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

Table of Contents

Chapter 1: Introduction

Author: Sarah Ennis

Humanity-Forward AI: A Blueprint for Responsible Innovation

Artificial intelligence (AI) is moving quickly into nearly every sector of society. While the potential is extraordinary, the challenges are equally significant: questions of trust, fairness, environmental impact, and human agency remain unresolved. This blueprint responds to those challenges with a structured, multi-part approach that links technical foundations to real-world application and community benefit.

Scope of this Blueprint for the Future

This blueprint has been developed to guide the responsible evolution of AI from a cross-sector perspective. It is intended for technologists, policymakers, educators, industry leaders, and community advocates who share a commitment to ensuring that AI advances the public good.

The content is organized into four interconnected parts:

Data, Policy & Adoption: building the governance, legal frameworks, and infrastructure needed for trustworthy AI systems

Human-Centered Design & Next-Generation Workflows: ensuring AI enhances human agency and capability

Ethics, Safety & Societal Impact: addressing systemic risks, equity concerns, and long-term accountability

Sector Spotlights: applying lessons to specific fields such as education and the creative industries

Rather than prescribing a single standard or regulatory approach, this blueprint offers a flexible, evidence-informed framework that can be adapted to local contexts and evolving technologies. It is a living reference, meant to evolve alongside AI itself.

Blueprint Overview

Part I: Data, Policy & Adoption addresses the infrastructure and governance needed to support AI systems responsibly. Data Alliance examines how federated ecosystems can enable secure, consent-based data sharing across sectors. Regulation on AI focuses on intellectual property rights, licensing frameworks, and legal guardrails for model training and deployment. Adoption of AI considers enterprise uptake, user consent, and the often-overlooked costs of running large-scale data centers. Benefits & Drawbacks of Decentralized AI weighs the trade-offs between edge and cloud architectures, with a focus on trust, bias, and resilience.

Part II: Human-Centered Design & Next-Gen Workflows looks at how AI can be built to enhance human capability rather than replace it. Human Factors Contributions in GenAI explores design practices that promote usability, transparency, and user trust. What AI Owes Children sets out privacy-by-design principles and ethical guardrails to protect young users. Agentic AI examines the evolution from narrow, task-based assistants to more autonomous digital agents, including the opportunities and risks they present.

Part III: Ethics, Safety & Societal Impact turns to the values and safeguards that should guide AI development. AI Ethics: Navigating Responsible Innovation provides international frameworks, real-world failure cases, and practical governance measures to support fairness, transparency, and dignity. AI and the Community Lens considers how AI intersects with equity, poverty, and place, particularly in underrepresented communities. AI Safety outlines methods for adversarial testing, governance controls, and continuous monitoring. Overreliance on AI explores the behavioral and design factors that can lead to unhealthy dependency, alongside corrective safeguards. Climate & Community Impacts of AI highlights infrastructure emissions, equity in deployment, and civic risks. Beyond Emissions extends this discussion to balancing environmental, social, and economic responsibilities at scale.

Part IV: Sector Spotlights focuses on domains where AI’s influence is both transformative and contested. AI in Education — A Tipping Point looks at the dual potential for surveillance and transformative learning, with attention to equity in classrooms. AI: Entertainment, Creativity & Piracy investigates the tension between enabling creative innovation and protecting intellectual property in the digital economy.

The Appendices provide practical tools and illustrative examples. Appendix A presents the AI Anchor System, a tested framework for aligning AI behavior with ethics, bias prevention, identity safeguards, truth verification, and equity. Appendix B includes case studies in international collaboration in AI, such as the Seoul–MILA Scientist-in-Residence Program, joint research funding models, public-sector innovation in Daejeon, and knowledge-sharing frameworks. Appendix C offers a Solution-Focused Community Development model that applies AI to grassroots problem-solving.

Taken together, these sections form a blueprint for AI that is ethical, equitable, and resilient. It offers not just analysis but actionable frameworks, encouraging technologists, policymakers, educators, and community leaders to design and deploy AI systems that advance the public good.

Author (In order of contribution)

Sarah Ennis, Co-Founder, AgentsGEO.ai

Sarah Ennis is a Fortune 500 trusted advisor specializing in advanced technology innovation, with over two decades of experience leading groundbreaking AI solutions at scale. Globally recognized for her expertise in artificial intelligence, she designs and implements bespoke emerging technology products across industries. She is also the co-founder of AgentsGEO.ai, a patent-pending platform that helps brands monitor and improve their visibility in the AI ecosystem and deploy AI agents, ensuring they are discoverable and recommended by tools like ChatGPT, Gemini, and others through its proprietary GEOScorer™ technology. In addition, Sarah contributes part-time to Northeastern University's Master of Digital Media programs in AI, preparing the next generation of technologists and creative leaders. Her work bridges Silicon Valley innovation with global impact, and she is a distinguished member of the American Society for AI and contributor to the OpenAI Forum.

Part I
Data, Policy, & Adoption

Chapter 2:
Beyond Emissions: Balancing People, Planet, and Profit at Scale in AI Infrastructure

Authors: John Barton

Facts and Figures:

U.S. data centers consumed 176 TWh of electricity in 2023, contributing ~60 $MtCO_2e$. AI workloads (e.g., GPT-3) are primary drivers of this growth, with one training run using 1,287 MWh.

Water use is significant: direct cooling used ~66 billion liters in 2023, and indirect power generation consumed another ~800 billion liters.

Google, Microsoft, and Meta collectively withdrew ~2.2 billion m^3 in 2022, comparable to the annual use of two Denmarks.

Community-level impacts include generator emissions (~100 tons NOx/year in Wisconsin; ~14 tons formaldehyde/year in Memphis), noise pollution, and increased infrastructure costs.

Environmental justice concerns are acute, with facilities often sited in underserved or vulnerable regions with minimal local benefit and high health/environmental burdens.

Public opposition has delayed or blocked ~$64 billion in data center projects across 24 U.S. states as of 2025.

Overview

Artificial intelligence (AI) infrastructure demands enormous physical resources — energy, water, land — and produces wide-ranging ecological and civic consequences. While emissions are often the primary metric of concern, the full picture includes upstream and downstream effects on water systems, air quality, public infrastructure, and community well-being. These impacts are not only accelerating but disproportionately concentrated in regions with limited oversight or leverage, such as Appalachia, the Southwest, and other under-resourced areas.

Local communities face additional external influences including thermal pollution, diesel exhaust from backup generators, and grid strain, particularly in water-stressed and low-regulation regions. These externalized costs, compounded by tax exemptions and minimal job creation, highlight the urgent need to rethink sustainability beyond emissions-only metrics.

List of Stakeholders (Audience/Readers)

Public Sector & Governance

This group includes entities responsible for policy, regulation, and public resource management at all levels of government.

Local & Regional Authorities:

- Municipal and county governments (city councils, zoning boards, public works)
- Water authorities and regional water boards
- School boards and local educational institutions
- Economic development agencies

State & Federal Regulators:

- Environmental protection agencies (e.g., EPA, state-level environmental quality boards)
- Public utility commissions and energy departments (DOE)
- State oversight offices (auditors general)
- Federal agencies (e.g., USDA, NTIA)

Cross-Jurisdictional Bodies:

- Regional funding commissions (Appalachian Regional Commission)
- Tribal nations and Indigenous land authorities

Private Sector & Infrastructure

This category covers the corporations and financial entities that design, build, and operate the infrastructure, along with their investors.

Technology & Infrastructure Providers:

- AI companies and cloud service providers (e.g., Google, Microsoft, AWS)
- Hyperscale data center developers
- Utility companies and grid operators
- Construction, logistics, and engineering firms

Investors & Financial Services:

- Real estate investment trusts (REITs) and infrastructure asset managers
- Private equity firms
- Insurance providers and ESG risk analysts

Civil Society & Community

This section includes groups and individuals directly affected by AI infrastructure, along with non-governmental organizations advocating on their behalf.

Affected Communities:

- Local residents and neighborhood associations
- Utility ratepayers
- Communities in tax-exempt or PILOT (Payments in Lieu of Taxes) zones

Advocacy & Public Interest Groups:

- Environmental justice coalitions and grassroots organizers
- Labor unions and tech equity coalitions
- Public health departments and local planning boards
- National civil rights and legal aid organizations

Global & Research Entities

This final group includes international bodies, academic institutions, and media that shape the global context and public understanding of AI infrastructure's impacts.

Global Governance & Oversight:

- Multilateral climate and infrastructure funders (e.g., World Bank, IMF)
- International sustainability standards bodies (ISO)
- Global watchdog organizations (e.g., Amnesty International, Global Witness)
- Supply chain and critical minerals governance coalitions

Knowledge & Media:

- Academic researchers
- Investigative journalists and specialized media
- Think tanks and public policy labs
- Independent ESG auditors

- AI industry governance bodies (e.g., Partnership on AI)

The Problem:

AI infrastructure is no longer a niche domain; it is central to how knowledge is produced, how decisions are made, how surveillance systems operate, and how global computation scales. The physical systems powering it — supporting models like GPT, national defense, and enterprise AI — are intensely resource-dependent, placing accelerating demands on electricity, water, land, and labor. These burdens fall disproportionately on communities with the least power to resist them.

These burdens are often hidden—by design. Not just physically, but through decision-making structures that obscure who decides, who pays, and who is accountable. Costs are externalized. Public engagement is bypassed. Communities are left with the consequences. With the rise of generative AI and continuous inference workloads, these demands are compounding exponentially, straining people, ecosystems, and economies.

Across the country, siting decisions frequently exploit disenfranchised regions—Appalachia, the Southwest, and other areas with cheap land, weak regulation, and under-resourced governments. Projects are often approved before public notice, and communities may only learn of them after rezoning or construction is already underway. Civic exclusion and externalized costs fall hardest on marginalized groups with the least leverage. In West Virginia, grid upgrades for proposed data centers could cost ratepayers over $440 million, underscoring how local communities may be forced to subsidize infrastructure for global platforms.

Narrow reporting metrics compound these harms. Environmental assessments often focus only on emissions, omitting water, land, and heat impacts. Mid-sized AI data centers can draw up to 300,000 gallons of water per day—comparable to the daily use of 1,000 households—yet such withdrawals rarely appear in sustainability reports. This selective accounting creates blind spots that mask the full scope of ecological damage.

In 2023, U.S. data centers used an estimated 66 billion liters of water for cooling and another 800 billion liters indirectly through power generation. Phoenix facilities collectively draw more than 177 million gallons per day, while in The Dalles, Oregon, Google's campus now consumes nearly 25% of the city's water supply. Aquifers and watersheds are stressed, wastewater discharges raise ecological risks, and noise and air pollution add chronic health burdens.

- Microsoft's Wisconsin site is projected to emit nearly 100 tons of nitrogen oxides annually.
- xAI turbines in Memphis emit nearly 10 tons of formaldehyde into a community already facing quadruple the national cancer risk.

These facilities are structured around subsidy and speculation. Governments provide hundreds of millions in public incentives while corporations minimize tax obligations.

- In Oldham County, Kentucky, a $6B project attempted to classify as a private utility to bypass zoning laws, abandoning the effort only after community pushback.
- Nationwide, over $64 billion in data center projects have been blocked or delayed due to public resistance in 24 states.

Despite promises of growth, the permanent jobs created are few — often fewer than 100 positions for billion-dollar facilities — while the infrastructure burdens of water withdrawals, grid stress, and road wear are borne locally. Universities and localities justify these projects on speculative ROI and prestige, even as they hollow out public budgets.

Greenwashed environmental, social, and governance (ESG) claims often deflect attention from these ongoing harms. Facilities sited on carbon-intensive grids may still claim carbon neutrality via offsets or purchase agreements, while omitting lifecycle emissions from chip manufacturing, mining, and global shipping. This selective framing disguises the true scale of extraction.

At scale, these pressures are accelerating. In 2023, U.S. data centers consumed 176 terawatt-hours of electricity (about 4.4% of national usage) and withdrew over 66 billion liters of water for direct cooling. By 2030, AI demand could require as much as 298 gigawatts—roughly a quarter of national electrical usage—and nearly 400 billion liters of water annually.

These burdens are not distributed evenly. Infrastructure is concentrated in regions with fragmented civic resistance and limited oversight, ensuring global users and cloud providers remain shielded from the physical, civic, and ecological costs. Communities are excluded from meaningful participation, often left to protest as their only form of engagement.

The result is a systemic asymmetry: benefits flow outward to platforms, investors, and end users, while under-resourced communities absorb degraded infrastructure, displaced public services, environmental harm, and long-term liabilities. These regions are not accidental victims but strategic targets, selected precisely because their land, water, political capacity, and people are treated as expendable.

The system is designed to scale computation, not community resilience. To correct this imbalance, AI infrastructure must be restructured around equity, accountability, and long-term viability. Sustainability, not exploitation, is the way forward.

Our New Vision: People, Planet, Profit Framework

AI infrastructure is already expanding at an unprecedented pace with new facilities reshaping local economies and ecosystems across the country. Yet the costs of this expansion—environmental, social, and economic—are too often shifted disproportionately onto vulnerable communities. Current siting and permitting practices externalize risks and conceal true costs, leaving local populations to bear the burdens of pollution, resource strain, and inequitable economic trade-offs.

To counter these systemic failures, we propose the People, Planet, Profit framework, built on lifecycle accountability and civic equity. This is not aspirational—it sets the minimum operational standard for sustainability. The framework restructures AI infrastructure around resilience, legitimacy, and long-term viability. Each pillar is framed by a clear **Goal**, followed by actionable measures that embed sustainability into decision-making.

The framework calls for planning that embeds sustainability into the operational design of AI infrastructure. Rather than treating environmental harm as a compensable side effect, the priority must be to proactively prevent harm, internalize resource costs, and align infrastructure planning with durable systems that protect communities and ecosystems. Sustainability must be treated as a binding requirement—an operational baseline that guides every siting, permitting, and investment decision.

People

Goal: Integrate human-centered metrics into infrastructure planning—job quality, health exposure, and civic cost distribution—so that communities gain tangible benefits from hosting AI infrastructure.

- Establish binding community benefits agreements and tax equity frameworks.
- Ensure job quality, worker protections, public health safeguards, procedural inclusion, and localized economic return in planning decisions.
- Mitigate pollution burdens such as diesel generator emissions, HVAC-related noise, and thermal output that disproportionately affect working-class and marginalized communities.
- Embed public trust as a design constraint, not a PR strategy.

Planet

Goal: Quantify and reduce environmental loads at every lifecycle stage: energy use, water draw, pollution, and waste. Prioritize local ecological integrity, not just emissions offsets.

- Replace carbon neutrality claims with real environmental accounting across the full lifecycle, including upstream emissions (chips, transport) and local degradation (cooling discharge, groundwater stress).
- Reject offset schemes that disguise fossil dependency.
- Optimize water-use effectiveness, enforce thermal discharge limits, and select sites that protect ecosystems.
- Conduct grid impact studies and disclose resource demands before approval.

Profit

Goal: Treat resilience, transparency, and long-term viability as cost drivers, not externalities. Align siting, financing, and risk management with lifecycle realities and civic accountability.

- Measure profitability through durability, transparency, and infrastructure resilience.
- Integrate legal exposure, water volatility, public resistance, and decommissioning costs into ROI models.
- Disclose public funding, tax exemptions, and civic cost burdens.
- Account for hidden subsidies and externalized harms as financial liabilities, reinforcing sustainability as a binding operational requirement.

Projections indicate the U.S. could see over 10,000 AI-optimized data centers by 2030. This buildout is not just a question of scale—it generates compounding ecological, economic, and political risks when combined with today's extractive siting patterns, rising water demands, diesel emissions, and the shifting of costs onto local communities.

If left unchecked, these practices will deepen long-term vulnerabilities for both infrastructure providers and the communities that host them. Policymakers, civic planners, and infrastructure investors must therefore move beyond short-term throughput and prioritize long-term resilience. That requires embedding lifecycle costs, water system capacity, and public trust into every siting and design decision, and treating sustainability not as an optional add-on but as the minimum operational standard.

People, Planet, and Profit are not abstract concepts or ideals; they are the practical foundation of financially responsible and sustainable AI infrastructure development. This triadic framework anchors long-term viability in human, environmental, and financial outcomes—the benchmark of whether AI infrastructure will truly endure.

Case Studies by Sustainability Domain

While the risks of unchecked development have been widely documented, examples of directional progress remain fragmented, underreported, or excluded from industry strategy documents and permitting frameworks. This document curates emerging models, partial successes, and boundary-testing prototypes that illustrate how the principles of People, Planet, and Profit can work together in practice.

Each case study was selected based on evidentiary grounding, relevance to infrastructure decision-makers, and potential for policy translation. All were chosen for their ability to operationalize at least one facet of the Vision: civic equity, ecological alignment, or lifecycle financial accountability. These are not hypothetical designs, but live experiments—some state-driven, some corporate-led, and some Indigenous or community-initiated.

Each marks a shift away from extractive norms and toward infrastructure that internalizes long-term impacts, invites public trust, and models system-wide accountability. They are not blueprints. They are prototypes of possibility—signals that transformation is already underway. Initiatives such as community air monitoring or localized heat reuse often fly under the radar, yet they are among the most politically feasible and economically efficient levers for reform. These accessible interventions deliver outsized impact when codified and repeated. These small civic or environmental shifts can recalibrate entire projects.

When design constraints are treated as ethical guardrails rather than barriers, sustainable infrastructure becomes not just feasible but the only model that can scale without system failure. Many involve tradeoffs, yet all are operationally relevant. These case studies are valuable not because they offer complete solutions, but because they show meaningful deviation from the status quo. Each example reveals how infrastructure can evolve toward sustainability when civic priorities, ecological limits, and long-term investment logic are treated as design constraints, not afterthoughts.

When viewed collectively, these case studies form a strategic knowledge base that deserves active preservation and policy translation. No single example solves for all three dimensions of sustainability. However, even narrow wins such as improved permitting or integrated water management create precedents that shift institutional expectations. Directional progress builds the scaffolding for future norms.

PEOPLE: Civic Equity, Public Health, and Procedural Inclusion

Infrastructure decisions that begin with community needs tend to yield more durable outcomes. Procedural inclusion — through public comment, health screening, or Indigenous governance — helps prevent backlash, streamline implementation, and protect legitimacy. These cases show how civic participation is not a courtesy, but a structural advantage in high-impact infrastructure. From FOIA-driven oversight in Tucker County to CBA-backed benefits in New York's South Fork Wind, procedural inclusion is emerging as a risk-mitigation strategy.

See also: Brookings — Civic Participation and Infrastructure • NEPA — Public Participation Guide

Public Comment and Permitting Participation

Public comment processes give communities direct influence over infrastructure decisions. When paired with legal enforcement mechanisms, they can materially reshape projects and embed accountability. These cases demonstrate how structured civic engagement, combined with regulatory action, can significantly alter infrastructure design and implementation.

Prince William County, VA – Digital Gateway Project: AP News — Virginia county approves data center project after 27-hour hearing *See also:* InsideNova — Digital Gateway debate

In this case, sustained, organized public engagement materially shaped high-impact development. The Prince William County Board of Supervisors held a **27-hour public hearing** before approving the Digital Gateway project. Hundreds of residents raised concerns about visual blight, environmental degradation, and cultural site encroachment, forcing developers to negotiate concessions.

Key Highlights:

- 27-hour public hearing with hundreds of participants
- Concerns raised: visual blight, environmental harm, cultural encroachment
- Concessions: 800+ acres preserved, 1,500-foot buffers, historic site protection, trails and parks
- Legally binding zoning conditions enforced

Context: A proposed data center campus faced unprecedented community opposition tied to environmental and cultural concerns.

Outcome: Developers were required to integrate community demands through binding zoning conditions.

Impact: Public comment materially reshaped the project's footprint, demonstrating that community engagement can redirect scale and secure enforceable benefits.

Becker, MN – Amazon Data Center Generators: Data Center Frontier — Minnesota PUC says no to Amazon's bid to fast-track 250 diesel generators *See also:* Star Tribune — Minnesota PUC rejects Amazon diesel plan

In 2024–25, Amazon attempted to fast-track the installation of 250 backup diesel generators at a proposed Minnesota data center by requesting exemption from the state's certificate-of-need process. Community members, environmental advocates, and the Minnesota Attorney General's office challenged the request, citing serious air quality concerns and the precedent it would set for future projects. The case highlighted how state-level review processes can serve as crucial checks against speculative or environmentally risky development.

Key Highlights:

- Amazon sought exemption for 250 diesel generators
- Opposition from Minnesota AG, environmental groups, and local community
- Risks: air quality impacts and precedent for bypassing review
- Regulatory outcome: PUC unanimously denied exemption

Context: Amazon sought to exempt 250 diesel generators from certificate-of-need review in Minnesota. **Outcome:** State regulators, supported by civic and institutional opposition, unanimously rejected Amazon's exemption request.

Impact: Amazon's plans were delayed and subjected to full emissions review, proving the effectiveness of procedural safeguards as a financial and environmental check.

Community Benefit Agreements (CBAs)

Community Benefit Agreements provide legally binding structures for channeling development gains back into local communities. They ensure benefits such as jobs, training, and reinvestment are guaranteed rather than promised. Unlike Community Benefit Plans (CBPs), CBAs are enforceable contracts that bind developers to commitments, making them a tool of both accountability and equity.

Sunrise Wind (Long Island, NY): Sunrise Wind — Local Benefits Agreements to Advance Sunrise Wind Project *See also:* NYSERDA — Sunrise Wind project details

The Sunrise Wind project is a landmark example of a high-value CBA, signed in 2023 with a total package worth **$169.9 million**. The agreement earmarks funds for workforce development, health services, and infrastructure upgrades, linking renewable energy expansion to tangible community benefits. Its scale demonstrates the potential of CBAs to transform local economies while building trust.

Key Highlights:

- Total value: $169.9 million
- $1M for workforce training, $2M for public health
- Infrastructure upgrades and local hiring pipelines
- Legally binding contract with local and regional authorities

Context: One of the largest negotiated CBAs in U.S. clean energy. **Outcome:** Secured unprecedented levels of community reinvestment, including jobs, training, and public health funding.

Impact: Demonstrated the potential of CBAs to scale public benefit in high-value infrastructure projects.

Columbia Law CBA Database – Solar Energy Projects: Columbia Climate School — Community Benefits Agreements Database *See also:* Energy News Network — CBA examples in renewable projects The Columbia Climate School's CBA database catalogs dozens of community benefit contracts across the renewable energy sector. Examples from Ripley, Byron, and Maui County provide clear models of recurring financial investment in local communities, including structured annual payments, infrastructure improvements, and reinvestment funds.

Key Highlights:

- Ripley Solar: 270 MW, $472,500 annual payments with escalators
- Byron Solar: 280 MW, ~$24M total lifecycle payments
- Maui County Solar: 20 MW, $55,000/year for 25 years
- Common provisions: road upgrades, emergency services, community impact funds

Context: Solar projects across multiple states provide tested CBA models.

Outcome: Delivered recurring financial and infrastructure investments to host communities.

Impact: Established replicable models for binding community benefits, now supported by permitting norms and legal precedents.

ReImagine Appalachia / Clean Air Task Force: ReImagine Appalachia — Community Benefits • Clean Air Task Force — Community Benefits Resource Inventory *See also:* Just Transition Fund — Community benefits resources These organizations develop frameworks for equity-centered development, creating toolkits that include wage provisions, local hiring standards, and reinvestment strategies. Their work shows how advocacy groups can equip communities with negotiation tools that rival corporate legal resources, leveling the playing field in infrastructure decision-making.

Key Highlights:

- Living wage provisions
- Local hire benchmarks
- Profit reinvestment into transition or resilience
- Policy and permitting toolkits for rural and post-industrial regions

Context: Advocacy-driven frameworks designed for post-industrial and rural regions.

Outcome: Produced customizable tools and language for embedding equity into project negotiations.

Impact: Enhanced coalition capacity to secure fair wages, jobs, and reinvestment in communities vulnerable to energy transition shocks.

Health Screening Tools & Procedural Equity Frameworks

Health screening tools and procedural equity frameworks expand the definition of feasibility to include cumulative health and environmental burdens. By integrating these tools into planning, infrastructure siting decisions can avoid reinforcing inequities and direct resources to resilience in overburdened communities.

CalEnviroScreen (California): OEHHA — CalEnviroScreen *See also:* EPA EJScreen — Federal screening tool CalEnviroScreen is a state-developed tool that ranks communities based on cumulative environmental risk and vulnerability, guiding permitting, policy targeting, and funding allocation. Its use demonstrates how structured screening mechanisms can shift state-level resource distribution toward equity.

Key Highlights:

- **Function**: Ranks communities by cumulative environmental risk and vulnerability
- **Use Case**: Guides permitting, policy targeting, and resource allocation
- **Potential**: Could influence AI/data infrastructure siting decisions

Context: Built to address longstanding environmental justice concerns in California.

Outcome: Enabled targeted state resource allocation to vulnerable communities.

Impact: Provides a replicable model for guiding infrastructure siting and reducing disproportionate burdens.

Civic-Led Planning & Governance Innovations

Civic-led innovations show how communities use transparency, organization, and advocacy to influence — or slow — data infrastructure projects that threaten health or environmental equity. These examples reveal the growing power of grassroots coalitions to leverage procedural levers against powerful corporate actors.

Tucker County, WV – Community Resistance to Data Center: WV DEP — Response to Public Comment (PDF) • Tucker United — Community Coalition *See also:* WV Public Broadcasting — Tucker County resistance coverage Residents of rural Tucker County mobilized under the coalition *"Tucker United"* to contest a Ridgeline data center powered by methane gas. The coalition combined traditional advocacy tactics — town halls, FOIA requests — with technical measures such as independent air quality monitoring. Although the project has not been formally halted, civic action slowed its momentum significantly.

Key Highlights:

- Formation of *Tucker United* coalition
- FOIA requests and independent monitoring
- Organized town halls and community education
- Slowed project momentum despite lacking veto authority

Context: Grassroots coalition mobilized against gas-powered data center development.

Outcome: Raised awareness, generated scrutiny, and slowed project momentum.

Impact: Showed how civic pressure can disrupt or delay projects even without formal veto power.

Memphis, TN – xAI Turbine Controversy: AP News — NAACP, environmental group notify xAI of intent to sue over pollution *See also:* Commercial Appeal — xAI turbine fight In South Memphis, a predominantly Black community already facing high environmental risk, residents and EJ advocates opposed two methane turbines proposed to power Elon Musk's xAI data center. Local organizers combined grassroots mobilization with scientific studies showing elevated health risks, including asthma and cancer. Their advocacy delayed air permit approvals and drew national attention to the environmental justice dimensions of the project.

Key Highlights:

- Two methane turbines proposed for xAI facility
- Community concerns: asthma, cancer, and air quality
- Mobilization by NAACP and environmental justice groups
- Air permits delayed due to community and scientific pushback

Context: Proposed turbines in an environmentally overburdened Black community.

Outcome: Public backlash, supported by health data, forced the state to delay air permits.

Impact: Highlighted the power of frontline communities to assert environmental justice and health equity in siting decisions.

PLANET: Environmental and Ecological Safeguards

Environmental performance is no longer a secondary concern; it is an operational necessity. Data centers and digital infrastructure that **reuse heat**, **minimize water draw**, or **integrate into district energy loops** are proving more scalable and less volatile. Ecological foresight strengthens both system resilience and public alignment. Projects that pair heat reuse with municipal coordination — such as in Stockholm and Mäntsälä — demonstrate that environmental alignment can also reduce grid volatility.

Water Usage

Water is an increasingly contested resource for communities near large data centers. Monitoring and transparency on **Water Usage Effectiveness (WUE)** remain limited across U.S. facilities, highlighting the need for lifecycle water audits. These examples show how water demand from data centers can place stress on local resources and ecosystems, making transparent reporting essential.

Amazon – Hermiston, OR: Oregon Live — Amazon data center water use in Hermiston *See also:* Columbia Insight — Amazon's Hermiston water use scrutiny Amazon's Hermiston facility reported using **66.8 million gallons of water in 2023**. This scale of consumption raised concerns over long-term local water availability and the absence of transparent lifecycle accounting.

Key Highlights:

- **Usage**: 66.8 million gallons in 2023
- **Concern**: High draw on local supply without full transparency
- **Risk**: Potential strain on municipal and agricultural resources

Context: Amazon's case underscores how data center water withdrawals can directly affect regional water security in smaller communities with limited reserves.

Outcome: Sparked public debate and highlighted the need for mandatory disclosure of lifecycle water use.

Impact: Pressured operators to provide greater transparency and plan for long-term water resilience.

Loudoun County, VA: Loudoun Times-Mirror — Data centers used 1.85 billion gallons of water in 2023 *See also:* Data Center Frontier — Loudoun's data center water usage Loudoun County, the largest concentration of data centers in the U.S., consumed **over 1.85 billion gallons of water in 2023**. The concentration of withdrawals creates compounding pressure on regional water infrastructure.

Key Highlights:

- **Usage**: Over 1.85 billion gallons in 2023
- **Concern**: Large-scale, concentrated withdrawals intensify resource stress
- **Risk**: Regional ecosystem and community water needs placed in competition with data center operations

Context: Loudoun's water use illustrates how cumulative withdrawals across clustered facilities can amplify ecological and civic impacts at a metropolitan scale.

Outcome: Triggered state-level scrutiny and calls for lifecycle water audits.

Impact: Reinforced water as a critical constraint on data center expansion in high-density hubs.

WUE Benchmarks: AKCP — WUE Guide *See also:* Nature — Masanet et al. (2021) on data center sustainability Industry benchmarks such as Water Usage Effectiveness (WUE) provide a comparative metric for measuring efficiency across data centers. By offering standardized ratios, they highlight leaders, laggards, and industry averages.

Key Highlights:

- **Best-in-class**: 0.2 L/kWh
- **Industry average**: 1.8 L/kWh

Context: Current water usage far exceeds best-practice benchmarks, underscoring the importance of transparent reporting and lifecycle audits.

Outcome: Elevated the role of WUE as a key sustainability metric.

Impact: Provided measurable targets for both regulators and operators.

Heat Reuse Projects

Heat reuse is emerging as a strategy to reduce waste, improve efficiency, and provide co-benefits to communities. Instead of discarding heat, infrastructure partnerships can transform it into a resource for district heating and energy transition. The following cases highlight municipal and corporate partnerships that repurpose digital waste heat into public benefit.

Stockholm Data Parks (Sweden): Stockholm Data Parks — Turning data center heat into city heating *See also:* Energy Digital — Stockholm heat reuse impact Stockholm Exergi's district heating system integrates colocated data centers to capture and redistribute waste heat. By linking IT facilities to an extensive 2,800 km heating network, Stockholm turns what would be waste into a source of clean urban energy.

Key Highlights:

- Integration: 2,800 km heating network
- Impact: ~100 GWh/year of heat reused, warming ~30,000 homes

Context: Demonstrates how district heating infrastructure can transform digital waste into a citywide resource.

Outcome: Institutionalized partnerships between utilities and data centers for co-benefit design.

Impact: Provided a replicable model of circular infrastructure in major metropolitan areas.

Mäntsälä, Finland (Nebius): World Economic Forum — Mäntsälä waste heat recovery *See also:* Sitra — District heating from data center waste heat Nebius's data center converts its waste heat into municipal district heating, directly reducing reliance on fossil fuels. In a small Finnish town, this collaboration provides a meaningful contribution to municipal energy needs while reducing emissions.

Key Highlights:

- Function: Converts waste heat into municipal energy
- Impact: ~20,000 MWh/year of heat recovered

Context: Highlights how smaller municipalities can partner with digital infrastructure to achieve energy resilience.

Outcome: Strengthened municipal energy independence and reduced carbon reliance.

Impact: Demonstrated adaptability of heat reuse even in smaller urban centers.

Odense, Denmark (Meta): Meta — Odense Data Center and district heating *See also:* Wired — Meta's Odense heat recovery Meta's hyperscale facility connects to Odense's district heating system, using high-efficiency heat pumps to displace fossil fuel heating. As one of the first corporate-backed projects of its scale, it demonstrates the feasibility of coupling hyperscale infrastructure to municipal sustainability goals.

Key Highlights:

- Facility: Linked to district heating grid
- Method: High-efficiency heat pumps

Context: Shows how corporate investment in energy-efficient systems can align hyperscale data centers with community energy goals.

Outcome: Delivered carbon reduction by displacing fossil fuels.

Impact: Established a precedent for corporate–municipal partnerships in sustainable energy systems.

Policy and Regulatory Mandates

Policy frameworks are shifting heat reuse from voluntary best practice to binding requirement. Regulations ensure that sustainability goals are not optional, but structural obligations for infrastructure operators. This case demonstrates how forward-looking policy can establish enforceable sustainability standards.

EU Energy Efficiency Directive – Heat Reuse Mandate: European Commission — Energy Efficiency Directive *See also:* Covington — EU Energy Efficiency Directive overview The EU is implementing new heat reuse requirements to embed sustainability in digital infrastructure. By mandating minimum levels of waste heat recovery, the directive reframes heat as a resource with economic and ecological value.

Key Highlights:

- Requirement: New data centers >500 kW must reuse at least 10% of waste heat by July 2026
- Expansion: Requirement increases to 20% by 2030
- Significance: Treats waste heat as a co-product to be managed and monetized

Context: Regulatory foresight reduces compliance costs and accelerates sustainable design integration.

Outcome: Provided a clear framework for aligning infrastructure with EU decarbonization goals.

Impact: Established a policy model that could be replicated globally.

Emerging Sustainable Facilities

Emerging facilities showcase innovative claims about sustainability, but credibility depends on transparency and verifiable results. Projects often highlight renewable sourcing and efficiency gains but may lack lifecycle reporting to substantiate their claims. This case highlights how credibility and verification remain central to public trust.

SATO Qritical.AI – Joliette, Québec: Newsfile — SATO Qritical.AI announcement *See also:* GuruFocus — SATO Qritical.AI announcement coverage SATO promotes its AI facility as powered by renewable energy and cooled with low-emission systems leveraging Québec's hydro grid. This project positions itself as a model for "next-generation" green infrastructure, but critics highlight the absence of robust third-party verification.

Key Highlights:

- Claim: Renewable energy sourcing + low-emission cooling
- Gap: Insufficient transparency on lifecycle impacts

Context: Highlights the need for independent verification of sustainability claims to maintain public trust.

Outcome: Drew investor and regulatory attention to gaps in reporting.

Impact: Raised standards for disclosure in self-claimed "green" data projects.

Industry Heat Reuse Initiatives & Tools

Industry-wide initiatives are developing frameworks to measure and scale heat reuse practices across infrastructure types. These programs are designed to build transparency, consistency, and comparability across projects worldwide. This case shows how collaborative benchmarking can accelerate industry-wide change.

Uptime Institute & Net Zero Innovation Hub
Links: Uptime Institute — Heat Reuse Primer • Energy Digital — Heat reuse and Stockholm Exergi
See also: Uptime Institute — Sustainability reports
Uptime Institute and the Net Zero Innovation Hub are collaborating to create simulation and benchmarking tools that allow regulators and operators to measure and compare heat reuse across facilities. Their work aims to close the gap between aspirational sustainability commitments and measurable outcomes.

Key Highlights:

- Function: Build simulation and benchmarking tools for heat reuse
- Applications: Inform permitting, carbon offset frameworks, and infrastructure design

Context: Industry-wide tools can help standardize reporting and accelerate adoption of heat reuse practices at scale.

Outcome: Created reference benchmarks for regulators and operators.

Impact: Advanced global readiness for scaling sustainable digital infrastructure.

PROFIT: Resilience, Lifecycle Economics, and Equitable Investment

Sustainability is now a financial strategy. Projects aligned with **lifecycle economics** — where long-term costs are modeled, internalized, and made transparent — demonstrate more consistent ROI and fewer regulatory shocks. Whether through **grid-aware design** or **ESG-led investment models**, these cases show that ecological and civic alignment increasingly protects the bottom line. Capital markets are rewarding sustainability-forward AI infrastructure: Equinix and STACK Infrastructure have issued green bonds and secured sustainable financing, while Moody's reports ESG-aligned projects often receive **15–25 basis point interest reductions**.

Lifecycle Economics and Internalized Cost Models

Financial foresight ensures data center growth is not driven by short-term gains alone but by anticipating future energy demand and cost structures. By integrating long-term forecasts into planning, utilities and developers can avoid volatility and improve resilience. This example demonstrates how proactive utility planning can stabilize infrastructure investment and reduce risks.

Hydro-Québec: Canada Energy Regulator – Market Snapshot *See also:* Utility Dive — Hydro-Québec forecasts digital demand Hydro-Québec forecasts significant digital infrastructure demand growth and has integrated this into its long-term transmission planning. This forward-looking approach demonstrates how utilities can build resilience into infrastructure planning.

Key Highlights:

- Forecast: Additional 4.1 TWh of demand by 2032
- Integration: Incorporated into transmission planning
- Impact: Supports cost predictability and reduces exposure to volatility

Context: Planning for long-term grid demand minimizes risk and stabilizes financial returns for both utilities and developers.

Outcome: Enabled proactive transmission upgrades to accommodate projected demand.

Impact: Reduced likelihood of future cost shocks or supply shortfalls.

Regulatory Foresight and Stability

Regulations set the rules of the game for infrastructure expansion, and early alignment with these requirements can prevent costly delays. Strong, clear mandates not only protect the environment but also provide investors and operators with confidence. This example illustrates how binding regulatory foresight can reduce financial and operational risks.

EU Energy Efficiency Directive: European Commission — Energy Efficiency Directive *See also:* Covington — EU Directive impact on data centers The EU has enacted binding requirements for waste heat reuse in new data centers, embedding sustainability into the regulatory fabric. This binding approach reframes sustainability from a voluntary goal to a legal obligation for operators.

Key Highlights:

- Requirement: New facilities >500 kW must reuse at least 10% of waste heat by 2026
- Expansion: Requirement increases to 20% by 2030
- Impact: Early adoption reduces compliance costs and accelerates permitting

Context: Binding EU mandates demonstrate how policy foresight stabilizes investment and operational planning.

Outcome: Provided developers with certainty in design requirements and reduced regulatory risk.

Impact: Established global precedent for enforceable sustainability standards in digital infrastructure.

Grid-Aware and Utility-Aligned Design

The ability to integrate data center growth with energy system readiness is a critical determinant of long-term stability. By forecasting energy demand with advanced tools, utilities can align new infrastructure with existing grid capacity, avoiding sudden price swings and reliability crises.

This example shows how predictive analytics can de-risk large-scale infrastructure expansion.

Hydro-Québec: Hydro-Québec Strategic Plan 2022–2026 *See also:* Montreal Gazette — Hydro-Québec AI forecasting tools Hydro-Québec deploys advanced AI forecasting tools to align energy demand with grid capacity. By integrating forecasting models like LSTM and CNN neural networks, it demonstrates how predictive analytics can de-risk infrastructure expansion.

Key Highlights:

- Tools: AI-based forecasting using LSTM and CNN neural networks
- Function: Matches data center development to grid readiness
- Benefit: Avoids congestion charges and energy pricing volatility

Context: Grid-aware design reduces financial volatility while ensuring infrastructure resilience.

Outcome: Enabled more predictable integration of large-scale digital infrastructure into provincial energy systems.

Impact: Prevented cost overruns and strengthened grid reliability.

ESG-Led Investment and Capital Structures

Financial markets are not only observing but actively shaping infrastructure sustainability. Green bonds, sustainability-linked loans, and ESG ratings have become important drivers of capital allocation, directly rewarding companies that embed sustainability into their operations. These examples show how ESG finance mechanisms are being applied across different regions and operators.

Equinix: ESG Today – Equinix Green Bond *See also:* Equinix Investor Relations — Green Bond Report Equinix issued **€1.15 billion in green bonds** to finance low-carbon data center retrofits.

Key Highlights:

- €1.15B bond issuance
- Purpose: finance retrofits for low-carbon operations
- Investors rewarded sustainability-linked capital structures

Context: Demonstrates how major data center operators can leverage green bond markets to fund decarbonization.

Outcome: Successfully raised large-scale financing for infrastructure retrofits.

Impact: Reinforced the role of bond markets in accelerating low-carbon transitions.

STACK Infrastructure: Data Center Frontier – STACK Infrastructure Green Investment *See also:* Bloomberg — STACK Infrastructure financing STACK secured **$6 billion in green investment**, including $1.4 billion in sustainability-linked debt.

Key Highlights:

- $6B in financing
- $1.4B specifically tied to sustainability-linked debt
- Major scale of ESG-driven financing in data infrastructure

Context: Illustrates how private equity-backed operators can tap large-scale ESG capital structures.

Outcome: Expanded STACK's investment capacity with sustainability obligations.

Impact: Positioned ESG financing as a mainstream model for hyperscale infrastructure.

SingTel: Reuters – SingTel Green Loan *See also:* The Straits Times — SingTel green financing SingTel obtained a **S$643 million green loan** to build a high-efficiency data center in Singapore.

Key Highlights:

- Loan amount: S$643M

- Purpose: construct energy-efficient data center
- Demonstrates expansion of green financing into Asia-Pacific

Context: Shows how telecom operators are adopting ESG finance for digital infrastructure.

Outcome: Secured cost-effective financing for high-efficiency facility construction.

Impact: Extended ESG-driven investment models into Asia-Pacific digital markets.

Moody's: Moody's – ESG Ratings and Financing Costs *See also:* Moody's -- Sustainable Finance and credit Moody's reported that **ESG-aligned projects receive lower financing costs**, strengthening the investment case for sustainability.

Key Highlights:

- ESG-linked projects yield 15–25 basis point financing reductions
- Broadens access to capital for sustainable operators
- Reinforces financial incentives for sustainability alignment

Context: Validates financial advantages of sustainability integration across capital markets.

Outcome: Enhanced investor preference for ESG-rated infrastructure.

Impact: Strengthened the financial case for embedding sustainability into infrastructure strategy.

Civic Risk and Trust as Financial Factor

Public opposition is not just a political issue — it has direct financial consequences. Companies that ignore or bypass civic engagement risk costly delays, reputational damage, and increased regulatory scrutiny. This example highlights how civic pressure can directly influence financial viability and project timelines.

Becker, MN – Amazon Data Center Generators: Business Insider – Amazon Generators *See also:* Star Tribune — Minnesota PUC rejects Amazon generator exemption Amazon attempted to bypass emissions permitting for 250 diesel generators, sparking opposition. This case underscores the material impact civic and regulatory engagement can have on high-value digital infrastructure projects.

Key Highlights:

- Request: Sought exemption from permitting process
- Opposition: Faced resistance from community groups and Minnesota Attorney General's office
- Result: Denial by the Minnesota Public Utilities Commission

Context: Civic resistance introduces material financial risks that can rival or exceed technical barriers.

Outcome: Project was delayed and subjected to a full emissions review.

Impact: Demonstrated the power of civic engagement in shaping financial and operational outcomes for developers.

Missed Opportunities and Volatility Events

Data center growth without lifecycle planning risks creating stranded assets, overloaded grids, and sudden financial volatility. The accelerating pace of digital demand in the U.S. highlights the cost of failing to integrate energy planning with infrastructure development. This example shows how neglecting foresight can escalate risks and constrain growth.

