeguards & Where to Find Them
		OMB M-24-10 for federal use The White House Deletion and trace request pipeline + encrypted telemetry logs
Misalignment & runaway autonomy	At the AI Seoul Summit (May 2024) 16 governments and 8 frontier labs agreed to joint red-team "stress tests," kill-switch R&D, and recall protocols for any model that shows unsafe emergent behavior: an implicit acknowledgment that the risk is real.	"Kill-switch" remote-weight revocation (part of Seoul commitments) GOV.UK Stage-gated capability release based on safety levels (ASL-2→ASL-4) Closed-scope sandboxes for agentic features (tool use, code execution)
Concentration of power & weak governance	The voluntary Frontier AI Safety Commitments Act (Seoul Summit, 2024) pertains to only a handful of dominant cloud and model providers. Critics note that regulators still lack audit or recall authority, leaving systemic risk in private hands.	Adopt an AI-Management System (ISO 42001) – board-level oversight, KPIs, audit rights ISO Publish model cards + incident reports (NIST AI RMF "Govern → Manage" functions) NIST Publications External whistle-blower and bug-bounty channels (Seoul commitment §III-3) GOV.UK

Safeguards and Standards to Know

Safeguards against these safety threats and the standards or policies that back them up are shown in the list below. One rarely needs to have *all* the controls in place, but every high-stakes AI deployment should be covered by at least one specific measure.

One of the sources cited for mitigating AI risk is the National Institute for Standards & Technology (NIST) for its work in ensuring trustworthy and responsible AI. A July 2024 NIST report, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," notes that, it "develops measurements, technology, tools, and standards to advance reliable, safe, transparent, explainable, privacy-enhanced, and fair artificial intelligence (AI) so that its *full commercial and societal benefits* can be realized without harm to people or the planet."

NIST, which has conducted fundamental and applied work on AI for more than a decade, also helps to fulfill the 2023 Executive Order on Safe, Secure, and Trustworthy AI. The agency, which resides under the U.S. Department of Commerce, established the U.S. AI Safety Institute and the companion AI Safety Institute Consortium to continue the efforts set in motion by the Executive Order to "build the science necessary for safe, secure, and trustworthy development and use of AI."

What Organizations Should Do Now (Checklist)

Here are several suggested process and systems guidelines for safeguarding your organization of it relies heavily on AI:

1. **Stand up governance first:** Create or plug into an AI-risk committee with legal, security, product and domain expertise. Use NIST AI RMF's "G-M-M-M" (Govern-Map-Measure- Manage) loop as your operating rhythm.

In addition, there are some quick steps that you can start right away for protection:

1. **Ship content-authenticity headers** on every AI-generated image or video you publish.
2. **Sign up for a multi-party red-team exercise** (see the NIST AI Safety Institute or an industry hackfest).
3. **Implement differential-privacy fine tuning** for any model that ingests user data.
4. **Draft an ISO 42001 "gap list."** Most orgs find 70% of requirements map to existing ISO 27001 or SOC-2 controls, so remediation is often modest.

2. **Map risk to use-case:** Inventory every current and planned AI component, tag it against the threats shown above and decide which standards apply (EU AI Act, ISO 42001, sector regs, etc.).
3. **Select layered controls:** For each threat, pick at least one *technical* control (filters, DP training, provenance tags) and one *process* control (red-team cadence, human-oversight checklist, audit log retention).
4. **Test before & after launch:** Run adversarial evaluations (jailbreak attempts, bias stress-tests, poisoning probes) before release and after every major model update. Seoul Summit signatories now publish test methodologies; use them.
5. **Monitor & log continuously**: Hook real-time anomaly detectors to model inputs and outputs and training metrics; store logs immutably for forensics and regulatory reporting.
6. **Prepare an incident-response & recall playbook**: Include a rapid rollback path (shadow-model, feature flag, or full weight revocation), external disclosure

templates, and a consumer-facing support plan.

7. **Audit & improve:** At least annually, benchmark controls against new research (e.g., updated OWASP LLM Top 10, NIST profiles) and tighten thresholds where attacks have succeeded.

Conclusion: Building & Deploying Trustworthy AI

To make and use AI responsibly in your organization, it would be wise to address the issues that we have highlighted in this chapter: AI safety, highlighting potential harm, defining key characteristics of trustworthy AI, and detailing specific threats and their safeguards. We've examined various adverse outcomes, from erroneous advice and biased systems to deep fakes and privacy breaches, alongside recent real-world examples. We have made the case for establishing robust governance, mapping risks to use cases, and implementing layered technical and process controls.

Continuous testing, monitoring, and a well-defined incident response plan are crucial for mitigating risks to your productivity and reputation. By adopting these proactive measures and leveraging resources such as the NIST AI Risk Management Framework and ISO 42001, organizations can confidently navigate the complexities of AI development and deployment, ensuring its full commercial and societal benefits are realized responsibly and without harm.

Author (In order of contribution)

Ann M. Marcus, Director, Ethical Tech & Communications, WeAccel
Ann M. Marcus is a Sonoma-raised, Portland-based communications strategist and ethical technology analyst focused on smart cities, community resilience, and public-interest innovation. She leads the Marcus Consulting Group and serves as director of ethical technology and communications at WeAccel.io, a public-good venture advancing mobility, communications, and energy solutions for communities. Ann has advised public and private organizations—including Cisco, the City of San Leandro, Nikon, AT&T, and InfoWorld—on trust-based data exchange, digital public infrastructure, resilience strategy, AI and more. Her current projects include a California senior evacuation program, a Portland robotics hub, and digital energy resource initiatives with utilities in Portland and the Bay Area.

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

View the Next Chapter