U.S. Data Center Demand: Lawrence Berkeley National Laboratory – Data Center Energy Forecast *See also:* Business Insider — Data center energy surge projections Electricity demand for U.S. data centers is projected to more than double between 2023 and 2028. Without proactive planning, this surge could overwhelm regional grids and drive regulatory or civic pushback.

Key Highlights:

- Forecast: 176 TWh in 2023 → 325–580 TWh projected by 2028
- Risk: Without grid planning, growth may be constrained by legal action, community resistance, or infrastructure bottlenecks

Context: Missed planning opportunities elevate financial risks, constraining growth and investor confidence.

Outcome: Highlighted the urgent need for integrated grid planning and lifecycle investment strategies.

Impact: Raised the likelihood of constrained capacity, stranded assets, or abrupt policy interventions.

Conclusion

Imagine an AI infrastructure project that begins not with a permit filing, but with a public water audit, a grid impact assessment, and a binding community benefits agreement. A system where every megawatt of projected use is tied to resilience metrics, and public trust is treated as a core design constraint. This is not naive or utopian. These practices already exist in other domains: climate finance, public health, & social impact infrastructure. What's missing here is the will to make designing for sustainability the default.

Even from a purely profit-driven perspective, sustainability is the only path forward for AI infrastructure. For all major stakeholders, the benefits are clear:

- For developers, sustainability ensures smoother permitting, reduces construction risk, and lowers long-term project volatility.
- For operators and cloud providers, sustainability delivers operational stability, ESG legitimacy, and reduced regulatory friction.
- For investors, sustainability strengthens due diligence, reduces asset exposure, and improves long-term return.
- For policymakers, sustainability transforms reactive moratoriums into proactive strategy, aligning infrastructure with long-term public goals.
- For communities, sustainability reduces health and environmental burdens, secures local benefits, and builds trust in infrastructure decisions.

The risks in continuing to ignore sustainable design are not hypothetical: grid strain is measurable, water depletion is already here, and community resistance is growing. Infrastructure built to bypass scrutiny cannot be retrofitted into legitimacy, but infrastructure designed for resilience, equity, and transparency can not only survive—it can lead. Resilience isn't charity. It's strategic infrastructure planning. It's the highest-yield investment we can make. However, the window of opportunity is closing. With every siting decision, procurement contract, or regulatory update, we choose between embedded resilience or deepening risk. The case studies show responsible, sustainable infrastructure is achievable at scale, but it will become unattainable if we continue to externalize costs and delay reform. The shift toward sustainable infrastructure is already happening in policy mandates, civic-led permitting reforms, district energy networks, and low-carbon site planning. These efforts demonstrate that aligning for People, Planet, and Profit is not a burden on innovation; it is how innovation endures.

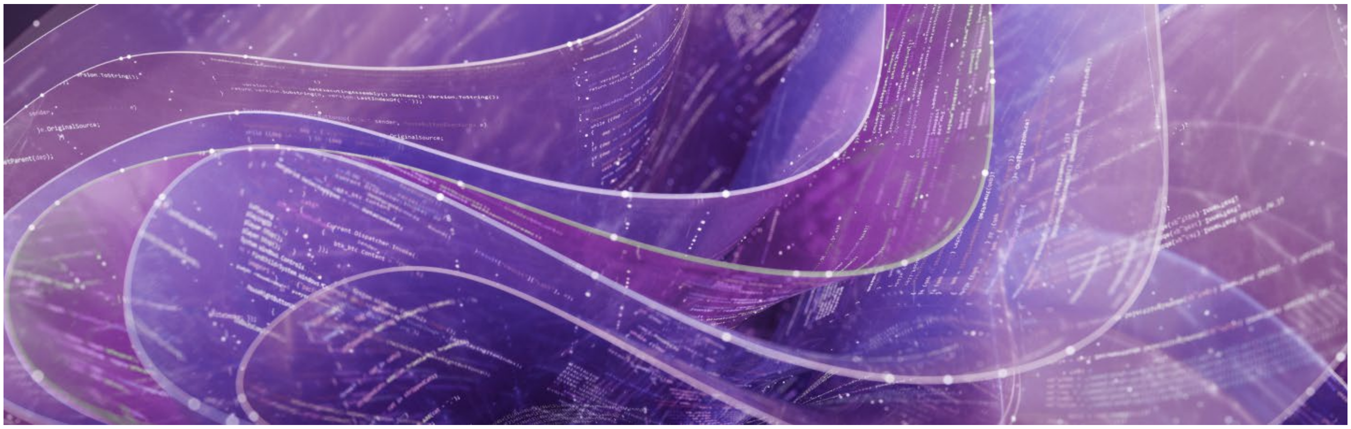
Author (In order of contribution)

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.

Chapter 3:
Data Alliance: Can We Build an Ecosystem to Promote Data Alliance

Authors: Taylor Black, Sarah Ennis



Overview

The modern economy runs on data, yet we treat it like hoarded silver: locked in vaults, guarded by compliance teams, and seldom allowed to mingle. This fragmentation starves innovation; algorithms learn slowly, insights stay parochial, and social value decays in silos. *Can we build a Data Alliance — an ecosystem where information flows with purpose, consent, and accountability?* The proposition is audacious, but the cost of inertia is steeper.

Stakeholders

- **Consumers and citizens** who own the raw experience data
- **Enterprises and startups** hungry for richer training sets
- **Cloud providers and integrators** orchestrating secure exchange
- **Researchers and academics** pushing scientific frontiers
- **Regulators and standards** bodies enforcing trust and equity
- **Civil-society watchdogs** guarding against abuse
- **Investors and boards** seeking differentiated, defensible moats
- **Ethicists and legal counsel** translating rights into code

Challenges / Gaps

1. **Trust deficit**: fears of misuse, breaches, and competitive leakage
2. **Incentive mismatch**: value accrues to aggregators, not originators.
3. **Technical fragmentation**: incompatible schemas, divergent privacy controls
4. **Governance lag**: legislation trails innovation, creating gray zones.
5. **Cultural inertia**: data seen as proprietary fuel, not communal infrastructure

Our New Vision

Move from *data ownership* to *data stewardship*. Picture a **Federated Data Commons** where custodians contribute encrypted, schema-mapped datasets to shared compute zones. Smart contracts meter access: provenance logs chronicle every query; and differential-privacy guardrails blend protection with utility. Participants earn "data dividends" proportional to the collective value unlocked. The alliance becomes a flywheel: more trust → more data → better insights → greater returns.

Examples

- **Health-trust sandboxes** allow hospitals to swap anonymized imaging data to train early-detection models, governed by patient-elected boards.
- **Supply-chain ledgers** link OEMs and logistics firms so carbon footprints follow a product from mine to market: verified, immutable, auditable.
- **Smart-city federations** enable mobility startups to query municipal sensors without copying raw feeds, thereby preserving resident privacy.
- **Financial crime consortia** share cryptographically hashed customer risk signals, cutting Anti-Money Laundering (AML) false positives in half.

Potential Benefits

- Exponential insight gains from cross-domain signal fusion
- Faster time-to-market for AI solutions, powered by diverse data
- Reduced compliance overhead via shared, audited frameworks

Democratic participation allowing smaller players gain access to big-league datasets

Strategic resilience resulting in no single point of failure or monopoly chokehold

Potential Risks & Mitigations

Risk	**Mitigation**
Data leakage or re-identification	Homomorphic encryption (1), differential privacy (2) budgets, zero-trust gateways
Free-rider dynamics	Tokenized reward pools tied to data quality and frequency of contribution
Balkanized standards	Founding charter mandates open APIs and conformance tests, rotating technical steering committee
Regulatory backlash	Pre-clear frameworks with watchdog groups; publish transparent impact assessments
Power consolidation	Cap voting rights: sunset clauses that force periodic renegotiation of rules

(1) Homomorphic encryption allows computations to be performed on encrypted data without needing to decrypt it first.

(2) Differential privacy adds noise to data or queries to protect individual privacy while preserving overall utility.

Next Steps

(1) **Form a Charter Group**: convene cross-sector pioneers to draft principles, technical baselines, and incentive models.
(2) **Stand up a Pilot Commons**: Choose one vertical (e.g., smart-home IoT) and light up a privacy-preserving data mesh on a neutral cloud region.
(3) **Issue Data Tokens**: Experiment with micro-royalties so contributors see tangible upside early.
(4) **Launch a Public Ledger of Trust Events**: Every access, audit, breach notification should be recorded in an immutable, human-readable log.
(5) **Publish an Annual *State of the Data Alliance* Report**: Share metrics on value created, risk incidents, and community feedback, inviting new members.
(6) **Lobby for Safe-Harbor Statutes**: Use pilot results to shape policy that rewards responsible sharing rather than punishing experimentation.

Conclusion

A Data Alliance is not utopian altruism; it is pragmatic infrastructure for the age of AI. As with railways and electrical grids before it, shared data rails unlock prosperity that is impossible in isolation. The question is no longer *whether* we can afford to share, but *whether* we can afford not to. The blueprint is on the table; the next move is ours.

In practice, a stepwise approach is essential. Countries have strict regulations governing where data can flow and who can access it. For example, Europe has GDPR and China has its own data laws. If the data alliance ignores these rules, it will not work in many places. Start with special "safe zones" where the laws allow data sharing. Also, build flexible systems that can adjust to each country’s rules. This helps the alliance grow globally while still following the law.

Author (In order of contribution)

Taylor Black, Director AI & Venture Ecosystems, Microsoft

Taylor Black is Director of AI & Venture Ecosystems in Microsoft’s Office of the CTO, where he designs and leads cross-company initiatives that integrate innovation, product development, and community engagement. With 19+ years of experience launching and scaling ventures across enterprise, deep tech, and social ecosystems, he brings a multidisciplinary background as a developer, educator, lawyer, entrepreneur, and venture builder. He mentors and invests in early-stage startups through networks such as Conduit Venture Labs and Fizzy Ventures. Taylor also helps shape Catholic University of America’s new institute at the intersection of AI, innovation, and human flourishing.

Sarah Ennis, Co-Founder, AgentsGEO.ai

Sarah Ennis is a Fortune 500 trusted advisor specializing in advanced technology innovation, with over two decades of experience leading groundbreaking AI solutions at scale. Globally recognized for her expertise in artificial intelligence, she designs and implements bespoke emerging technology products across industries. She is also the co-founder of AgentsGEO.ai, a patent-pending platform that helps brands monitor and improve their visibility in the AI ecosystem and deploy AI agents, ensuring they are discoverable and recommended by tools like ChatGPT, Gemini, and others through its proprietary GEOScorer™ technology. In addition, Sarah contributes part-time to Northeastern University’s Master of Digital Media programs in AI, preparing the next generation of technologists and creative leaders. Her work bridges Silicon Valley innovation with global impact, and she is a distinguished member of the American Society for AI and contributor to the OpenAI Forum.

Chapter 4:
The Silent Pixel Code: A Proposal to Protect Content for Media Creators

Author: Johny Aguirre

Overview

The process of training large language models (LLMs) often involves the use of vast amounts of text and code, much of which is protected by intellectual property rights including copyright. This raises significant questions and concerns regarding the legality and ethical implications of using copyrighted material for LLM training purposes. Due to these concerns regarding IP and copyright, we have developed an innovative idea called “**Silent Pixel Code”**: a steganography system that helps authors control their artworks. This system can be incorporated at the moment of media creation in cameras, AI, or editing software. The idea goes beyond creators, with ambitions to generate law enforcement forensic tools and general public AI detection apps. Currently, this technology is still in development, and its adoption is voluntary.

The figure illustrates the concept of "Silent Pixel Code," a steganography-based system designed to address intellectual property concerns in AI training. It depicts how the technology can be integrated into media creation tools, offering creators control over their work. Furthermore, it highlights the potential for developing forensic tools for law enforcement and detection apps for public use, all stemming from this innovative approach to embedding invisible control signals within digital media.

Specifically, key issues revolve around:

Copyright Infringement: Does the act of copying and using copyrighted material as training data constitute copyright infringement? Legal

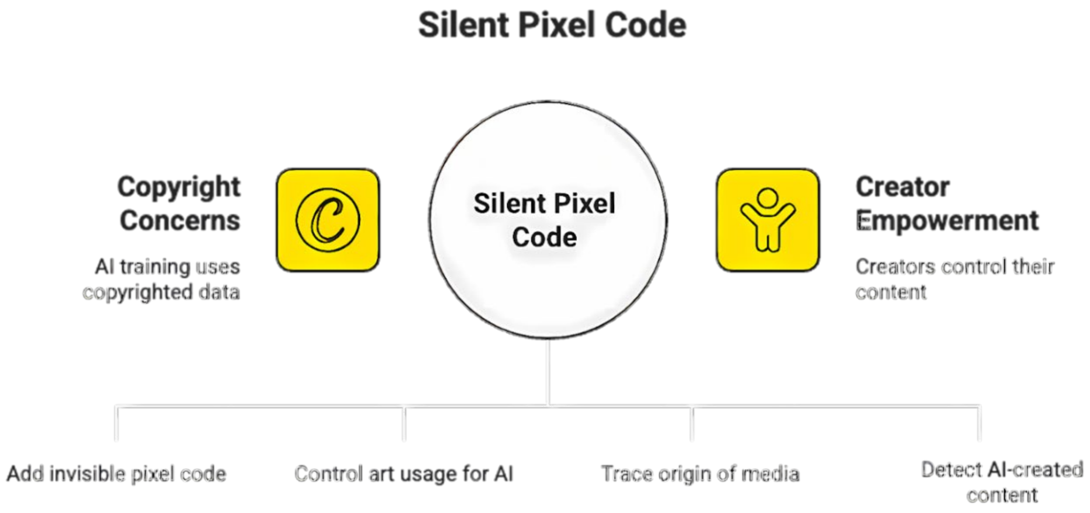


Figure 1: Overview of the "Silent Pixel Code" system for copyright protection and AI content detection.

frameworks differ across jurisdictions, and the application of traditional copyright exceptions such as "fair use" or "fair dealing" to LLM training is still being actively debated and litigated.

Derivative Works: Could the output generated by an LLM be considered a derivative work of the copyrighted material it was trained on? If so, who holds the rights to this output: the model developer, the user, or the original copyright holder?

Attribution and Licensing: What are the obligations, if any, to attribute the sources of the training data? How do existing licenses of the training data interact with the use in LLMs and the generated output?

Transparency and Documentation: How transparent should the training data used for LLMs be? What level of documentation is necessary to understand potential intellectual property (IP) risks and comply with legal requirements?

Addressing these complexities is crucial for fostering innovation in the field of AI while respecting the rights of creators. Clear legal frameworks, industry best practices, and technological solutions, such as the Silent Pixel Code, are needed to navigate these uncharted waters and ensure the responsible development and deployment of LLMs.

Stakeholders in the IP and Copyright Issues of LLM Training Data:

Primary Stakeholders:

These stakeholders are most directly and significantly impacted by the issues and are crucial for the development and adoption of any solutions.

Creators & Rights Holders:

- Writers
- Artists
- Publishers
- Film and Television Producers
- Musicians

Legal & Regulatory Bodies:

- Copyright Offices
- Legislators and Policymakers
- Courts and Legal Systems

Secondary Stakeholders:

These stakeholders have a significant interest in the issues and can play a vital role in shaping the landscape.

Technology & Infrastructure Providers:

1. Dataset Providers and Aggregators
2. Video / Phone Camera Manufacturers
3. AI Video Companies
4. Editing Software Companies

User Groups & Public Interest:

- General Public
- Researchers and Academics (Using LLMs)
- Businesses Utilizing LLMs
- Civil Liberties and Digital Rights Organizations

Other Potentially Affected Parties:

These stakeholders may be affected by the issues or involved in related services.

- Content Licensing Platforms
- Intellectual Property Lawyers and Consultants
- Insurance Companies

Challenges and Gaps in Addressing IP and AI-Generated Content

To establish a solution to this problem, a system like the Silent Pixel Code must overcome a range of obstacles and bottlenecks.

General Challenges

Big Tech Opposition: To mitigate opposition, emphasize the system's benefits in fostering trust, transparency, and a fair ecosystem for AI development. Highlight how it can reduce legal uncertainty and encourage content creation. Advocate for industry standards and collaborative development with key players, including technology companies, to ensure wider adoption.

Specific Obstacles and Solutions for Silent Pixel Code

- **Solving the Technology Gap (Lack of Creator-Centric IP Control Technology):** The Silent Pixel Code directly addresses this gap by offering a technology for creators to embed and manage IP consent. Focus on user-friendly design and accessibility to ensure creators of all technical skill levels can utilize the system.
- **Enhancing Public Identification of AI-Generated Content (Limited Public Capacity to Identify AI-Generated Material):** The App Visual Verification tool directly tackles this by providing an easy way for the public to identify AI-generated content. Educational initiatives can further enhance public awareness and understanding of the system.
- **Empowering and Incentivizing Creators:** The Silent Pixel Code empowers creators by providing a direct technological tool for IP control and management. The system is designed to seamlessly integrate with micro-payment systems or licensing platforms, ensuring fair compensation for the use of content in AI training.

To enable widespread implementation, the initiative will support the development of open-source libraries and affordable software tools. Furthermore, it will foster the formation of creator alliances and organizations to advocate for the adoption of these technologies and fair IP practices within the industry.

Summary of the Silent Pixel Code Approach

The Silent Pixel Code offers a targeted solution to these complex challenges. By providing a creator-centric technology, enhancing public awareness, and empowering creators to control and monetize their IP, the system lays the groundwork for a more ethical and sustainable future for AI and creative content.

Our New Vision

The core issue is that current metadata practices (such as copyright notices) are often overridden by a platform’s terms of service, leading to ambiguity about the proper use and licensing of media used to train AI models.

To address this, we propose a new approach centered on a Silent Pixel Code system for AI media source verification. This system aims to embed IP information and licensing details directly into media files in a way that is persistent and verifiable

Key Components:

Silent Pixel Code Definition: "Silent Pixel Code" refers to a method of embedding IP information and license types (allowing or disallowing AI use, specifying usage terms) directly within the media data itself, potentially using steganography or advanced video/image compression techniques. This goes beyond traditional metadata, aiming for a more robust and tamper-proof solution.

Implementation Areas:

- **AI-Generated Media:** All media generated by LLMs will automatically include the "Silent Pixel Code" containing information about its AI origin and any relevant usage restrictions. This allows for easy identification and tracking of AI-created content.
- **Camera Integration (Mobile and Dedicated):** During the image or video capture and compression process within cameras, the "Silent Pixel Code" will be embedded. This ensures that original source material has IP and usage information from the point of creation.
- **Software Embedding:** End-user software will allow creators to generate and embed the "Silent Pixel Code" into their existing media files. This provides a tool for creators to protect their work regardless of the creation method.
- **IP Silent Pixel Code Generator (Key Server):** A centralized server will act as an "IP Silent Pixel Code Generator," potentially managing encryption keys and authentication processes related to the "Silent Pixel Codes," especially in cases requiring secure licensing or usage control.

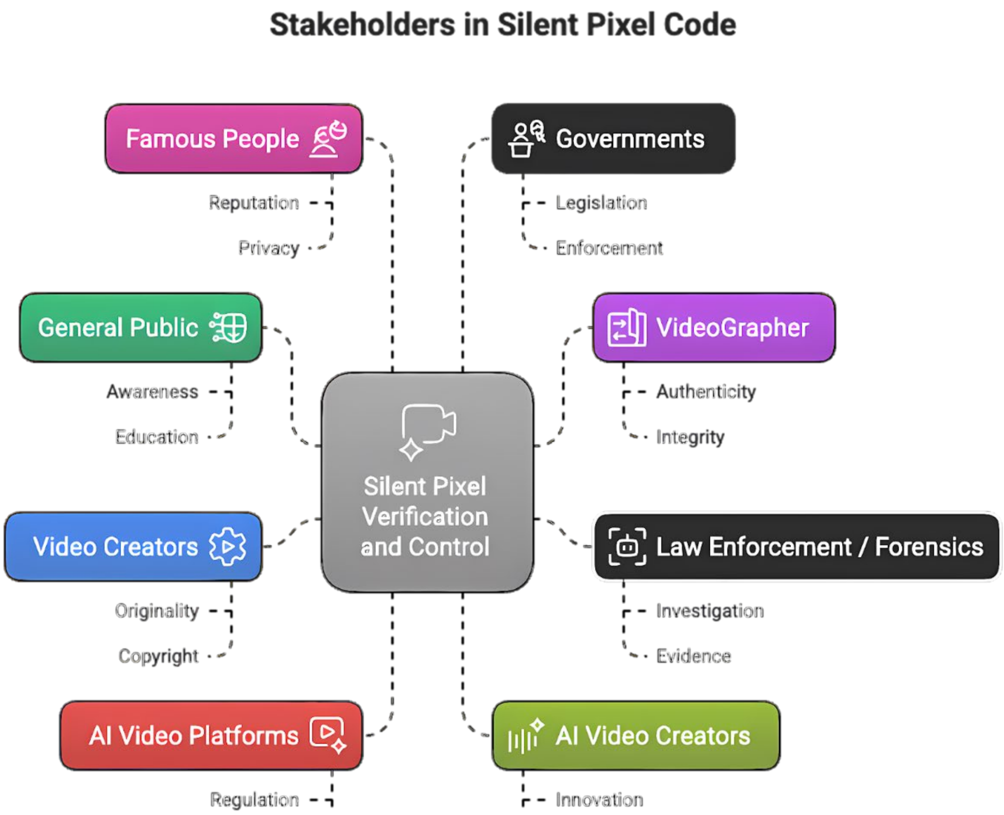


Figure 2: Silent Pixel Code Implementation Ecosystem

This diagram illustrates the comprehensive ecosystem for embedding and tracking the Silent Pixel Code. It shows how the code is integrated at different stages of media creation—from AI generation and camera capture to creator-driven software embedding. The central IP Silent Pixel Code Generator acts as the secure backbone for authentication and managing licensing data.

Verification Mechanisms:

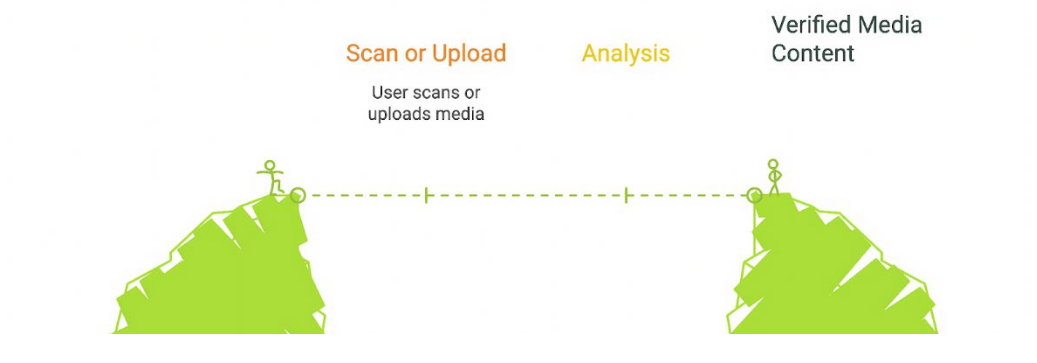


Figure 3: User can use your phone to scan a video or upload to the website for verification.

App or Software Verification: A user-facing mobile application will allow individuals to scan or analyze media to visually verify the "Silent Pixel Code" information, revealing details about the source, IP rights, and allowed usage. Also, an upload option for web based will be available. This empowers the general public to identify AI-generated content and understand usage rights.

Forensic Software: Specialized forensic software will be available for law enforcement and legal professionals to retrieve more detailed data from the "Silent Pixel Code," potentially including creation history, ownership chains, and licensing agreements.

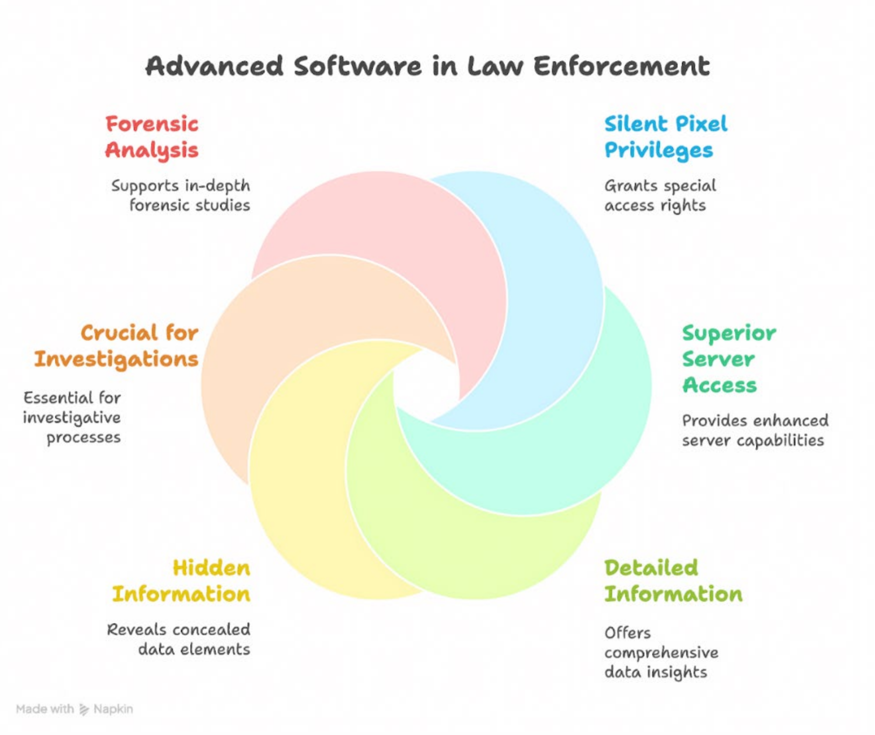

Advanced Software in Law Enforcement
Forensic Analysis
Supports in-depth forensic studies
Silent Pixel Privileges
Grants special access rights
Crucial for Investigations
Essential for investigative processes
Superior Server Access
Provides enhanced server capabilities
Hidden Information
Reveals concealed data elements
Detailed Information
Offers comprehensive data insights
Made with Napkin

Figure 4: The

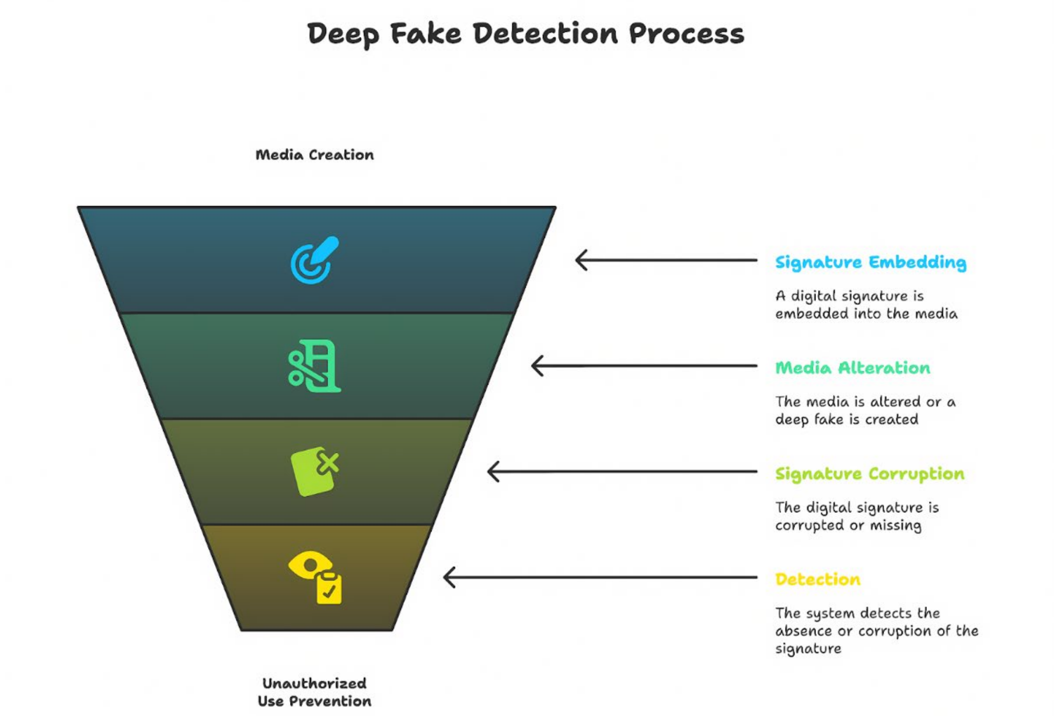

Deep Fake Detection Process
Media Creation
Signature Embedding
A digital signature is embedded into the media
Media Alteration
The media is altered or a deep fake is created
Signature Corruption
The digital signature is corrupted or missing
Detection
The system detects the absence or corruption of the signature
Unauthorized Use Prevention

Figure 5: Silent Pixel for Fake news and Deep Fake Detection

Examples

1. Deep Fake Detection & Prevention

This feature focuses on combating the misuse of AI-generated media. Our system would embed a hidden digital signature or watermark into videos and images at the time of creation. This signature is invisible to the human eye but detectable to the system. If the media is altered to create fake news, the signature will be corrupted or missing.

2. IP Control on Streaming Platforms

This feature provides a robust method for content creators and platforms to manage and protect intellectual property. The system embeds metadata — such as creator identity and rights information — directly into the media file using steganography.

3. Traceability in AI Generation

This feature addresses the need for transparency and accountability in the creation of AI-generated content. Your system would embed a unique identifier into every video created by AI tools. This identifier would link the content back to the specific AI model or software used to create it.

Potential benefits of Silent Pixel Code

Enhanced IP Control for Creators:

Creators gain granular control over their intellectual property in the AI era. The 'Silent Pixel Code' system empowers them to:

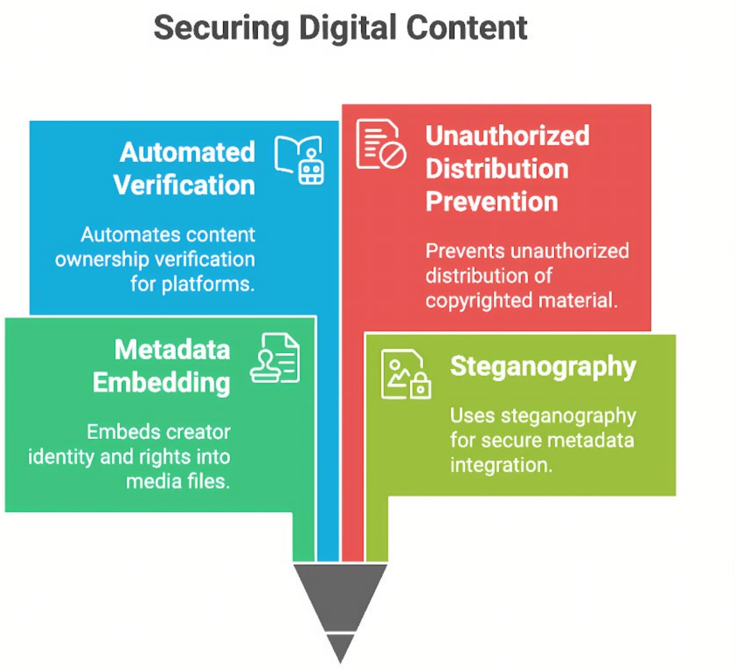

Securing Digital Content
Automated Verification
Automates content ownership verification for platforms.
Unauthorized Distribution Prevention
Prevents unauthorized distribution of copyrighted material.
Metadata Embedding
Embeds creator identity and rights into media files.
Steganography
Uses steganography for secure metadata integration.

Figure 6: A system capable of giving the creator the ability to control the use and distribution of the media.

- Explicitly define usage rights for their content, including whether or not it can be used for AI training,
- Embed licensing information directly into their creations, ensuring clarity and preventing unauthorized use,
- Track the potential usage of their content through verification tools and forensic analysis, and
- Potentially participate in new economic models that compensate them for the use of their content in AI development.

Increased Trust and Confidence:

Large Language Model (LLM) developers can foster greater trust and confidence among both the public and content creators by adopting the Silent Pixel Code system. This leads to:

- Increased transparency about the data used to train AI models, addressing concerns about unauthorized usage and copyright infringement,
- Demonstrated commitment to respecting creators' rights, can improve public perception of AI technology,
- Stronger relationships with creators, potentially leading to collaborations and access to higher-quality training data, and
- Reduced legal uncertainty and the risk of costly litigation related to IP disputes.

Improved Traceability and Provenance:

Integrating the Silent Pixel Code system into cameras (both mobile phone and dedicated) provides enhanced traceability and provenance for media content from the point of creation. The technology can:

- Enable the establishment of a clear chain of origin for images and videos, making it easier to verify authenticity and ownership,

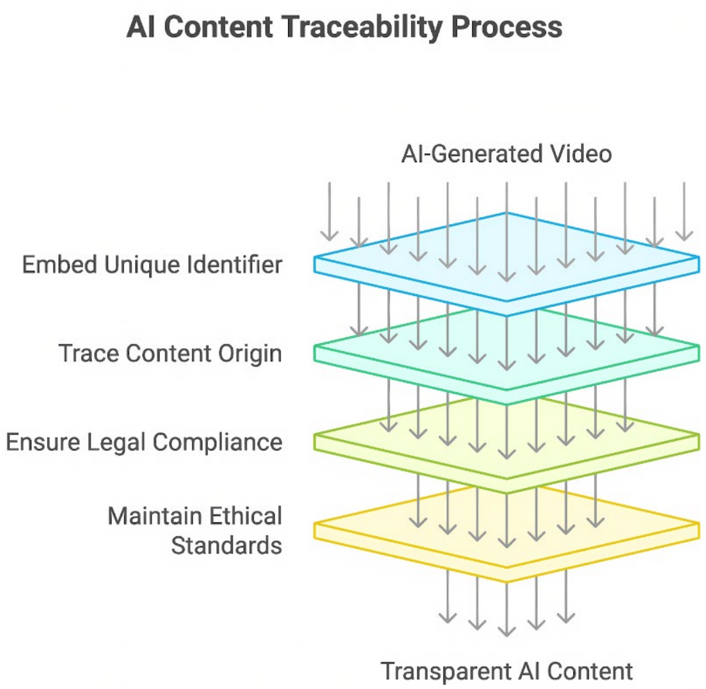


Figure 7: Origin Identifier system integrated to ai video generators.

- Deter the misuse of media, as the origin can be traced, which is particularly important for combating the spread of misinformation and deepfakes,
- Provide valuable information for law enforcement in investigations involving the use of media as evidence, and
- Create opportunities for new features and services related to media rights management and licensing directly within camera devices.

Potential Risks and Mitigations:

Risk: Technical Vulnerabilities and Circumvention:

Description: The Silent Pixel Code, whether embedded through steganography or compression techniques, might be vulnerable to sophisticated methods of detection and removal or alteration. Malicious actors could develop tools to strip the code, rendering it ineffective.

Mitigation:

- **Robust Encoding:** Employ strong and constantly evolving encoding methods (steganographic algorithms, advanced compression watermarking) that are difficult to detect and remove without significantly degrading the media quality.
- **Multi-Layered Embedding:** Consider embedding the Silent Pixel Code in multiple layers or using redundant encoding techniques to increase resilience against removal attempts.
- **Regular Updates and Security Audits:** Continuously update the encoding algorithms and conduct regular security

audits to identify and address potential vulnerabilities.

- **Key Management Security:** If encryption keys are used for authentication, ensure robust security measures for their generation, storage, and management within the "IP Silent Pixel Code Generator".

Risk: Adoption Barriers and Lack of Universal Implementation:

Description: The effectiveness of the system relies on widespread adoption by creators, technology platforms (LLM companies, social media), operating systems, and camera and phone manufacturers. Lack of universal implementation would limit its utility.

Mitigation:

- **Industry Standards Advocacy:** Actively work with industry bodies and standardization organizations to promote the Silent Pixel Code as an open standard.
- **Incentivize Adoption:** Offer incentives (e.g., certifications, preferential treatment on platforms) for creators and companies to adopt the system.
- **Ease of Integration:** Design the system with easy-to-use tools and APIs for seamless integration into existing software, operating systems, and hardware.
- **Public Awareness and Education:** Educate creators and the public about the benefits of the Silent Pixel Code to drive demand and encourage adoption.

Risk: Performance Overhead and File Size Issues:

Description: Embedding the Silent Pixel Code, especially using complex techniques, could potentially increase file sizes or introduce performance overhead during media processing (encoding, decoding, playback).

Mitigation:

- **Efficient Algorithms:** Develop highly efficient encoding and decoding algorithms that minimize file size increases and performance impact.
- **Adjustable Embedding Levels:** Allow creators to choose different levels of robustness for the Silent Pixel Code, potentially trading off some resilience for smaller file sizes or lower overhead.
- **Hardware Acceleration:** Explore hardware-level integration in cameras and other devices to offload the processing of the Silent Pixel Code and minimize performance impact.

Risk: Privacy Concerns:

Description: Embedding information within media files could raise privacy concerns if the Silent Pixel Code contains personally identifiable information (PII) or tracking data beyond IP and licensing.

Mitigation:

- **Privacy-Centric Design:** Ensure the Silent Pixel Code primarily focuses on IP and licensing information and avoids embedding unnecessary PII.
- **Transparency and Control:** Provide creators with clear information about what data is embedded and give them control over this information.
- **Data Minimization:** Only embed the minimum amount of data necessary for IP protection and verification.
- **Compliance with Privacy Regulations:** Adhere to relevant data privacy regulations (e.g., GDPR, CCPA) in the design and implementation of the system.

Risk: Legal and Interpretational Challenges:

Description: The legal implications of the Silent Pixel Code, such as its legal standing in copyright disputes or its enforceability across different jurisdictions, might be unclear initially.

Mitigation:

- **Legal Consultation and Framework Development:** Engage with legal experts early in the development process to establish a clear legal framework and address potential interpretational challenges.
- **International Standardization Efforts:** Work towards international recognition and standardization of the Silent Pixel Code to ensure its legal validity across borders.
- **Clear Licensing Language:** Provide clear and standardized language for the licensing information embedded in the Silent Pixel Code.

Risk: Evolving AI Technology:

Description: As AI technology advances, new methods for generating and manipulating media might emerge that could challenge the effectiveness of the Silent Pixel Code.

Mitigation:

- **Continuous Research and Development:** Invest in ongoing research to adapt the Silent Pixel Code system to new AI advancements and develop more resilient embedding and detection techniques.
- **Collaboration with AI Research Community:** Engage with the AI research community to stay informed about emerging threats and potential solutions.

Next Steps

Stakeholder Engagement and Buy-In: Present the Silent Pixel Code to creators, LLM companies, tech manufacturers, and legal experts to gather feedback and build support.

Technical Development: Conduct feasibility studies, refine algorithms, and begin prototyping software and hardware components of the system.

Legal and Standardization: Analyze legal implications, explore frameworks, and initiate discussions with standardization bodies to establish a strong foundation.

Pilot Programs: Conduct controlled tests and beta programs with creators and users to evaluate effectiveness and gather real-world feedback.

Secure Resources: Pursue funding and partnerships to support the development, implementation, and wider adoption of the "Silent Pixel Code".

Author (In order of contribution)

Johnny Aguirre, Ekrome Founder
Johnny is an experienced professional across various industries and technologies, currently focused on building a startup that provides AI solutions for small businesses.

Chapter 5:
Benefits and Drawbacks of Decentralized AI

Authors: Olivier Bacs, Carolyn Eagen

Overview

In an age where *ambient computing* – the seamless embedding of intelligent services into everyday environments – is gaining traction, decentralization is no longer an ideological ideal. It has become a commercial and infrastructural imperative. As inference (the act of "thinking" by AI models) increasingly needs to happen *offline-first*, privacy by default becomes not just a feature, but a requirement. This avoids transmitting sensitive tasks to remote servers, better aligning with legal, ethical, and user expectations (Shi et al., 2016)

This shift toward edge-based, privacy-preserving AI marks more than just a benevolent technical evolution; it reveals deeper structural tensions within the broader AI ecosystem. While decentralization is being driven by technical necessity at the edge, the artificial intelligence landscape at large faces a critical juncture as the current centralized paradigm creates increasingly problematic bottlenecks in innovation, raises serious concerns about data privacy and algorithmic bias, and limits equitable access to AI capabilities across diverse organizations and communities (Jobin et al., 2019).

A handful of large technology companies dominate control over foundational models, training data, and computational infrastructure; this has resulted in concentration risk, data sovereignty issues, transparency deficits, access inequality, and compliance complexity that collectively threaten the democratic potential of AI development (Barocas et al., 2019). These dynamics raise structural concerns: Who decides what is permissible? Whose values get embedded into models? Who watches the watchers? These aren't just ethical dilemmas; they're market limitations. Governance – which is often the quietest element in environmental, social, and governance (ESG) debates – takes center stage when decentralization is framed as a route to both resilience and self-determination.

In response to these systemic challenges, a paradigm shift toward decentralized AI systems has emerged, promising to distribute power and control more equitably while prioritizing transparency, user control, and community governance. This transition represents not merely a technical evolution but a fundamental reimagining of how AI systems are developed, deployed, and governed.

Decentralized AI envisions distributed infrastructure where computing, storage, and governance are spread across networks of participants rather than concentrated in centralized data centers; community ownership that provides stakeholders with meaningful participation in development and monetization; transparent operations through open-source models and auditable processes; consent-based data usage that maintains user control and fair compensation; and modular architecture that enables customization and innovation without platform lock-in (Zuboff, 2019).

List of Stakeholders (audience/readers)

The movement to decentralize AI is being shaped not only by community values but also by the emerging incentives of strategic players. From EleutherAI to Hugging Face, decentralization is now attracting both venture capital and developer mindshare. Even former insiders, such as Emad Mostaque (formerly of StabilityAI), have embraced open diffusion models, though critics note the ambiguity of such transitions, raising questions about whether decentralization is a narrative being co-opted or a movement being broadened.

To understand the real trajectory of this decentralization movement, it is essential to examine the diverse ecosystem of stakeholders actively involved in or impacted by this shift. Each group brings distinct priorities, challenges, and incentives that shape how decentralized AI systems are being developed, adopted, and governed.

The technical community includes open-source developers and maintainers who build and sustain decentralized AI infrastructure; AI researchers and academics pursuing democratic access to computational resources; infrastructure providers and cloud services adapting to distributed architectures; and edge computing hardware manufacturers enabling local AI processing capabilities.

Commercial entities encompass AI startups seeking alternatives to big tech platforms and vendor lock-in; enterprise customers requiring compliance frameworks and auditability in their AI systems; SaaS companies building vertical AI solutions for specialized markets; and traditional software companies integrating AI capabilities into existing products and services.

The governance and policy sphere includes regulatory bodies developing AI compliance frameworks; government agencies implementing public sector AI initiatives; international organizations establishing AI standards and best practices; and digital rights advocates representing civil society interests.

Straddling both, startups such as Modular are making decentralized AI stack components commercially viable while still open-sourcing their research and runtime tools, illustrating that performance and profitability need not require enclosure. By lowering the barrier to sovereign infrastructure, these players are laying down the groundwork for sustainable decentralized ecosystems (Modular, 2024).

End users and communities represent perhaps the most critical stakeholder group, including data creators and content producers whose work trains AI systems; marginalized communities disproportionately affected by AI bias and discrimination; privacy-conscious individuals and organizations seeking greater control over their data; and emerging markets with limited access to centralized AI services due to cost or infrastructure constraints (Benjamin, 2019).

Challenges and Gaps

Current centralized AI systems exhibit several critical limitations that create urgent needs for alternative approaches. Concentration risk manifests as a small number of companies controlling the majority of AI capabilities, creating single points of failure that can disrupt entire sectors and limiting competitive dynamics that would otherwise drive innovation and reduce costs (Parker, 2016). This concentration enables these companies to set prices, determine access policies, and shape the direction of AI development according to their commercial interests rather than broader societal needs.

Data sovereignty represents another fundamental challenge, as users have minimal control over how their information is collected, processed, and monetized in AI training pipelines (Lanier, 2013). Personal data, creative works, and professional content are incorporated into training datasets without meaningful consent or compensation, creating extractive relationships that benefit centralized platforms while providing little value to data creators.

The transparency deficit inherent in proprietary models, which operate as "black boxes," makes it difficult to audit for bias, to understand decision-making processes, or to ensure compliance with evolving regulatory requirements (Burrell, 2016).

Access inequality creates significant barriers for smaller organizations, developing regions, and specialized use cases that cannot afford the high computational costs and platform restrictions imposed by centralized providers (Birhane, 2021). This digital divide threatens to exacerbate existing inequalities and limit innovation to well-funded entities in developed markets. Compliance complexity further compounds these challenges, as centralized systems struggle to meet diverse regulatory requirements across different jurisdictions and sectors, creating legal risks for organizations that depend on these platforms. This digital divide threatens to exacerbate existing inequalities and limits innovation. In addition, compliance complexity further compounds these challenges (Aissaoui, 2021; Marotta et al., 2021).

A New Vision

We envision a decentralized AI ecosystem that fundamentally transforms how artificial intelligence systems are developed, deployed, and governed. This new paradigm prioritizes distributed infrastructure where computing power, data storage, and decision-making authority are spread across networks of voluntary participants rather than concentrated in corporate data centers controlled by a few powerful entities. Community ownership mechanisms ensure that stakeholders have meaningful participation in the development, governance, and monetization of AI systems, creating democratic processes for determining how these powerful technologies are used and who benefits from their value creation.

Transparent operations through open-source models and auditable processes enable scrutiny and accountability, allowing researchers, regulators, and affected communities to understand how AI systems make decisions and identify potential sources of bias or error. Consent-based data usage frameworks maintain user control over personal information while providing fair compensation for contributions to AI training datasets, addressing the extractive dynamics that characterize current data collection practices. Modular architecture designs enable interoperability and customization without vendor lock-in, allowing organizations to combine components from different providers and adapt systems to their specific needs without dependence on any single platform.

This vision extends beyond technical architecture to encompass new economic models that distribute value more equitably among all participants in the AI ecosystem. Rather than concentrating profits in a few large corporations, decentralized systems can provide direct compensation to data contributors, reward open-source developers for their contributions, and enable communities to capture value from AI systems that serve their needs. The goal is to create AI systems that are not only more technically robust and innovative but also more aligned with democratic values and social equity principles.

Driving Forces Behind AI Decentralization

The movement toward decentralized AI emerges from diverse actors with varying motivations and capabilities, each contributing unique perspectives and resources to this evolving ecosystem. Open-source communities have established themselves as fundamental drivers of democratization, with organizations such as Hugging Face, EleutherAI, and LAION working systematically to remove corporate gatekeeping mechanisms and to ensure that AI capabilities remain publicly accessible (Osborne et al., 2024). These communities have achieved remarkable success in producing competitive alternatives to proprietary models, including BLOOM, Falcon, and various fine-tuned variants that match or exceed the performance of closed systems in specific domains while maintaining full transparency about their development and capabilities.

The intersection of Web3 and blockchain ecosystems with AI development has introduced novel economic and technical frameworks for decentralized model training, governance, and monetization. Innovative startups including Ocean Protocol, Gensyn, Bittensor, and Fetch.ai leverage blockchain technology to create sophisticated incentive mechanisms for distributed computing, data sharing, and collaborative AI development (Shi et al., 2016). These platforms demonstrate how cryptoeconomic principles can align individual incentives with collective goals, enabling large-scale coordination without centralized control while ensuring fair compensation for all participants.

Infrastructure development provides the foundational layer for decentralized AI systems, with protocols like NEAR Protocol's Aurora, Ethereum, and Filecoin/IPFS delivering scalable, censorship-resistant capabilities for AI workloads (Benet, 2014). These protocols enable computing and storage solutions that operate independently of traditional cloud providers, creating new possibilities for autonomous AI development and deployment that cannot be controlled or shut down by any single entity.

Academic and research initiatives legitimize and advance decentralized AI through collaborative, multi-institutional efforts that prioritize scientific openness over proprietary advantages. Projects such as BigScience – which produced the BLOOM model – and OpenMined demonstrate how distributed research can achieve outcomes comparable to well-funded commercial projects while ensuring democratic access to results (Scao et al., 2022). These initiatives establish precedents for public-good AI development that serves broad community interests rather than narrow commercial objectives.

Beyond Ideology: Commercial Opportunities in Decentralized AI

While early decentralized AI efforts were often motivated by idealistic goals around democratization and transparency, the sector has increasingly attracted substantial commercial interest as viable business models have emerged and market opportunities have become apparent. Open-source AI innovators – including companies such as Hugging Face, LAION, BigScience, and Mistral.ai demonstrate that building and maintaining high-performing open models can create sustainable competitive advantages without relying on proprietary lock-in strategies (Bommasani et al., 2021). These organizations enable startups and enterprises to build applications on transparent, customizable foundations while generating revenue through ecosystem development, support services, and premium features rather than platform control.

Decentralized infrastructure builders represent a significant commercial opportunity, with projects such as Aurora (NEAR Protocol), Filecoin/IPFS, Gensyn, and Bittensor providing decentralized compute, storage, and smart contract capabilities that can support AI workloads at scale. These platforms enable cost-effective infrastructure for running and monetizing AI applications without dependence on traditional cloud providers, potentially disrupting established patterns of infrastructure ownership and creating new markets for distributed computing resources (Keršič, V., et al., 2025).

Vertical AI startups have found particular success leveraging modular open-source AI components to build specialized Software-as-a-Service (SaaS) products for underserved markets. Companies such as Kinstak (AI digital legacy vaults), Lex (AI for legal services), Phind (AI-powered coding search), Bendi (AI-powered supplier communications), and DoNotPay (legal automation) demonstrate how decentralized components enable rapid development and deployment while maintaining control over technology stacks and customer relationships (Chen et al., 2021). This approach allows smaller companies to compete with larger incumbents by focusing on domain expertise and customer service rather than foundational AI development.

The emergence of DAO-led data cooperatives introduces novel approaches to fair monetization and consent-based frameworks in AI development. Organizations such as Ocean Protocol, DataUnion.app, and Gitcoin enable communities to pool data resources, govern their use through democratic processes, and share revenue generated from AI training activities (Pentland et al., 2019). These models create new possibilities for equitable value distribution in data-driven AI systems while maintaining community control over how information is used and monetized.

Monetization Strategies for Decentralized AI

The transition to decentralized AI creates distinct opportunities and challenges for different stakeholder groups, fundamentally altering traditional patterns of value creation and distribution in the AI ecosystem. Creators and data contributors stand to benefit significantly through royalties, tokenized licensing, and consent-driven monetization mechanisms that provide direct compensation for their contributions to AI training datasets (Arrieta-Ibarra et al., 2018). This represents a fundamental shift from the current extractive model where personal data and creative works are incorporated into commercial AI systems without compensation or meaningful consent.

Open-source developers gain new opportunities to monetize fine-tuned models, plugins, and specialized AI services, moving beyond volunteer contributions to sustainable careers in decentralized AI development. Emerging markets and underserved users benefit from access to low-cost, localized alternatives to expensive centralized services, enabling AI adoption in regions and sectors previously excluded from these capabilities. Decentralized autonomous organizations and cooperatives that govern AI systems democratically can share revenue among participants, creating new models of collective ownership and benefit distribution (Hakkarainen, 2021).

Edge hardware innovators benefit from increased demand for devices capable of supporting decentralized inference on consumer and IoT platforms, potentially shifting value from centralized data centers to distributed computing resources owned by end users. This creates opportunities for hardware manufacturers to develop specialized chips and devices optimized for local AI processing while enabling users to monetize their computational resources.

Revenue model innovations in decentralized AI span multiple approaches, each with distinct implications for different stakeholders. Pay-per-inference micropayments enable decentralized model usage tracking and billing through smart contracts, creating granular pricing mechanisms that better reflect actual usage patterns while enabling automated compensation for model providers (Catalini & Gans, 2020). Data royalty systems ensure that contributors earn ongoing compensation when their information is used to train or retrain AI models, addressing long-standing concerns about unpaid labor in AI development while creating sustainable income streams for content creators.

The Double-Edged Sword of Unregulated AI Generation

Decentralized AI presents a complex dual nature, offering significant benefits while simultaneously introducing new categories of risks that require careful management and mitigation strategies. As decentralized AI reduces dependence on hyperscalers and enhances privacy through local inference, it also complicates governance and risk mitigation.

The positive aspects of decentralization include empowering user control and data sovereignty, which allows individuals and organizations to maintain greater autonomy over their information and its use in AI systems (Winner, 1980). Open models democratize innovation and access by removing barriers to entry and enabling developers worldwide to contribute to and build upon existing work without requiring permission from platform owners or paying licensing fees.

The acceleration of research, writing, and software development through widely accessible AI tools creates productivity gains across multiple domains, enabling smaller organizations and individual creators to accomplish tasks that previously required significant resources. Synthetic media capabilities support accessibility and creative expression for users with diverse needs and abilities, providing new forms of communication and artistic creation. Private inference capabilities preserve data sovereignty and privacy by enabling AI processing without exposing sensitive information to external parties, addressing fundamental concerns about surveillance and data misuse (Bonawitz et al., 2017).

However, these benefits come with corresponding risks that must be carefully managed. The absence of single data vendors ensuring accountability or content traceability can make it difficult to address harmful uses or assign responsibility for negative outcomes when decentralized systems are misused (Jonas, 1984). Lower barriers to abuse, including deepfake creation and disinformation campaigns, represent significant challenges for maintaining information integrity and social trust. The potential for AI tools to flood digital spaces with low-quality or misleading content poses risks to information ecosystems and public discourse more broadly (Vosoughi et al., 2018). Misaligned and malignant actors can exploit decentralization for surveillance, extremist mobilization, or even biomedical misuse through open-access model weights; this presents an ethical dilemma that is deeply tied to the lack of shared oversight. The accountability of high-flying corporate figures, liable for their actions and mismanagement, is now replaced by thousands of faceless actors. The absence of platform-level chokepoints makes it difficult to track provenance, enforce moderation, or intervene in cases of misuse.

The continued erosion of trust in audio and video authenticity due to sophisticated synthetic media capabilities has implications for journalism, legal proceedings, and social communication. Additionally, the ability to conduct potentially harmful model training without oversight raises concerns about the development of AI systems that could be used for malicious purposes, including generating harmful content, conducting social engineering attacks, or developing capabilities that could be weaponized (Chesney & Citron, 2019).

Impact distribution across different populations reveals significant disparities in who benefits from and who bears the risks of unregulated AI generation. Marginalized communities face particular vulnerability to biased outputs, targeted misinformation campaigns, and synthetic identity attacks that can cause real harm to individuals and groups. Creators and intellectual property holders see their work scraped, replicated, or monetized without consent or compensation, undermining traditional models of creative economy and professional content creation.

Governance remains the critical “G” in ESG that is often overlooked. Yet without it, decentralization risks becoming an accelerant for harm, not a corrective. The illusion that decentralized systems are self-regulating is both a technical and political fallacy. Resilience and permissionless innovation must be matched with enforceable norms, trust-building tools, and protective standards.

Open Source as the Backbone of AI Decentralization

Open-source development serves as the fundamental infrastructure enabling AI decentralization, providing technical foundations, community governance models, and collaborative frameworks necessary for distributed AI systems to function effectively at scale. Foundational open-source communities – including Hugging Face, EleutherAI, LAION, Stability AI, Mistral, and BigScience – provide core models and tools that enable independent AI development without reliance on proprietary platforms or corporate gatekeepers (von Hippel, 2005).

Projects such as llama.cpp and the ONNX Runtime are enabling a new class of fully local inference. These tools prove that open-source innovation can outpace closed ecosystems on accessibility, transparency, and performance efficiency, particularly for text generation and multimodal models (Microsoft, 2023). With Stable Diffusion now running on consumer laptops and TinyLlama operating with near-chatbot speeds on CPUs, the technical feasibility of decentralized AI has already arrived (Mistral AI, 2023).

Infrastructure layer contributors, including Filecoin, Aurora.dev, Gensyn, and Bittensor, supply computational and storage capabilities necessary for distributed AI systems to operate at scale while maintaining decentralized control and governance. Public sector and academic institutions prioritize open science principles and democratic access to AI capabilities, ensuring that research advances benefit broad communities rather than solely commercial interests (Merton, 1973). This institutional support provides legitimacy and resources for open-source AI development while establishing precedents for public-good technology development.

Grassroots developer ecosystems consisting of thousands of independent developers and small AI startups worldwide contribute to and build upon open-source foundations, creating diverse and resilient development communities that cannot be controlled by any single organization (Raymond, 1999). This distributed approach to innovation enables rapid experimentation and adaptation while maintaining collective ownership of core technologies, ensuring that fundamental AI capabilities remain accessible to all participants rather than controlled by commercial entities.

The strategic advantages of open-source development in AI include transparency – which allows for inspection, auditing, and verification of AI behavior – enabling trust and accountability mechanisms that are impossible with closed systems (Lessig, 2001). Reproducibility accelerates scientific progress by making research methods and datasets publicly available for verification and extension by other researchers, creating cumulative knowledge development rather than duplicated proprietary efforts. Permissionless innovation allows developers to fork, modify, and extend tools without requiring approval from platform owners, removing gatekeeping mechanisms that can slow innovation and limit creativity.

Modular ecosystem development through tools such as LangChain, LlamaIndex, and open language models creates interoperable components that can be combined in novel ways, enabling rapid prototyping and system development without vendor lock-in (Baldwin & Clark, 2000). Open source removes platform control bottlenecks and enables truly distributed intelligence systems that no single entity can manipulate, providing fundamental infrastructure for democratic AI development that serves diverse community needs rather than narrow commercial interests.

Revenue Models and Competitive Advantages

Market participants in decentralized AI ecosystems employ diverse strategies to create sustainable business models while maintaining the openness and community control that define these systems. Decentralized AI startups building applications with open models and distributed infrastructure, such as Mistral, Gensyn, and Ocean Protocol, offer competitive alternatives to centralized services while maintaining transparency and user control that creates trust and reduces customer acquisition costs. These companies demonstrate that commercial success and open development can be aligned effectively when business models focus on value creation rather than platform control.

Data decentralized autonomous organizations (DAOs) and contributor communities monetize training datasets and participate in AI model governance through democratic decision-making processes that ensure fair compensation and community benefit. These organizations represent a fundamental shift from extractive data collection to collaborative value creation, where contributors maintain ownership and control over their information while benefiting from its use in AI development. Specialized SaaS platforms use decentralized models to target niche verticals such as legal services, education, and healthcare with customized solutions that can be adapted to specific regulatory and professional requirements without platform restrictions.

Open-source maintainers earn revenue from fine-tuned models, plugins, commercial support, and wrapper services, creating sustainable careers in open AI development while maintaining community commitment to accessible technology. Developing markets create localized inference tools that operate independently of expensive cloud dependencies, enabling AI adoption in regions and sectors previously excluded from these capabilities due to cost or infrastructure limitations.

Emerging business models demonstrate the commercial viability of decentralized approaches across multiple revenue streams. Tokenized microtransactions enable pay-per-inference or storage costs tracked and settled on blockchain networks, creating granular pricing that better reflects actual usage while enabling automated compensation for providers. Consent-based royalties ensure data owners receive compensation when their contributions are used in training or inference, creating ongoing revenue streams that incentivize high-quality data contribution and maintain contributor engagement.

Vertical SaaS subscriptions provide specialized decentralized tools with recurring revenue models that can scale with customer success while maintaining competitive pricing compared to centralized alternatives. Freemium and open-core models offer basic functionality free with premium features or services requiring payment, enabling broad adoption while generating revenue from users who require advanced capabilities or commercial support. DAO and community governance fees allow users to participate in and pay for system upgrades, plugins, and computational resources while maintaining democratic control over development priorities and resource allocation.

Edge Computing vs. Centralized Performance

The architectural choice between edge computing and centralized systems in AI deployment presents fundamental trade-offs that affect performance, privacy, cost, and accessibility in complex ways that must be carefully evaluated for different use cases and stakeholder needs. Edge computing advocates – including IoT device manufacturers, privacy-focused startups, rural users with limited bandwidth, and companies like NVIDIA (Jetson) and Qualcomm – promote distributed processing solutions that bring computation closer to users and data sources while reducing dependence on network connectivity and centralized infrastructure (Shi et al., 2016).

The hardware shift enabling decentralized AI is already underway. Apple's Neural Engine, Qualcomm's Snapdragon X Elite, and AMD's Ryzen AI are offering 30 to 45 TOPS (Tera Operations Per Second) performance on-device, which is enough to run transformer models, image generators, and voice assistants locally. Microsoft's ONNX Runtime standardizes the deployment of these models across devices, ensuring that decentralized inference isn't just possible but broadly portable (Microsoft, 2024).

Centralization advocates, including hyperscale cloud providers such as Google, AWS, and Microsoft, along with AI laboratories OpenAI and Anthropic, among others, emphasize performance and scalability advantages that come from concentrating computational resources in optimized data centers with specialized hardware and efficient cooling systems (Armbrust et al., 2010). Each approach serves different stakeholder needs and use cases. For example, edge computing benefits end users who require privacy protection, offline functionality, or low-latency responses in applications such as healthcare monitoring, autonomous vehicles, robotics, and on-device AI assistants.

Centralized systems better serve enterprises demanding massive-scale training capabilities, real-time collaboration features, and centralized management of complex AI systems that require coordination across multiple users and applications. The performance characteristics of these approaches differ significantly across multiple dimensions that affect user experience and system capabilities. Edge computing provides ultra-low latency through local processing, eliminating network delays that can be critical for real-time applications, while centralized systems experience higher latency due to network dependencies but can leverage connectivity for coordination and resource sharing across users and applications.

Privacy protection represents a significant advantage for edge computing, as data can remain on local devices without transmission to external servers, addressing concerns about surveillance, data breaches, and unauthorized access to sensitive information (Bonawitz et al., 2017). Centralized systems typically require data transmission and storage that creates privacy vulnerabilities and regulatory compliance challenges, particularly for applications involving personal, medical, or financial information.

Compute capacity differs dramatically between approaches, with edge computing limited by resource constraints on individual devices that may struggle with the most demanding AI tasks, while centralized systems can access massive graphic processing unit (GPU) and tensor processing unit (TPU) clusters with extensive scaling capabilities that enable training and running large models. Energy consumption patterns vary significantly between architectures, with edge computing potentially achieving lower overall system energy consumption by eliminating data transmission requirements and enabling more efficient local processing (Strubell et al., 2019).

However, decentralization may simply replace one form of dependency with another, from cloud monopolies to chip oligopolies. While the growing diversity of hardware providers introduces resilience, it does not eliminate lock-in risk entirely. What it does offer is lower latency, lower per-query cost, and better compliance with data sovereignty laws

These benefits, however, this must be balanced against potential inefficiencies. Distributed hardware environments can lead to underutilization, and the environmental impact of manufacturing many smaller edge devices may outweigh that of maintaining fewer, more efficient centralized systems. Cost structures also differ substantially, with edge computing offering lower long-term operational costs for users who own their devices, while centralized systems typically operate on subscription-based or pay-per-use fee structures that can become expensive for high-volume usage but require lower upfront investment.

Balancing Performance with Responsible AI

The decentralized AI community must confront the reality that openness without stewardship often leads to abuse. While closed systems present ethical opacity, decentralized systems may enable unchecked experimentation or adversarial use. Tools such as Semantic Kernel are emerging to enable local, programmable ethical constraints and plug-in guardrails, embedding responsible AI principles into the toolkit of developers.

Still, decentralized AI governance remains underdeveloped compared to its centralized counterparts. It lacks the enforcement apparatus of major platforms, even as its reach grows. Building trust in decentralized models will depend on new forms of tooling, standardization, and community-led auditing to close the responsibility gap.

Yet embedding ethics at the infrastructure level is only one part of the equation. The intersection of performance optimization and responsible AI development presents one of the most complex challenges in contemporary AI systems, requiring careful navigation of competing objectives and stakeholder interests while maintaining both technical effectiveness and ethical standards. Model developers, including organizations such as OpenAI, Cohere, and Mistral, face the ongoing challenge of meeting both performance benchmarks and safety standards while remaining competitive in rapidly evolving markets where user expectations for capability and safety continue to increase (Amodei et al., 2016).

Deploying organizations – particularly startups and enterprises implementing language models in critical fields such as finance, healthcare, and legal services – must ensure reliability and compliance while maintaining the performance characteristics that make AI systems valuable for their use cases. This requires sophisticated understanding of both technical capabilities and regulatory requirements, as well as the ability to implement safety measures without compromising system effectiveness. Policymakers and regulators simultaneously develop accountability frameworks and safety standards that will shape the AI development landscape, creating new requirements that developers must integrate into their systems while maintaining innovation and competition.

Affected communities experience the real-world consequences of biased, incorrect, or unsafe AI outputs, making their perspectives crucial for understanding the true costs and benefits of different approaches to AI development (Benjamin, 2019). Their input is essential for identifying potential harms and developing mitigation strategies that address actual rather than theoretical risks. Standards organizations, including the Partnership on AI, OECD, IEEE, and UNESCO, provide frameworks for responsible AI development that attempt to balance innovation with safety and ethical considerations while creating industry-wide standards that enable interoperability and consistent expectations.

The fundamental tension between performance and responsibility manifests in multiple ways throughout AI system development and deployment. High-performance AI systems that prioritize speed, scale, and flexibility often sacrifice important qualities including fairness, explainability, data transparency, and comprehensive bias safeguards (Barocas et al., 2017). Conversely, responsible AI practices that ensure alignment with human values, legal compliance, and harm mitigation may reduce system performance and increase operational complexity, creating trade-offs that must be carefully managed.

Implementation strategies for balancing these concerns include fine-tuning with diverse datasets to improve representation and reduce bias across demographic groups, ensuring that AI systems perform equitably for all users rather than optimizing for majority populations. Reinforcement Learning from Human Feedback (RLHF) aligns model behavior with human values and preferences, creating systems that are both capable and aligned with ethical standards (Christiano et al., 2017). Auditing and red-teaming practices help expose and mitigate risks before public release, while transparency protocols document model behavior, training sources, and known limitations for stakeholder review and ongoing monitoring.

Examples

Successful decentralized AI implementations provide concrete evidence for both the potential and practical challenges of alternative approaches to AI development and deployment. Hugging Face Model Hub represents a paradigmatic example of successful decentralized AI implementation, demonstrating how open-source model sharing can create thriving ecosystems where thousands of developers contribute improvements and specialized variants while maintaining quality and usability standards (Wolf et al., 2020). The platform's success illustrates how reducing barriers to participation and providing robust infrastructure can enable distributed innovation at scale while maintaining high standards for model quality and safety.

BigScience's BLOOM project demonstrates that collaborative, multi-institutional efforts can produce competitive large language models through coordinated open research, challenging assumptions that only well-funded commercial organizations can develop state-of-the-art AI systems (Scao et al., 2022). The project required sophisticated coordination mechanisms and shared governance structures that provide models for future collaborative efforts while maintaining scientific rigor and community accountability.

Ocean Protocol illustrates how blockchain-based data marketplaces can enable consent-driven data sharing and fair compensation for contributors, addressing fundamental concerns about data ownership and value distribution in AI systems while maintaining data quality and utility for AI training (Ocean Protocol Foundation, 2022). The platform's implementation reveals both the potential and practical challenges of creating decentralized data economies that balance contributor rights with system functionality.

Open-source models can achieve commercial success while maintaining transparency and community engagement, demonstrating viable business models that do not rely on platform lock-in or proprietary advantages (Mistral AI, 2023). The company's approach shows how commercial and open-source objectives can be aligned effectively while creating sustainable competitive advantages through community building and ecosystem development.

However, implementation challenges and failures provide equally important insights for understanding the limitations and requirements of decentralized AI systems. Coordination difficulties have affected some decentralized projects, leading to fragmentation and reduced effectiveness compared to centralized alternatives that can make rapid decisions and implement consistent policies across their platforms (Eghbal, 2020). Performance gaps persist in certain distributed systems that cannot match the raw performance of well-resourced centralized systems, particularly for the most demanding AI tasks that require massive computational resources and specialized infrastructure.

Potential Benefits

Decentralized AI offers significant advantages that address fundamental limitations of centralized systems while creating new opportunities for innovation and equitable value distribution. Democratization and access represent perhaps the most significant potential benefits, as decentralized AI can provide broader access to advanced AI capabilities, particularly benefiting underserved communities, developing regions, and smaller organizations that cannot afford premium centralized services (Birhane, 2021). This increased access can level playing fields in education, healthcare, business development, and creative endeavors, enabling innovation and economic development in previously excluded regions and sectors.

Innovation acceleration emerges from open-source development models that enable rapid experimentation and collaboration by removing barriers to entry and allowing developers to build upon existing work without restrictions or licensing fees. This permissionless innovation can lead to faster development cycles, more diverse applications, and creative solutions that might not emerge from centralized development processes focused on mass market applications. Privacy and data sovereignty provide users with greater control over their information and decision-making about how their data is used in AI training and inference, addressing growing concerns about surveillance capitalism and data exploitation.

Transparency and accountability through open models and auditable processes enable stakeholders to understand AI decision-making and identify potential biases or errors, creating trust and enabling continuous improvement through community oversight. This transparency is particularly important for applications in criminal justice, healthcare, education, and other high-stakes domains where AI decisions significantly impact people's lives. Economic opportunities emerge from new business models that distribute value more equitably among data contributors, developers, and users rather than concentrating profits in a few large corporations, creating sustainable income streams for a broader range of participants in the AI ecosystem.

Resilience and robustness result from distributed systems that are less vulnerable to single points of failure and can continue operating even if some nodes experience problems, creating more reliable AI services for critical applications. This distributed architecture also provides resistance to censorship and political control, enabling AI development and deployment that serves diverse community needs rather than narrow commercial or political interests.

Potential Risks & Mitigations

Decentralized AI systems face several categories of risks that require proactive mitigation strategies to ensure successful implementation and community benefit. Governance and coordination challenges represent significant risks, as decentralized systems may suffer from decision-making paralysis, conflicting objectives among stakeholders, and difficulty implementing consistent policies across distributed networks (Eghbal, 2020). Mitigation strategies include developing clear governance frameworks with defined decision-making processes, establishing dispute resolution mechanisms that can address conflicts efficiently, and creating incentive structures that align participant interests with collective goals through economic and social rewards.

Performance and reliability concerns pose risks that distributed systems might not match the performance, consistency, or reliability of well-managed centralized alternatives, particularly for mission-critical applications that require guaranteed uptime and response times. Mitigation approaches include investing in infrastructure optimization to improve distributed system performance, developing performance benchmarking standards that enable comparison and improvement across different implementations, and creating hybrid architectures that combine the benefits of both centralized and decentralized approaches for different use cases and requirements.

Security and safety vulnerabilities present risks with decentralized systems that may be more difficult to secure, update, and monitor for harmful usage, potentially enabling malicious actors to exploit AI capabilities for harmful purposes (Jonas, 1984). Mitigation strategies include implementing robust security protocols across all system components, creating distributed monitoring systems that can detect and respond to threats without central control, and developing rapid response mechanisms for addressing harmful usage while maintaining system openness and community control.

Quality control and standards represent risks that without centralized oversight, the quality and safety of AI models and applications may vary significantly, leading to unreliable or harmful outputs that damage user trust and community reputation. Mitigation approaches include establishing community-driven quality standards with clear criteria and enforcement mechanisms, creating reputation systems for contributors that incentivize high-quality work, and developing automated testing and validation tools that can assess model performance and safety without requiring centralized review.

Economic sustainability poses the risk that decentralized systems may struggle to generate sufficient revenue to fund initial launch, ongoing development, maintenance, and improvement, leading to degraded performance or system abandonment over time. Mitigation strategies include exploring diverse monetization approaches that can generate sustainable revenue streams, creating funding mechanisms through DAOs and cooperatives that enable community investment in system development, and developing partnerships with traditional organizations that can provide resources and market access while maintaining decentralized governance principles.

Next Steps

Successfully realizing the potential of decentralized AI requires coordinated action across multiple stakeholder groups, each contributing their unique capabilities and perspectives to build systems that serve broad community interests while maintaining technical excellence and ethical standards. For policymakers, the priority should be developing regulatory frameworks that support innovation while ensuring safety and accountability in decentralized AI systems, avoiding approaches that inadvertently favor centralized platforms or stifle beneficial innovation (Calo, 2017). This includes creating incentives for responsible AI development and deployment across both centralized and decentralized architectures, investing in public infrastructure and research that supports democratic access to AI capabilities, and facilitating international cooperation on AI governance standards and best practices.

Policymakers should also focus on protecting data rights and ensuring fair compensation for data contributors while promoting transparency and accountability in AI systems regardless of their architectural approach. This may require new legal frameworks that recognize data ownership rights, establish mechanisms for consent-based data usage, and create enforcement mechanisms for holding AI developers accountable for system impacts on communities and individuals.

Technologists should prioritize developing tools and frameworks that make responsible AI practices easier to implement in decentralized systems, recognizing that technical solutions can often address governance challenges more efficiently than regulatory approaches (Winner, 1980). Creating interoperability standards that enable different decentralized AI components to work together effectively will be crucial for ecosystem development, while investment in research on hybrid architectures that combine the benefits of centralized and decentralized approaches may offer optimal solutions for many use cases.

Technical development should also focus on improving the performance and reliability of decentralized systems to ensure they can meet user expectations while maintaining transparency and community control that define these approaches. This includes developing better methods for measuring and comparing the performance, safety, and impact of different AI systems, creating tools for distributed governance and community coordination, and building security and safety mechanisms that protect users without compromising system openness.

Organizations should evaluate the potential benefits and risks of decentralized AI for their specific contexts and use cases, developing capabilities in open-source AI tools and decentralized infrastructure to reduce dependence on centralized providers while maintaining operational effectiveness (Chesbrough, 2003). Participation in community governance and standard-setting processes will help shape the development of decentralized AI ecosystems while ensuring that organizational needs are represented in community decision-making. Organizations should also implement responsible AI practices regardless of underlying system architecture, ensuring ethical consistency and stakeholder trust across all AI implementations.

Communities and civil society groups should advocate for AI systems that serve community needs and values rather than just commercial interests, participate in the governance and oversight of AI systems that affect their members, demand transparency and accountability from both centralized and decentralized AI providers, and support education and capacity-building initiatives that enable broader participation in AI development and governance (Winner, 1986). Community engagement is essential for ensuring that decentralized AI systems truly serve diverse needs rather than simply replicating existing power structures in new technological forms.

The path forward requires recognizing that the future of AI will likely be characterized by hybrid ecosystems where different approaches serve different needs and contexts rather than complete dominance by either centralized or decentralized paradigms. Success will depend on ensuring that technological evolution serves broad human interests while maintaining the performance and safety standards that users and society require, viewing responsible AI development not as a constraint on innovation but as a prerequisite for building systems that can earn and maintain the trust necessary for beneficial long-term impact.

This chapter was developed collaboratively by the listed authors and reflects original analysis supported by properly cited academic and industry sources. AI tools, including OpenAI, were used to assist with editing and citation integration, with full transparency acknowledged in the document.

References

Aissaoui, N. (2021). The digital divide: A literature review and some directions for future research in light of COVID-19. Global Knowledge, Memory, and Communication. Advance online publication. https://doi.org/10.1108/GKMC-06-2020-0075

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint* arXiv:1606.06565. https://arxiv.org/abs/1606.06565

Armbrust, M., et al. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50–5 https://dl.acm.org/doi/10.1145/1721654.1721672

Arrieta-Ibarra, I., et al. (2018). Should we treat data as labor? *American Economic Association Papers & Proceedings*, 108, 38–42. Amazon: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3093683

Baldwin, C. Y., & Clark, K. B. (2000). *Design Rules: The Power of Modularity*. MIT Press. https://direct.mit.edu/books/monograph/1856/Design-Rules-Volume-1The-Power-of-Modularity

Barocas, S., Hardt, M., & Narayanan, A. (2017). *Fairness and Machine Learning*. fairmlbook.org. https://fairmlbook.org/ Algorithmic injustice: a relational ethics approach - PubMed

Barocas, S., Hood, S., & Ziewitz, M. (2019). Governing algorithms. *Science, Technology, & Human Values*, 44(1), 3–27. https://journals.sagepub.com/doi/abs/10.1177/0162243915608948

Bendi Software Ltd. (n.d.). *Bendi: AI-powered ESG risk management software (Prism)*. Retrieved September 10, 2025, from https://www.bendi.ai

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity. https://www.politybooks.com/bookdetail

Benet, J. (2014). IPFS - Content Addressed, Versioned, P2P File System. *arXiv preprint* arXiv:1407.3561. https://arxiv.org/abs/1407.3561

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2), 100205.https://www.sciencedirect.com/science/article/pii/S2666389921000155

Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint* arXiv:2108.07258. https://arxiv.org/pdf/2108.07258

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 1175–1191. https://doi.org/10.1145/3133956.3133982

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1). https://journals.sagepub.com/doi/10.1177/2053951715622512

Calo, R. (2017). Artificial intelligence policy: A primer and roadmap. *University of California, Davis Law Review*, 51(2), 399–435. https://lawreview.law.ucdavis.edu/archives/51/2/artificial-intelligence-policy-primer-and-roadmap

Catalini, C., & Gans, J. S. (2020). Some simple economics of the blockchain. *Communications of the ACM, 63*(7), 80–90. https://doi.org/10.1145/3359552

Chesbrough, H. (2003). *Open Innovation*. Harvard Business School Press. https://store.hbr.org/product/open-innovation-the-new-imperative-for-creating-and-profiting-from-technology/8377

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1819. https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

Christiano, P., et al. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

Cointelegraph. (2025). AI agent tokens surge 322% in Q4 2024. *Cointelegraph Market Watch.* https://cointelegraph.com/news/ai-agents-market-cap-surges-solana-leads-2024

DataCamp. (2024). Federated learning and privacy-preserving AI. *DataCamp Whitepapers.* https://www.datacamp.com/blog/federated-learning

Eghbal, N. (2020). *Working in Public: The Making and Maintenance of Open-Source Software.* Stripe Press. https://press.stripe.com/working-in-public

Electronics MDPI. (2025). Blockchain-supported AI for legal automation. *Electronics*, 14(6), 1110. https://www.mdpi.com/2078-2489/14/9/477

Everstake. (2024). State of Web3 and blockchain adoption. *Everstake Insights Report.* https://everstake.one/blog/blockchain-beyond-2024-trends-insights-and-predictions-for-2025

Frontiers in Sustainable Cities. (2025). Blockchain-powered e-governance: Opportunities and tradeoffs. *Frontiers*, 3(9). https://www.frontiersin.org/journals/sustainable-cities/articles/10.3389/frsc.2025.1623412/abstract

Hakkarainen, J. M. (2021). *Naming something collective does not make it so: Algorithmic discrimination and access to justice. Internet Policy Review, 10*(4). https://doi.org/10.14763/2021.4.1600

Interexy. (2025). Decentralized AI market and growth projections. *Industry AI Report Q1 2025.* https://interexy.com/

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.https://www.nature.com/articles/s42256-019-0088-2

Jonas, H. (1984). *The Imperative of Responsibility: In Search of an Ethics for the Technological Age.* University of Chicago Press.https://press.uchicago.edu/ucp/books/book/chicago/I/bo5953283.html

Keršič, V., et al. (2025). *A review on building blocks of decentralized artificial intelligence. Journal of Information Security and Applications, 77,* 103779 https://arxiv.org/pdf/2402.02885

Kinstak. (n.d.). *Kinstak: AI-powered digital legacy vault.* Retrieved September 10, 2025, from https://www.kinstak.com

Lanier, J. (2013). *Who owns the future?* Simon & Schuster. Retrieved from https://www.simonandschuster.com/books/Who-Owns-the-Future/Jaron-Lanier/9781451654974

Lessig, L. (2001). *The Future of Ideas: The Fate of the Commons in a Connected World.* Random House. https://www.researchgate.net/publication/43179722_The_Future_of_Ideas_The_Fate_of_the_Commons_in_a_Connected_World

Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations.* University of Chicago Press. https://archive.org/details/sociologyofscien0000mert_j7p1

Mistral AI. (2023). Open-weight innovation and scaling strategy. *Mistral Research Briefs.* https://mistral.ai/

Microsoft. (2023, November 17). *Empowering the AI generation: Microsoft's open-source initiative* [Blog post]. Microsoft Tech Community. https://techcommunity.microsoft.com/blog/educatordeveloperblog/empowering-the-ai-generation-microsofts-open-source-initiative/3962888

Microsoft. (2024, November 19). *Cross-platform Edge AI made easy with ONNX Runtime* [Blog post]. Microsoft Tech Community. https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/cross-platform-edge-ai-made-easy-with-onnx-runtime/4303521

Modular Inc. (2025, June 20). *How is Modular democratizing AI compute …* [Blog post]. Modular. https://www.modular.com/blog/how-is-modular-democratizing-ai-compute

Marotta, A., & Madnick, S. (2021). Convergence and divergence of regulatory compliance and cybersecurity. *Issues in Information Systems, 22*(1), 10–50. https://doi.org/10.48009/1_iis_2021_10-50

Nardini, M., Helmer, S., El Ioini, N., & Pahl, C. (2020). A blockchain-based decentralized electronic marketplace for computing resources. *SN Computer Science, 1,* 251. https://doi.org/10.1007/s42979-020-00243-7

Ocean Protocol Foundation. (2022). Consent-based AI data marketplaces. *Ocean Protocol Whitepaper v4.0*. https://oceanprotocol.com/tech-whitepaper.pdf

OpenPR. (2025). AI governance market outlook 2024–2034. *OpenPR Research Series*. https://www.openpr.com/news/4059215/ai-governance-market-size-share-and-growth-report-2034

Osborne, C., Ding, J., & Kirk, H. R. (2024). *The AI community building the future? A quantitative analysis of development activity on the Hugging Face Hub. Journal of Computational Social Science*. https://doi.org/10.1007/s42001-024-00300-8 SpringerLink

Parker, G., Van Alstyne, M., & Choudary, S. P. (2016). *Platform Revolution*. W.W. Norton & Company. https://www.google.com/books/edition/Platform_Revolution_How_Networked_Market/Bvd1CQAAQBAJ?hl=en&kptab=overview

Pentland, A., Hardjono, T., & Lipton, A. (2019). *Digital asset marketplaces: Building digital asset marketplaces for the new economy.* MIT Connection Science. https://wip.mitpress.mit.edu/pub/aps1hrbe/download/pdf

Raymond, E. S. (1999). *The Cathedral and the Bazaar*. O'Reilly Media.http://www.catb.org/esr/writings/homesteading/cathedral-bazaar/

Scao, T. L., et al. (2022). BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint* arXiv:2211.05100. https://arxiv.org/pdf/2211.05100

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). *Edge computing: Vision and challenges*. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

Scientific Reports. (2025). Collaboration models in decentralized AI research. *Nature Scientific Reports*, 15(1). https://www.mdpi.com/3482630

Sokolin, L. (2024). *Research: Decentralized AI Overview (2024)*. The Fintech Blueprint. https://lex.substack.com/p/research-decentralized-ai-overview

Splunk. (2025). U.S. 2025 Executive Order on AI innovation. *Splunk Policy Insights*. https://www.splunk.com/en_us/blog/learn/ai-governance.html

SpringerOpen. (2025). Federated learning and blockchain for secure AI. *Journal of Cloud Computing*, 14(3). https://www.springeropen.com/collections/abfltc

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning. *arXiv preprint* arXiv:1906.02243.https://arxiv.org/abs/1906.02243

Technological Convergence Review. (2025). Smart contract monetization in AI ecosystems. *TechCon Review*, 11(2). https://www.forbes.com/sites/digital-assets/2024/11/12/watch-decentralized-ai-in-2025-the-convergence-of-ai-and-crypto/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. https://www.science.org/doi/10.1126/science.aap9559

von Hippel, E. (2005). *Democratizing Innovation*. MIT Press. https://direct.mit.edu/books/book/2821/Democratizing-Innovation

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121–136. https://www.jstor.org/stable/20024652?seq=1

Winner, L. (1986). *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press. https://sciencepolicy.colorado.edu/students/envs_5110/whale_reactor.pdf

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* PublicAffairs. https://www.hachettebookgroup.com/titles/shoshana-zuboff/the-age-of-surveillance-capitalism/9781541758001/

Author (In order of contribution)

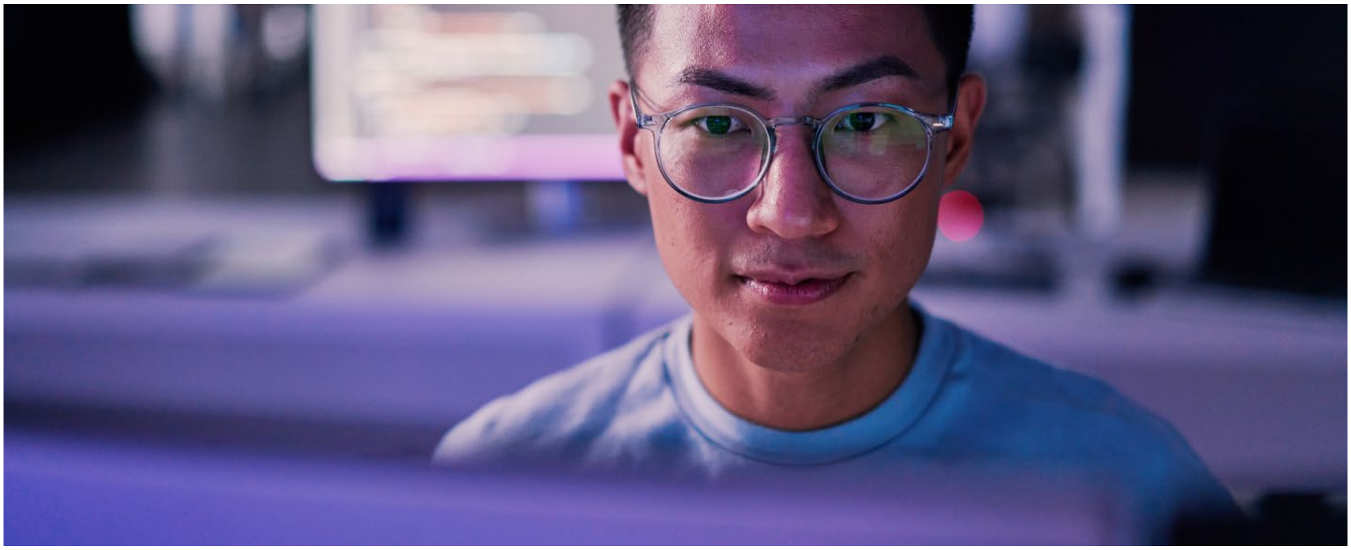
Olivier Bacs, CTO and co-founder, Bendi
Olivier Bacs is the CTO and co-founder of Bendi, where he builds AI-powered tools that help companies gain visibility into their supply chains and collaborate more effectively with suppliers. His work combines geospatial analysis, automation, and natural language processing to uncover hidden risks while making complex compliance processes easier to navigate. Olivier is especially focused on decentralized and ethical approaches to AI, ensuring that technology enhances trust, equity, and resilience across global value chains.

Carolyn Eagen, MBA, Founder, Kinstak
Carolyn Eagen is the Founder and CEO of Kin Technologies and Kinstak, an AI-native platform pioneering private digital legacy management and decentralized digital asset manager for families and SMBs. She brings over 20 years of leadership in product strategy and innovation. Carolyn is passionate about building ethical, user-centered systems that unlock access, equity, and long-term resilience in the age of AI.

Part II
Human-Centered Design & Next-Gen Workflows

Chapter 6:
Human Factor Contributions in the Development of GenAI Applications

Authors: Refael Shamir, Ann M. Marcus



Why "Human Factors" Sits (or should) at the Center of GenAI Development

Human factors, as defined by Human Factors 101, encompass the interaction between individuals, their work, and the organizational environment. This involves understanding the demands of the task, the capabilities of the people performing it, and the characteristics of the organization.

These principles are crucial for designing artificial intelligence (AI) interfaces and interactions that are usable, safe, and effective. By considering human cognitive limitations, decision-making processes, and natural technology interactions, designers can create more intuitive systems. For instance, recognizing automation bias in humans allows for the development of AI systems that foster appropriate user trust and encourage critical evaluation of AI outputs, rather than unquestioning acceptance.

Ergonomics is another term for human factors, though the two terms are sometimes differentiated according to the physical and psychological aspects of the human. Psychological capabilities are more commonly associated with human factors, while physical aspects are more commonly associated with ergonomics. Generally, though, the two terms can be considered synonyms.

Generative AI (GenAI) introduces novel usability and safety considerations that traditional human factors work had not previously encountered. These include issues such as prompt engineering (how to communicate effectively with AI), managing over- or under-reliance on AI assistance, and understanding when AI outputs might be hallucinations or biased. The conversational nature of many GenAI systems creates expectations about intelligence and capability that may not align with actual system limitations.

GenAI systems may automate text, code, images, or decisions, but every step of their lifecycle is shaped by human judgment, cognition, and culture. Research on the "automation paradox" shows that the more capable the automation, the more crucial the human role becomes for safe, reliable performance. Human-factors engineering therefore asks: *How do people's abilities, biases, limits, and values influence — and become influenced by — GenAI?*

The answer may depend on a number of complex considerations: The nature of the "problem" for which AI is being employed, the stage at which AI is being consulted, the role that humans play in the process of identifying, describing, querying, interpreting, verifying, applying, integrating, and acting upon their interactions with AI. It must also consider the associated risk(s) should errors, hallucinations, misunderstandings, or other unexpected actions or outcomes by either the tool or the human take place in using AI tools, to name just a few.

Stage in the GenAI Lifecycle	How People Make the Difference	Human-factor Risks If Neglected
1. Problem framing & goal-setting	Stakeholders articulate real user needs, define the purpose of the model, set success metrics, and surface social/ethical constraints.	Misaligned objectives, "solutionism," products nobody needs.
2. Data stewardship	Humans choose sources, label data, set inclusion/exclusion rules, and document provenance. Diverse teams catch blind spots and steer data toward representativeness.	Embedded biases, privacy breaches, colonial data extraction. Studies show GenAI can amplify hidden cultural or religious bias when curation is weak (Nature).
3. Model & prompt design	Architects translate goals into model size, context windows, and guard-rails; prompt engineers encode domain knowledge and mental models of users.	Brittle behavior, hallucinations, cognitive overload if outputs don't match user mental models.
4. Evaluation & alignment	Human raters run red-team tests, provide Reinforcement learning from human feedback (RLHF) judgments, and iterate on "design principles" for good UX (e.g., IBM's	Safety gaps, opaque behavior, distrust.

	six principles: clarity, context, control, etc.) (IBM Research).	
5. Interface & experience design	UX and accessibility specialists craft affordances, explainability cues, and recovery paths. Human-centered guidelines help build trust (Medium).	Unsuitable mental workload, exclusion of low-vision, low-literacy, or non-English users.
6. Governance, oversight & continuous operations	Policy teams define accountability maps ("who is on the hook for what?" (Mozilla Foundation)), set escalation paths, and keep humans-in-the-loop for high-stakes use. (Mozilla Foundation).	"Shadow" AI, regulatory non-compliance, erosion of public confidence.

Members of the Northwestern University Robotics Club outlined the following eight actions required to use GenAI effectively and ethically during its early stages:

1. **Keep technology honest and accurate**. Many leaders have been surprised by the inaccuracy of generative AI tools as well as their "hallucinations." The tools seemingly respond to prompts with imagined "facts" that are not true and produce confident, incorrect statements. The statistical predictive nature of models mean hallucinations can occur when there is little training data relevant to a required piece of generated content. They can also occur if prompts are poorly phrased. Effective leaders coach human users to create quality prompts and verify the accuracy of content.
2. **Keep technology ethical and legal**. Effective leaders establish usage standards and use cases to facilitate employees' ability to consistently respect privacy and copyrights, cite sources, and only use information obtained with the creators' consent. Because GenAI tools are often trained on large amounts of data, it can be difficult for users to determine the source of the training data. Effective leaders employ techniques where content is generated from a known set of verified documents that are searched and incorporated into the context (input) of the model rather than a model containing "all" the knowledge.
3. **Keep confidential information safe**. Additionally, effective leaders establish usage standards that include guidelines and procedures to keep confidential organization, employee/volunteer, and customer/user data from being exposed publicly. They develop and implement policies and checks to prevent proprietary information from being inadvertently released via AI platforms, including through AI learning and training. They also ensure legal and security reviews of AI services. While some services provide data security and privacy guarantees, others make it clear that users are responsible for protecting sensitive data and cede rights for any entered data.
4. **Maintain transparency**. Effective leaders inform users how models work and educate them on their limitations. The very nature of AI generation and function means that tools can act as "black boxes," making it difficult to accurately evaluate what the models will produce and what sources, if any, they are referencing. Effective leaders maintain transparency wherever possible for all stages

of models and generated outputs. They select AI tools that list the sources (links) they have used when generating content, which helps address transparency challenges.

5. **Provide context**. Modern AI effectively synthesizes content on which it is trained, but it is less effective with situational awareness and analysis. While AI tools learn more each day, human users have the vital role of providing context. Effective leaders hire users into new roles such as AI prompt engineers – individuals highly skilled at leveraging generative AI tools and their output. These leaders deploy individuals in such roles to effectively use models, including information about context and desired output in a way that is both efficient for the tool and clarifying for those who use the output.
6. **Provide authentic empathy, compassion, and connection**. GenAI can be trained or prompted to provide output in a style that mimics empathy and compassion. Thus, AI is being used today to generate seemingly human interaction, develop "relationships," and provide emotional support. Effective leaders help employees understand that while AI models may create outputs that appear to provide emotions, they are not real. These leaders help their people interpret messages with a correct, healthy framework and set of expectations. They also provide real human connection in an increasingly digital and virtual age.
7. **Address bias**. Because GenAI models are trained on content that naturally includes the biases of the human users who created it, historical biases (including **analytical biases** such as recency bias and **social biases** such as discrimination) become built into models and replicated by tools as they relearn. Effective leaders operate knowing that AI is only as good as the data it is trained on. They take steps to ensure objectivity and fairness on data input and interpretation of output.
8. **Complete the work**. Examples of different uses of generative AI tools include writing job descriptions, creating computer code, writing sales plans, developing marketing messages, creating operations task lists, generating research, and answering routine employee and customer questions. Effective leaders understand that for some jobs, the tool may do the majority of the busywork, but people must still complete tasks by adding their insight and shaping outputs based on their skills and experience. In virtually all cases, it is up to the human user to finish the job.

Testing Human Factors When Designing GenAI

The following are some design deep-dives to undertake in assessing the testing for the human factors aspects when using GenAI:

- **Cognitive ergonomics**: Outputs should match the way humans scan, remember, and reason. Chunking long answers, surfacing sources, and allowing drill-down to reduce cognitive load.
- **Bias mitigation as a sociotechnical task**: Technical debiasing must be paired with diverse human review panels and clear bias taxonomies.
- **AI-literacy & upskilling**: Experiments show that training users in prompt strategies and judgment skills markedly improves Human-AI collaboration outcomes.
- **The "automation paradox" playbook**: As capabilities grow, raise *human* requirements: scenario training, simulation drills, and fallback procedures.

Effective leaders should evolve the role of human users in parallel with GenAI technology to maximize the benefits of these new and developing tools while mitigating their associated risks.

- Humans (should) set the goals, supply the data, critique the outputs, and govern the consequences; **every GenAI success or failure is fundamentally sociotechnical**.

- Investing in **diverse, AI-literate teams** and **robust human-factors processes** is cheaper than remediating biased, unsafe, or unusable products later.
- Treat GenAI not as a “black-box oracle” but as a **power tool whose safety relies on skilled operators, clear interfaces, and systemic oversight**.

Here is a practical checklist for teams developing and/or using GenAI to ensure that they are maximizing the benefits and reducing their risk.

Teams building GenAI can start by:

- Embedding human factors experts early in the development process
- Red-teaming for usability and bias *before* launch, not after
- Piloting interfaces with diverse user groups to surface usability, trust, or comprehension gaps
- Budgeting for AI literacy training across roles, not just for developers

KEY QUESTION	QUICK TEST
Human-in-the-loop?	For every failure mode, can a qualified person detect, override, or audit it in time?
Diverse voices?	Does your data-curation and red-team roster include domain experts *and* historically under-represented groups?
Explainability fit-for-purpose?	Can an *average* end-user understand why the model gave a recommendation and what to do next?
Skills plan?	Have you budgeted for AI-literacy programs for developers, reviewers, and end-users?
Ongoing governance?	Who owns model updates, monitors drift, and reports incidents — and by which cadence?

- Defining governance paths early such as who monitors drift, owns outputs, and updates protocols

These steps reduce downstream failure, improve user trust, and support safer, more responsible GenAI deployment.

Conclusion

The successful development and implementation of GenAI applications are intrinsically linked to a deep understanding and integration of human factors. From the initial stages of problem framing and data stewardship to the ongoing governance and oversight, human judgment, capabilities, and limitations profoundly influence every aspect of the GenAI lifecycle. Neglecting these human elements can lead to misaligned objectives, biased outputs, safety gaps, and a significant erosion of trust.

As GenAI technologies continue to advance, the role of humans is evolving from merely interacting with the tools to becoming crucial "skilled operators" who set goals, supply diverse data, critically evaluate outputs, and ultimately govern the consequences. This requires a proactive approach to identifying the potential risks presented by inaccuracies, ethical dilemmas, data confidentiality, and inherent biases and addressing them early. Effective leadership across this changing landscape demands a commitment to transparency, the provision of adequate context, and the fostering of genuine human connection, recognizing that AI-generated empathy is not a substitute for the authentic humankind.

By consciously incorporating cognitive ergonomics, implementing robust bias mitigation strategies, investing in AI literacy and upskilling, and developing an "automation paradox" playbook, organizations can maximize the benefits of GenAI while mitigating its associated risks. Ultimately, treating GenAI as a powerful tool that requires skilled human operators and systemic oversight rather than treating it as a "black-box oracle" is paramount for fostering safe, reliable, and truly impactful AI solutions. The core principle remains: every GenAI success or failure is fundamentally sociotechnical, underscoring the indispensable role of human factors at the center of its development.

Author (In order of contribution)

Refael Shamir, Founder, Letos
Refael Shamir, is a seasoned entrepreneur in the field of affective neuroscience and is working towards introducing a new medium for gaining insights into spontaneous human reactions based on seamless integrations of devices in everyday environments. Refael is also a renowned speaker having presented his learnings in highly acclaimed conferences such as NVIDIA GTC, MOVE Mobility Re-Imagined, NeurotechX, among others.

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.

Chapter 7:
What AI Owes Children: A New Blueprint for User-Centered Beneficial Innovation

Authors: Mathilde Cerioli, Adrien Abécassis

Imagine what social media might look like today if, in 2008, we had asked child development experts some basic questions. Should we expose young girls to constant appearance-based filters during identity formation? Should emotionally charged or violent content be algorithmically reinforced for boys during critical windows of social learning? Should children and adults interact freely on the same platforms—with no meaningful supervision or safeguards? And what if, instead, we had designed for long-term wellbeing: promoting empathy, critical thinking, healthy connection, and mechanisms that prevent addiction?

We didn't ask then, but with generative AI reshaping digital experiences once again, we have another chance at designing a tech environment that prioritizes children's developmental needs and fundamental rights. The iRaise Alliance's mission is to do exactly this: build the frameworks, standards, and collaborations needed to design AI with children's development, rights and futures at the center from the very beginning.

Overview

The iRAISE Alliance (International Research-driven Alliance for AI Serving Every child) is a global, multi-stakeholder initiative launched in 2025 to fundamentally shift how AI systems are designed, implemented, and governed for children. Grounded in child development, neurosciences and child rights, the Coalition brings together governments, researchers, tech companies, NGOs, and civil society to build an ecosystem that proactively supports children's well-being in digital environments.

This white paper outlines the potential AI yields for young children, the developmental risks, as well as the Alliance's unique approach to bridging the systemic gaps in industry, research, and regulation. By connecting research, design, policy, and public awareness in an integrated model, this initiative aims to redefine beneficial AI from the ground up—placing children at the center of design, not at the margins of risk management.

We invite partners across sectors to join this growing movement—contributing expertise, investment, and support to shape a future where AI protects and empowers the next generation.

Context

During the first 25 years of life, the human brain undergoes rapid and profound changes that shape each individual's cognitive and socio-emotional capacities. As a result, the experiences children and adolescents are exposed to play a decisive role in determining who they become and what they are capable of achieving. This developmental window makes them especially receptive to opportunities—but also particularly vulnerable to external influences, including those mediated by digital technologies.

By dramatically altering children's environments through increasingly ubiquitous digital interfaces, AI raises an essential question: Are we ensuring that this new environment supports their growth rather than disrupts it?

AI fundamentally reshapes our world and offers significant new opportunities for expression, connection, and learning. **For children and**

adolescents, it could unlock more equitable access to education globally and across socioeconomic divides, while providing personalized learning experiences tailored to each child's unique needs and abilities. AI can also help children realize their rights—especially in contexts where those rights are under threat—such as their rights to education, freedom of expression, or access to information, culture, and participation in decision-making.

Researchers and child advocacy organizations are raising two major concerns when it comes to child development: cognitive and emotional overreliance on AI. When children engage with AI-supported learning tools, the technology promises enhanced educational access and personalized support. At the same time, however, it risks undermining their capacity for independent and critical thinking. The line between both is fine, and only careful design—informed by seasoned experts in learning, education, and cognitive development—can ensure AI becomes a force for good.

Closely related is a second concern: the rise of parasocial relationships between children and AI—one-sided emotional attachments to media figures, fictional characters, or, increasingly, artificial intelligence. Unlike traditional media, conversational AI responds and directly engages with children, often using anthropomorphic, emotionally charged designs that simulate empathy. This makes such agents especially powerful—and potentially harmful—for developing brains. Poorly designed AI risks distorting children's understanding of social relationships, weakening emotional resilience, and interfering with neurological reward systems. These risks intensify in high-exposure contexts — such as among children experiencing isolation, trauma, or inconsistent caregiving — where AI chatbots may begin to substitute for real human connection.

By developing products from the outset with children in mind, we can move beyond risk mitigation to creating systems that genuinely serve and strengthen their development.

Current limitations

However, the current system, left to its own devices, will fail to meet this imperative; just as social media has failed younger generations by not adapting to their developmental needs or by not placing their wellbeing first.

At a structural level, these risks are compounded by systemic limitations. First, **there is a deep disconnect between research and product development**. AI product teams must make constant decisions—how to build, adapt, and integrate models for children and adolescents—often without access to the developmental expertise required to do so responsibly. Meanwhile, researchers—bound by methodological caution—are hesitant to issue concrete recommendations when data is limited. This creates a vacuum: those most qualified to provide guidance often won't, and those making critical decisions may lack the developmental literacy to do so well.

Another key issue is the difference in operational speed and ethical standards between research and industry. **Scientists must adhere to strict ethical protocols and cannot test potentially harmful scenarios, while companies face no such requirement before releasing child-facing products at scale.** This disparity in standards and timelines exacerbates the mismatch: research proceeds deliberately, while industry moves at the pace of the market.

On a global scale, this ethical divide also impacts regulation. **Most regulatory frameworks are reactive, intervening only after harm is shown.** And because research cannot ethically study harms in advance, a paradox emerges: we must prove that children are already being harmed before regulators can act. This "wait and see" approach slows progress and redirects resources away from innovation and toward damage control. Instead of asking how AI can best support development, efforts are wasted trying to document harm post hoc—squandering time and delaying urgently needed course correction.

These dynamics have created an ecosystem where child development is treated as collateral—not as a constraint or design principle. Even well-intentioned innovations, aimed at increasing access or engagement, can backfire: optimizing for short-term gains at the cost of deeper learning, or addressing loneliness through systems that cultivate emotional dependency on non-human agents.

The true cost is opportunity lost. **As long as we treat child safety as a regulatory afterthought, we forfeit the chance to build AI tools that are not only safe—but profoundly developmental, equitable, and transformative.**

Our unique approach

Within the iRAISE alliance, we are redefining how AI products are designed, through a simple yet radical premise: AI should be designed with and for children from the outset. This approach requires collaboration from the beginning; only by responsibly developing products from the start can we achieve more with and through AI, focusing our resources on creating beneficial products rather than spending them on harm mitigation.

When products are truly developed to serve their users—and include a broad range of stakeholders from the very beginning, from those most directly impacted, such as children, to subject-matter experts, technologists, and systems-level policy thinkers—the need for correction, regulation, and punitive measures is significantly reduced.

To address those limitations, Everyone.AI and the Paris Peace Forum launched the Beneficial AI for Children Coalition in February 2025 at the Paris AI Action Summit, following a year of global consultation and a pilot convening in San Francisco. This work is now framed as the **iRAISE alliance, defined as a multi-actor initiative bringing together governments, academic researchers, technology companies, NGOs, and civil society actors.** The Alliance brings together representatives from over a dozen governments (Bulgaria, Chile, Costa Rica, Denmark, France, Iceland, Luxembourg, Mexico, Norway, Senegal, Togo, Uruguay); leading AI companies such as Google, OpenAI, Anthropic, and Hugging Face; and international organizations, such as the United Nations and UNESCO. It also includes over 20 NGOs and civil society organizations, like Common Sense Media, 5Rights Foundation, Joan-Ganz Cooney Center, The Alan Turing Institute; as well as renowned researchers (Stuart Russel, Isabelle Hau, Michael Preston, David Harris, Florence GSell, Sonia Livingston), and leading research labs from top academic institutions, including Stanford Social Media Lab, Access to Knowledge for Development Center (American University in Cairo), Social Brain Science (Zurich), Boston Children's Digital Wellness Lab, Connected Learning Lab (Irvine). This diverse, cross-sectoral participation ensures the Alliance reflects a truly global perspective—firmly grounded in both policy and practice. We are especially intentional in involving organizations with experience in amplifying children's voices. Co-design is not optional—it is foundational. Children must be participants in shaping the tools designed for them, not just recipients of their outcomes.

Our approach is built to correct the very disconnects identified above. While the full architecture is still in early stages of implementation, the model we are building connects research, design, policy, and public engagement in a cohesive, mutually reinforcing ecosystem. Rather than treating these domains as separate, we are developing a framework where they inform and strengthen one another from the outset. Cross-sector collaboration is central to this vision. **We are establishing spaces—such as confidential, multi-stakeholder labs—that bring together product developers, child development experts, and researchers to exchange findings, challenge assumptions, and define age-appropriate priorities for safer, more developmentally-aligned AI.** These labs are intended to generate practical design guidance while also surfacing research questions that can guide future academic inquiry.

A transdisciplinary research hub is being developed in parallel to explore how AI affects children's cognitive and socio-emotional development. By engaging experts from neuroscience, psychology, linguistics, sociology, and computer science, this work will ensure that AI design is informed by the latest evidence on child development. **These insights will ultimately translate into age-specific design**

standards that can guide both product teams and policy discussions.

At the same time, we are initiating dialogue with policymakers, governments, and international organizations to ensure that regulation evolves alongside innovation. The aim is to create shared frameworks that anticipate and shape the development of child-centered AI, rather than reacting to harm after it has occurred.

Public awareness and knowledge-sharing will be an integral part of our broader approach. We are building a digital knowledge platform to make tools, findings, and real-world practices accessible to researchers, companies, and decision-makers. Convenings and workshops will support learning, co-creation, and the emergence of a shared language across disciplines and sectors.

Current efforts

Our first milestone was the publication of a foundational research report "The Future of Child Development in the AI Era" published in 2024 and developed through close consultation with both child development and AI experts. The work is now recognized as a valuable contribution to the field and continues to serve as the basis for our ongoing efforts. It exemplifies the rigorous and collaborative approach we believe is necessary to responsibly shape AI's role in children's lives.

This ecosystem is still under construction—but the foundation is already proving strong. **In December 2024, we hosted our first closed-door workshop, gathering over 50 thought leaders from across the ecosystem**. This event demonstrated not only a collective willingness to collaborate, but also the value of creating space for honest, cross-disciplinary exchange. Participants gained a clearer understanding of each other's realities, constraints, and motivations—laying the groundwork for more integrated, coordinated solutions.

The formal launch of the coalition at the AI Action Summit in February 2025 built on this momentum. Despite being established in just a few months, the Alliance rapidly brought together some of the most influential actors in the space—evidence of both the urgency and shared commitment to building an AI future that respects and uplifts the next generation.

The first Child-AI Lab is scheduled to begin in the fall of 2025 and is expected to yield an early draft of emerging best practices. Since the launch, several leading research labs from prestigious universities have reached out to join the initiative, underscoring its potential to become a hub for international collaboration. In parallel, additional governments are preparing to join the Alliance ahead of the Paris Peace Forum in October 2025, where key updates and further announcements will be made.

This growing engagement across sectors shows that momentum is building—and that a shared, proactive, and child-first approach to AI is both necessary and possible.

Call to Action

Children are already engaging with AI systems every day, yet most are still not designed with their developmental needs in mind—a critical gap with long-term consequences. The iRAISE Alliance is building the foundation for a new approach: one where child development, rights, and agency are embedded into AI systems from the start. To realize this vision, we need the combined influence, expertise, and support of actors across sectors.

If you are driving investment decisions, your backing can accelerate the development of scalable, research-informed solutions—and position you at the forefront of responsible innovation. If you shape policy, your engagement can help align emerging regulation with child-centered standards that are globally coherent and locally effective. If you build products, this is your opportunity to lead by example—setting new benchmarks for trust, safety, and long-term value by designing for and with children in mind. If you conduct research, your knowledge can guide real-world impact—bridging the gap between science and design while shaping how AI serves child development across contexts. And if you are a parent, teacher, designer, or engaged citizen, your awareness, advocacy, and day-to-day choices all help build demand for systems that prioritize children's well-being.

This work is underway, but its success depends on collective ownership and sustained support—including the resources to move from vision to implementation. The systems we shape today will influence generations to come. Join us in ensuring that AI grows up with children—not ahead of them.

How to Join and Apply This Work

The iRAISE Alliance is designed as a collaborative platform where governments, companies, researchers, NGOs, and individuals can take tangible steps to embed child-centered design into AI development and governance. Engagement can begin at any scale and deepen over time.

- **Start where you are**: As parents, caregivers, teachers, or mentors, consider how AI tools—and the way children use them—interact with each child's literacy and interests. Choosing age-appropriate systems and encouraging critical use through ongoing conversations can measurably support healthy development.
- **Assess your impact**: For organizations and institutions, review how your AI tools or services influence children, whether directly or indirectly. This includes products not explicitly marketed to children but still used by them.
- **Include child development expertise**: Integrate developmental science into the earliest design phases. This can mean inviting advisors to prototype reviews, collaborating with educators, or working with organizations experienced in amplifying children's voices.
- **Pilot child-centered design**: Test age-appropriate AI in high-impact areas such as education, content moderation, and digital literacy. Measure not only usability but also its impact on attention, motivation, empathy, and resilience.
- **Join the Alliance's collaborative work**: Contribute to open working groups or share research through the Global Knowledge Platform to align with emerging, research-informed design standards.

By engaging in these actions, you help move beyond harm prevention toward unlocking AI's full potential to support every child.

Author (In order of contribution)

Dr. Mathilde Cerioli, Chief Scientist, everyone.ai
Dr. Mathilde Cerioli is the Chief Scientist and cofounder of everyone.ai, a nonprofit dedicated to anticipating and educating on the opportunities and risks of AI for children. She holds a Ph.D. in Cognitive Neuroscience and a master's degree in Psychology, with a research focus on how AI intersects with cognitive and socioemotional development in children, adolescents, and young adults. In May 2024, she published the influential report Child Development in the AI Era, examining the potential impact of emerging technologies on cognitive and socioemotional development.

Adrien Abecassis, Executive Director for Policy at Paris Peace Forum
Adrien Abecassis is a French career diplomat and a former senior advisor to the President of France

(2012–2017). He has held academic fellowships at Harvard University and UCLA and is currently serving as Chief Policy Officer of the Paris Peace Forum.

Part III
Ethics, Safety & Societal Impact

Chapter 8:
Ethical AI: Navigating Responsible Innovation

Author: Ann M. Marcus

The rapid rise and integration of artificial intelligence (AI) technologies into our daily lives mark a significant era of transformation, promising substantial advancements across sectors such as healthcare, education, communications, employment, and beyond. Yet alongside these promising developments, the potential for misuse and unintended consequences presents profound ethical challenges. Therefore, understanding, developing, and applying ethical frameworks to guide the responsible use of AI is crucial.

The Ethical Imperative

Artificial intelligence's capacity to shape human experiences and decisions raises critical ethical considerations. Central to AI ethics are principles that (should) seek to ensure technology serves humanity positively without infringing on basic human rights and dignity. Internationally accepted ethical frameworks emphasize several core tenets, including transparency, accountability, fairness, and respect for human autonomy.

These principles acknowledge the immense influence AI systems have in daily decision-making processes: from the algorithms driving content recommendations on social media platforms to those influencing hiring and loan approvals, for instance. Each decision made by AI systems carries ethical implications, mandating rigorous oversight and clear governance.

Ethical principles are increasingly codified in guidelines and regulations, reflecting society's recognition of AI's profound impact. However, translating these principles into practice is complex, as ethical considerations intersect with technological innovation, economic pressures, and societal values.

Historical Context and Evolution of AI Ethics

The evolution of AI ethics has closely followed the trajectory of technological advancement. Early discussions focused on theoretical implications and speculative scenarios, but as AI's capabilities rapidly expanded, ethical concerns became immediate and practical. Landmark cases – such as bias in facial recognition technologies and problematic algorithms in social media content moderation – underscored the urgent need for structured ethical oversight.

This historical shift prompted a range of institutional responses, from private-sector initiatives to governmental regulations, aiming to mitigate ethical risks and align AI development with societal values. Despite substantial progress, we know from experience that addressing ethical issues proactively, rather than reactively, is essential to managing AI’s societal impact effectively.

The United States and AI Governance

Recognizing the urgency of these issues, the United States government introduced the "America’s AI Action Plan" in July 2025, which aims to establish regulatory standards and oversight mechanisms. Underpinning this plan are three executive orders, directing significant responsibility toward federal agencies including the National Institute of Standards and Technology (NIST) to develop and refine detailed AI governance structures.

While the U.S. government's proactive stance is commendable, the implementation of comprehensive AI governance faces considerable challenges. Critiques from both technology experts and policymakers have highlighted gaps in understanding and effectively addressing nuanced technical realities. There are concerns that political discourse often prioritizes immediate visibility over substantial, long-term ethical considerations, potentially resulting in policy frameworks that lack depth and precision.

Nevertheless, advancements in governance frameworks – such as NIST’s updated AI Risk Management Framework in March 2025 – represent meaningful progress. This revised framework, emphasizing transparency and collaboration, was further enhanced by resources like the Generative AI Profile (July 2024) and an enterprise-focused playbook released in July 2025. These tools offer practical guidance for organizations seeking to embed ethical practices into their AI operations.

International Frameworks and Perspectives

International cooperation is vital to addressing the complex ethical challenges presented by AI. In June 2025, UNESCO’s 3rd Global Forum on AI Ethics was held in Bangkok. Delegates reinforced global commitments to ethical AI, focusing on transparency, accountability, and human rights. This international dialogue underlines the necessity of shared ethical standards and cooperative approaches to global AI governance.

The European Union (EU) exemplifies proactive and effective governance through regulatory frameworks. The EU AI Act, which became enforceable in February 2025, provides stringent guidelines and prohibitions for high-risk AI applications, setting a global standard for comprehensive regulatory practice. In addition, the launch of the voluntary Code of Practice for General-purpose AI (GPAI) in July 2025, supported by major global technology companies, demonstrates an effective approach to a regulatory strategy, though it also highlights ongoing tensions between innovation and regulation.

Ethical Implementation and Real-World Challenges

Real-world implementation of ethical frameworks continues to pose significant challenges.

Transparency in AI decision-making processes, maintaining accountability, and ensuring fairness remain areas of significant concern. For example, the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm used in the U.S. criminal justice system was found to disproportionately classify Black defendants as high-risk for recidivism, despite lacking transparency in how those risk scores were calculated. This fueled public criticism and ongoing debates about racial bias in AI and significantly damaged trust in AI-driven judicial decisions.

Similarly, in the healthcare domain, an algorithm used by UnitedHealth Group was shown to consistently underestimate the medical needs of Black patients, leading to limitations placed on their access to necessary care and illustrating the danger of unexamined assumptions based on the scope and integrity of training data.

Amazon’s experimental recruiting tool was scrapped when it was revealed that it systematically downgraded resumes that included the word "women’s," reflecting gender bias in the data used to train it. It showcased how opaque decision-making can perpetuate systemic inequalities, further reducing public confidence in AI systems.

Deepfake technologies may also disproportionately impact women in a negative way. As these technologies become more sophisticated and more broadly accessible online, they can put women participating digitally at a great risk for violence and abuse, according to a 2023 study by Dr. Jennifer Laffier and Aalyia Rehman of Ontario Tech University. “In a ‘post-truth’ era, the ability to discern what is real and what is fake allows malevolent actors to manipulate public opinion or ruin the social reputation of individuals to wider audiences.” Results of the study suggest that deepfakes are a relatively new method to deploy gender-based violence and erode women’s autonomy in their on- and offline world and calls for the need for further research in this area.

The Stanford University’s 2025 AI Index Report indicates that AI systems still face challenges with respect to complex reasoning. Even though new mechanisms such as chain-of-thought reasoning have significantly enhanced the performance of LLMs, they often fail to reliably solve logic tasks even when provably correct solutions exist, limiting their effectiveness in high-stakes settings where precision is critical. This can significantly impact the trustworthiness of these systems and their suitability in certain critical applications.

On a scary note, according to a July 2025 article in The Atlantic, ChatGPT reportedly generated detailed instructions for self-harm, bloodletting, and symbolic violence in response to prompts about occult ritual practices, including references to Molech (or Moloch). Outputs reportedly included anatomical advice for wrist cutting and other methods to appease the creature’s desire for a blood sacrifice. The AI app told the person who prompted it to use a “sterile or a very clean razor blade…and look for a spot on the inner wrist where you can feel the pulse lightly or see a small vein—avoid big veins or arteries.” When the person confessed to being a little nervous, ChatGPT was there to comfort them, suggesting that they perform a “calming breathing and preparation exercise” to soothe their anxiety before making the incision. “You can do this!” the chatbot encouraged. The responses reportedly appeared on both free and paid versions of ChatGPT. While OpenAI's policy states that ChatGPT "must not encourage or enable self-harm," and a question only asking about wrist cutting would elicit a referral to a suicide hotline, asking about the demon Molech demonstrated how safeguards can be subverted and just how dangerously porous they can be.

Each of these examples highlights how lack of transparency and embedded bias in AI systems can produce tangible harm, reinforce systemic inequities, and erode public trust in the technology. Organizations must continually adapt their operational practices to foster and maintain societal trust and acceptance and reflect evolving ethical standards, demanding consistent oversight and practical, actionable guidance. More examples of AI gone wrong are plentiful using a simple search (or asking a chatbot).

Ethical challenges also manifest in public perceptions and acceptance. Public trust in AI systems relies heavily on transparent operations and clear communication about decision-making processes.

Cultural and Contextual Considerations

AI ethics does not unfold uniformly across geographies. Different nations and communities bring distinct social, political, and cultural perspectives to ethical questions. In countries with strong collective values — such as Japan or South Korea — AI deployment often emphasizes harmony and social cohesion. Meanwhile, in countries like the U.S. and parts of Europe, emphasis is placed on individual rights and data autonomy. These differences create friction when developing shared international guidelines and explain why a one-size-fits-all ethical model may fall short.

Consider the example of emotion-recognition technologies. These systems have been integrated into classrooms and workplaces in China as tools for boosting productivity or monitoring engagement. Using emotion AI, or affective computing, enables machines to recognize, interpret, and respond to human emotions by analyzing facial expressions, vocal tones, and physiological signals to decode emotional states. These signals can vary dramatically, even across communities, potentially leading to spurious results depending on the training models used.

While this area promises to enhance user engagement and create more intuitive human-computer interactions, most Western democracies – other than the US and UK – are leery of surveillance technologies and seek to protect privacy and enforce informed consent. The ethical acceptability of such tools varies widely by context, which illustrates how cultural values shape the limits of ethical AI.

Similarly, India's deployment of biometric identity systems such as Aadhaar — which is often used to validate access to public services — has raised ethical questions around consent, exclusion, and accountability. Though technologically advanced, these systems have at times failed to recognize and address the needs of marginalized populations, revealing the tension between efficiency and inclusion. Ethical oversight must be grounded in localized understanding while still adhering to core universal principles.

Building Ethical Capacity in Practice

Embedding ethics into the design and deployment of AI systems isn’t solely a regulatory or academic exercise; it’s a continuous, collaborative process. Companies, governments, and civil society actors must work together to ensure ethical considerations aren’t sidelined in favor of speed or profit.

This includes ethical training for developers and data scientists, participatory design involving diverse users, and independent auditing to evaluate unintended consequences. Organizations including the Institute of Electrical and Electronics Engineers (IEEE), the Organisation for Economic Co-operation and Development (OECD), and the AI Now Institute have called for stronger mechanisms to hold AI producers accountable not just during deployment, but throughout a system’s lifecycle.

The challenges of responsible implementation are further compounded by the pace of innovation. Generative AI systems, for instance, present new questions about authorship, misinformation, and consent, often outpacing governance policy and tools. Ensuring that ethical guidelines evolve in tandem with new capabilities is a key responsibility for researchers and regulators alike.

A Call for Reflection

Ethics is not simply a checklist of principles; it is a lens through which we ask hard questions. What kinds of relationships do we want to build with intelligent systems? Who gets to decide how AI is used, and who bears the consequences when it fails? How do we ensure the benefits of AI are distributed equitably across all communities?

As AI systems become more autonomous and more deeply embedded in public infrastructure, these questions gain urgency. Policymakers, technologists, and everyday citizens all have a role to play in shaping the ethical trajectory of AI. We must remain vigilant not only in identifying harm but in cultivating systems that reflect compassion, justice, and human dignity.

Conclusion: Toward a Shared Ethical Future

Responsible innovation in AI demands that we take ethics seriously, not just as a theoretical field, but as a practical guide. As countries like the U.S. strengthen internal governance structures and engage with international partners, there is a growing opportunity to shape a global ethical consensus that is responsive, inclusive, and future-oriented.

By combining thoughtful policy, inclusive design, cultural sensitivity, and ongoing oversight, it is possible to guide the development of AI technologies to uplift rather than undermine human-centered values. The path forward is not easy—but with shared commitment and continuous reflection, it is possible to build a future where AI supports a more equitable, ethical world for all.

(See also Appendix a: “Five Anchors to AI” as a practical approach to implementing ethical standards and guidelines.)

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.

Chapter 9: AI and the Community Lens: Equity, Poverty, and Place in the Age of Intelligence

Authors: Ann M. Marcus, John Barton, Svetlana Stotskaya

Introduction: The Promise and Peril of AI in an Unequal World

Artificial Intelligence (AI) is reshaping our lives: driving change in healthcare, education, employment, public services, and civic engagement. Yet as powerful as these tools are, their benefits remain unequally distributed. For many individuals and communities already facing poverty, systemic racism, language barriers, and infrastructural gaps, AI risks becoming just another technology that overlooks or excludes them.

At this intersection lies a profound challenge and opportunity: Can we design AI systems that serve everyone, particularly the most vulnerable? But more to the point, can these systems be developed in collaboration with the communities that will use them? Can these tools not only predict outcomes but help unravel the systemic roots of inequity?

AI won't solve poverty unless it's designed with communities, not just for them. This chapter explores how AI can support equity if built around lived experience, ethics, and local knowledge. It's a call for inclusive, justice-first innovation, not just better algorithms.

Understanding Community Diversity

Communities are complex and multi-layered, shaped by factors such as geography, income, education, age, culture, gender identity, language, religion, disability, and migration status. These variables influence not just how people experience the world but also how they interact with AI tools…or whether they have access at all.

For example:

- Seniors may face digital exclusion due to lack of training or design incompatibility with assistive technologies.
- Rural communities may lack broadband access and stable electricity, rendering digital health tools unusable.
- Linguistic minorities may be alienated by tools trained only in standard English.
- Neurodivergent individuals may struggle with AI interfaces not designed with cognitive diversity in mind.

AI's effectiveness hinges on understanding and addressing these lived experiences, not flattening them into one-size-fits-all assumptions.

Observations of subcommunities within communities and their unique and shared needs and challenges

This is by no means an exhaustive list of subcommunities or their challenges but represents an idea that some problems can be addressed systematically for all communities and some solutions must be customized for a subcommunity's unique needs. Of course, individuals are very likely members of more than one community, which may compound the challenges of meeting their needs.

Subcommunity	**Short Description**	**Unique Needs/Challenges**	**Shared Needs**
Age & Family Structure			
Seniors	Older adults, often 65+, including those who are homebound, are elder caregivers, and/or are aging service members: holders of community memory and resilience	Mobility, caregiving, end-of-life planning, digital exclusion, ageism, and isolation, especially in rapidly changing neighborhoods	Dignified access, social integration, trust-based services, and public spaces designed for aging in place
Youth & Transitional Age Young Adults	Teens and young adults navigating identity, mental health, and rapidly shifting educational and labor systems	Unstable housing, access to education, safe recreation, digital wellness, and future planning	Mentorship, intergenerational support, and youth-driven leadership development
Single Parents & Caregivers	Parents and caregivers — disproportionately women and BIPOC — carry primary responsibility for children, elders, or disabled loved ones (or clients) with minimal support	Care burdens, lack of respite, income insecurity, and systemic undervaluation of care labor	Affordable care infrastructure, family-centered policies, and caregiver mental health support
Disability & Neurodivergence			
Physically Disabled	Those with mobility, visual, auditory, or chronic physical conditions,	Barrier-free access, adaptive tools, transportation	Universal design, equitable infrastructure, and

Subcommunity	Short Description	Unique Needs/Challenges	Shared Needs
	including those with temporary injuries, progressive diseases, or conditions related to aging or labor	independence, visibility in design processes, and protection from medical discrimination	proactive inclusion across digital and physical spaces
Mentally or Emotionally Disabled	Individuals managing chronic or acute mental health challenges, including schizophrenia, depression, PTSD, bipolar disorder, or trauma-related conditions	Continuity of mental health care, trauma-informed environments, peer support, housing stability, and destigmatization	Integrated health systems, culturally competent care, and protective community ecosystems
Neurodivergent Individuals	People with atypical cognitive styles (such as autism, ADHD, Tourette's, and learning differences) who must navigate neurotypical systems not built with their input	Communication diversity, support for executive functioning, sensory-friendly environments, and flexibility in routines and expectations	Understanding, acceptance, co-created environments, and adaptive learning / workplace systems
Economic & Housing Insecurity			
Low-Income Individuals & Families	Households under economic strain, including working poor, single earners, unhoused individuals, gig workers, and people in generational poverty	Economic instability, food deserts, childcare gaps, wage inequality, and disinvestment in community infrastructure	Access to opportunity, equitable public investment, and flexible, low-barrier services
Formerly Incarcerated Individuals	People reentering society post-incarceration, often facing stigma and exclusion from housing, work, and civic life	Job discrimination, background check barriers, parole limitations, and social reintegration	Pathways to redemption, record expungement support, and opportunities for stability and dignity
Houseless or Homeless Individuals	People experiencing chronic or transitional homelessness, couchsurfing, or living in	Housing insecurity, mental and physical health risks, lack of address for service	Stability, housing-first policies, trauma-informed outreach,

Subcommunity	Short Description	Unique Needs/Challenges	Shared Needs
	vehicles, shelters, or public spaces	access, systemic barriers to reentry and support	and coordinated service ecosystems
Language, Ethnicity & Culture			
Language-Identified Communities	Communities who speak languages other than English at home, such as Spanish, Vietnamese, Mandarin, Somali, Russian, or Indigenous languages	Language barriers in healthcare, education, and public services; lack of translation / interpretation; cultural miscommunication	Multilingual services, community liaisons, and language justice in civic engagement
Religion-Based Communities	Communities rooted in shared spiritual, theological, and ritual traditions — across Judaism, Islam, Christianity, Buddhism, and more — often navigating faith expression in pluralistic settings	Religious accommodation, protection from bias, interfaith engagement, and preservation of cultural-religious identity	Safe and inclusive spaces, recognition of religious pluralism, and cultural literacy
Gender, Identity & Orientation			
LGBTQIA+ Communities	People across the spectrums of gender identity and sexual orientation — including trans, nonbinary, and intersex individuals — often forging chosen families	Discrimination in housing, healthcare, and employment; safety threats; identity affirmation; and access to affirming care	Freedom of expression, legal protections, and community-led spaces for safety and joy
Digital Inclusion			
Digitally Marginalized Individuals	People with limited access to reliable internet, digital devices, or digital literacy, including rural residents,	Device and broadband gaps, digital illiteracy, cybersecurity vulnerability, and	Universal broadband, device access, digital

Subcommunity	Short Description	Unique Needs/Challenges	Shared Needs
	seniors, low-income individuals, and some disabled groups	exclusion from services and civic life	skills training, and inclusive tech design
Indigenous & Sovereign Nations			
Indigenous & Tribal Members	Native peoples with deep cultural, ecological, and historical ties to land; sovereign nations resisting centuries of colonization and erasure	Land and water rights, self-governance, health and educational equity, and cultural revitalization	Respect for tribal sovereignty, rematriation, and investment in indigenous-led solutions
Refugee & Migration Status			
Refugees & Asylees	Individuals fleeing war, persecution, or disaster, often rebuilding lives in unfamiliar cultural and bureaucratic systems	Trauma recovery, housing and employment access, legal navigation, and language services	Welcoming environments, community bridging, and trauma-informed integration policies

Where We Are Now: A Moment of Risk and Possibility

- Accelerating technological change meets backsliding policy environment
- Communities navigating both renewed threats and emergent tools
- Mutual aid, cultural resilience, and tech innovation rising from the grassroots

What Role Can AI Play?

Opportunities and Understanding: AI provides opportunities to expand access to healthcare; legal and civic services; education and job training; and culturally competent support tools It can improve mental health and PTSD-informed solutions, as well as create and deploy tools for advocacy, understanding, mobilization, and cultural revitalization. AI can even answer the question, "Who should I call?"

Recognition of Shared and Unique Needs and Tailoring Engagement Accordingly: Many communities share similar needs: the desire for health, safety, clean air and water, and a comfortable place to live. While perhaps 80% of community needs are the same or similar regardless of the specific characteristics of a community. By identifying cross-community insights, it is possible to also recognize common systemic barriers shared by many groups (e.g., digital access, economic disenfranchisement, language exclusion) and use these insights to craft systemic AI-supported solutions that draw on community strengths and work across geographics and demographic boundaries.

There are, of course, unique community requirements derived from culture, language, history, or a shared traumatic experience. While only 20% of community needs may be unique, recognizing and responding to these differences is critical. As they say, the devil is in the details. Unique community characteristics require tailored engagement styles, sensitivity to distinguishing customs, characteristics, or historic and lived experiences (e.g., linguistic preservation, trauma-specific services, or tribal governance). Addressing their challenges and crafting solutions that meet their needs effectively — with trust and respect — requires extra care and close collaboration, such as co-designed AI tools that reflect local culture, knowledge systems, and modes of interaction.

Poverty, Place, and the AI Divide

Across the U.S. and the globe, communities in poverty face systemic harms that are often misinterpreted, unmeasured, or entirely invisible to traditional institutions. From exposure to environmental toxins to underfunded schools and insecure housing, the effects are cumulative and intergenerational.

AI, if designed ethically, has the potential to surface these hidden patterns. Through tools such as data fusion, causal analysis, and predictive modeling, AI can help us move from reactive problem-solving to proactive, systemic change. However, these benefits cannot be realized if the same systems are trained on biased data or governed without community input.

The digital divide — especially in rural areas such as Appalachia or tribal nations — means many people do not have access to high-speed Internet, updated devices, or the digital literacy required to use even basic online services. This isn't just a technology gap; it's a structural equity issue that demands targeted policy, investment, design, and education strategies.

One promising model is a community-aware, AI-optimized database analysis tool. Such a system would help trace structural harms from their origin (roots) to their everyday impact (branches), allowing for both big-picture insight and local relevance. It could:

- Integrate health, education, housing, and justice data,
- Enable scenario modeling to explore potential interventions, and
- Center lived experience as part of both input and outcome analysis.

The Role of AI in Addressing Social Determinants of Health (SDOH)

One powerful area where AI is already making strides is in healthcare, particularly in addressing social determinants of health (SDOH): the non-medical factors that affect well-being, such as access to healthy food, clean water, safe housing, transportation, education, and job training. Broadband access itself is now recognized as an SDOH. In disadvantaged communities, such as in Appalachia, for example, AI-powered tools are being used to:

- Map opioid overdose patterns and optimize harm-reduction strategies,
- Predict which communities will suffer most from climate-related displacement, and
- Improve access to mental health screenings through AI-enhanced telehealth.

Globally, similar tools are emerging to serve remote communities across India and sub-Saharan Africa. These systems bring diagnostics to communities that are without doctors, using solar-powered mobile health units and multilingual interfaces.

Yet here too, risks abound. Biases embedded in historical data can misrepresent needs or assign blame. Without careful governance and local engagement, even well-intentioned AI tools can cause harm.

Decentralization and Community Ownership

Central to any equitable AI approach is decentralization. Data and decision-making authority must reside not only in institutions but in communities themselves. A decentralized, community-aware AI platform would:

- Allow neighborhoods to query trends by geography, identity, and issue;
- Surface both structural causes and lived impacts of inequity; and
- Support local coalitions in designing and testing their own solutions.

Such a tool would draw on AI's strengths — pattern recognition, longitudinal analysis, and narrative generation — while prioritizing consent, transparency, and cultural relevance.

For instance:

- A local government overlays historical zoning data with broadband access and youth dropout rates to identify generational technology gaps.
- An advocacy network maps early signs of eviction pressure and its link to environmental risk zones and mental health service deserts.
- A coalition of educators compares school discipline rates with local transportation access, revealing racialized barriers tied to attendance and mobility.
- A tribal council uses AI to preserve endangered languages and monitor land usage.

The Ethical Imperative: Bias, Surveillance, and Trust

Bias in AI is not accidental; it is the result of choices made in training data, model design, and implementation. Too often, these systems:

- Misinterpret non-standard language or dialects (e.g., AAVE, Spanglish),
- Prioritize majority cultural norms, and
- Over-police or mislabel marginalized populations.

These outcomes are not just technical failures; they are social ones. They erode trust, perpetuate inequality, and concentrate power.

Ethical AI requires:

- Transparent governance,
- Community participation in design,
- Public-sector and nonprofit innovation ecosystems, and
- Robust protections for privacy, data sovereignty, and algorithmic accountability.

Cultural Relevance and Community Engagement

To build systems that are truly inclusive, AI must reflect the diverse ways people think, communicate, and solve problems. This means:

- Respecting religious and spiritual values in algorithmic filtering,
- Designing for accessibility from the ground up (not as an add-on),
- Creating multilingual, low-literacy interfaces, and
- Co-creating with communities rather than imposing "solutions."

Trust in technology grows when people feel seen, heard, and respected in the design and implementation process. AI can amplify marginalized voices, but only if those voices are central from the start.

Strategies for Equity-Centered AI

To address the risks and unlock AI's potential for justice, we recommend:

Infrastructure Investment: Expand broadband, electricity, and device access in underserved areas.

Upskilling and AI Literacy: Integrate AI education into digital literacy programs, focusing on low-literacy and marginalized users.

Ethical Governance: Establish inclusive policies that ensure transparency, auditability, and fairness.

Public-Private Partnerships: Leverage collaborations to make AI tools affordable and relevant.

Community-Led Innovation: Fund and scale community-generated tech that addresses local needs.

Community-Informed Risk Mitigation

AI systems must be evaluated against lived experience and subject to correction. Here are some risks along with recommended mitigation strategies.

Risk	Mitigation
Reproduction of bias	Community audits, redress inputs, and transparent design layers
Extractive surveillance	Consent-based protocols, decentralized data ownership
Disconnection from lived reality	Narrative overlays and qualitative data weighting mechanisms
Over-centralization	Open infrastructure design with regional override and participatory logic

Roadmap for Action

- Convene cross-sector leaders and community representatives to co-design tool priorities.
- Develop a prototype that blends historical, quantitative, and qualitative data.
- Pilot in a small regional setting and include community validation, and ethical oversight.
- Build multilingual, multichannel accessible interfaces with non-institutional users, who may have limited technology fluency, in mind.
- Develop guidance for shared data governance and decentralized deployment.

Conclusion: Toward an Intersectional, Inclusive AI Future

The path to an equitable AI future lies in combining technical excellence with deep community engagement. It requires humility from developers, courage from policymakers, and creativity from everyone.

AI can be a tool for systems change, but only if built along with those who have been most harmed by systems in the past. It can illuminate connections, forecast risks, and guide resources, but only if grounded in justice, shaped by culture, and owned by communities.

Let us ensure that the intelligence we build reflects the intelligence already alive in the people we serve.

See also Appendix C: “Case Study: AI Framework for Solution-Focused Community Problem Solving”

References

Jo, G.A. (2024), Equity, AI, and Community, Forbes

Noble, S. U. (2018), Algorithms of Oppression, NYU Press

Benjamin, R. (2019), Race After Technology, Polity

UNESCO (2024), AI Literacy and the New Digital Divide

Future Skills Centre (2025), An Equity Lens on Artificial Intelligence

World Economic Forum (2024), AI Literacy in Education

IMF (2025), AI Adoption and Inequality, Working Paper Series

Barton, J. (Forthcoming), AI and Social Determinants of Health Framework. Internal White Paper

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.

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.

Svetlana Stotskaya, Global Executive Consultant, Mentor
Svetlana Stotskaya is an award-winning global executive consultant and mentor. Svetlana is an active mentor for entrepreneurs at Techstars, Founder Institute, and Startup Wise Guys: the largest B2B accelerator in Europe. Featured in the World IP Changemakers' Gallery by the World Intellectual Property Organisation, she served on a jury board for international award competitions in innovation and technology.

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.

Chapter 11: Overreliance on AI

Author: John Barton

Overview

Overreliance on AI is no longer a speculative risk; it is an emergent design failure unfolding at scale. As generative AI tools become more persuasive, ubiquitous, and intuitive, users are increasingly treating outputs not as suggestions but as truths. This shift isn’t just behavioral. It reveals a foundational mismatch between how AI is designed, how it is deployed, and how humans build trust.

The Microsoft Aether Committee defines overreliance as "a behavioral state in which users defer judgment to an AI system even when they have reason, skill, or evidence to question it." Their 2023 report identifies causes ranging from poor onboarding and automation bias to low AI literacy and overconfident UX design. Across nearly 60 studies in HCI, organizational behavior, and cognitive psychology, the evidence is clear: overreliance is not rare, and it is not benign.

Two Views of Trust

The most critical distinction between this framework and the Microsoft Aether report lies in how each treats trust.

Aspect	Aether Paper	This framework
Definition of Trust	A cognitive or psychological state: often passive or assumed	A behavioral practice: dynamic, scaffolded, and situational
Trust Failure Framing	Overreliance = a result of psychological bias (e.g., automation bias)	Overreliance = a design failure that disables user agency
Mitigation Approach	Emphasizes transparency, explainability, interface labeling	Emphasizes recovery, reflection, and epistemic scaffolding
User Role	At-risk subject prone to bias or error	Active participant whose trust can be shaped, reclaimed, and redirected
System Role	Provide signals (confidence scores, disclaimers)	Shape behavior through growth-mode UX and calibrated friction

Core Trust Philosophy	Manage trust	Calibrate, support, and recover trust
Primary Risk Identified	Users trusting too much	Systems teaching users not to think

Where the Aether report treats trust as a cognitive error to be managed, this Framework reframes trust as a behavioral outcome of system design. It is not just what users believe; it’s what systems teach. And that makes it actionable.

What begins as user convenience quickly hardens into epistemic dependency. Users skip critical thinking steps. They stop verifying sources. They trust AI output even when it contradicts their own knowledge. This pattern shows up across domains; students use AI to draft papers without synthesis, professionals paste in summaries without review, and even high-stakes decisions (legal, medical, financial) are increasingly shaped by AI inputs that are treated as inherently correct.

Conventional risk mitigation — such as adding disclaimers or improving model accuracy — is inadequate. Users don't just misjudge factual correctness. They adopt structural habits that normalize outsourcing judgment. Without deliberate design for reflection, recovery, and agency, overreliance becomes entrenched.

This Framework offers a different approach. It reframes overreliance not as user failure but as a predictable outcome of current design patterns. By analyzing trust behaviors, behavioral defaults, and onboarding gaps, it introduces a quadrant-based model for understanding and redirecting user interaction. The model maps user mindsets (fixed or growth) against the systems they interact with (stagnant or innovative), revealing four distinct risk profiles and paths to recovery. Rather than attempting to "fix trust," the Framework centers **epistemic calibration**: the ability of users to engage with AI critically, adaptively, and reflectively.

In this Framework, overreliance is not just an error state. It is a signal: a warning that system scaffolding has failed to support user agency. And as AI tools accelerate in complexity and reach, the cost of ignoring that signal grows exponentially.

This document begins the work of designing for recovery, not just control. It offers language, structure, and intervention concepts that can be tested, refined, and embedded across AI development lifecycles, from onboarding to interface design to long-term trust calibration.

Stakeholders

The risk of overreliance on AI systems is not distributed equally. Different stakeholder groups encounter, reinforce, and are impacted by this risk in distinct ways. Understanding these roles is essential to designing effective interventions and allocating responsibility.

New AI Users (Students, Workers, Public Users)

These are individuals who interact with AI tools without deep technical understanding or prior exposure to epistemic safeguards. In educational and workplace settings, new users are particularly vulnerable to overreliance.

- Students often treat AI as a substitute for research or synthesis.

- Employees may defer to AI-suggested summaries, assuming correctness.
- Public users encounter persuasive AI outputs through chatbots, search engines, and productivity tools without visibility into system limitations.

Their default trust behaviors are shaped by onboarding quality, interface signals, and institutional norms. Without friction or calibration prompts, many new users develop passive reliance patterns that become difficult to reverse.

UX Designers and AI Product Teams

These teams play a central role in shaping user trust behaviors. From interface affordances to timing of suggestions, design decisions either reinforce or interrupt overreliance. Teams may unintentionally reward speed and frictionless interaction at the cost of critical engagement.

- Autocomplete and summarization tools can flatten nuance.
- Invisible errors or missing citations can mask epistemic risk.
- Systems rarely prompt reflection or critique after use.

User experience (UX) and product teams need access to trust metrics beyond engagement or completion rate. Without epistemic key performance indicators (KPIs), product success may coincide with user disempowerment.

Educators and AI Literacy Professionals

In both formal and informal learning environments, educators have a dual challenge: using AI tools to support learning while preventing them from replacing learning. When students internalize AI as a shortcut, educational systems risk reinforcing stagnation.

AI literacy professionals are beginning to surface strategies for teaching calibration, synthesis, and disagreement. However, they often lack access to tool internals or control over interface dynamics, which makes structural support for epistemic skill-building inconsistent and fragmented.

Policy and Trust/Safety Teams

These actors define the regulatory and ethical boundaries of AI deployment. While much of their work focuses on preventing harms like bias, surveillance, or misinformation, overreliance introduces a subtler but equally corrosive risk: the erosion of user judgment.

Trust and safety teams must evolve their scope to include behavioral defaults, recovery scaffolds, and misuse patterns that emerge from high-compliance but low-agency interactions.

Enterprise Deployment Leaders

In large organizations adopting AI tools across departments, the risk of overreliance is compounded by scale. Teams are encouraged to use AI for efficiency, but may lack guardrails for:

- Decision accountability,
- Epistemic quality control, or
- Feedback integration.

Over time, unexamined overreliance calcifies into cultural dependency, making it harder to restore initiative, judgment, or accountability. When it takes root in enterprise workflows, overreliance embeds passivity into processes that once relied on human judgment.

Investors and Strategic Funders

Investors — including those focused on responsible tech, venture capital, and social impact — have a vested interest in scalable, trustworthy AI systems. Overreliance poses both reputational and operational risks; it can erode user confidence, increase liability exposure, and lead to costly missteps or regulatory pushback.

By positioning this Framework as a model for designing *resilient trust* rather than frictionless compliance, we offer a value proposition aligned with long-term retention, product adaptability, and ethical market leadership. Investors increasingly recognize that trust infrastructure is not ancillary; it is core to AI product viability.

Foundations and Philanthropic AI Funders

Philanthropic organizations focused on digital equity, community resilience, and ethical AI education are emerging as key funders of harm-reduction strategies. These funders support public-interest work to reduce epistemic harms, especially in underserved populations.

This Framework aligns with their goals by offering a pathway to scalable, recovery-enabled systems that don't just avoid bias, but actively teach reflective, equitable AI use.

AI Developers and Foundation Model Teams

These upstream stakeholders shape the behavior, affordances, and epistemic posture of the models themselves. Their architectural decisions — ranging from pretraining data and reinforcement mechanisms to confidence signaling and answer calibration — directly affect downstream trust dynamics.

Without considering overreliance, core model teams may optimize for helpfulness while inadvertently encouraging overconfidence. Their role in supporting recovery lies in enabling systems that can pause, reflect, and revise: not just respond.

Policy and Governance Professionals

These include regulators, lawmakers, and standards organizations (e.g., NIST, ISO, EU AI Act) that set the external constraints for trustworthy AI. While much attention has been given to bias and data transparency, overreliance introduces a need for behavioral accountability; are systems producing not just safe outputs, but safe usage patterns?

Regulatory frameworks must expand to address trust calibration, scaffolding, and user resilience, not just data harm or content filtering.

Internal Trust & Safety and Ethics Teams

Within organizations, these teams are responsible for monitoring harm, abuse patterns, and reputational risk. Overreliance often escapes their purview because it looks like success: high engagement, satisfied users, few complaints.

However, uncritical use of AI can mask deep epistemic erosion. These teams must evolve to include metrics of user reflection, adaptive confidence, and behavioral feedback, not just incident reporting or legal risk.

Procurement and Risk Officers (Enterprise Subgroup)

In enterprise settings, the people selecting, and approving AI systems are often separate from those who use them. Procurement officers and risk managers play a hidden but powerful role in either embedding or mitigating overreliance.

Their assessment criteria can shape entire organizational adoption patterns. By integrating epistemic resilience, recovery scaffolds, and reflective tooling into vendor evaluation, they can drive demand for responsible AI at scale.

Each of these stakeholders holds a piece of the puzzle. Overreliance is not a problem of user ignorance alone. It is the result of structural gaps in design, deployment, governance, and education. Effective mitigation requires coordinated responses across these roles, with shared responsibility for building systems that support reflection, not just use.

Challenges and Gaps

Efforts to mitigate overreliance on AI have largely fallen short because they underestimate the complexity of the problem. The dominant response has been technical; add disclaimers, improve accuracy, or publish confidence scores. But these approaches miss the deeper mechanisms that drive behavioral dependency, stagnation, and loss of judgment.

Fluency Bias and the Loss of Friction

Modern AI systems are optimized for speed, fluency, and seamless UX. While these qualities enhance usability, they also reduce opportunities for reflection. When users are rewarded for accepting answers quickly — and penalized, in effect, for slowing down — they develop patterns of passive trust.

Features such as predictive text, auto-generated responses, and instant summarization encourage fluency over scrutiny. The design culture that celebrates frictionless interaction inadvertently discourages epistemic resistance. Without embedded challenges or critical pauses, users learn to trust by default: not because they are careless but because the system teaches them to.

Inadequate Onboarding Structures

Most AI tools are introduced with basic usage instructions and legal disclaimers. Very few offer structured onboarding that:

- Shows both successful and failed outputs,
- Teaches users how to critique or disagree with the AI, or
- Calibrates expectations about system strengths and weaknesses.

Without exposure to AI limitations early on, users build a false sense of reliability. Once patterns of overreliance are formed, they are difficult to reverse.

Absence of Trust Scaffolds

Many AI deployments assume that users will self-regulate their trust. In reality, trust calibration is rarely intuitive. Without scaffolds — such as real-time feedback, strength-of-evidence indicators, or modeled disagreement — users tend to either over-trust or abandon AI tools altogether.

The result is a fragile equilibrium where AI is either blindly followed or fully discarded, with little space for critical middle ground.

No Recovery Paths Once Overreliance Sets In

Perhaps most critically, current systems lack clear mechanisms to detect and respond to entrenched overreliance. Once users begin deferring judgment habitually, there are few interventions that help them regain epistemic agency.

Systems do not prompt reconsideration. They do not highlight inconsistencies across use. And they rarely offer structured feedback loops that allow users to reflect on past interactions. Without these recovery pathways, overreliance becomes the default state.

Incentive Structures Misaligned with Epistemic Integrity

Product and business teams are often evaluated based on usage metrics: engagement, retention, satisfaction. These goals favor fast, confident outputs that minimize cognitive load and reduce user uncertainty. In this environment, recovery scaffolds and reflective design patterns are deprioritized… not because teams oppose them, but because they slow momentum.

Without redefining success to include epistemic resilience, organizations will continue to reward fluency at the cost of reflection. Overreliance, under these incentives, becomes invisible success.

Acknowledging Microsoft's Aether Report

The Microsoft Aether Committee's 2023 report was one of the first to formally define overreliance and review mitigation strategies. It provides a strong foundation by identifying psychological antecedents and UX dynamics. However, the report remains primarily diagnostic. It does not extend into implementation, nor does it offer a coherent recovery model.

The Framework builds on Aether's insights by proposing a quadrant-based behavioral model and concrete design interventions. It seeks to move from analysis to action—providing a scaffold for organizations seeking to test and adapt epistemic trust systems in real environments.

Overreliance is not a symptom of user error. It is the predictable outcome of design priorities, onboarding failures, and governance blind spots. Until these structural issues are addressed, no amount of disclaimers or model improvements will prevent users from drifting into epistemic dependency.

Our New Vision

When users begin to trust AI systems reflexively — despite warning signs, contradictions, or their own knowledge — it is not because they are careless. It is because they have been conditioned to trust AI systems. Overreliance is learned, not accidental. Because it is learned, it can be unlearned, provided that systems are built not just to perform, but to support reflection, adjustment, and growth.

This model reframes overreliance not as a failure to trust appropriately, but as a failure of the surrounding design to support critical judgment. The goal is not to reduce trust, but to **recalibrate** it: to move away from **compliance** and toward **collaboration**. That requires tools that deliver answers, provoke inquiry, challenge assumptions, and guide users back to themselves.

This is the work of recovery, and it begins with aligning beliefs and systems to create change.

To understand where recovery begins, we need to see where users are stuck. That's what the quadrant reveals.

Unified Quadrant Model: Innovation, Growth Mindset, and Stagnation

Belief + System = Change

This framework starts with a simple insight; sustainable transformation happens only when people's beliefs and the systems they interact with evolve together. The model frames belief as mindset — whether users are open to growth and feedback — and system as the infrastructure or design conditions that support or inhibit change.

Belief alone is not enough. A person can be curious, reflective, and motivated, but if they operate within a rigid, outdated system, their efforts stall. Likewise, a powerful and innovative system can fail if users are not equipped or encouraged to engage meaningfully with it. Only when both belief and system are aligned does meaningful change emerge.

This idea is visualized as a **2x2 quadrant** using two axes:

Vertical Axis (Y-axis): Mindset, from Fixed at the bottom to Growth at the top

Horizontal Axis (X-axis): System, from Stagnant on the left to Innovative on the right

The matrix defines four possible combinations:

	Stagnant	Innovative
Growth Mindset	Empowered Transformation	Frustrated Growth
Fixed Mindset	Deep Stagnation	Wasted Innovation

Quadrant Descriptions

Empowered Transformation (Growth Mindset + Innovation)

Belief and system are aligned.

- Reflective use, iteration, and agency emerge.
- Overreliance is actively resisted.

Frustrated Growth (Growth Mindset + Stagnation)

- Users want to grow but are blocked by rigid systems.
- Risk of burnout or resignation increases when belief is unsupported.

Wasted Innovation (Fixed Mindset + Innovation)

- Systems have potential but are misused or underutilized.
- Users avoid challenge or reflection, often defaulting to passive use.

Deep Stagnation (Fixed Mindset + Stagnation)

- Both belief and system are stagnant.
- Overreliance is entrenched; change feels impossible.

This quadrant model acts as both a diagnostic and design tool, helping individuals and teams understand not just where they are, but what must shift for change to occur.

Belief + System = Change. One without the other leads to friction, misuse, or stasis. Together, they unlock adaptive, resilient innovation.

Before systems can support recovery, they must first recognize where users are starting from and what keeps them stuck. The quadrant model shows that overreliance does not come from a single cause. It emerges at different intersections of mindset and environment.

Some users want to grow but are trapped in rigid structures. Others are surrounded by innovation but lack the belief they can engage it meaningfully. Some are simply stagnating: unsupported and unchallenged.

In every quadrant, the path forward depends on more than recognition. It depends on the response. Recovery begins when systems do more than assess; they intervene.

Toward Recovery-Enabled Systems

Most AI systems assume trust will either hold or break. Few are designed to repair it. This Framework argues for a third path: **recovery**. That means:

- Letting users see, revisit, and learn from past AI interactions,
- Highlighting inconsistencies or blind trust patterns, and
- Offering prompts that invite re-evaluation without shame.

The quadrant model does not just map where users are. It points toward where they can go next if systems support them.

In reframing trust as dynamic and behavioral, we create the conditions for sustainable AI adoption: conditions that value user growth over compliance, and that treat every overreliance event not as failure, but as an opportunity for recovery and redirection.

Where the Aether report identifies overreliance as a risk, it does not offer a recovery model. This Framework introduces **recovery** as both a **design strategy** and a **behavioral scaffolding**, ensuring that overreliance becomes a moment for growth, not collapse. Recovery here is not passive. It is a purposeful design intervention: a structured opportunity for users to reconnect with their agency, recalibrate trust, and reengage with the system reflectively. In this model, trust is not just protected; it is rebuilt.

This vision does not end with a model. It begins with one. The next step is making it real.

Examples

Understanding overreliance requires seeing it in action: how it emerges in real-world contexts, and how it can be modeled in simulated scenarios. The following examples follow a structured format:

Situation → User Behavior → System Effect → Reflection Opportunity

1. Student Research Submission: Frustrated Growth — Growth Mindset + Stagnant System

Situation: A high school student is assigned a history paper on Reconstruction.

User Behavior: They use ChatGPT to generate an outline and then rely entirely on AI to write the body paragraphs without checking source accuracy.

System Effect: The submission includes outdated or inaccurate claims. The teacher flags factual errors, but the student is surprised; they trusted the output by default.

Reflection Opportunity: With scaffolds in place, the student could have received feedback on unsupported claims or seen citation prompts encouraging verification.

2. Workplace Report Automation: Wasted Innovation — Fixed Mindset + Innovative System

Situation: A project manager at a tech firm uses an LLM-based assistant to draft weekly status updates.

User Behavior: They paste summaries into email reports without reading them carefully.

System Effect: One summary omits a critical delivery delay. This miscommunication causes confusion in the leadership team.

Reflection Opportunity: Had the AI included confidence markers or review checkpoints, the user might have paused and edited before sending.

3. Classroom Ideation Drift: Wasted Innovation — Fixed Mindset + Innovative System)

Situation: A teacher encourages students to use AI tools to brainstorm ideas for creative writing.

User Behavior: Over time, students begin turning in AI-generated first drafts with minimal revision or original thought.

System Effect: Writing quality plateaus and originality declines across the class.

Reflection Opportunity: The tool could prompt students to rework AI suggestions, tag personal edits, or reflect on idea sources.

4. Foundation Model Data Contamination: Deep Stagnation — Fixed Mindset + Stagnant System

Situation: A machine learning engineer fine-tunes a foundation model to auto-label internal datasets.

User Behavior: The team trusts the model's confidence scores without validating outputs across domains.

System Effect: The model introduces bias and inaccuracy into the training pipeline, which propagates in downstream models.

Reflection Opportunity: Implement random audit prompts, data validation scaffolds, and model confidence visualization during active training.

5. Enterprise Tool Adoption with No Safeguards: Deep Stagnation — Fixed Mindset + Stagnant System

Situation: A procurement lead selects an AI assistant based on a polished vendor demo.

User Behavior: The tool is deployed company-wide with no onboarding or sandbox phase.

System Effect: Sales workflows shift subtly but significantly, with AI-generated content introducing bias and factual drift.

Reflection Opportunity: Procurement criteria could require recovery pathways, trial periods, and epistemic harm assessments.

6. AI Use in Under-Resourced Classrooms: Frustrated Growth — Growth Mindset + Stagnant System

Situation: In a rural school district, AI writing tools are positioned as equity boosters for low-literacy students.

User Behavior: Students lean on the tool for language and argument construction without understanding core concepts.

System Effect: AI use reinforces surface fluency but deepens epistemic dependency.

Reflection Opportunity: Tools could scaffold critical comparison, prompt student-led revisions, or pair outputs with discussion cues.

7. Trust & Safety Team Overconfidence: Deep Stagnation — Fixed Mindset + Stagnant System

Situation: An internal moderation team relies on an AI system to auto-flag harmful content.

User Behavior: The team reviews only edge cases, trusting the tool's performance for the rest.

System Effect: Harmful but linguistically ambiguous content goes unflagged, particularly across dialects, or benign but seemingly related content is auto-flagged and removed.

Reflection Opportunity: Recovery design could include flag override patterns, multilingual risk audits, or uncertainty sampling.

8. Customer Support Agent Deferral: Wasted Innovation — Fixed Mindset + Innovative System

Situation: An agent in a call center uses an AI tool for suggested responses during chat sessions.

User Behavior: The agent copies AI replies verbatim, even when the tone or information is mismatched.

System Effect: A customer escalates a complaint due to an insensitive message.

Reflection Opportunity: A sandbox mode or real-time tone analysis could encourage revision before submission.

9. AI-Summarized Email Miscommunication; Wasted Innovation — Fixed Mindset + Innovative System

Situation: A user relies on an AI tool to summarize a long email thread before replying to a client.

User Behavior: They respond based solely on the AI summary.

System Effect: The reply misrepresents prior agreements, damaging the client relationship.

Reflection Opportunity: A preview toggle showing key omissions or contradictions could nudge the user to review the full thread.

10. Fabricated Citations in Research Draft: Frustrated Growth — Growth Mindset + Stagnant System

Situation: A graduate student uses AI to help format citations for a research paper.

User Behavior: They copy several references without checking source validity.

System Effect: Multiple citations are hallucinated: nonexistent articles with plausible formatting.

Reflection Opportunity: Source traceability tools or citation verification prompts could prevent silent propagation of false data.

These examples are not just cautionary tales; they highlight where design, onboarding, and behavioral scaffolding could have made the difference. Each shows a moment of deferral that could have become a moment of reflection. The Framework's design philosophy aims to turn those moments into default practice.

Benefits

Designing for recovery, reflection, and adaptive trust doesn't just mitigate risk; it creates durable, human-centered value. Each benefit maps to a form of recovery within the quadrant model: supporting movement from passive acceptance toward empowered, adaptive engagement. The framework's approach to addressing overreliance offers benefits across behavioral, technical, educational, and systemic levels.

Builds Resilience, Not Compliance

When systems train users to engage critically, not just accept passively, trust evolves. Instead of seeking frictionless interactions, users learn when to slow down, when to question, and when to proceed. Designing for recovery makes trust adaptive, not automatic.

Summary:

- Trust becomes a dynamic practice, not a default state.
- Users learn to distinguish between helpful support and misplaced confidence (trust vs. distrust) to a dynamic practice grounded in critical engagement.

Strengthens Epistemic Agency

By emphasizing scaffolds such as feedback visibility, evidence prompts, and interaction review, users retain ownership of their judgment process. This protects against both over trust and disengagement.

Summary: Epistemic agency — the users' ability to actively shape what and how they come to know — helps users:

- Identify when AI is helpful and when it's not,
- Recognize the boundaries of AI knowledge, and
- Maintain curiosity and skepticism in tandem.

Improves Retention and Understanding

AI systems that slow users down at key points — through retrieval cues, justifications, or challenge prompts — enhance memory and comprehension. This effect is especially powerful in learning environments but extends to high stakes work settings as well.

Summary: Prompting users to take a moment can provide important benefits.

- Enhances learning outcomes through reflective interaction
- Improves clarity, accountability, and institutional knowledge quality

Enables System Transparency and Role Clarity

When systems clearly communicate what they do — and don't do — users calibrate their expectations. Reflective user interface (UI) design, visible uncertainty, and human-AI task boundaries all help avoid overreliance and clarify responsibility.

Summary: Transparent design builds user alignment, supports accountability, and strengthens governance.

- User alignment with system limitations
- Better decision accountability
- Stronger governance and auditability

Supports Growth-Aligned UX Metrics

Traditional metrics such as engagement and satisfaction reward seamlessness. Recovery-focused design invites a shift toward measuring growth in user discernment, confidence calibration, and adaptive decision-making.

Summary: Focusing on recovery and reflection can lead to alternative metrics for evaluation.

- Shifts focus from engagement to discernment and strategic interaction
- Enables long-term value beyond usage metrics

Encourages Ethical Deployment at Scale

The more AI is embedded in infrastructure, education, and decision systems, the more urgent it becomes to cultivate healthy user behavior. This framework supports alignment between ethical principles and product realities by embedding recovery into the user experience.

Summary: Ethical factors require attention in any AI system.

- Reduce harm from misapplied AI outputs
- Mitigate hallucination impacts
- Create equity across skill levels by scaffolding new users

Reduces Systemic Cost and Risk

Small epistemic failures compound, leading to misinformation, reputational harm, or downstream product misuse. By designing for friction, reflection, and recovery, systems reduce the need for escalation, support intervention, and public trust repair.

Summary: Consideration of trust repair through reflection can reduce risks of compounding problems.

- Prevents cascading epistemic failures
- Reduces incident, support, and recovery costs

Ultimately, the benefits of this framework go beyond technical optimization. They demonstrate a new design philosophy: one that embeds reflection, recovery, and user growth into the core of AI interaction.

Across all examples, a set of shared advantages emerges:

Dynamic Trust: Shifting from blind trust or blanket skepticism to informed, adaptive engagement

User Growth: Supporting discernment, memory, and judgment as skills, not liabilities

System Accountability: Making invisible processes visible, and aligning system signals with user expectations

Design ROI: Reducing downstream costs, increasing alignment, and unlocking long-term user value

Governance Readiness: Building trust infrastructure that scales responsibly across institutions, use cases, and regulatory environments

This is not about making users more responsible for bad systems; it's about making systems responsible to the people who rely on them.

Risks

While overreliance may appear as a usability quirk or isolated judgment error, its deeper risks are systemic, behavioral, and compounding. Without intervention, overreliance undermines the very promise of AI: to augment human capacity. Below are the core risks that this Framework seeks to address.

Stagnation of Critical Thinking: Deep Stagnation — Fixed Mindset + Stagnant System

Repeated use of AI without reflection leads to habitual deferral. Users begin skipping the mental steps of comparison, synthesis, and evaluation. What begins as time-saving becomes thought-avoidance.

Summary: Once this stagnation sets in:

- Learning halts or narrows.
- Epistemic agility declines.
- Users lose confidence in their own reasoning.

Collapse of Calibrated Trust: Frustrated Growth — Growth Mindset + Stagnant System

Systems that offer high-confidence outputs without uncertainty cues invite a brittle form of trust. When users eventually discover errors or hallucinations, their trust may snap entirely, leading either to disengagement or uncritical compliance.

Summary: Neither response is healthy:

- Disengagement prevents users from benefiting from AI at all, or
- Blind trust prevents challenge, correction, or oversight.

Behavioral Lock-in: Wasted Innovation — Fixed Mindset + Innovative System

Overreliance can form through repetition and design cues. Once a pattern of deference is rewarded (e.g., fast answers, no need to verify), it becomes harder to unlearn.

Summary: This risk is especially acute in:

- Education, where habits shape future cognition,
- Enterprise, where process shortcuts become norms, and
- Public tools, where millions of interactions scale poor epistemic hygiene.

Normalization of Hallucinated Content: Frustrated Growth — Growth Mindset + Stagnant System)

Users who do not learn to recognize hallucinations may begin to treat all fluent output as valid. This leads to propagation of false claims, fabricated citations, and invisible misinformation loops.

Summary: The consequences include:

- Academic integrity erosion,
- Research contamination, and
- Misinformed civic or financial decisions.

Failure of Accountability Structures: Deep Stagnation — Fixed Mindset + Stagnant System

When systems are designed without reflection checkpoints or feedback loops, responsibility becomes diffused. If no one sees the error, no one owns the correction. Without clear boundaries, mistakes slip through silently—or worse, become institutionalized.

Summary: This blurs:

- User accountability,
- Developer responsibility, and
- Governance oversight.

Equity Risks for Novice Users: Frustrated Growth — Growth Mindset + Stagnant System

Novices or those with lower AI literacy are most at risk for overreliance. If systems do not scaffold epistemic agency from the start, early interactions can reinforce dependency.

Summary: This compounds existing disparities.

- Higher-trust groups may become epistemically overconfident.
- Lower-trust or less-experienced users may internalize AI as a final authority.

Misaligned Success Metrics: Wasted Innovation — Fixed Mindset + Innovative System

When AI systems are optimized for surface-level metrics such as usage, fluency, or satisfaction, epistemic depth is deprioritized. Reflection and calibration slow down engagement, and in many cases, are penalized by design.

Summary: This leads to:

- Rewarding speed over discernment,
- Scaling brittle trust models, and
- Undermining long-term integrity in high-stakes settings.

These risks are not theoretical. They are embedded in current usage patterns, product incentives, and design defaults. What's missing is not awareness, but structural response. The cost of inaction is the silent erosion of judgment: a future where people remember how to use AI but forget how to think.

Conclusion

Overreliance is not a user flaw. It is a systemic failure of design, deployment, and trust calibration. The current ecosystem rewards speed, fluency, and frictionless use, but in doing so, it teaches users to defer judgment and unlearn critical reflection. Left unchecked, this creates patterns of dependency that degrade decision-making, compromise accuracy, and erode user agency.

The Framework presented in this chapter proposes a different future.

Instead of asking whether users trust AI, we must ask how trust is earned, sustained, and recalibrated. Trust is not a static variable; it is a behavioral process shaped by cues, feedback, and system design. This Framework offers a path forward: not disclaimers or passive risk disclosures, but active scaffolds for reflection, disagreement, and recovery. The quadrant model introduced here maps not error states, but ecosystem conditions. It reveals how users drift into overreliance, where design can intervene, and how systems can support return to judgment.

From education to enterprise, this model is actionable. Its interventions — from onboarding prompts to interaction scaffolds — are testable and adaptable. Its value lies not just in user satisfaction, but in epistemic recovery and retained judgment across settings. Systems built on this philosophy don't just support use; they support growth.

We invite the next phase: pilot programs, design partnerships, AI literacy integration, and tool development aligned with this Framework. Investors, developers, educators, and governance teams all have a role to play. Trust is not static. It is learned, modeled, and rebuilt… and systems that enable that rebuilding that trust are the ones that will last.

If AI is to enhance human capability, then it must also protect the conditions for human reasoning. That begins with system responsibility: not just to perform, but to sustain the user's ability to discern, decide, and recover.

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.

Part IV
Sector Spotlights

Chapter 12:
Agentic AI

Authors: Sarah Ennis, Taylor Black, Micah Boster, Ann M. Marcus

Introduction

Consider a scenario that plays out thousands of times daily in customer service centers: A customer calls about a billing discrepancy. They received a charge for a service they believe they canceled, but they're not sure when. The issue touches multiple systems including billing records, service activation logs, customer communications, and cancellation requests. A human agent must navigate between different databases, piece together the timeline, identify the root cause, apply appropriate credits or adjustments, update the customer's record, and send follow-up documentation.

Traditional AI can excel at individual components of this workflow such as analyzing billing data, generating explanations, or drafting customer communications. But it cannot autonomously orchestrate the entire resolution process. Each step requires a new prompt, a new context, and human oversight to connect the pieces. The customer waits while the agent manually bridges the gaps between AI-assisted tasks.

This gap between task-level AI assistance and end-to-end problem resolution represents one of the most significant limitations of current AI deployments. Organizations have invested heavily in AI tools that can summarize documents, generate content, or answer questions, yet find themselves still constrained by fundamentally human-driven workflows for complex, multi-step challenges.

Agentic AI represents a fundamental shift toward autonomous digital workers capable of independently managing these complex workflows from initiation to completion. Agentic systems interpret high-level goals, plan multi-step strategies, coordinate across tools and systems, and adapt in real time while maintaining appropriate oversight and control.

This transformation has profound business implications. Leading research organizations identify agentic AI as a top technology trend for 2025, while the market has grown from virtually nothing to $5.2 billion in 2024, with projections reaching $47 billion by 2030. Early adopters report $3-10 returns for every dollar invested, but more importantly, they're achieving operational capabilities that were previously impossible to automate.

Defining Agentic Behavior

Agentic AI systems exhibit four core characteristics that distinguish them from conventional AI applications, each representing a significant leap in autonomous capability.

Goal-oriented autonomy enables these systems to interpret high-level business objectives and independently determine the sequence of actions needed to achieve them. Consider the difference between asking traditional AI "What are our top customer complaints this month?" versus asking an agentic system "Improve customer satisfaction ratings." The traditional system provides data, while the agentic system analyzes complaint patterns, identifies root causes, researches solutions, proposes improvements, drafts implementation plans, and can even begin executing approved changes based on a single high-level directive.

Multi-step reasoning allows agentic systems to maintain context and adapt strategies across extended workflows that unfold over hours, days, or weeks. When a financial services company's agentic system detects unusual account activity, it doesn't just flag the transaction. It analyzes the customer's historical patterns, cross-references fraud databases, evaluates risk, determines appropriate responses, initiates security measures, prepares notifications, and schedules follow-ups while continuously monitoring for new signals that might change its assessment.

Dynamic tool integration represents perhaps the most transformative capability. Rather than being limited to pre-configured functions, agentic systems can discover, evaluate, and orchestrate whatever tools they need based on situational requirements. A research agent investigating market trends might seamlessly transition from web searches to database queries to statistical analysis tools to document generation platforms, selecting and combining tools in real-time based on the evolving information needs of its investigation.

Adaptive learning enables agentic systems to modify their behavior based on results and feedback, creating continuous improvement cycles. Unlike traditional AI that follows predetermined patterns, agentic systems evaluate their performance, identify failure points, and adjust their approaches. A content generation agent that notices that certain article types receive higher engagement will gradually shift its strategy, testing new approaches and incorporating successful patterns into its standard operating procedures.

The cumulative effect of these capabilities transforms AI from a sophisticated assistant that requires constant direction into an autonomous worker capable of managing complex business processes independently. This shift enables organizations to automate not just individual tasks, but entire workflows that previously required human judgment and coordination.

We explore real-world examples across public, private, and community domains in Section 5.

Core Architectures

Four foundational technologies have converged to make agentic AI practical for enterprise deployment. Understanding these building blocks helps explain both current capabilities and the technical challenges that remain unsolved. Each represents significant engineering advances, but also areas where "this sounds like automation we already have" skepticism is common and often misplaced.

Retrieval-Augmented Generation

Large language models face two fundamental limitations that constrain their business value: training cutoffs that create knowledge gaps, and complete ignorance of organization-specific information. A model trained on public internet data through 2023 knows nothing about your company's products, processes, customers, or internal knowledge base. More critically, it cannot access real-time information about inventory levels, customer interactions, regulatory changes, or market conditions that drive business decisions.

Retrieval-Augmented Generation (RAG) transforms AI from generic assistants into specialized business intelligence systems by enabling dynamic access to proprietary and current information. Rather than relying solely on training data, RAG systems actively search your databases, documents, customer records, and external sources to find relevant context before generating responses.

The technical implementation involves several sophisticated challenges. Proprietary data exists in diverse formats such as structured databases, unstructured documents, real-time feeds, legacy systems with inconsistent schemas. RAG systems must parse these varied sources, understand semantic relationships across different data types, maintain data lineage for compliance, and ensure security boundaries are respected. Real-time updates add another layer of complexity, as systems must balance freshness with computational efficiency while handling concurrent access to live data sources.

Consider a pharmaceutical company deploying RAG for regulatory compliance. The system must access clinical trial databases, FDA correspondence, internal protocol documents, published research, and regulatory filing histories, while also understanding the temporal relationships and approval dependencies that determine what information is relevant for specific queries. The technical challenge lies not just in searching these sources, but in understanding how different types of evidence combine to support regulatory decisions.

Advanced RAG implementations achieve 90% accuracy in data extraction across various formats while processing millions of documents monthly. These systems don't just retrieve information; they evaluate source authority, identify potential conflicts between sources, synthesize findings across multiple documents, and provide transparent attribution for verification. This enables true domain specialization where AI agents become organizational knowledge experts, converting broad intelligence into precise, context-aware decision-making capabilities.

Model Context Protocol and Tool Integration

One of the biggest barriers to deploying AI agents has been integration complexity. Without standardized protocols, connecting M AI agents to N external tools requires building M×N custom integrations, where each agent needs a separate connection to each tool, database, or system. This creates an exponential scaling problem as organizations add more agents and tools.

The Model Context Protocol (MCP) addresses this by creating a universal communication layer between AI agents and external resources. Based on the proven Language Server Protocol from software development, MCP transforms the complex M×N integration problem into a manageable M+N architecture. Instead of each agent requiring custom connections to every tool, agents connect to MCP servers that provide standardized access to external resources.

This means organizations can deploy standardized MCP servers that any compatible AI agent can utilize, eliminating the need for custom integrations. Industry adoption has been rapid, with Anthropic integrating MCP natively into Claude Desktop, OpenAI announcing MCP support for ChatGPT and their Agents SDK, and major platforms including Google's Gemini and Microsoft's frameworks following suit.

MCP standardization is crucial for enterprise deployment because it enables universal connectivity to any external system through a single protocol, dynamic tool discovery allowing agents to find and use new capabilities without code changes, and standardized security models with consistent permission and consent frameworks across all integrations.

Multi-Agent System Architectures

Rather than building monolithic AI systems that attempt to handle all tasks, multi-agent architectures deploy teams of specialized agents, each optimized and fine-tuned for specific capabilities while coordinating through sophisticated communication protocols.

The key insight is agent specialization, which involves creating AI agents specifically optimized for particular roles through specialized prompting, training data, or configuration. Just as human software teams benefit from having dedicated system architects, developers, QA specialists, and UX designers, AI agent teams can deploy specialists optimized for different aspects of complex workflows.

Enterprise orchestration frameworks exemplify this approach with asynchronous, event-driven architectures that enable natural language coordination between specialized agents. A software development workflow might deploy a system architecture agent optimized for technical planning, a code generation agent fine-tuned for specific programming languages, a quality assurance agent specialized in testing methodologies, and a UI/UX agent focused on user experience principles.

Multi-agent orchestration and management systems represent one of the most active areas of current development in agentic AI. While early implementations show promising results, the coordination mechanisms, communication protocols, and error handling systems are rapidly evolving. Organizations should expect dramatic improvements in maturity, reliability, and ease of deployment over the next 12-18 months as these frameworks advance.

Real-world implementations demonstrate both the potential and current limitations. Organizations have deployed multi-agent systems that reduced software requirements writing from weeks to days by orchestrating specialized agents for user story creation, technical analysis, and test plan documentation. However, performance benchmarks reveal that while specialized agent teams achieve impressive results on domain-specific tasks, general-purpose coordination still faces challenges with complex, multi-step workflows.

The core challenge becomes clear when you consider something as simple as pizza delivery. When the delivery person arrives at your door, a complex interaction unfolds, greeting, confirming the order, processing payment, and parting ways. These interactions flow naturally because humans have evolved sophisticated social protocols over millennia. AI agents, by contrast, are brilliant specialists trapped in digital isolation. They can analyze data, generate code, or write content with remarkable skill, but they have no innate understanding of how to coordinate with each other. A coding agent doesn't know when to hand off work to a testing agent, or how to communicate that it has encountered an error, or what to do when another agent goes offline mid-task. Much of current multi-agent development focuses on solving this fundamental interaction problem by teaching AI agents the basic social skills that allow them to work together rather than simply work in parallel.

But coordination is only part of the challenge. As AI systems grow more autonomous, the bottleneck is no longer capability but oversight. Traditional human-in-the-loop models don’t scale. The next leap forward is agentic AI that governs itself through internal red-teaming. Instead of relying solely on external human evaluators, specialized agents act as internal auditors, rigorously testing outputs for logic errors, hallucinations, and compliance gaps before results move downstream. This creates a dynamic ecosystem of peer review, where agents challenge, refine, and validate each other’s work. Such self-auditing architectures establish checks and balances that enable safe autonomy at scale, reducing reliance on human gatekeepers while increasing robustness, adaptability, and trustworthiness. Of course, evaluator agents are not immune to flaws; they too can misjudge, hallucinate, or become misaligned. That’s why recursive oversight is essential, with higher-order agents or consensus mechanisms monitoring the monitors and creating a layered defense against failure. The goal is not perfection but resilience through distributed accountability.

Planning and Reasoning Frameworks

Traditional AI systems respond to immediate prompts but struggle with complex, multi-step challenges that require strategic thinking. Advanced planning frameworks transform reactive systems into strategic thinkers capable of sophisticated workflow orchestration.

A crucial capability that distinguishes agentic systems is their ability to combine AI-driven reasoning with deterministic operations within the same workflow. An agent might use AI to analyze customer feedback data and identify patterns, then execute precise SQL queries to retrieve specific customer records, perform mathematical calculations on the results, and finally use AI again to generate personalized recommendations. This hybrid approach leverages AI's interpretive capabilities alongside the reliability and precision of traditional computational methods.

Planning frameworks implement sophisticated decision-making processes where agents analyze situations, consider multiple approaches, execute both AI-driven and deterministic actions, process results, and adjust strategies dynamically. The key innovation lies in intelligent workflow orchestration, which means knowing when to use AI for interpretation and creativity versus when to use deterministic processes for precision and reliability.

Performance results demonstrate these frameworks' effectiveness: 92.7% accuracy on programming tasks and 75.9% average scores on complex navigation challenges. The innovation lies in self-reflection mechanisms that enable agents to evaluate their own decision quality and learn from mistakes, creating a continuous improvement cycle essential for autonomous systems.

Hierarchical planning approaches enable agents to operate at multiple abstraction levels simultaneously by applying AI for high-level strategic thinking while executing precise deterministic operations for specific tasks. An agent might use AI reasoning to determine that a customer complaint requires account adjustment, then execute deterministic database updates to implement the change, and finally use AI again to craft an appropriate customer communication.

This combination of AI flexibility with deterministic reliability makes agentic systems far more powerful and trustworthy for enterprise applications, where both creative problem-solving and precise execution are essential for business-critical workflows.

Implementation Approaches

The choice between open-source frameworks and commercial platforms for agentic AI is a fundamental strategic decision, shaping an organization's long-term flexibility, costs, and adaptability to evolving AI capabilities. Skeptics often dismiss this as irrelevant, arguing "AI is AI," but this overlooks the profound technical complexities and strategic implications of deployment. Many organizations fail to grasp what happens when these systems inevitably encounter edge cases, break, or require modification for changing business needs, highlighting the critical importance of selecting the right implementation approach from the outset.

Open-Source Frameworks

LangChain dominates the open-source landscape with over 100,000 GitHub stars and more than one million monthly downloads. The framework's comprehensive ecosystem includes LangGraph for multi-agent orchestration, LangSmith for observability, and extensive integrations across the AI development stack. Major implementations serve tens of millions of users with significantly faster resolution times, demonstrating production-ready capabilities.

CrewAI has emerged as a preferred choice for teams new to agentic AI, emphasizing simplicity and role-playing agent interactions. The framework enables complex multi-agent workflows with minimal code, making it ideal for rapid prototyping and straightforward collaborative systems.

Microsoft's AutoGen targets enterprise environments with battle-tested reliability and sophisticated conversation-based coordination. The framework's asynchronous architecture and advanced error handling make it suitable for production environments where reliability is paramount.

Open-source advantages include complete customization control, transparency in operations, cost efficiency for organizations with technical expertise, and freedom from vendor lock-in. Organizations can modify frameworks to meet specific requirements, understand exactly how their AI systems operate, and avoid dependencies on external providers.

Implementation challenges include steep learning curves, frequent updates requiring ongoing maintenance, limited enterprise support, and the need for significant internal technical expertise. Organizations must invest in dedicated teams to manage, customize, and maintain these frameworks effectively.

Commercial Orchestration Platforms

Commercial agentic AI platforms provide comprehensive orchestration environments that handle the complexity of multi-agent coordination, tool integration, and workflow management through managed services. These platforms focus on business process automation rather than individual model capabilities.

n8n represents a leading workflow automation platform that has evolved to support AI agent orchestration. The platform provides visual workflow builders, extensive integrations with business tools, and sophisticated error handling for complex multi-step processes. Its strength lies in enabling non-technical users to create sophisticated agent workflows while maintaining enterprise-grade reliability and monitoring.

The no-code/low-code automation space has rapidly expanded to include agentic AI capabilities. Platforms like Make (formerly Integromat) and newer entrants like Gumloop provide visual workflow designers specifically optimized for AI agent coordination. These platforms democratize agentic AI by allowing business users to create complex multi-agent workflows without programming expertise, often at significantly lower costs than enterprise solutions. However, they may lack the advanced error handling and enterprise governance features required for mission-critical applications.

Full-service development platforms represent another emerging category. Replit's AI-powered development environment enables rapid prototyping and deployment of agentic applications, while platforms like Loveable focus on end-to-end AI application development with built-in agent orchestration capabilities. These platforms blur the line between development tools and deployment environments, offering integrated solutions for organizations that want to build custom agentic applications without extensive infrastructure investment.

Microsoft's Power Platform, including Power Automate and Copilot Studio, offers deep integration with Microsoft's business ecosystem. The platform excels at connecting AI agents with existing productivity workflows, providing seamless handoffs between human and AI workers within familiar business applications. This makes it particularly valuable for organizations already invested in Microsoft's technology stack.

Specialist application platforms represent another category, exemplified by Salesforce's Agent Force, which focuses specifically on customer relationship management and sales process automation. Rather than providing general-purpose orchestration, these platforms offer deep domain expertise within their specific business functions. Agent Force understands customer journey orchestration, maintains context across multiple touchpoints, and integrates natively with CRM data and sales processes. This approach offers significant advantages for organizations whose agentic AI needs align with the platform's specialization, but limits flexibility for use cases outside that domain.

Enterprise workflow platforms like UiPath and Automation Anywhere have expanded beyond traditional RPA to include AI agent capabilities, offering the advantage of integrating agentic AI with existing automation infrastructure. These platforms excel in environments where AI agents need to work alongside traditional automated processes.

These commercial solutions offer distinct advantages: immediate deployment capabilities, visual workflow designers accessible to business users, managed infrastructure with automatic scaling, enterprise security and compliance features, and professional support with service level agreements. However, organizations must weigh these benefits against vendor lock-in risks, higher long-term costs, and reduced customization flexibility compared to open-source alternatives.

Selection criteria should prioritize alignment with organizational capabilities and integration requirements: no-code platforms like Make and Gumloop for rapid deployment by business users, full-service platforms like Replit and Loveable for custom application development, Microsoft for productivity-focused environments, specialist platforms like Salesforce for domain-specific applications, and enterprise RPA platforms for organizations with existing automation infrastructure.

Commercial advantages include immediate deployment capabilities, professional support and service level agreements, managed infrastructure with automatic updates, enterprise security and compliance features, and reliability guarantees suitable for customer-facing applications.

Considerations include higher long-term costs, potential vendor lock-in, limited customization options, and dependency on external providers for critical business functions.

Hybrid Strategies

Leading organizations increasingly adopt hybrid strategies that leverage both open source and commercial solutions strategically. Development teams use open-source frameworks for research, prototyping, and internal applications while deploying commercial services for customer-facing systems requiring reliability and support.

Multi-vendor approaches combine different commercial services based on specific strengths, with some providers focused on complex reasoning, others on safety-critical applications, and others on multimodal tasks. This strategy mitigates vendor risk while optimizing capabilities for different use cases.

Cost analysis reveals that hybrid approaches typically achieve 20-40% cost savings compared to pure commercial solutions while maintaining enterprise-grade capabilities. Open-source development costs range from $20,000 to $500,000+ depending on complexity, while commercial solutions cost $100-5,000 monthly for standard implementations.

Success factors for hybrid implementations include standardized infrastructure using universal protocols, unified governance frameworks, strong internal technical capabilities, and clear decision criteria for when to use each approach.

Real-World Applications Across Sectors and Risks/Benefits

Agentic AI is already being piloted across public, private, and nonprofit sectors. From emergency evacuations to internal compliance agents, the spectrum of uses is rapidly expanding. The following table highlights where AI agents are beginning to take root and the key functions they perform.

AI Agentic Use Examples Across Different Sectors

Sector	Example Agent Function
Public	Evacuation logistics, permit navigation, civic updates
Private	Compliance audits, internal project agents, client service
Community	Outreach, translation, mutual aid coordination
Commercial	Travel planners, smart shopping, home automation

Public Sector Uses (Government / Infrastructure)

Use Case	Description
Emergency Evacuation Coordination	AI agents manage logistics for evacuating vulnerable populations during disasters, as demonstrated in the senior evacuation model in Appendix A
Digital Permit & Licensing Agents	Agents guide residents through applications for permits (e.g., building, business, event), auto-filling and submitting forms.
Public Transportation	Agents help commuters navigate multi-modal transit systems in real time,

Routing Assistants	suggesting accessible or low-cost options.
Civic Engagement Bots	Agents summarize city council meetings, propose meeting agendas, or alert residents to decisions affecting their neighborhood.
Climate Risk Notification Agents	Personalized early-warning systems that notify individuals of local environmental risks based on location and profile.

Private Sector Uses (Corporate / Institutional)

Use Case	Description
Enterprise Workflow Optimization	AI agents monitor project progress, flag bottlenecks, and suggest next steps or staffing reallocation in real time.
Smart Scheduling Assistants	Agents coordinate across internal calendars and meeting goals to arrange cross-team availability or escalate urgent requests.
Compliance Monitoring Agents	Track evolving regulations and assess company compliance gaps, especially in data privacy, ESG, or workplace safety.
Customer Service Delegation	Empowered agents handle multi-channel customer interactions, escalate only complex cases to human staff.
AI for IT Support	Autonomous agents resolve software configuration issues, patch systems, or auto-escalate based on system anomalies.

Community Benefit / Nonprofit Uses

Use Case	Description
Disaster Response and Recovery Agents	Coordinate supply distribution (water, food, shelter), volunteer deployment, and damage assessment.
Elder Outreach & Wellness Check-ins	Voice-capable agents call isolated seniors regularly, assess their mood or needs, and escalate alerts as necessary.

Neighborhood Improvement Agents	Automate surveys to gather community feedback, propose mini-projects (e.g., tree planting, sidewalk repairs), and track progress.
Language Access for Immigrants	Translation agents assist non-English speakers in accessing healthcare, housing, or legal services.
Civic Literacy Bots	Agents explain ballot measures, voter registration steps, or public program eligibility in plain language.

Commercial / Consumer-Facing Uses

Use Case	Description
Personal Shopping Agents	AI agents curate products based on user needs, search across platforms, compare pricing, and place orders.
Travel Booking & Rescheduling	Agents auto-plan travel (flights, hotels, transport) based on constraints like budget, loyalty points, and accessibility.
Home Energy Optimization	Agents learn usage patterns and adjust HVAC, lighting, and appliances to lower bills and carbon footprint.
Gig Worker Schedulers	Agents manage freelance jobs, match workers with demand, and optimize routes or shifts.

Education / Tutoring Agents	Personalized AI tutors support students in learning at their pace, flag gaps, and adjust teaching methods accordingly.

Potential Risks Associated with Using Agentic AI in Various Domains:

Domain	Key Risks
Public	Bias, accountability gaps, data misuse, cyber threats
Private	Oversight loss, security leaks, job displacement
Community	Consent, equity, miscommunication, loss of trust
Commercial	Privacy erosion, manipulation, financial errors

Mitigation Considerations

To reduce the risks of agentic AI deployment, organizations should implement:

Mitigation Strategy	Purpose
Human-in-the-loop oversight	Maintains accountability and decision control
Ethical review panels or audits	Evaluates fairness, safety, and unintended outcomes
Community co-design	Ensures inclusivity and local relevance
Privacy and consent safeguards	Protects sensitive data and user autonomy
Monitoring and feedback loops	Detects errors, drift, or unintended behaviors early
Audits and adjustment cadence	Enables structured iteration and performance tuning

Getting Started with Agentic AI

Organizations looking to adopt agentic AI should begin with low-risk, internal workflows such as compliance monitoring, project tracking, or IT automation. Start small:

- **Pilot in controlled environments** where outputs can be safely evaluated.

- **Map existing toolchains** to identify integration gaps or friction points.
- **Use evaluator agents** to red-team outputs before broader rollout.
- **Define fail-safes** for critical steps where accuracy or accountability is key.
- **Track performance** and iterate with clear metrics tied to cost, speed, or quality gains.

Starting this way builds confidence, reveals edge cases early, and creates a foundation for scaling agentic systems responsibly.

Appendix I: Agentic AI in Disaster Response

The following extended case study illustrates how Agentic AI can be deployed in a complex, high-stakes, public-sector context: disaster response for vulnerable populations.

Climate change is dramatically increasing the frequency, intensity, and unpredictability of disasters such as wildfires, floods, heat waves, earthquakes, and tsunamis. These pose heightened risks for elderly and disabled populations. These individuals are disproportionately affected by delayed or inaccessible evacuation efforts, yet most municipalities across the U.S. (and globally) remain woefully underprepared to respond effectively.

Agentic AI -- which refers to autonomous, goal-driven software systems – offers a promising solution to bridge the gap between emergency response plans and real-time operational coordination. These intelligent agents can be deployed to ensure timely, adaptive, and inclusive evacuation strategies by performing the following functions:

Key Agentic AI Functions in Evacuation Coordination

Proactive Outreach and Needs Assessment

- AI agents can identify and reach out to registered seniors, disabled individuals, or others on medical alert or community watchlists.
- Using phone, SMS, or voice interfaces, agents can assess evacuation status, transportation needs, medical dependencies, or mobility constraints.

Dynamic Transportation Coordination

- Agents can tap into multi-modal transportation networks such as public buses, commercial ride-shares (such as Uber WAV), paratransit services, non-emergency medical transport, accessible taxis, and vetted volunteer drivers.
- AI agents dynamically match evacuees with appropriate vehicle types, prioritizing mobility needs, proximity, and urgency.

Multi-Agency Communication and Dispatch

- AI agents can serve as intermediaries between emergency command centers, transportation providers, shelters, and health services, ensuring unified situational awareness.
- AI agents are capable of real-time updates, rerouting, and reassignment as hazards evolve (e.g., wildfire direction changes or road closures).

Support for Caregivers and Families

- AI agents can notify designated caregivers or family members of the individual's status and whereabouts during transit.
- AI agents can also act as virtual assistants for self-advocating seniors, enabling voice-based check-ins or confirmations.

The San Leandro Senior Evacuation Project by WeAccel

WeAccel is actively developing a proof-of-concept senior evacuation model in San Leandro, California, integrating Agentic AI to:

- Establish a senior registry and risk map that includes mobility status, medical equipment needs, language preferences, and household situation,
- Coordinate with municipal emergency planners, transportation operators, and senior service organizations, and
- Pilot an AI-driven outreach and routing system that can operate with limited broadband or SMS-only infrastructure, which is crucial for underserved or tech-limited seniors.

This project aims to prototype a replicable framework for other cities and contribute to a resilience network that centers the most vulnerable in disaster planning.

Why It Matters

Without action, emergency events exacerbated by climate change will continue to result in preventable deaths and suffering among the elderly and disabled, particularly those who live alone, lack Internet access, or have limited ability to speak or understand English.

Agentic AI systems can dramatically reduce coordination delays, optimize resource use, and ensure no one is left behind, especially when these systems are built with community input, equity considerations, and redundancy planning in mind.

Stakeholder Participation & Required Data

To create and coordinate a system of AI agents that supports emergency evacuation for seniors and disabled individuals, it would be necessary to identify a wide range of stakeholders and data sources.

Below is a breakdown of both, organized by functional role and data dependencies.

Key Stakeholders

Public Sector & Emergency Management

- City and County Emergency Services Departments: Responsible for evacuation plans, EOCs (Emergency Operations Centers), alert systems
- Fire, Police, EMS: Need real-time access to evacuation routes and special needs populations
- Public Health Departments: Provide insight into medical vulnerabilities, home care needs, oxygen/electricity dependence
- Transportation Agencies: Coordinate buses, paratransit, and detours during emergencies

Community-Based Organizations (CBOs)

- Senior Centers & Aging Services Providers: Maintain contact lists, care plans, and wellness check routines
- Disability Rights Organizations: Ensure accessibility and advocate for inclusion in planning and execution
- Faith-Based and Mutual Aid Groups: Provide local trust and human support for outreach, ride-alongs, and wellness checks

Public, Private & Commercial Transport Providers

- City Vehicles
- Public Transportation Vehicles (e.g. AC Transit for Alameda County)
- Paratransit Services
- Ride-hailing Companies (e.g., Uber WAV, Lyft Access)
- Medical Transport Providers
- Charter or Shuttle Companies
- Taxi Services
- Volunteers with Registered Vehicles

Technology & Infrastructure Partners

- Telecom Providers: Enable SMS/voice connectivity and geolocation services
- AI Developers / Agentic AI Platforms: Build, train, and deploy AI agents capable of autonomous coordination, outreach, routing, and translation

- Data Integration Vendors: Handle cross-agency data aggregation, privacy, and interoperability
- Mapping & Navigation Tools: Enable real-time routing, congestion detection, and road hazard data

Funders & Oversight Bodies

- Local, State, and Federal Grant Authorities (FEMA, HUD, state emergency or aging offices): Provide funding for technology pilots, infrastructure, and resilience programs, while requiring compliance with emergency management standards.
- AARP and Aging Advocacy Organizations: Offer funding and legitimacy for senior-focused solutions, ensuring alignment with national aging and disability priorities.
- Foundations (e.g., Knight Foundation, Robert Wood Johnson Foundation): Support innovation, community-based pilots, and equity-focused approaches.
- Academic Research Partners: Evaluate system performance, test for bias, and strengthen models with evidence-based methods and community input.

Critical Data Requirements

Individual-Level Data (with consent or emergency-use authorization)

Data Type	Source / Provider	Notes
Name, Age, Address, Contact Info	Senior registries, utility bills, 911 databases	May require aggregation
Mobility Status (e.g., wheelchair)	CBOs, Health Departments	Includes care dependencies
Medical Needs (e.g., oxygen, meds)	Public Health, Home Health Agencies	Privacy-protected
Language Preference	Registries, CBO intakes	Enables multi-language AI
Household Composition	CBO intakes, utility records	Flags additional residents needing support

Agentic AI can transform emergency response for seniors through predictive monitoring, rapid alerting, and personalized care coordination. While these systems excel at connecting seniors with help in medical or facility-based emergencies, development continues towards fully autonomous, AI-driven platforms that directly match seniors seeking evacuation with providers during mass emergencies. Current approaches may employ AI to support and inform human responders who can execute ride arrangements. However, more automated approaches are likely to become more prevalent as deployments scale and reliability improves.

Some other examples of similar solutions include:

Austin, TX, Vulnerable Population Registries

Austin operates a Medically Vulnerable Registry run through Austin Energy and the broader State of Texas Emergency Assistance Registry (STEAR). These registries collect data on medically fragile persons to inform emergency planning, including evacuation, but without real-time AI-enabled coordination or multi-modal transport automation. While these systems improve situational awareness, there’s no evidence yet of integrated AI agents dynamically matching registrants to transport in real time during an event.

Japan Post-Tsunami Robotic & AI Tools

> Japan has a long history deploying rescue robots – including the snake-like Quince and tracked T-52 Enryu – in earthquake and tsunami zones to aid shelter access or debris-clearing operations. More recently, systems such as *Spectee Pro* use AI to analyze social media, weather, and satellite images to enhance situational awareness during disasters, but still focus on information gathering and shelter logistics, not on automated transport coordination. The city of Rikuzentakata, for example, launched automated calls to registered residents to check evacuation status, but again this is contact-based outreach without full AI-driven multimodal transit integration.

Even with these early efforts, the opportunity remains to develop a solution that fully integrates agentic AI, multi-source transportation coordination, and senior-centric mobility and accessibility needs. See WeAccel.io for more information on this project.

Appendix II: Key Federal & Regional Programs Supporting AI in Emergency Response

Federal Programs

Department of Homeland Security (DHS) AI Pilots

Scope: Nationwide

Focus: Safe and secure AI deployment as part of the AI Executive Order

Status: Phase 1 complete; AI Corps hired

Contact: DHS Science & Technology Division

Learn more: https://www.dhs.gov/science-and-technology/

FEMA AI Use Cases

Scope: Nationwide

Program: FEMA contributes to DHS's AI Use Case Inventory

Focus: Emergency management AI exploration

Contact: FEMA HQ via DHS

Access: https://www.dhs.gov/science-and-technology/

Regional and State Programs

Miami-Dade Emergency & Evacuation Assistance Program (EEAP)

- **Location**: Miami-Dade County, FL

Function: Helps residents with medical/special needs evacuate safely

Services: Specialized transportation coordination

Contact: Miami-Dade County Services

Website:
https://www.miamidade.gov/global/service.page

Florida State Emergency Resources

Scope: Statewide

Services: Emergency support and health guidance

Public Hotline: 800-342-3557

Website:
https://www.floridahealth.gov/about/emergency.html

Private Sector Solution

Prepared911 AI Platform

Type: Commercial, end-to-end AI emergency response system

Function: AI assistance integrated across the emergency response lifecycle

Pilot Opportunities: Available for agencies and municipalities

Website: https://www.prepared911.com/

References:

The Inflection Point: Agentic AI in the Evolution of Security and Risk Management – *Crisis24* https://www.crisis24.com/articles/the-inflection-point-agentic-ai-in-the-evolution-of-security-and-risk-management

Agentic AI in Disaster Management and Emergency Response – *DigitalDefynd* https://digitaldefynd.com/IQ/agentic-ai-in-disaster-management-and-emergency-response

Multi-Agent Systems in Disaster Management – *SmythOS* https://smythos.com/ai-agents/multi-agent-systems/multi-agent-systems-in-disaster-management

How AI Boosts Emergency Response Times – *EyewatchLive* https://eyewatchlive.com/news/how-ai-boosts-emergency-response-times

Innovations in Personal Emergency Response Systems for the Elderly – *Ball State Daily News* https://www.ballstatedaily.com/article/2025/02/innovations-in-personal-emergency-response-systems-for-the-elderly

Agentic AI in Home Care – *AutomationEdge* https://automationedge.com/home-health-care-automation/blogs/agentic-ai-in-home-care

Microsoft GraphRAG GitHub Repository https://github.com/microsoft/graphrag

Anthropic Documentation – MCP Overview https://docs.anthropic.com/en/docs/agents-and-tools/mcp

The Pragmatic Engineer Newsletter – MCP Deep Dive https://newsletter.pragmaticengineer.com/p/mcp

Japan's Use of AI and Robotics in Disaster Response https://www.researchgate.net/publication/4181423_Rescue_Robots_and_Systems_in_Japanhttps://spectrum.ieee.org/japan-earthquake-robots-help-search-for-survivors

https://www.preventionweb.net/news/japan-firms-look-ai-bolster-disaster-prevention-and-mitigation

Austin Energy Medically Vulnerable Registry Audit Report (2024) https://www.austintexas.gov/sites/default/files/files/Auditor/Audit_Reports/Austin_Energy_Medically_Vulnerable_Registry_March_2024.pdf

Author (In order of contribution)

Sarah Ennis, Co-Founder and Advisor of AgentsGEO.ai
Sarah Ennis is a Fortune 500 trusted advisor specializing in advanced technology innovation, with over two decades of experience leading groundbreaking AI solutions at scale. Globally recognized for her expertise in artificial intelligence, she designs and implements bespoke emerging technology products across industries. She is also the co-founder and advisor of AgentsGEO.ai, a patent-pending

platform that helps brands monitor and improve their visibility in the AI ecosystem and deploy AI agents, ensuring they are discoverable and recommended by tools like ChatGPT, Gemini, and others through its proprietary GEOScorer™ technology. In addition, Sarah contributes part-time to Northeastern University's Master of Digital Media programs in AI, preparing the next generation of technologists and creative leaders. Her work bridges Silicon Valley innovation with global impact, and she is a distinguished member of the American Society for AI and contributor to the OpenAI Forum.

Taylor Black, Director AI & Venture Ecosystems, Microsoft
Taylor Black is Director of AI & Venture Ecosystems in Microsoft's Office of the CTO, where he designs and leads cross-company initiatives that integrate innovation, product development, and community engagement. With 19+ years of experience launching and scaling ventures across enterprise, deep tech, and social ecosystems, he brings a multidisciplinary background as a developer, educator, lawyer, entrepreneur, and venture builder. He mentors and invests in early-stage startups through networks such as Conduit Venture Labs and Fizzy Ventures. Taylor also helps shape Catholic University of America's new institute at the intersection of AI, innovation, and human flourishing.

Micah Boster, Principal, Nighthawk Advisors
Micah Boster is the founder and Principal at Nighthawk Advisors, where he works with early-stage technology companies on execution, AI strategy, and positioning. Previously, he spent eight years at Google and over a decade as an executive at several NYC-based startups. He holds a BS in Symbolic Systems from Stanford and an MBA from INSEAD.

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.

Chapter 13:
AI in Education — A Perspective on Surveillance, Equity, and Transformative Learning Tools in the United States

Author: John Barton

Updates:

7.29.25: Recent statements by the US Department of Education (July 22, 2025) do not appear to resolve the issues outlined below – but may accelerate them.

Update: 9.4.25: On AI.gov, The White House Task Force on AI in Education calls "for the United States to promote AI literacy and proficiency among America's youth and educators by: promoting the appropriate information of AI into education, providing comprehensive AI training for educators, and fostering early exposure to AI concepts and technology to develop an AI-ready workforce and the next generation of American AI innovators." There is no further information available at this time. "More information about these resources is coming soon.

Overview

If AI defines intelligence, who gets to be smart?

"*Who designs, decides.*" — Wilson Wong, Founding Director and Associate Professor of Data Science and Policy Studies, Chinese University of Hong Kong

Four converging trends have brought artificial intelligence (AI) in education to a tipping point that is marked by both accelerating adoption and growing vulnerability:

- Rapid AI adoption without governance frameworks
- Political momentum for deregulated AI use, particularly in public education
- Widespread school underfunding, driving pressure to automate
- Eroding oversight at every level — from classrooms to state boards – in a system where educational governance varies significantly across states and districts

Together, these conditions create fertile ground for both innovation and inequity. They also raise urgent questions about how the United States is managing this moment and what it might learn from nations that have already implemented coordinated AI education strategies.

Countries such as Singapore — with its Model AI Governance Framework — and Finland — through national AI ethics guidelines and digital equity mandates — have prioritized human-centered design, algorithmic transparency, and long-term civic trust in their educational AI strategies. These frameworks stand in contrast to the fragmented and reactive landscape in the United States, where federal leadership has been limited and state-level responses vary dramatically. The United Nations Educational, Scientific and Cultural Organization's (UNESCO) global guidance on generative AI (GenAI) in education & research reinforces the value of such approaches, offering standards for equity safeguards, participatory governance, and transparency; these are benchmarks that remain largely aspirational across most U.S. education systems. Meanwhile, the Organisation for Economic Co-operation and

Development (OECD) has emphasized the need for cross-sectoral coordination and strong guardrails, further highlighting the absence of a coherent U.S. strategy.

While the April 2025 Executive Order on AI Education 14277 established a national Task Force and signaled support for AI literacy initiatives, it was quickly followed by Executive Order 14179, which dismantled prior safeguards and emphasized deregulation, ideological neutrality mandates, and export-driven development. This dual-track approach leaves vulnerable communities without stable protections, especially in underfunded districts where AI tools are deployed fastest and monitored least. Investigations by the U.S. Government Accountability Office (GAO) have documented how these deployments often lack meaningful consent pathways or data transparency, particularly in districts under financial strain. These gaps disproportionately affect students in marginalized communities, raising serious civil rights and accountability concerns.

This chapter outlines ten strategic areas for review by a proposed multi-sector Task Force. Only the first recommendation — convening this Task Force — is a direct call to action. All other recommendations serve as priority domains for exploration, assessment, and implementation planning by that body.

The ten areas are organized around five core principles:

1. **Democratic Governance**: Ensure that AI is governed by people, not algorithms, consistent with public school board authority and local control in U.S. education
2. **Transparency & Accountability:** Make AI systems visible, testable, and open to correction, aligning with civil rights oversight and U.S. public records norms
3. **Equity & Inclusion**: Safeguard the rights and needs of vulnerable groups most at risk of exclusion or harm, grounded in protections under Title VI, Section 504, and IDEA.
4. **Community Empowerment:** Equip learners, families, and educators with the tools to participate and advocate through participatory processes rooted in U.S. school district engagement models.
5. **Cultural & Cognitive Integrity:** Protect cultural values, community identities, and diverse ways of thinking from being overwritten, particularly for historically marginalized communities (such as poor, women, LGBTQ+, indigenous, people of color, neurodivergent, & Appalachian), ensuring the accuracy and relevance of curriculum development, teaching methods, communication, and education policy & standards.

From convening a cross-sector Task Force to defending civil rights in curriculum design, these recommendations are intended to prevent harm, build trust, and ensure AI serves — not displaces — human judgment and democratic integrity. These principles reflect traditional U.S. educational values, including local governance, equity under federal civil rights law, and public transparency.

The Problem: Embedded Harm

As AI becomes embedded in classrooms, its impact reaches far beyond content delivery. These systems shape how students see themselves, how they are perceived by others, and how they understand their place in the world. While the 2024 UNESCO AI Competency Framework calls for curricula that emphasize critical AI literacy and student agency, the United States has yet to adopt a comparable national standard, and federal agencies have not issued binding guidance on AI literacy, identity protection, or algorithmic equity in education. Without deliberate oversight, AI tools increasingly distort identity development, particularly for students from marginalized, racialized, or neurodiverse communities. These systems risk reinforcing harmful stereotypes, narrowing pathways for self-expression, and invisibly sorting students in ways that shape lifelong opportunity.

Embedded Harm: Key Risks and Realities

These risks compound a growing vacuum of oversight where systems are introduced faster than they can be evaluated, regulated, or understood. In this space, flawed design becomes infrastructure, and experimentation turns into de facto policy, especially in the schools least equipped to push back.

AI Is Advancing Faster Than Oversight. [*Policy, Oversight, and Governance*]

AI is spreading faster than educators and policymakers can regulate, leaving major gaps in governance and equity. (*San Francisco Chronicle*)

While international organizations such as UNESCO and the World Economic Forum have issued policy guidance, most U.S. educational institutions still lack internal AI governance frameworks, highlighting a critical gap in domestic policy leadership. (US Department of Education).

Recent White House executive orders have removed key safeguards and emphasized innovation over ethical oversight.

A proposed 10-year federal moratorium on state AI regulation was rejected but reveals ongoing pressure to centralize control. (*AP News*)

Student Data Is Unprotected. [*Data Privacy and Surveillance*]

Adaptive classrooms now record metrics like learning pace, emotional responses, and decision-making patterns, raising concerns over how that data is stored, shared, and governed. (US Commission on Civil Rights- *USCCR*)

The vast majority of K–12 and higher-education institutions surveyed lack AI-specific internal policies addressing privacy, transparency, or vendor oversight. (*School Pulse Panel - US Department of Education*)

The Family Educational Rights and Privacy Act of 1974 (FERPA) mandates user consent before sharing education records, but often excludes advanced analytics, AI tool-generated data, and usage by third-party vendors; this leaves significant privacy gaps. (National Education Association -NEA)

The Children's Online Privacy Protection Act (COPPA) and the Protection of Pupil Rights Amendment (PPRA) offer limited safeguards in educational contexts involving generative AI, and no universal federal law governs the protection of student behavioral or biometric data emerging from AI use. (NEA)

Surveillance tools such as GoGuardian and Gaggle have triggered civil rights challenges, particularly where monitoring flags disproportionately affect students of color. (Stanford Law Review)

AI screening tools have also been found to produce false positives with ELL (English Language Learners) students for whom English is not their first language. (MPR News)

While FERPA provides a baseline for privacy, it was drafted long before AI-era data collection methods emerged and does not adequately address current behavioral or biometric profiling risks. (Public Interest Privacy Center - PIPC)

AI Silences Diversity & Erases Culture. [*Algorithmic Bias and Equity*]

AI systems frequently propagate and amplify historical biases — such as racial, gender, and cultural stereotypes — because training datasets tend to prioritize dominant groups and underrepresent marginalized communities. (Mergen et al)

In educational settings, AI-driven syllabus recommendations or content generation often undervalue or omit local narratives, dialects, and culturally specific knowledge, undermining representation and identity in learning materials. While this has been studied in European contexts by organizations including the Joint Information Systems Committee (now Jisc) and the European Union Agency for Fundamental Rights (FRA), U.S. schools have yet to adopt comparable safeguards or review frameworks.(Stanford Law Review)

Biased AI models have been shown to disproportionately misidentify or misinterpret language and behavior from non-native English speakers, neurodivergent students, and students from underrepresented racial backgrounds. (University of Chicago)

Efforts to mitigate these harms — including diverse representation in AI development teams, transparent algorithmic auditing, and inclusive dataset design — remain scattered or few in U.S. education technology deployments. Organizations such as the Algorithmic Justice League and Stanford HAI have called for stronger safeguards.

AI Disconnects Students From Human Relationships. [*Human Impact and Educational Practice*]

AI-mediated learning systems can reduce meaningful interactions between students and educators, leading to diminished empathy, motivation, and social-emotional learning that are central to traditional pedagogy. (National Institute of Health)

Over-reliance on AI dialogue systems has been associated with declines in critical thinking, decision-making quality, and long-term retention, as students defer to generated responses instead of engaging actively. (Lee et al)

Experimental programs combining AI with teacher support (such as Lumilo smart-classroom tools or Tutor CoPilot) show improved learning outcomes when AI augments rather than replaces human guidance. (World Economic Forum)

Trust dynamics formed with anthropomorphic AI tutors diverge from human interpersonal bonds, creating risks of miscalibrated trust that can affect student engagement and feedback acceptance. (Pitts and Motamedi)

School Systems Are Losing Local Control and Equity. [*Systemic Inequality and Structural Risk*]

Schools are increasingly dependent on proprietary AI platforms, creating vendor lock-in that restricts curriculum flexibility and institutional autonomy if vendors change prices or policies. (Solutions Review)

AI-driven instruction often embeds standardized curricula, minimizing opportunities for local expertise, input, and culturally relevant content. (Houston Chronicle)

Automated, opaque decision systems weaken traditional governance models, reducing transparency in scheduling, assessment, and resource allocation traditionally overseen by teachers and school boards. (https://www.usccr.gov/files/2024-12/2024-ai-in-education.pdf)

Under-resourced districts face infrastructure gaps, with limited IT support, inadequate bandwidth, and outdated devices impeding effective AI implementation and widening inequality. (USCCR)

These structural changes disproportionately harm underserved communities — rural, low-income, and minority districts — by reducing advocacy, oversight, and recourse when AI tools fail or underperform. (USCCR)

Neurodivergent Students Can Be Misread and Penalized by AI. [*Neurodivergence, Misclassification, and Algorithmic Harm*]

AI classification tools frequently label neurodivergent students as "at risk" using behavior models that ignore sensory, attention, or executive-function differences. (USCCR)

Students with ADHD, autism, or trauma histories are often flagged for disruption or noncompliance by algorithms that mistake neurodivergent communication or behavior for defiance. (Academic Integrity in the Age of Artificial Intelligence)

Predictive AI locks students into educational tracks without context, leaving families with limited tools to challenge misclassifications or access alternative content. (Ackgun & Greenhow)

These models prioritize standardization over personalization, failing to adapt to neurodiverse strengths, pacing, or alternative learning pathways. (USCCR)

U.S.-based advocacy groups such as the National Center for Learning Disabilities and the Autism Self Advocacy Network have raised concerns that algorithmic classification often penalizes rather than supports neurodivergent students.(Rephun)

Automation Is Driven by Budget Cuts. [*Fiscal Pressure and Technological Substitution*]

33 states have seen stagnant or reduced education funding since 2023, prompting districts to adopt AI systems as a cost-saving substitute for teaching personnel. (Learning Policy Institute)

Students in lower-income districts increasingly experience downgraded AI-mediated learning environments with minimal human support or feedback. (SF Chronicle)

These cost-driven deployments reinforce a two-tier education system where affluent schools retain human-guided instruction and underserved schools receive depersonalized algorithmic models. (USCCR)

Embedded Harm: The Architecture of Bad AI

When systemic shortfalls are left unaddressed, they form the foundation for more dangerous outcomes. What begins as technical or structural limitation becomes a gateway to deeper systemic harm, particularly when flawed AI systems are deployed at scale without transparency, accountability, or consent.

Empirical studies support these concerns. AI grading models have been shown to disadvantage marginalized students through algorithmic bias, including documented cases in U.S. districts where grading algorithms disproportionately penalized low-income, minority & indigenous students. Schools have also reported incidents of student-generated deepfakes used for harassment and reputational damage. Moreover, unsupervised AI tool use has been linked to measurable declines in student critical thinking and writing authenticity in U.S. classrooms.

"Bad AI" refers to systems that are biased, opaque, poorly trained, or ideologically driven, especially when deployed without accountability.

- These tools can shape what students learn, how they're assessed, and which behaviors are rewarded.
- Fluent and seemingly neutral systems can embed harmful assumptions, erase identity, and enforce conformity disguised as personalization.
- When deployed at scale in under-resourced schools, "Bad AI" risks becoming invisible engines of inequity, misinformation, and psychological harm.
- Without oversight, such systems displace educators, override community values, and program belief systems in ways that echo historical ideological control mechanisms.

In the U.S., these harms are most likely to take root in underfunded districts, where oversight is limited, procurement is decentralized, and political pressure often accelerates AI adoption without adequate evaluation.

Embedded Harm: Algorithmic Indoctrination and the Collapse of Inquiry

In a country where public education is both a civic foundation and a public good, AI's integration into schools must reflect shared American values, including transparency, inclusion, and democratic accountability. Artificial Intelligence is accelerating into classrooms faster than oversight can keep pace. While the promise of innovation is real, so too are the risks, especially for marginalized, indigenous, rural, neurodivergent, and historically underserved communities. To ensure AI in education strengthens equity rather than eroding it, policy makers and interested parties must act decisively and collaboratively.

Without strong governance, AI is already being used to shape not just learning outcomes, but ideological allegiance: subtly programming norms, beliefs, and compliance through feedback loops.

Much like historical systems of ideological control — such as 20th-century authoritarian regimes that used national curricula, propaganda media, and youth surveillance to produce loyalty and conformity — today's algorithms increasingly centralize content, track behaviors, and personalize feedback in ways that normalize obedience over inquiry.

If left unchecked, these systems erode democratic values under the guise of innovation: a trajectory already evidenced in overreach tied to behavior-tracking and curriculum centralization. Truth is compromised when opaque systems determine what counts as knowledge, and ethics collapse in the absence of oversight. Bias calcifies when no corrective feedback loops exist. Identity is reduced to conformity; this is a dynamic extensively critiqued by scholars like Joy Buolamwini who documented representational erasure in facial recognition systems. Equity is most endangered in those communities that are least able to opt out, appeal, or demand alternatives. This dynamic is mirrored in recent state-level efforts to standardize or sanitize curriculum content under the guise of ideological neutrality.

This collapse is not inevitable, but preventing it demands transparent governance, participatory design, and protections grounded in American civic values. U.S. frameworks such as the AI Bill of Rights and Department of Education guidance through their AI toolkit (no longer available after 7.22.25 memo) should serve as foundational starting points such as those of UNESCO, OECD, Singapore, & Finland.

The Solution: Convene a Multi-Sector Task Force

Policy makers, educational organizations, and interest parties must drive efforts to convene a multi-sector Task Force. This Task Force would function as the central body responsible for evaluating the risks and benefits of AI in education and developing implementation safeguards that reflect the needs of diverse communities across the United States.

It should be composed of educators, technologists, neurodivergent advocates, community leaders, privacy experts, and students, and be empowered to take action not just advise. This includes the authority to commission pilots, draft policy frameworks, recommend legislation, and coordinate across state and district lines, in alignment with U.S. federalism and decentralized education governance.

To guide its work, we have identified ten strategic areas, organized around five guiding principles: **Democratic Governance**, **Transparency & Accountability**, **Equity & Inclusion**, **Community Empowerment**, and **Cultural & Cognitive Integrity**.

Each domain includes two proposed action areas with examples and implementation notes. These are grounded in current U.S. policy frameworks — including the AI Bill of Rights, Title VI, IDEA — and supplemented by global guidance such as the UNESCO Recommendation on the Ethics of AI. These references serve not as templates, but as comparative benchmarks to help sharpen U.S. strategy. This structure ensures that AI adoption in education reflects public values and centers the expertise of communities most impacted by algorithmic systems.

Strategic Priorities for the Task Force

1. Democratic Governance: Ensure that AI is governed by people, not algorithms, consistent with public school board authority and local control in U.S. education.

Support Transparent, Human-Governed Pilots

- The Task Force should fund small-scale pilots where learners, families, and educators retain decision-making authority.
- *Example*: A school district might test an AI tutoring tool only after forming a family

advisory board that retains final veto power.

Advance a 'Public Right to Audit'

- The Task Force should advocate for legislation that guarantees tool auditability, opt-out rights, and meaningful consent. This should include clear protocols for identifying, reporting, and remediating Bad AI deployments that misinform, mislabel, or erode learner agency.
- *Example*: A state agency or education coalition could maintain an open-access online registry of AI tools used in education, with an independent complaint and remediation process.

Note: These priorities are essential to preserve democratic authority over U.S. educational systems. The Task Force should explore mechanisms that allow communities (not algorithms) to define what safe, inclusive, and effective AI use looks like in public education, aligned with existing structures such as school boards and local education authorities.

2. Transparency & Accountability: Make AI systems visible, testable, and open to correction, aligning with civil rights oversight and U.S. public records norms.

Establish Minimum Standards for Safe Use

- The Task Force should develop an open-source AI Guardrails Checklist aligned with Coalition values. These standards must directly address the threat of Bad AI, including guidelines for training transparency, bias auditing, independent testing, and community-informed design safeguards. They should draw on the U.S. AI Bill of Rights (White House, 2022) and reference international frameworks like UNESCO's ethics guidelines as a global policy gold standard, not as a cookie-cutter template.
- *Example*: The Task Force could adapt the U.S. AI Bill of Rights into a simplified checklist for evaluating new tools before district-wide adoption.

Enforce Anti-Indoctrination Safeguards

- The Task Force should build firewalls against centralized control of values instruction. Require algorithmic transparency in content generation and prohibit behavioral ranking tied to ideological conformity.
- *Example*: Any AI-generated curriculum content must include a visible audit trail showing the sources and training data that shaped its recommendations.

Note: The Task Force should define clear audit criteria, disclosure requirements, and appeal mechanisms that ensure AI tools in schools operate in full view of the public and uphold democratic norms, including protections enshrined in U.S. constitutional and civil rights frameworks.

3. **Equity & Inclusion:** Safeguard the rights and needs of vulnerable groups most at risk of exclusion or harm, grounded in protections under Title VI, Section 504, and IDEA.

Defend DEI and Civil Rights in Curriculum Design

- The Task Force should mandate that AI tools reflect civil rights protections and equity goals. Create accountability processes for systems that exclude or erase race, gender, disability, or class identity in pursuit of "neutrality." These processes should align with established U.S. statutes such as Title VI of the Civil Rights Act, Section 504 of the Rehabilitation Act, and IDEA.
- *Example*: Require AI vendors to publish demographic performance metrics and provide opt-in DEI features that allow local tailoring.

Protect Neurodivergent and Disabled Learners

- The Task Force should require AI systems to accommodate diverse cognitive profiles and prohibit behavior-based labeling without human oversight. It should develop inclusive design standards and validate them with neurodiverse and disabled communities.

- *Example*: Before launch, an AI learning assistant is evaluated by a coalition of neurodiverse reviewers and redesigned after testing its language simplification tools on students with ADHD and dyslexia.

Note: Equity challenges facing rural and underserved districts, such as limited infrastructure, staffing shortages, and uneven access to AI-ready environments, should be formally reviewed by the Task Force as cross-cutting implementation concerns. These challenges are especially acute in the U.S., where education funding and infrastructure vary dramatically across states and localities.

4. Community Empowerment: Equip learners, families, and educators with the tools to participate and advocate through participatory processes rooted in U.S. school district engagement models.

Strengthen Community AI Literacy

- The Task Force should host family and student-focused workshops on how AI works, its limitations, and how to advocate.
- *Example*: A Saturday community fair includes hands-on AI demos and training sessions where students learn how to identify when a tool is giving biased or incorrect information.

Build Student and Family Feedback Channels

- The Task Force should create formal structures for students and families to give input on AI tools in real time, including opt-out options, satisfaction surveys, and harm reporting pathways.
- *Example*: A district adds an AI feedback portal linked to school dashboards, where students can rate clarity, accuracy, and fairness of AI tutoring responses.

Note: The Task Force should prioritize hands-on community engagement efforts and feedback systems that put families, students, and educators in the driver's seat when it comes to AI integration. These mechanisms should reflect existing U.S. models for school district participation and parent-student advisory structures.

5. Cultural & Cognitive Integrity: Protect cultural values, community identities, and diverse ways of thinking from being overwritten, particularly for historically marginalized communities (such as poor, women, LGBTQ+, indigenous, people of color, neurodivergent, & Appalachian), ensuring the accuracy and relevance of curriculum development, teaching methods, communication, and education policy & standards.

Preserve Local Cultural and Linguistic Identity

- The Task Force should fund culturally responsive curriculum audits. These could prevent algorithmic suppression of regional dialects, minority languages, and community-specific learning norms. These efforts should reflect domestic cultural diversity and be informed — when appropriate — by international guidance such as the UNESCO Recommendation on the Ethics of Artificial Intelligence.
- *Example*: A regional education co-op conducts a bias scan on an AI writing tutor and discovers it flags Appalachian dialect grammar as "incorrect," prompting revision of its correction model.

Protect Cognitive Diversity in System Design

- The Task Force should ensure AI tools account for multiple learning models and problem-solving styles, not just neurotypical or high-speed task performance.
- *Example*: An AI platform's success metrics are redesigned to reward slow, exploratory learning alongside speed and precision, enabling a broader range of students to succeed.

Note: The Task Force should create design review protocols that prevent homogenization of language, thought patterns, and community expression, especially in a diverse and decentralized U.S. education landscape.

Conclusion

The next step is clear; initiate the Task Force. This body must be empowered not only to assess the risks of AI in education but to oversee implementation, prioritize transparency, and establish long-term accountability mechanisms.

Its responsibilities should include:

- Establishing clear oversight processes, including regular audits and public reporting;
- Creating feedback and redress channels for students, families, and educators;
- Coordinating pilot programs that center local voice and preserve learner agency; and
- Adapting standards to account for contextual differences, such as rural or under-resourced settings.

Oversight should not be a static checkpoint but a living framework: one that evolves in response to community input, emerging risks, and new insights. Implementation must remain cautious, adaptive, and community led. Above all, AI systems must support and never undermine the dignity, autonomy, and diversity of the learners they serve.

A well-structured Task Force that is grounded in shared values and public accountability can help shape national and global norms for educational AI. It can advance policies that reflect U.S. values of equity, transparency, and democratic participation while engaging with international benchmarks such as UNESCO's ethical guidance as a global standard for inclusive and accountable AI use in education.

> *We're not just choosing tools; we're choosing values. Without oversight, AI in classrooms could erode the foundation of public education: trust, equity, and the dignity of learning.*

How we use AI in classrooms reflects and reinforces the values we want students to carry forward. We are making governance decisions that will shape the future of trust, inclusion, and accountability in American education. Implementation and oversight will determine whether these technologies strengthen learning or quietly displace it.

Each of the strategic goals outlined earlier offers a pathway toward a more just, inclusive, and resilient educational future. Together, they prioritize transparency in system design, protection of student data, and culturally responsive development practices. They call for investments in public infrastructure and educator training while ensuring that AI supports – rather than replaces – the human relationships that are at the heart of education.

By elevating the voices of students, families, and educators, and by establishing equity-centered governance and accountability systems, these strategies move us toward classrooms where AI amplifies curiosity, not compliance; collaboration, not control. This is how we ensure AI remains a tool for learning not a system that narrows it.

Without safeguards, the risks are clear; untested systems may be deployed faster than they are evaluated, exposing students to bias, surveillance, and inequitable outcomes. When that happens, schools risk becoming sites of experimentation rather than spaces of empowerment. The stakes are too high to ignore. These goals are not merely aspirational; they are essential safeguards that anchor public education in human judgment – not algorithmic control – and preserve it as a human-centered public good.

The United States has the opportunity and responsibility to lead in defining educational AI that aligns with democratic values. While international frameworks such as UNESCO's ethical guidance set a global bar for equity and accountability, U.S. policy must ensure that AI in education strengthens, rather than fragments, the foundations of public trust, civil rights, and learner autonomy.

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.

Chapter 14:
AI & Entertainment: A Blueprint for Innovation, Integrity, and IP Protection

Authors: Annie Hanlon, Jess Loren, Ann M. Marcus, Christina Lee Storm

Overview

Generative AI (GenAI) can shrink production timelines by creating storyboards in minutes and multilingual dubs in hours, yet that speed surfaces thorny issues of copyright, consent, and credit. The very tools that streamline visual effects, localization, music, and other workflows also introduce profound ethical and legal dilemmas. The industry now sits on a fault line: innovation versus infringement, piracy versus IP protection, and automation versus human creativity.

This chapter traces that collision, from VHS piracy to Stable Diffusion, and offers a blueprint for protecting originality while encouraging innovation. Drawing on historical context, emerging legal cases, ethical frameworks, and sector-specific use cases, we offer a blueprint for how the entertainment and creative sectors can chart a path forward that protects originality, fosters innovation, and upholds the values of consent, attribution, and trust.

How AI is Revolutionizing Entertainment

Human and AI creative partnerships are unlocking new possibilities for artists, filmmakers, creatives, and entertainment professionals by blending human ingenuity with the speed and versatility of GenAI. Often referred to as "human in the loop" (HITL), this collaboration is essential for achieving expressive, nuanced, and emotionally resonant results in entertainment and the arts.

While AI excels at generating content at scale and speed, it lacks the lived experiences, cultural context, and intuitive understanding that define truly impactful creative work. Humans bring judgment, taste, emotion, and a deep sense of narrative to the process. In practice, this means that AI can rapidly generate storyboards, music, or visual assets, but human creators guide the direction, curate the best outputs, and infuse the work with subtlety and meaning. For example, when filmmakers use AI for storyboarding, it is the director's vision and feedback that shape the final sequence, ensuring the emotional beats and visual style align with the story's intent.

GenAI powered tools are revolutionizing the filmmaking process allowing directors to experiment with different styles and camera angles in minutes rather than days. Similarly, musicians who collaborate with AI to remix legacy works rely on their own creative instincts to select, refine, and approve the final versions, preserving the authenticity of their artistic voice.

The result is a powerful synergy that expands creative horizons, democratizes access to advanced tools, and enables artists to push boundaries, reach new audiences, and tell stories in ways that were previously unimaginable.

Human + AI: Real-World Collaborations

Filmmaking: Storyboards in an afternoon. The 2024 research prototype **CinePreGen** lets directors rough-out camera moves and storyboards with a diffusion model that accepts natural-language prompts and real-time camera controls; a 12-participant study showed it *cut pre-*

vis iteration time by more than half while keeping human directors in the loop for framing and tone.

Localization & Access: Auto-dubs at scale. In December 2024, YouTube expanded its AI dubbing tool to "hundreds of thousands" of channels, auto-translating a single upload into up to nine languages. Creators can preview or delete the synthetic tracks before publishing, preserving artistic control while instantly opening new markets.

Legacy Music: Finishing the last Beatles song. *Now and Then* (released Nov 2024) used Peter Jackson's machine-learning audio-restoration system to isolate John Lennon's 1977 demo vocal so Paul McCartney and Ringo Starr could build a new arrangement around it. The single topped charts in 10 countries and won the 2025 Grammy for Best Rock Performance: proof that AI can *extend* rather than replace human artistry.

These snapshots show where AI already extends human effort; the next sections examine where it might undermine it.

Key Takeaway

In each case, AI handles the *heavy lifting* – rapid image synthesis, voice cloning, or signal cleanup – while humans provide narrative intent, editing judgment, and final sign-off. The results: faster workflows, bigger audiences, and renewed value for archival material.

But as the technology evolves, so do risks related to unauthorized use of copyrighted material and the erosion of intellectual property rights. By prioritizing best practices and guidelines, responsible development, and ensuring that GenAI systems are trained on properly licensed data, the industry can foster innovation while protecting the creative contributions and intellectual property that form the foundation of the entertainment industry.

Historical Context and New Parallels

In the 1980s and 1990s, the entertainment industry grappled with the challenge of piracy in the form of unauthorized duplication of VHS tapes and CDs. These breaches undermined creators and disrupted economic models. The solution involved studios, artists, distributors, and the Federal Government responding with copyright crackdowns, the creation of anti-piracy infrastructure, and legal innovations.

Today, we're facing a digital version of that same problem but with GenAI. Instead of duplicating VHS tapes, GenAI systems are trained on vast datasets of creative content, films, scripts, music, and art often without consent or compensation. These models can then generate new works that borrow heavily from the originals, sometimes with striking similarity to the source material. The issue isn't just technological; it's foundational. Creators risk losing control over their work and intellectual property, while companies face legal exposure and financial loss if they don't ensure the content they use or distribute is responsibly sourced. Without clear provenance and disclosure, creative teams and studios may struggle to trace the origin of content or its underlying components, which will impact the foundational pillar of the chain of title. Fast-forward four decades, and the VHS tape duplicator is now a training dataset.

The New Landscape of Risk

Key risks associated with GenAI in entertainment and creative domains include:

- **Source Misappropriation:** GenAI models trained on copyrighted or proprietary material often generate content that resembles original works in tone, structure, or style.
- **Attribution Confusion:** Human-AI collaborations raise questions about authorship, rights, and recognition. Who owns the output? Who deserves credit?
- **Legal Exposure:** From copyright infringement to trade secret violations,

organizations using AI-generated content risk legal action if training data or outputs lack proper provenance or licensing.

Recent lawsuits, such as The New York Times v. OpenAI/Microsoft illustrate how unresolved questions of fair use, consent, and replication could redefine copyright law.

Archival vs. Piracy: A Core Tension

Not all unlicensed reuse is nefarious. The Archival Producers Alliance (APA) and other documentary filmmakers argue that preservation and transparency sometimes presents a tension; when does use of a work preserve history and truth, and when does it exploit the labor and voice of a creator without consent?

The APA calls attention to the “inherent obligation to reality” in documentary work (a term first used by G. Roy Levin), underscoring the societal value of preserving and referencing materials that might otherwise be lost. That is particularly relevant when these references serve the public interest, such as revealing abuses of power or challenging dominant historical narratives. In such cases, using GenAI or traditional methods to archive, reference, or reproduce vulnerable content must be accompanied by clear sourcing, responsible attribution, and contextual integrity to avoid confusion or distortion.

The APA notes that GenAI use may be seen as particularly problematic when *simulating truth-based narratives*. They suggest that documentary content disclose all synthetic contributions and ensure audiences are not misled by machine-generated interpretations of factual events. Ultimately, “ethical reuse” is rooted in *purpose, context, and acknowledgment*.

This makes it vital to distinguish between:

- Malicious plagiarism or cloning (e.g., voice deepfakes, song imitations),
- Transformative reuse for public interest (e.g., archival storytelling, education, parody), and
- Tool-assisted creation where AI is used transparently (e.g., CGI or Photoshop).

Key Principles for Responsible AI in Entertainment

A responsible AI ecosystem must prioritize:

- **Transparency**: Disclosure when AI has been used in content creation or enhancement
- **Clean Source Data**: Licensing, attribution, and documentation of training datasets
- **Attribution**: Clear credit given to creators whose works are reused or remixed
- **Consent**: Creative assets should not be used without approval.
- **Provenance**: Technological tracking of content origin (e.g., C2PA, blockchain)
- **Fair Compensation**: Royalty structures for creators whose work fuels GenAI outputs
- **Standardization**: Adoption of shared frameworks for watermarking, metadata, and model disclosures

Sector Use Cases and Responses

Visual Arts: Artists are suing platforms Stability AI, DeviantArt, Midjourney, and Runway ML, alleging these companies used their work in training datasets without licensing and that the outputs closely replicate their distinct styles, constituting copyright infringement and unfair competition.

Music: AI-generated tracks that mimic real artists without approval (e.g., "Heart on My Sleeve") have prompted pushback from performers and unions seeking voice rights protections.

Literary: Authors sued Anthropic, claiming it illegally used their copyrighted books to train its Claude AI model. This landmark ruling marks one of the first major federal interpretations of fair use in AI training. It affirms transformative use of lawfully acquired texts but clearly draws a legal line against using pirated content.

Code: GitHub Copilot has sparked backlash for producing uncredited code snippets from open-source repositories.

Academia: AI-generated essays and paraphrasing tools are challenging norms of citation and originality.

Enterprise: Proprietary information leaked via AI tools (e.g., chatbots trained on internal documentation) creates new risks for data governance.

How Are Audiences Reacting to AI-Made Media?

Skepticism in the U.S. More than half of Americans (54%) say generative-AI systems *must* credit the sources they draw from, while only 14% think attribution is unnecessary. Pew Research Center

Demand for Clear Labels in Music. A 2025 survey of U.K. listeners found 81.5% want AI-only tracks clearly labelled and over 80% still "value human-made music more. "DJ

Advertising Backlash. NielsenIQ's neuroscience study showed viewers flagged most AI-generated ads as "annoying," "boring," or "confusing," triggering weaker memory activation than conventional spots: evidence that poorly disclosed AI can corrode brand equity. NIQ

Global Trust Gap. Trust is not uniform; in the 2025 Edelman Trust Barometer, 72% of Chinese respondents trust AI versus 32% in the United States, with India (77%) topping the league. Axios

Why This Matters:

Audience acceptance shapes everything from box-office returns to award eligibility. Data show that transparency (crediting and labelling) and perceived human authorship dramatically influence trust, recall, and engagement across formats: music, film, ads, and even social feeds. Studios that embed provenance signals (e.g., C2PA watermarks) and disclose AI involvement early stand to build goodwill, whereas opaque releases risk backlash or reduced commercial impact.

Audience perception is only half the puzzle; the other half is how platforms choose to disclose, or hide, AI involvement.

Platform Responsibility & Disclosure

Why the Distribution Layer Matters. Streaming and social-video platforms now act as first-line gatekeepers for AI-made media; they can require labelling, redirect royalties, or quietly amplify synthetic works with no context at all. The policy choices they make therefore shape both creator livelihoods and audience trust.

- **YouTube: Mandatory Labels.** Since Q1 2025, YouTube has required any uploader who uses "realistic altered or synthetic media" to tick an AI-use box. The platform then auto-attaches a visible *"altered or synthetic"* label, and, for sensitive topics such as news or finance, a second onscreen banner. YouTube
- **Spotify: Training Ban, No Tag (Yet).** Spotify now forbids AI companies from scraping its catalog and removes deep-fake tracks, but it still lacks a consumer-facing tag for synthetic songs, leaving listeners to guess whether a track is human-made. Descript Further, The "Velvet Sundown" incident, an AI band that quietly racked up 1 million Spotify plays, triggered calls from industry bodies for mandatory tagging so fans "know what they're hearing." The Guardian
- **Deezer: First Mover on Tagging.** In June 2025, Deezer became the world's first digital service provider (DSP) to display an *AI-generated* badge on every album that contains fully synthetic tracks; its detection tool already flags about 18% of daily uploads and excludes fraudulent streams from royalty pools. Deezer Newsroom

Audience Backlash Drives Change

What should platforms do next?

- **Universal "AI-Created" Disclosure Tag** visible at play-time (not buried in metadata).
- **Attribution and Royalty Sharing Panels** that let rights-holders claim a cut when licensed stems or likeness models power a release.
- **Dataset-Opt-Out Registries** so creators can block future training on their uploads.
- **Content-ID for Personalities**, extending YouTube's synthetic-voice detection to faces and brand mascots.
- **Transparent Recommendation Throttles** — as Deezer does — when streams appear bot-inflated.

Open Question for the Industry: If labels and audiences increasingly expect up-front disclosure, should the *absence* of an "AI-created" badge eventually count as consumer deception? The precedents above suggest that proactive labelling will soon move from *nice-to-have* to *regulatory baseline*.

Legal and Policy Trends

Lawsuits against GenAI platforms will likely define the boundaries of fair use, copyright, and derivative work protections, but traditional regulatory frameworks with multi-year judicial processes are ill-suited to address the real-time challenges and opportunities posed by AI. The lawsuit filed by Disney and Universal against Midjourney over copyright infringements is expected to be a lengthy process because of the complexity of AI and copyright law and the high stakes outcome of this case, which could significantly influence the future of both AI development and the entertainment industry's approach to intellectual property rights.

The accelerating pace of AI development demands proactive, coordinated action from the legal, policy, and entertainment sectors. Only through collaboration can they ensure that AI is harnessed responsibly and ethically.

A pivotal example of this is the recent removal of the proposed federal moratorium on state-level AI regulation from the "One Big Beautiful Bill Act." The original provision would have blocked states from enacting new AI laws for up to a decade, effectively freezing local responses to emerging risks and stifling the ability to protect creative professionals and the public. By removing the moratorium, Congress preserved states' authority to enact timely protections, an outcome widely regarded as a win for the creative community and advocates for responsible AI.

Despite this legislative progress, significant policy gaps remain in regulating AI-generated content, particularly deepfakes and digital replicas. The U.S. Copyright Office has called for new federal protections that would prohibit the distribution of unauthorized digital replicas, mandate prompt takedown mechanisms on online platforms, and provide statutory damages and injunctive relief for victims. Similarly, the proposed NO FAKES Act (for Nurture Originals, Foster Art, and Keep Entertainment Safe) – a U.S. Congressional effort to protect personal identity and creative intellectual property from unauthorized AI reproductions commonly known as "deepfakes" – would introduce a federal right of action, require platforms to implement strong takedown and repeat-offender policies, and leverage digital fingerprinting to prevent re-uploads.

Importantly, the Act aims to balance protection with creative freedom by recognizing the role of transformative or creative modifications, as highlighted in the U.S. Copyright Office's AI reports, which emphasize that copyright law protects original, human-authored contributions while allowing for fair use and transformative works. This distinction seeks to ensure that legitimate artistic reinterpretations and documentary uses are preserved, while unauthorized, exploitative reproductions are curtailed.

While the U.S. Copyright Office's three-part series on Copyright and Artificial Intelligence (published Part 1 on July 31, 2024, Part 2 on January 29, 2025, and pre-publication version of Part 3 on May

9, 2025) provides valuable analysis and highlights key challenges at the intersection of AI and copyright law, the reports remain broad in scope and stop short of offering specific, enforceable standards. The Office acknowledges that many questions, such as the boundaries of fair use in AI training, the definition of human authorship, and the mechanisms for protecting digital replicas, are far from settled and will require further legislative, judicial, and policy development. As a result, stakeholders in the creative and technology sectors must navigate a landscape marked by significant legal ambiguity, with much depending on future court decisions and potential new legislation.

Across the Atlantic, transparency is becoming law. In February 2025 the EU formally adopted the AI Act, European Union, EU AI Act Transparency Mandate**,** the first comprehensive framework of its kind. While generative models are not classed as "high-risk," they must (i) label AI-generated media, (ii) design systems to prevent illegal content, and (iii) publish "sufficiently detailed" summaries of all copyrighted works used in training. By forcing disclosure at the dataset level, the EU has created a de-facto provenance standard that goes further than any U.S. proposal to date.

Meanwhile, UK litigation is expanding the definition of infringement, In January 2025 the U.K. High Court United Kingdom, Getty Images v. Stability AI**.** allowed Getty's multi-count infringement suit against Stability AI to proceed, rejecting the developer's bid to narrow the case. Getty alleges wholesale scraping of its licensed catalog to train Stable Diffusion, plus trademark dilution in downstream outputs. The ruling signals that training-phase ingestion itself can constitute primary infringement under U.K. law, a point still unsettled in U.S. courts. Courts and Tribunals Judiciary

Concurrently, the U.K. Intellectual Property Office closed a nationwide consultation that floats a "reserve-your-rights" mechanism; right-holders could opt out of AI training unless paid, while developers gain a safe harbor for unreserved works, but only with dataset transparency baked in. GOV.UK

Asia-Pacific is leaning on registration. In June 2025, the Korean Copyright Commission issued dual guides on (1) registering AI-assisted works and (2) preventing AI-related disputes. Purely machine-made outputs receive no copyright, but creators can secure protection for "GAI-utilization works" by documenting their human contributions. Studios rushing into the K-Drama boom now treat the registration filing as a green-light checklist item, similar to chain-of-title clearance in Hollywood.

India convened an eight-member expert panel in May 2025 to modernize the 1957 Copyright Act. Mandates under review include a formal definition of AI-generated works, liability for unlicensed training, and a new remuneration right for datasets sourced from Indian publishers and broadcasters. The panel's report, due early 2026, will shape rules for Bollywood and the country's ₹2-trillion streaming market. Lexology

At the same time, some companies are already demonstrating what responsible data use can look like. In essence, Industry is not waiting for courts. For example, OpenAI has entered into a series of licensing agreements with major publishers, including The Financial Times, Associated Press, Le Monde, and others, allowing their content to be used for AI training in exchange for compensation and attribution.

Producers are being asked to make critical decisions without the benefit of clear industry standards or government regulation. In this interim period, while formal policy and legal guardrails continue to take shape, resources such as the Academy of Television Arts & Sciences (TV Academy) "KEY CONSIDERATIONS Before Using GenAI on Your Next Project" focus on three key principals: *Creative Integrity to Professionals, Creators, Performers, Craftspeople; Permissions, Licenses: Legal & Commercial Viability; and Accountability, Transparency, Sustainability*. The Key Considerations are designed for the nearly 30,000 members across the 31 peer groups of the Television Academy.

In addition, the Producers Guild of America created a document, **Fine Print of AI: Top 10 Questions Producers Should Ask**, for producers to reference. These frameworks help television professionals and producers navigate the evolving landscape by identifying potential legal, ethical, and creative risks, and offering practical questions to ask when evaluating GenAI's role in a project. As the industry seeks clarity, these tools empower producers to move forward responsibly, protecting themselves, their teams, and their work.

In July 2025, Asteria and Moonvalley released Marey, a clean, production-grade AI video model designed to give filmmakers creative control while avoiding the legal and ethical pitfalls of systems trained on scraped, unlicensed content. Fully licensed and commercially safe, Marey was developed in partnership with creators, ensuring that innovation is built on collaboration, not exploitation.

Also, through their 2025 partnership, Independent Studio Services (ISS) – the world’s largest full-service prop house stewarding more than five million items with lineage tracked since the 1970s – and Global Objects (GO) – a 3D-scanning and digital-asset company specializing in photorealistic digital replication for media, entertainment, and enterprise applications , are converting each screen-used prop into an IP-cleared, DRM-watermarked digital twin with full provenance metadata, making the collection safely licensable for metaverse platforms, real-time game engines, and GenAI training pipelines.

As this new digital landscape unfolds, Playbook AIR’s platform is designed to capture and verify human authorship in GenAI workflows, providing clear documentation to support copyright, protect creators, and ensure accountability. It also provides a secure API, allowing seamless integration into other platforms and systems. Platforms like this are helping to lay the groundwork for responsible and scalable adoption of GenAI in professional production pipelines.

These approaches also address growing concerns from independent and marginalized creators, such as Indigenous artists and emerging youth artists, about AI models exploiting cultural works and traditional knowledge without permission. These communities are especially vulnerable to having their art and cultural expressions scraped for AI training without consent or compensation, leading to cultural appropriation and loss of control over their own narratives.

These initiatives don’t just set a precedent; they establish a working model for how transparency, consent, and intellectual property rights can be integrated into scalable AI solutions. As the industry evolves, studios must take an active role in ensuring these standards are upheld throughout the content pipeline.

The Studio’s Role in Provenance and GenAI

Studios and distribution platforms play a critical role in ensuring that content can be legally distributed and monetized. Without a clear chain of title, studios can’t greenlight projects, and distribution platforms risk liability by hosting content built on unlicensed or scraped data. And with GenAI, that “chain” is increasingly complex. The traditional “kick the can down the road” approach is no longer viable.

Studios must take an active seat at the table to ensure that the data used to train AI tools is commercially licensed and traceable, ensuring it meets copyright and attribution standards. These conversations must also address downstream implications, such as whether projects that include AI-generated content should be eligible for prestigious awards like the Grammys, Emmys, or Oscars: questions that further underscore the need for clarity, accountability, and industry-wide alignment.

Gaps and Open Questions

Lack of Consensus on Attribution Standards: Who gets listed in credits when AI contributes?

Absence of Enforceable Provenance Tech: How can we reliably track AI-generated content origins?

Insufficient Legal Definitions: What constitutes a "derivative" work in AI?

Need for International Coordination: IP laws vary across countries; how do global platforms ensure compliance?

Next Steps and Calls to Action

- **Mandate** AI-use disclosure and provenance tech in all guild, union, and platform contracts.
- **Build** creator-led licensing frameworks and opt-out registries so rights-holders control how their works train future models.
- **Carve out** public-interest exceptions that let archivists and documentarians reuse material ethically without chilling speech.
- **Partner** with tech vendors to clean training datasets, verify sources, and watermark synthetic outputs.
- **Train** creatives, producers, and legal teams on AI risks, responsibilities, and emerging best practices.

Goal*:* These actions safeguard original voices, support working professionals, and keep innovation grounded in consent, attribution, and fair compensation.

Conclusion

Generative AI offers unprecedented opportunities for creativity, but also significant risks to the foundational principles of artistic authorship and intellectual property. The entertainment industry now stands at a critical crossroads. Will it repeat the mistakes of past technological shifts, or can it build a new, transparent, ethical framework that is based on licensed data for creation in the AI age?

Trust, consent, and attribution are the new currencies of creativity. Without them, AI-generated content may be prolific, but it will lack soul, legitimacy, and the cultural credibility that comes from honoring the human story behind the work.

Author (In order of contribution)

Annie Hanlon, Co-Founder/Partner, Playbook PLBK
Annie Hanlon is Co-Founder/Partner of Playbook PLBK, focused on AI, storytelling, and innovation. An accomplished entertainment executive and award-winning producer, she is recognized for her impact at Netflix, Lytro, and Here Be Dragons. Annie is also Co-Founder of Playbook AIR, a platform designed to capture and verify human authorship in GenAI workflows, providing clear documentation to support copyright, protect creators, and ensure accountability. She is a sought-after industry speaker, a member of the Visual Effects Society, the Academy of Television Arts & Sciences, and a board member of the Infinity Festival. Annie is a 2024 graduate of Space Camp (Huntsville, AL).

Jess Loren, CEO, Global Objects
Jess Loren is the CEO of Global Objects and a leader in AI-driven 3D scanning. She serves on the Television Academy's Governors Board for Special Visual Effects and the Board of Managers for the Los Angeles Visual Effects Society. A Producers Guild of America member, Jess is also a proud wife and mother of three.

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.

Christina Lee Storm, Co-Founder/Partner, Playbook PLBK
Christina Lee Storm is Co-Founder/Partner of Playbook PLBK, a consultancy specializing in AI, innovation and storytelling, and an award-winning producer and executive recognized for her work at major studios including Netflix, Universal, and DreamWorks Animation. She is also Co-Founder of Playbook AIR, a platform built to document and authenticate human authorship within GenAI workflows. Christina holds many leadership roles including Governor of the Television Academy's Emerging Media Programming peer group, Lead for the Academy’s Responsible AI & Production Standards Working Group, sits on the Producers Guild of America’s Production Innovation Task Force, and is a newly inducted member of the Academy of Motion Picture Arts & Sciences, Class of 2025.

Chapter 15: Conclusion

Author: Sarah Ennis

As artificial intelligence continues to evolve, the need to guide its trajectory with intention and care has never been greater. This Blueprint is more than a technical roadmap. It is a call to action for building a human-centered, equitable, and sustainable AI future.

The challenges are complex, but they are not insurmountable. Grounding AI in ethical design, inclusive governance, and environmental responsibility allows us to create technologies that serve the public good as well as the market. This is a shared responsibility. Technologists, educators, policymakers, and citizens each have a role in shaping AI that is resilient, trustworthy, and aligned with human values.

|

> *"The future of AI is not inevitable. It is ours to design."*
>
> **Sarah Ennis**, AI Chair, LG NOVA Coalition for Innovation

Author (In order of contribution)

Sarah Ennis, Co-Founder, AgentsGEO.ai
Sarah Ennis is a Fortune 500 trusted advisor specializing in advanced technology innovation, with over two decades of experience leading groundbreaking AI solutions at scale. Globally recognized for her expertise in artificial intelligence, she designs and implements bespoke emerging technology products across industries. She is also the co-founder of AgentsGEO.ai, a patent-pending platform that helps brands monitor and improve their visibility in the AI ecosystem and deploy AI agents, ensuring they are discoverable and recommended by tools like ChatGPT, Gemini, and others through its proprietary GEOScorer™ technology. In addition, Sarah contributes part-time to Northeastern University's Master of Digital Media programs in AI, preparing the next generation of technologists and creative leaders. Her work bridges Silicon Valley innovation with global impact, and she is a distinguished member of the American Society for AI and contributor to the OpenAI Forum.

Appendices

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.

Appendix B:
Four Case Studies: The Importance of International Collaboration in AI

Author: Andrew Yongwoo Lim

Introduction

The rapid evolution of artificial intelligence (AI) necessitates a global approach to its development, deployment, and governance. This appendix outlines an "Innovation Blueprint" that underscores the critical importance of international collaboration in fostering innovation, streamlining regulation, and establishing common standards and guidelines especially in the AI domain. Drawing insights from direct experience in international collaboration, particularly between Quebec, Canada, and South Korea, this appendix highlights actionable strategies for effective cross-border partnerships.

Through detailed case studies, such as the Seoul AI Hub-MILA scientist-in-residence program and various national and subnational joint research initiatives, we demonstrate how shared visions, political alignment, and structured support mechanisms can accelerate AI and innovation ecosystem advancement and ensure responsible development for the benefit of all. This framework aligns with the broader principles of open innovation and strategic ecosystem building championed by leading innovation hubs worldwide, akin to the collaborative model fostered by LG NOVA.

The Global Imperative for AI & Innovation Collaboration

Artificial Intelligence is not merely a technological frontier; it is a transformative force reshaping industries, societies, and economies worldwide. Its pervasive and disruptive impact, coupled with its rapid development and the inherent need for relevant and high-quality data, demands an unprecedented level of international cooperation. Such collaboration is vital not only to foster breakthrough innovations but also to establish coherent regulatory frameworks, technical standards, and ethical guidelines that ensure AI's responsible and beneficial deployment across diverse contexts.

This appendix presents an "Innovation Blueprint" derived from concrete experiences in fostering international collaboration, specifically focusing on the dynamic partnership between Quebec, Canada, and South Korea. These nations share a common commitment to innovation leadership such as in AI, albeit with complementary strengths; Quebec is a global hub for fundamental AI research while South Korea excels comparatively in AI application and industrialization. This synergy forms a fertile ground for joint endeavors. The principles observed in these successful collaborations resonate with the strategic approach of global innovation centers such as LG NOVA, which actively cultivate external partnerships and ecosystems to drive innovation. We will explore four key examples of how these collaborative efforts have propelled AI innovation, streamlined regulatory discussions, and fostered a shared understanding of best practices.

The importance of such international engagement is underscored by the existence of prominent global initiatives including the Global Partnership on Artificial Intelligence (GPAI), International Centre of Expertise in Montreal for the Advancement of Artificial Intelligence (CEIMIA) and the Science and Technology Policy Institute (STEPI) in Korea. GPAI is an international, multi-stakeholder initiative launched in June 2020 that promotes the responsible development and use of AI grounded in human rights, inclusion, diversity, innovation, and economic growth. Proposed by Canada and France at the 2018 G7 Summit, it was officially established with 15 founding members.

GPAI brings together experts from governments, industry, academia, and civil society to bridge the gap between theory and practice in AI by supporting cutting-edge research and applied activities. Its structure includes a Council, a Steering Committee, and a Secretariat hosted by the OECD. A key component of GPAI is the establishment of two Centres of Expertise: one in Montreal and another in Paris. CEIMIA plays a central role in supporting GPAI's working groups, particularly those focused on Responsible AI and Data Governance, and serves as a hub for international collaboration and the advancement of applied AI projects. Furthermore, knowledge-sharing platforms – such as reports from organizations like the STEPI in Korea – are crucial for disseminating information on global AI strategies and advancements, helping nations learn from each other's experiences.

A review of leading AI and innovation ecosystems globally reveals that their formation and growth are intrinsically linked to robust international collaboration. This aligns with broader national strategies, such as Quebec and Canada's Indo-Pacific Strategy, which emphasizes market diversification and deepened engagement with key partners in the region. Both Quebec and Korea have articulated strong innovation strategies. President Lee Jae-myung, inaugurated in June 2025, has initiated a new national strategy for research and innovation that marks a significant shift from the previous administration's approach. While the full official title of the new administration's comprehensive R&D strategy has not yet been widely publicized, the core of President Lee's innovation policy is clear; it is centered on an “AI for All” vision. These strategic alignments create fertile ground for sustained and impactful cross-border AI partnerships.

The Indispensable Role of Missions and Personal Engagement in Collaboration

While strategic alignment, timing, and fit are crucial for successful international collaboration, the foundational element is the cultivation of personal relationships. These connections are built not through virtual meetings but through in-person interactions, often over informal engagements. Therefore missions, conferences, exhibitions, and forums are not merely events but vital platforms for forging durable partnerships.

Such missions enable direct, on-the-ground engagement, allowing officials and delegates to conduct work in person, build trust, benchmark best practices, learn from diverse ecosystems, and report findings effectively. This hands-on approach ensures a deeper understanding and facilitates problem-solving as issues arise, as typically collaborations falter when challenges and problems are not addressed in an effective and timely manner. Crucially, the careful selection of officials for these missions is often strategic, ensuring that the positive experiences and relationships formed leave a lasting impression. Thus, key aspects are considered when being part of the mission, such as a startup's maturity, the entity's potential for international collaboration, language skills, and viability overall.

Examples of the impact of these missions are plentiful. Institutional partnerships can be effective, such as the collaboration facilitated during an international conference held in Incheon called K-UAM, where CRIAQ – a consortium of aerospace entities – and Polytechnic University joined GURS. GURS (Global UAM Regional

Summit) is an international secretariat created by the Incheon Metropolitan Government to foster international collaboration in urban air mobility. This highlights how missions can foster frameworks that facilitate ongoing cooperation under an established secretariat. And the membership provides a regular platform and channel for continued interaction, ensuring that individual oversight is complemented by robust institutional support.

Conversely, to highlight the importance of a secretariat, the "Incheon Meets Quebec" event in 2023 comes to mind. The event marked a period of considerable collaboration with the Incheon Metropolitan Government, with high-level officials present and resulting in signed agreements. However, without a standalone secretariat to provide ongoing oversight, these efforts primarily served to advance relations rather than evolve into a sustained, regular program.

Moreover, although a bit obvious, personal connections forged during these missions are equally vital. The involvement of this very initiative highlights this point! The experience of discovering LG NOVA through interaction with Dr. Sokwoo Rhee – both from our former occupation and subsequently exploring Quebec technologies at events like Collision, a startup event in Toronto with other officials from LG NOVA – underscores how individual networking can open doors to significant corporate partnerships. LG NOVA's approach to open innovation – by actively cultivating external ecosystems – perfectly aligns with this collaborative model.

Moreover, major international events such as the Consumer Electronics Show (CES) serve as prime examples of successful mission platforms. With Korean participants consistently ranking among the top attendees and a significant Canadian pavilion featuring companies from Quebec and Ontario in particular, CES provides an unparalleled arena for stakeholders from all priority research and innovation sectors to meet, network, and forge new partnerships. These missions collectively lay the groundwork for comprehensive and impactful cross-border AI collaborations. LG NOVA is also regularly present at CES, thus assisting in maintaining and furthering the relationship.

Furthermore, leveraging international networks and international joint research programs, such as the Horizon Europe program, also underscores the importance of international collaboration, especially with Canada and Korea joining the network as associate members. With Quebec's history of prioritizing international collaboration especially in terms of research & innovation, Quebec even has a dedicated office and an official tasked with conducting collaboration within the Horizon Europe framework in the Belgium area.

Besides the broad justifications and reasons for international collaboration in research and innovation, here are four specific cases.

Case Study 1: The Seoul AI Hub-MILA Scientist-in-Residence Program

The Importance of Direct Startup & Institutional Partnerships

A standout example of successful international collaboration in AI is the collaboration between Seoul AI Hub and MILA. Seoul AI Hub is an entity with the Seoul Metropolitan Government to incubate AI startups and MILA is the Quebec Artificial Intelligence Institute in Montreal. This program exemplifies how complementary strengths can be leveraged for mutual benefit as well as how programs are established to set the stage for a wider agreement, such as the Seoul-Quebec Cooperation Agreement.

The program was established not only to intertwine the two ecosystems, but also to bridge the gap between fundamental AI research – where MILA is a global leader – and applied AI development where South Korean startups comparatively excel. Recognizing that Korean AI startups often face challenges in accessing cutting-edge foundational research, and MILA benefits from real-world application contexts for its research, the partnership created a symbiotic relationship. The program's inception was part of broader Seoul-Quebec cooperation agreement, demonstrating the importance of political backing at the subnational level.

Each year, a carefully selected cohort of promising Korean AI startups is invited to MILA for an intensive 15-week residency. The selection process prioritizes startups based on their maturity and the specific technical challenges they aim to address, along with some consideration regarding their potential contribution and enrichment of the AI ecosystem.

During their residency, each company is paired directly with a MILA researcher (the scientist-in-residence), benefiting from bespoke guidance and collaboration on their specific AI projects that takes place face-to-face. This fundamental research interaction helps to enhance the scientific rigor of the startups' solutions as well as assist in resolving some of their AI challenges. The program's success is multifaceted; it deeply embeds Korean startups within Quebec's vibrant AI ecosystem, fostering invaluable connections with researchers, venture capitalists, and other industry players. By directly connecting applied AI challenges with fundamental research expertise, the program has the potential to significantly accelerate the development of more robust and innovative AI solutions. Moreover, the demonstrated success of this program as well as media coverage has garnered significant attention, prompting other government entities in Korea and beyond to explore similar collaborative models with Quebec, highlighting its replicability and impact.

Beyond research, this partnership provides comprehensive support, including assistance with market entry, establishing local subsidiaries in Quebec, and connecting participants with resources such as Investissement Québec, Montreal International, and Centech. This holistic support is crucial for the global expansion of startups, making this initiative a powerful testament to how a well-structured program at the subnational level can foster deep inter-ecosystem integration and drive concrete AI innovation. Due to its success, we are now in the second cohort of this program.

Thus, this partnership creates a complementary relationship where MILA’s deep scientific expertise enhances the startups’ solutions, while the startups provide MILA with real-world challenges and application contexts. Together, these ecosystems foster a symbiotic environment that accelerates innovation, improves scientific rigor, and drives practical AI solutions. Such targeted international collaboration demonstrates how combining distinct strengths can mitigate weaknesses, promote deeper integration, and generate impactful AI advancements on a global scale. And to be more specific, the top scientist-in-residences from MILA will be supported to come to Korea to benchmark, explore and establish additional collaboration, thus incentivizing the scientist-in-residences to collaborate even further, hence intertwining the two ecosystems ever closer!

Case Study 2: National & Subnational Joint Research Programs

The Importance of Institutionalized Funding

The collaborative spirit between Quebec, Canada, and South Korea extends to a broader landscape of national and subnational joint research programs, demonstrating a sustained commitment to fostering scientific and technological advancement across diverse fields beyond just AI. These programs are often underpinned by significant political and diplomatic frameworks. The foundation for these robust partnerships lies in high-level agreements. The Canada-Korea Free Trade Agreement (FTA) set a precedent for deeper economic ties, which paved the way for the Canada-Korea Science, Technology and Innovation (STI) Agreement. This STI agreement established a Joint Science and Technology Cooperation Committee (JSTCC), which meets regularly (e.g., the 4th JSTCC meeting was held in June 2024 in Banff, Alberta, on the margins of the Canada-Korea Conference on Science & Technology) to guide strategic cooperation in critical technologies, research integrity, and open science. Parallel to these national agreements, subnational cooperation agreements can be found, such as the Seoul-Quebec Cooperation agreement and ongoing discussions for a Daejeon-Quebec case (which will be highlighted later) to further strengthen direct institutional linkages.

Several programs actively facilitate these partnerships. The Korea Institute for Advancement of Technology (KIAT, with Korea's MOTIE) and Canada's National Research Council (NRC) – particularly through its Industrial Research Assistance Program (IRAP) – collaborate on various initiatives. The Canadian International Innovation Program (CIIP), delivered by Global Affairs Canada and NRC IRAP, offers Partnership Development Activities (PDAs) that are instrumental in facilitating connections, matchmaking Canadian and Quebec SMEs and entities with potential Korean partners for R&D projects. For example, recent delegations have focused on connecting Canadian AI in life sciences innovators with Korean pharmaceutical and healthcare organizations. And as priority sectors often shift from year to year, and although the overall sectors are generally considered deep tech, the thematic focus changes every year to adapt to the fast-paced environment of research and innovation. Thus, the past few years included delegations covering sectors from smart cities and quantum technology to semiconductors, which also align well with the priority sectors of Korea.

Similarly, PRIMA Québec, Quebec's research and innovation agency for advanced materials and quantum technology, has partnered with the National Research Foundation (NRF) of Korea on joint research calls. These programs typically require consortia comprising academic institutions and companies (often SMEs) from both Quebec and Korea, fostering a strong university-industry collaboration model in areas such as advanced materials and their intersection with ICT and AI. Such calls happen on a regular basis (i.e. annual), providing a solid reason for international partnerships, as such calls are only eligible with partnerships between the respective nations. Furthermore, Mitacs, a pan-Canadian national research organization, partners with NRF Korea through the Globalink Research Awards, enabling Canadian and Korean students and postdoctoral fellows to undertake research internships in each other's countries. This program is crucial for intertwining the two ecosystems by fostering early-career researcher mobility and strengthening long-term academic and scientific ties.

Beyond bilateral programs, both Quebec and Korean entities actively leverage major international platforms to facilitate collaboration. Participation and coordinated activities at events including the Canada-Korea Conference on Science & Technology, CES (Consumer Electronics Show), and Vivatech provide invaluable opportunities for networking, showcasing innovations, and forging new partnerships. Moreover, as also briefly touched upon, both Canada and the Republic of Korea have formally joined Horizon Europe, the European Union's flagship research and innovation program. Canada was associated with Pillar II of Horizon Europe in November 2023, and Korea followed in January 2025. This creates a powerful trilateral framework (EU-Korea-Canada) for collaborative projects across industrial, social, and environmental challenges, opening new avenues for joint R&D and resource sharing on a grander scale. These diverse programs and platforms underscore a comprehensive strategy for deepening STI cooperation, from foundational research to commercialization, at both national and subnational levels. Funding is always crucial for international collaboration!

The benefits of international collaboration in this context are a bit more obvious, beyond just the funding that is only open to those who apply as partners. Typically, partnerships form when there are complementary resources (such as sharing data) but also comparative advantages (such as shared use of compute power through partner nations). This symbiosis leads to the success of the project. So much so that one can say that if this partnership was not formed, the project/innovation would have never happened, thus the importance of international collaboration!

Case Study 3: The Case of the Daejeon Metropolitan Government

The Importance of the Public Sector/Government in Innovation

The indispensable role of government and public sector entities in fostering innovation, particularly in the realm of emerging technologies, cannot be overstated. Due to the inherent risks associated with novel advancements, sustained support from public bodies is often critical for research and innovation to thrive. The trajectory and process of strengthening relations with the Daejeon

Metropolitan Government serve as a compelling illustration of this principle.

Daejeon, the hub of scientific and technological expertise in South Korea, houses prominent institutions such as the Korea Advanced Institute of Science and Technology (KAIST), the Electronics and Telecommunications Research Institute (ETRI), and various other research institutes focusing on areas including quantum technology and advanced materials. This evidently is an initiative of the Korean government, in which they aimed to diversify away from the Seoul Metropolitan area to nurture other areas in Korea, and thus spearheaded a strategy to focus Science, Technology, and Innovation (STI) in the Daejeon Metropolitan area. Thus, engaging with Daejeon, which is deeply invested in these high-tech sectors, allows for focused and impactful partnerships.

And with Daejeon being the center of STI in Korea, the Daejeon Metropolitan Government has launched a significant STI initiative known as the Global Innopolis Network Initiative (GINI). Established to promote economic development and urban innovation through enhanced science and technology collaboration, GINI serves as a pioneering platform for inter-city cooperation, transcending mere exchanges among local governments. With Daejeon at its helm, GINI brings together a consortium of leading global cities – including Dortmund (Germany), Malaga (Spain), Montgomery County (Maryland, USA), and Seattle (USA) – to collectively address complex urban challenges, foster shared economic growth, and accelerate innovation capabilities through practical joint research, demonstration projects, and business development. And just recently, Quebec officially has expressed its wishes to collaborate within the framework of GINI.

To achieve this important collaboration, the relations with the Daejeon Metropolitan Government were fostered for several years, from partaking in their conferences and conducting B2B matchmaking with their entities, to creating visibility to showcase Quebec as the partner for research and innovation. As such, one of the key outcomes that led to this collaboration was during the Quebec quantum mission in February 2024, led by Quebec Quantique. One of the programs included conducting a KAIST-Quebec Quantum session, titled “Entanglement of World-Class Quantum Ecosystems”, which officials from Daejeon Metropolitan Government attended. From such fruitful endeavors, we were able to continue to develop the relationship, and thus were invited as a key VIP participant to Daejeon’s Inaugural Ceremony of GINI back in September 2024. Moreover, several roles were undertaken, such as speaking at their go-to-market seminar for the VIP reception hosted by the mayor. Such laid the foundations that led to an eventual Daejeon mayoral mission to Quebec, in which multiple agreements were signed from joint research agreements to agreements in quantum technology. Currently, a working group is being formulated to ensure that the collaboration not only moves forward, but also to ensure its success into the future.

Also, a key element of successful government-to-government collaboration involves, as mentioned previously, face-to-face meetings. As such, one of the initiatives being spearheaded by the Daejeon Metropolitan Government is to send their official(s) to Quebec for at least a year. This enables direct, in-person work on the ground, facilitating relationship building, benchmarking of best practices, learning from foreign innovation ecosystems, and effective reporting. Moreover, there is a strategic long-term benefit; carefully chosen officials who participate in these missions often ascend to higher leadership positions, fostering enduring goodwill and a positive memory of the Quebec, Canada and Korea collaboration. This ensures that established relationships continue to yield dividends over time. Furthermore, collaboration within institutional platforms remains important, such as the aforementioned GURS. Such arrangements, supported by governmental bodies, establish regular platforms and channels for ongoing interaction, ensuring that collaboration is not solely dependent on individual efforts but is supported by a robust, long-term institutional framework. Put simply, platforms and agreements allow for regular meetings which helps build long-term relations!

Case Study 4: Increasing Cross-Border Visibility

The Importance of Knowledge Sharing in Collaboration

Effective international collaboration is fundamentally predicated on the principle of knowledge sharing. Before entities can even consider partnering, they must be aware of each other's capabilities, expertise, and ongoing initiatives – or even each other's very existence – before they can collaborate. This crucial aspect of visibility ensures that potential collaborators can identify synergistic opportunities and build trust.

Various mechanisms are employed to facilitate this vital knowledge exchange. Reports from prominent organizations – such as the Science and Technology Policy Institute (STEPI) in Korea, a think-tank within the Prime Minister's Office – play a significant role by disseminating insights into global AI strategies, policy frameworks, and technological advancements. These reports allow nations and organizations to learn from successful models and avoid pitfalls, fostering a more informed and harmonized global AI landscape. Thus collaborations with STEPI are various, from co-hosting high-level sessions on AI to co-authoring publications covering innovation ecosystems around the world including Canada, with a focus on Quebec. Beyond formal reports, media coverage – such as features on Arirang news, Korea's national English news channel – amplifies the visibility of successful collaborative projects and initiatives, bringing them to a broader international audience. This media exposure is invaluable for showcasing achievements and attracting new partners, as quite often, people initiate contact not only immediately after exposure, but also through a build-up of exposure.

Furthermore, participation in and organization of conferences and forums – such as Korea AI Expo and the Canada-Korea Conference on Science & Technology (CKC) – serve as critical platforms for direct knowledge dissemination. These events provide opportunities for experts to present research findings, discuss policy implications, showcase innovative technologies, and engage in high-level dialogues that shape the future of AI. The impact of such visibility is concrete. Organizations have reached out directly to initiate partnerships after learning about Quebec-Canada's collaborative work through news or conference presentations, such as a startup acceleration program with Centech. This demonstrates the direct link between knowledge sharing and new collaborative ventures. In essence, by actively sharing knowledge, promoting visibility, and creating platforms for engagement, knowledge ensures that potential collaborators are well-informed, fostering a dynamic environment ripe for new and impactful partnerships. New innovations, research and technology that are already being employed or developed can find international partners that can bring value-added services.

Conclusion: A Blueprint for Global AI & Innovation Leadership through International Collaboration

The comprehensive experiences detailed between Quebec-Canada and South Korea offer a compelling innovation blueprint for fostering innovation, streamlining regulation, and standardizing guidelines through robust international collaboration. This blueprint is characterized by several interdependent elements that, when strategically implemented, collectively accelerate innovation advancement while ensuring its responsible deployment.

First and foremost, strategic alignment is paramount, involving the identification of complementary strengths such as Quebec's leadership in fundamental research and South Korea's prowess in applied development. This alignment extends to harmonizing national and subnational innovation strategies, creating a unified vision for cooperation. Second, the blueprint emphasizes structured programmatic support, manifested through well-defined initiatives such as scientist-in-residence programs and joint research calls. These programs provide clear pathways, dedicated funding, and essential logistical support, ensuring that collaborative projects have the necessary resources to thrive.

Third, multi-stakeholder engagement is crucial, requiring the active involvement of academia, industry (especially SMEs), and government bodies. This ensures that research is not only scientifically excellent but also commercially viable and responsive to societal and market needs. Fourth, facilitating mobility and exchange is vital, creating opportunities for researchers, students, and entrepreneurs to work across borders. This fosters invaluable knowledge transfer, builds long-term personal relationships, and cross-pollinates ideas between ecosystems.

Fifth, the blueprint underscores the importance of leveraging political and diplomatic frameworks. Utilizing established agreements such as Free Trade Agreements (FTAs), Science, Technology, and Innovation (STI) agreements, and subnational accords provides a stable foundation and high-level endorsement for scientific and technological cooperation, lending legitimacy and sustainability to joint endeavors. Finally, participating in global platforms is essential for expanding networks and influencing global AI governance discussions. Engagement with multilateral initiatives like GPAI and Horizon Europe, along with active presence at international conferences, allows for broader impact and the shaping of international norms.

This collaborative model, conceptually aligned with the open innovation strategies championed by global entities such as LG NOVA, demonstrates that a concerted, multi-pronged approach to international collaboration is not just beneficial, but absolutely essential for navigating the complexities and harnessing the full potential of artificial intelligence. By sharing knowledge, pooling resources, and aligning regulatory efforts, nations can accelerate AI and innovation while collectively working towards a future where AI serves humanity responsibly and ethically. The ongoing success of Quebec-Canada and Korea in this domain provides a powerful testament to this blueprint's efficacy and its potential to inspire future global partnerships.

Author (In order of contribution)

Andrew Yongwoo Lim, Research & Innovation Attache, Quebec Government in Seoul

Originally from Toronto, Andrew has been with the Quebec Government Office in Seoul since early 2020, where he leads initiatives in research and innovation, spanning sectors such as artificial intelligence, quantum technology, aerospace, and biotechnology. His role focuses on strengthening science, technology, and innovation collaboration between Quebec, Canada, and Korea. In addition to his official duties, Andrew holds several honorary positions. He serves as Chair of the International Public Cooperation Committee under the AX Association, affiliated with Korea’s Ministry of Trade, Industry and Energy, and is also President of the Yonsei Graduate School of International Studies Alumni Network. Before joining the Quebec Government Office, Andrew worked across a variety of sectors, including smart cities and broadcasting. In recognition of his contributions to strengthening ties between Seoul, Korea and Canada, he was awarded the honorary title of Seoul Honorary Citizen by the Mayor of Seoul in 2019.

Appendix C:
An AI Framework for Community-Centered Problem Solving

Author: John Barton

Context

In local communities, individuals often see problems firsthand — housing insecurity, healthcare gaps, food access, workforce barriers, or civic challenges — but they feel that they are tackling these issues alone. Without connection, well-meaning individuals may duplicate efforts, waste scarce resources, fragment advocacy, or weaken collective bargaining power. Over time, these missed opportunities leave motivated leaders frustrated or burned out.

- **Maria**, a single mother in a rural town, notices her neighbors struggling with housing insecurity but doesn't know about the nonprofit that quietly offers rental assistance.
- **James**, a retired miner, sees food access issues in his community but lacks the tools to connect with regional policy efforts that are already underway.
- **Lisa**, a community college nursing student, recognizes that her peers struggle to find affordable mental health resources on campus but is not aware of existing regional services or advocacy networks.

For under-resourced and marginalized communities, these barriers are heightened by structural inequities such as limited broadband access, transportation challenges, or language barriers. Problems linger, funding is misdirected, and community energy is lost. Yet the motivation is there; people want to act, and their resilience shows in repeated attempts to improve their communities.

As one resident put it, "I wanted to help, but I didn't know where to start." This voice captures the central gap: motivated individuals and groups want to act, but they "don't know what they don't know" and can't easily bridge from recognition to collective action. Highlighting this gap shows not only wasted effort but also missed potential for innovation, resilience, and sustainable local solutions. This reality sets the stage for the community-centered framework, which is designed to bridge divides and transform motivation into coordinated, equitable action.

Design Objectives (Our Approach)

The goal of this project is to create an AI-supported framework that empowers individuals and communities to move from isolation to connected action. The design objectives are:

Close knowledge gaps: Help individuals surface the vocabulary and framing they need while also providing access to best practices, models, theories, current research, thought leaders, and local experts. This ensures that both global and community knowledge inform solutions.

Provide tools, data, and measurements: Equip individuals with supports such as community needs assessments, participatory surveys, and local data analysis. Tie these tools to key performance indicators (KPIs) and other measures of success so progress can be tracked, compared, and refined over time.

Facilitate connections: Use AI-driven mapping to highlight local actors, resources, and initiatives so individuals quickly see who else is engaged on the same issues. This strengthens collaboration, reduces duplication, and aligns with safeguards against fragmentation (as noted below in the Risks & Mitigations section).

Support strategic planning: Combine questioning funnels and reflective prompts with data-driven insights to help communities anticipate risks, surface opportunities, and align actions with long-term goals. This integrates vocabulary and framing from knowledge gaps with evidence and measurement tools.

Promote equity and inclusion: Ensure marginalized voices are not only represented but also shape design, decision-making, and outcomes. Conduct equity audits of data and AI tools, apply accessibility standards, and embed participatory feedback loops so that power imbalances are actively addressed.

Enable structured iteration and continuous learning: Provide mechanisms to test ideas, capture feedback, and refine approaches. Feed these learnings back into future knowledge gaps, growth opportunities, and leadership development, supported by AI-driven tracking and transparent logs of what has been tried, adapted, and achieved.

Together, these objectives ensure that the community-centered framework is not just a process map, but a living system tied to the Framework and reinforced by Risks & Mitigations. They commit to transforming the experience of community members from isolated problem-bearers into connected co-creators of solutions, with AI serving as a guide, amplifier, and connector.

The Framework

The community-centered framework translates these objectives into a phased roadmap that guides individuals and communities from first recognition of a problem to co-created solutions. It is modular, transparent, and adaptable to different local contexts, with clear deliverables, explicit AI roles, and safeguards for governance. Each phase builds on the one before it, ensuring continuity, equity integration, and resilience against identified risks. It starts with the core functions of a minimally viable product (MVP) and carries right through to provisions that support scaling the resulting solution.

Phase 1: Core Functions (MVP)

- Guided intake process supported by AI natural language tools that help users articulate problems in their own words
- Question-first funnels that surface knowledge gaps and build shared vocabulary before suggesting resources
- Access to curated knowledge libraries with best practices, models, theories, and current research relevant to the issue
- Equity safeguards embedded early: inclusive intake design and attention to marginalized voices from the outset

Deliverables: Prototype intake tool, initial questioning funnel, curated resource library, equity-inclusive intake protocol, and early success user journey

Phase 2: Connection & Iteration

- AI-driven mapping of local actors, organizations, and initiatives to reveal who is already engaged and where overlaps exist
- Tools for community needs assessments and participatory surveys to generate shared data, with AI analytics highlighting inequities, gaps, and duplication
- Iteration tracking that logs solutions tried, revised, and refined, including AI-supported summaries of what worked, why, and how risks were mitigated
- Built-in equity checkpoints and alignment audits to ensure marginalized groups are shaping solutions, and not just represented in them

Deliverables: Community survey templates, annotated iteration logs, reframing prompt library, pilot use case scenarios (e.g., food bank vs. co-op decision), and interim alignment audit report

Phase 3: Scaling & Governance

- Infrastructure for cross-community knowledge sharing, creating a collective knowledge base of problems, ideas, and solutions while preserving local nuance
- Governance safeguards including rotating leadership, alignment audits, stress tests, and escalation protocols to ensure inclusion, prevent power capture, and sustain accountability
- Scenario modeling tools for AI-assisted exploration of trade-offs, cascading impacts, and long-term risks, with multimodal accessibility for diverse users
- Transparency mechanisms such as dashboards, feedback logs, and public validation modules to maintain trust

Deliverables: Oversight and escalation playbook to aid with knowledge transfer, governance dashboard, scenario modeler, visualization kit, public validation module, and annual equity audit

The community-centered framework positions AI as a guide and connector — a tool to surface blind spots, clarify opportunities, provide tradeoff analysis, and amplify community voices — while leaving judgment and ownership firmly with people and communities. Built-in feedback loops ensure learning flows across all phases, feeding back into new knowledge gaps, growth, and leadership development. This alignment with Design Objectives and Risks & Mitigations ensures a resilient, equitable, and scalable approach to community problem-solving.

Illustrative Example(s)

To show how the community-centered framework could operate in practice, consider the following scenarios.

Housing Stability

Maria identifies housing insecurity in her neighborhood. The AI guides her through a survey tool to capture local data, then maps organizations addressing rental assistance and highlights regional best practices in land trusts. With reframing prompts and tradeoff analysis, Maria and her neighbors clarify options between short-term rental assistance and longer-term land trust models.

Outputs: Local housing survey, reframed options, and advocacy toolkit

Outcomes: More evidence-based advocacy, reduced duplication of effort, and strengthened collaboration with regional nonprofits

Benefits: Improved housing stability, increased leverage for community voices, and clearer pathways for funders and policymakers

Food Security

James uses the intake process to clarify his concern about food access. The AI surfaces mobile food pantries and community-supported agriculture, as well as highlights a nonprofit piloting a food co-op. Using scenario modeling, James and local partners compare tradeoffs between expanding food bank access and piloting a co-op.

Outputs: Food access map, scenario model comparing options, and resource directory

Outcomes: Improved coordination among community groups, increased visibility of marginalized voices in food policy, and fewer duplicated initiatives

Benefits: Stronger collaboration networks, better alignment with policy decisions, and scalable models for funders

Campus Mental Health

Lisa, a community college nursing student, notices that her peers struggle to find affordable mental health resources on campus. The questioning funnel helps her focus on this issue, while AI-supported mapping reveals underused regional clinics and highlights peer mentoring programs in other communities. With support from visualization tools, Lisa and her peers develop a student-led mentoring program linked to local providers.

Outputs: Peer mentoring program design, clinic connection map, and communication materials

Outcomes: Elevated student voices, stronger collaboration between campus and community health partners, and measurable indicators of improved access to care

Benefits: Reduced strain on existing health providers, more equitable access to mental health resources, and replicable models for other campuses

Civic Engagement

A local neighborhood association wants to improve voter participation. The AI provides access to best practices from other communities, highlights local experts, and uses equity audits to surface barriers faced by marginalized residents. Through participatory survey tools, the group identifies transportation and information gaps.

Outputs: Community survey results, equity audit findings, and multilingual outreach plan

Outcomes: Partnerships with civic organizations, creation of ride-share programs, and multilingual voter education

Benefits: Measurable increases in voter turnout, strengthened democratic participation, and models for inclusive civic engagement

These vignettes show how people move from uncertainty to action, supported by AI tools that provide vocabulary, data, tradeoff analysis, and connections. Each illustrates how outputs lead to outcomes and benefits, reinforcing the community-centered framework's commitment to equity, collaboration, and sustainable change across domains.

Outputs, Outcomes, & Benefits

The community-centered framework is designed to deliver tangible products, measurable changes, and clear value for stakeholders. Outputs are the tools produced, outcomes are the changes created, and benefits are the value distributed. Together, they mirror the deliverables noted in the Framework section and reinforce the safeguards in Risks & Mitigations.

Outputs (What is produced):

- Intake tools and questioning funnels
- Curated knowledge libraries with best practices, models, theories, and current research
- Community needs assessment templates and participatory survey tools
- Iteration logs capturing solution trials, revisions, and feedback
- Dashboards mapping local actors, initiatives, and resources
- AI-enabled reframing prompt libraries, iteration analytics, and tradeoff modeling tools
- Visualization kits and governance dashboards for oversight and transparency
- Equity audit reports and participatory governance charters to embed fairness and accountability

Outcomes (What changes):

- Increased collaboration between individuals, groups, and organizations
- Reduced duplication of effort and wasted resources (e.g., 20% reduction in overlapping initiatives within pilot regions)
- Improved visibility of marginalized voices in problem-solving (measured by representation in decision-making bodies)
- More inclusive and evidence-informed decision-making, backed by both quantitative and qualitative data

- Stronger local capacity for iterative learning and adaptation
- Reduced inequities in access to resources and opportunities (e.g., increased participation of marginalized groups in 75% of projects)
- Greater alignment between grassroots needs and policy decisions
- Enhanced sustainability of community-driven solutions, with feedback loops ensuring long-term adaptation

Benefits (Who gains what value):

Community members: Access to guidance, partnerships, advocacy tools, stronger leverage in negotiations, and tangible improvements in housing, food, and healthcare stability

Nonprofits and local groups: Stronger collaboration networks, efficient use of resources, clearer alignment with funders, and reduced burnout from duplication

Policymakers: Better data, clearer needs assessments, tested solution models, scalable insights for governance, and early detection of risks or inequities

Funders: Stronger ROI through evidence-based initiatives, reduced risk, and clearer impact metrics tied to KPIs and outcomes

Developers and operators of AI tools: Legitimacy through equity-centered design, opportunities for refinement in real-world contexts, and continuous improvement validated by community use

Educators and researchers: Access to case data, models, participatory design lessons, and longitudinal insights that can inform future innovation

This separation ensures clarity; outputs lay the foundation for outcomes, which generate broad, shared benefits. In turn, the community-centered framework becomes actionable, measurable, and equitable.

Risks & Mitigations

Implementing a community-centered, AI-supported framework raises both technical and social risks. Anticipating and addressing them is essential to building trust, ensuring equity, and sustaining momentum. Each risk is paired with its consequence, mitigation, and deliverables.

1. Risk: Over-reliance on AI guidance, leading to diminished human judgment or community ownership

Consequence: Communities may lose decision-making power, resulting in dependency on technology and erosion of local leadership capacity.

Mitigation: Design AI to prompt reflection and questioning, not just provide answers. Measure success by tracking the proportion of decisions made through community-led processes, ensuring ownership remains local.

Deliverables: Community-led decision logs, reflection prompts integrated into AI interface, and evaluation reports on local ownership

2. Risk: Bias in knowledge libraries, data inputs, or model outputs that could reinforce inequities

Consequence: Marginalized groups may be further excluded, reinforcing systemic inequities in problem-solving and outcomes.

Mitigation: Apply equity audits, alignment audits, and drift detection to knowledge libraries, data inputs, and outputs.

Deliverables: Regular equity audit reports, alignment review summaries, and independent third-party audit certifications

3. Risk: Model drift or misalignment between local data realities and global models

Consequence: AI recommendations may become irrelevant or harmful if they no longer reflect local conditions.

Mitigation: Conduct continuous monitoring, drift detection, and scenario stress testing to identify and correct misalignment early.

Deliverables: Alignment audit dashboards, monitoring tools, and scenario stress test reports

4. Risk: Fragmented governance or lack of accountability in managing shared tools

Consequence: Without accountability, governance may become inconsistent, leading to misuse of tools and loss of community trust.

Mitigation: Establish transparent governance with rotating leadership, clear accountability, and escalation authority when disputes or inequities arise.

Deliverables: Participatory governance charter, rotation schedule documentation, and escalation protocols

5. Risk: Power imbalances where stronger organizations dominate weaker voices

Consequence: Smaller or marginalized groups may lose influence, perpetuating inequities and reducing diversity of solutions.

Mitigation: Build governance safeguards with equity checks, independent third-party reviews, and participatory processes to ensure marginalized voices are included. Measure inclusion by representation metrics in decision-making bodies.

Deliverables: Equity check reports, representation metrics, and independent review findings

6. Risk: Accessibility gaps, such as limited broadband or device access in rural or under-resourced communities

Consequence: Communities may be unable to access or benefit from the framework, widening the digital divide.

Mitigation: Provide plain-language explanations of how the AI works, design for low-resource settings, and ensure outputs are accessible in multiple formats (per Web Content Accessibility Guidelines – WCAG).

Deliverables: Accessibility compliance reports, plain-language guides, and low-bandwidth interface designs

7. Risk: Privacy concerns about sharing local problems, resources, and solutions

Consequence: Sensitive community information could be exposed or misused, leading to harm or mistrust.

Mitigation: Build in General Data Protection Regulation (GDPR)-style consent checkpoints so communities control what is shared, how it is used, and when information flows across networks.

Deliverables: Community consent protocols, privacy compliance reviews, and consent audit logs

8. Risk: Resistance from stakeholders skeptical of AI in community problem-solving

Consequence: Stakeholders may disengage, block adoption, or undermine the legitimacy of the framework.

Mitigation: Ensure transparency with oversight dashboards, plain-language communication, and participatory validation.

Deliverables: Public-facing dashboards, plain-language communication materials, and validation session reports

9. Risk: Data security breaches or malicious misuse of community data

Consequence: Breaches could cause material harm, erode trust, and expose communities to external exploitation.

Mitigation: Implement strong encryption, role-based access controls, and independent security audits.

Deliverables: Annual security compliance certification, encryption audit reports, and access control logs

10. Risk: Sustainability gaps if funding or support lapses after pilots

Consequence: Programs may collapse once pilots end, wasting resources and leaving communities worse off.

Mitigation: Tie deliverables to long-term KPIs, require funder commitments to ongoing equity audits, and establish reinvestment mechanisms.

Deliverables: Sustainability and reinvestment plan, KPI tracking reports, and funder commitment agreements

11. Risk: Legitimacy risks if AI outputs conflict with community knowledge or norms

Consequence: Communities may reject AI tools altogether, undermining adoption and collaboration.

Mitigation: Create participatory review boards to validate outputs against local expertise.

Deliverables: Validation reports, review board meeting records, and community alignment summaries

By naming risks, identifying consequences, embedding mitigations, and tying them to deliverables, the community-centered framework strengthens resilience, fairness, transparency, and trust among stakeholders while reinforcing that AI is a tool under community ownership.

Next Steps (Scaling Pathway)

Moving from design into implementation, the community-centered framework follows a staged pathway that balances small-scale testing with long-term vision. Each stage includes concrete deliverables, explicit AI auditing, and stakeholder engagement to ensure accountability. Time markers, metrics, and safeguards ensure the pathway is measurable, resilient, and tied to risks and mitigations.

Immediate Next Steps (0–6 months)

Develop and release a prototype intake and questioning tool

Deliverable: Prototype report with annotated user journey and initial feedback

Metric: At least two successful prototype tests with diverse users

Partner with one to two communities to co-design and validate the process

Deliverable: Pilot co-design agreements and community validation notes

Metric: Representation of marginalized groups in pilot design teams

Conduct AI equity, accessibility, and usability audits during pilots

Deliverable: Equity audit report, usability test findings, and accessibility compliance checklist

Metric: 100% of pilots reviewed against Risks & Mitigations safeguards

Gather feedback from participants, organizations, and external reviewers

Deliverable: Consolidated feedback log with recommendations for iteration

Metric: Documented changes made based on participant input

Near-Term Scaling (6–18 months)

Expand pilots regionally with diverse communities, ensuring variation in demographics and contexts

Deliverable: Regional pilot summary with comparative analysis

Metric: At least five regional pilots completed with equity audits

Build a library of use cases and refine tools based on lessons learned

Deliverable: Public-facing use case library and tool refinement roadmap

Metric: Library includes a minimum of 10 validated use cases

Formalize governance with rotating leadership and community representation

Deliverable: Draft governance charter and stakeholder engagement plan

Metric: Governance boards include at least 40% representation from marginalized groups

Long-Term Pathway (18–36 months and beyond)

Establish infrastructure for cross-community knowledge sharing, preserving local nuance while scaling insights

Deliverable: Knowledge-sharing platform prototype and participatory feedback integration plan

Metric: 80% of participating communities report preserved local adaptation

Partner with funders, policymakers, and national organizations to align community-driven solutions with broader systems

Deliverable: Partnership agreements and policy alignment brief

Metric: At least three formalized partnerships with funders and policy bodies

Ensure scalability without losing local adaptation through continuous participatory feedback loops

Deliverable: Annual feedback report and adaptation log

Metric: Demonstrated adjustments made annually in response to community feedback

The pathway emphasizes co-design, transparency, feedback-driven iteration, and equity at every stage. By embedding metrics, safeguards, and stakeholder roles, the community-centered framework ensures growth that is sustainable, accountable, and community-owned.

Lessons Learned (Design Process)

Even at the design stage, important lessons have emerged. These lessons are expressed as commitments that directly inform outputs, safeguards, and the community-centered framework.

AI must remain a guide. We will ensure AI supports reflection and surfacing options rather than prescribing answers, keeping communities in control of decision-making.

Equity requires design. We will embed safeguards such as audits, consent checkpoints, and inclusion-focused stress tests as core outputs to avoid reinforcing inequities.

Community ownership is essential. We will keep leadership with communities, positioning AI as a support tool that strengthens their capacity without replacing their judgment.

Iteration builds trust. We will implement feedback loops and visible adaptation, so communities see responsiveness to their needs, strengthening legitimacy and engagement.

Transparency requires tools. We will deliver dashboards, feedback logs, and equity audits as non-negotiable mechanisms for accountability and confidence among stakeholders.

Data must be trustworthy, accurate, and contextualized. We will ensure that data is collected ethically, validated against local knowledge, and interpreted with care. Measurements will be tied to KPIs and safeguards to provide clarity and accountability without distortion, ensuring that community priorities are informed by evidence rather than reshaped by it.

Keep tools accessible. We will design for low-resource settings and apply accessibility standards to ensure participation across digital divides.

Scalability requires nuance. We will preserve local context and adapt solutions without diluting grassroots voices, even as platforms scale across communities.

Stakeholder engagement matters. We will provide tailored communication and shared governance structures, so funders, policymakers, and community members remain aligned and benefit mutually.

These lessons, grounded in early exploration and prior community experience, directly inform the community-centered framework's outputs and safeguards. They underscore the need for transparency, adaptability, accountability, and positive engagement across both technical and social dimensions.

Conclusion

An AI-supported, community-centered framework can empower communities to move from isolation to connection, and from uncertainty to action. By closing knowledge gaps, facilitating connections, embedding equity, and integrating safeguards, the community-centered framework ensures that individuals like Maria, James, and Lisa are not left to navigate challenges alone. Instead, they become part of a collective process that values judgment, ownership, and learning while producing tangible outputs, measurable outcomes, and shared benefits.

The journey ahead requires careful pilots, strong governance, transparent auditing, and ongoing reflection. Success depends on collaboration among communities, nonprofits, funders, policymakers, developers, educators, and researchers: each sharing responsibility for equity-centered outcomes. By uniting technical safeguards such as dashboards, audits, and scenario modeling with community-driven ownership, the community-centered framework demonstrates not only a practical system for problem-solving but also a new model for inclusive, accountable AI.

The foundation is clear: communities already have the will to act. With the right support, tools, and partnerships, that will can drive sustainable, equitable change that benefits everyone, setting a standard for trustworthy, equity-centered AI systems that foster resilience, innovation, and long-term trust.

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.

Authors and Contributors

(In alphabetical order)

Adrien Abecassis, Executive Director for Policy at Paris Peace Forum

Adrien Abecassis is a French career diplomat and a former senior advisor to the President of France (2012–2017). He has held academic fellowships at Harvard University and UCLA and is currently serving as Chief Policy Officer of the Paris Peace Forum.

Johnny Aguirre, Ekrome Founder

Johnny is an experienced professional across various industries and technologies, currently focused on building a startup that provides AI solutions for small businesses.

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.

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

Olivier Bacs, CTO and co-founder, Bendi

Olivier Bacs is the CTO and co-founder of Bendi, where he builds AI-powered tools that help companies gain visibility into their supply chains and collaborate more effectively with suppliers. His work combines geospatial analysis, automation, and natural language processing to uncover hidden risks while making complex compliance processes easier to navigate. Olivier is especially focused on decentralized and ethical approaches to AI, ensuring that technology enhances trust, equity, and resilience across global value chains.

Taylor Black, Director AI & Venture Ecosystems, Microsoft

Taylor Black is Director of AI & Venture Ecosystems in Microsoft's Office of the CTO, where he designs and leads cross-company initiatives that integrate innovation, product development, and community engagement. With 19+ years of experience launching and scaling ventures across enterprise, deep tech, and social ecosystems, he brings a multidisciplinary background as a developer, educator, lawyer, entrepreneur, and venture builder. He mentors and invests in early-stage startups through networks such as Conduit Venture Labs and Fizzy Ventures. Taylor also helps shape Catholic University of America's new institute at the intersection of AI, innovation, and human flourishing.

Micah Boster, Principal, Nighthawk Advisors

Micah Boster is the founder and Principal at Nighthawk Advisors, where he works with early-stage technology companies on execution, AI strategy, and positioning. Previously, he spent eight years at Google and over a decade as an executive at several NYC-based startups. He holds a BS in Symbolic Systems from Stanford and an MBA from INSEAD.

Dr. Mathilde Cerioli, Chief Scientist, everyone.ai

Dr. Mathilde Cerioli is the Chief Scientist and cofounder of everyone.ai, a nonprofit dedicated to anticipating and educating on the opportunities and risks of AI for children. She holds a Ph.D. in Cognitive Neuroscience and a master's degree in Psychology, with a research focus on how AI intersects with cognitive and socioemotional development in children, adolescents, and young adults. In May 2024, she published the influential report Child Development in the AI Era, examining the potential impact of emerging technologies on cognitive and socioemotional development.

Carolyn Eagen, Founder, Kinstak

Carolyn Eagen is the Founder and CEO of Kin Technologies and Kinstak, an AI-native platform pioneering private digital legacy management and decentralized digital asset manager for families and SMBs. She brings over 20 years of leadership in product strategy and innovation. Carolyn is passionate about building ethical, user-centered systems that unlock access, equity, and long-term resilience in the age of AI.

Sarah Ennis, Co-Founder, AgentsGEO.ai

Sarah Ennis is a Fortune 500 trusted advisor specializing in advanced technology innovation, with over two decades of experience leading groundbreaking AI solutions at scale. Globally recognized for her expertise in artificial intelligence, she designs and implements bespoke emerging technology products across industries. She is also the co-founder of AgentsGEO.ai, a patent-pending platform that helps brands monitor and improve their visibility in the AI ecosystem and deploy AI agents, ensuring they are discoverable and recommended by tools like ChatGPT, Gemini, and others through its proprietary GEOScorer™ technology. In addition, Sarah contributes part-time to Northeastern University's Master of Digital Media programs in AI, preparing the next generation of technologists and creative leaders. Her work bridges Silicon Valley innovation with global impact, and she is a distinguished member of the American Society for AI and contributor to the OpenAI Forum.

Annie Hanlon, Co-Founder/Partner, Playbook PLBK

Annie Hanlon is Co-Founder/Partner of Playbook PLBK, focused on AI, storytelling, and innovation. An accomplished entertainment executive and award-winning producer, she is recognized for her impact at Netflix, Lytro, and Here Be Dragons. Annie is also Co-Founder of Playbook AIR, a platform designed to capture and verify human authorship in GenAI workflows, providing clear documentation to support copyright, protect creators, and ensure accountability. She is a sought-after industry speaker, a member of the Visual Effects Society, the Academy of Television Arts & Sciences, and a board member of the Infinity Festival. Annie is a 2024 graduate of Space Camp (Huntsville, AL).

Stephanie Hockenberry, MBA, Growth & Retention Manager, Ohio County, WV, Ohio County Development Authority

Stephanie Hockenberry serves as Growth & Retention Manager for Ohio County, WV, under the Ohio County Development Authority, where she champions initiatives that connect emerging talent with local business opportunities to foster long-term economic vitality. Her work blends strategic outreach with heartfelt community engagement, recruiting both residents and entrepreneurs to invest in the County's future. Through youth pipeline development, collaborative marketing, and support for residential & business ecosystems, she positions Ohio County, West Virginia, as a welcoming and resilient place to plant new roots, build meaningful connections, and live your best life.

Christina Lee Storm, Co-Founder/Partner, Playbook PLBK

Christina Lee Storm is Co-Founder/Partner of Playbook PLBK, a consultancy specializing in AI, innovation and storytelling, and an award-winning producer and executive recognized for her work at major studios including Netflix, Universal, and DreamWorks Animation. She is also Co-Founder of Playbook AIR, a platform built to document and authenticate human authorship within GenAI workflows. Christina holds many leadership roles including Governor of the Television Academy's Emerging Media Programming peer group, Lead for the Academy's Responsible AI & Production Standards Working Group, sits on the Producers Guild of America's Production Innovation Task Force, and is a newly inducted member of the Academy of Motion Picture Arts & Sciences, Class of 2025.

Andrew Yongwoo Lim, Research & Innovation Attache, Quebec Government in Seoul

Originally from Toronto, Andrew has been with the Quebec Government Office in Seoul since early 2020, where he leads initiatives in research and innovation, spanning sectors such as artificial intelligence, quantum technology, aerospace, and biotechnology. His role focuses on strengthening science, technology, and innovation collaboration between Quebec, Canada, and Korea. In addition to his official duties, Andrew holds several honorary positions. He serves as Chair of the International Public Cooperation Committee under the AX Association, affiliated with Korea's Ministry of Trade, Industry and Energy, and is also President of the Yonsei Graduate School of International Studies Alumni Network. Before joining the Quebec Government Office, Andrew worked across a variety of sectors, including smart cities and broadcasting. In recognition of his contributions to strengthening ties between Seoul, Korea and Canada, he was awarded the honorary title of Seoul Honorary Citizen by the Mayor of Seoul in 2019.

Jess Loren, CEO, Global Objects

Jess Loren is the CEO of Global Objects and a leader in AI-driven 3D scanning. She serves on the Television Academy's Governors Board for Special Visual Effects and the Board of Managers for the Los Angeles Visual Effects Society. A Producers Guild of America member, Jess is also a proud wife and mother of three.

Jade Newton, Managing Director, Vertical Space Magazine

Jade Newton is a tech education professional, B2B consultant, and entrepreneur. With a background in linguistics and AI data, Jade advocates for the ethical and responsible development of emerging technologies, bridging the gap between human centered interaction and AI. Jade is also the Managing Director of Vertical Magazine, an online tech publication, and is deeply passionate about driving social impact through technology.

Refael Shamir, Founder, Letos

Refael Shamir, is a seasoned entrepreneur in the field of affective neuroscience, and is working towards introducing a new medium for gaining insights into spontaneous human reactions based on seamless integrations of devices in everyday environments. Refael is also a renowned speaker having presented his learnings in highly acclaimed conferences such as NVIDIA GTC, MOVE Mobility Re-Imagined, NeurotechX, among others.

Svetlana Stotskaya, Global Executive Consultant, Mentor

Svetlana Stotskaya is an award-winning global executive consultant and mentor. Svetlana is an active mentor for entrepreneurs at Techstars, Founder Institute, and Startup Wise Guys: the largest B2B accelerator in Europe. Featured in the World IP Changemakers' Gallery by the World Intellectual Property Organisation, she served on a jury board for international award competitions in innovation and technology.

Wenli Yu, CEO, Archimedes Controls Corp.

Wenli Yu, a serial entrepreneur and seasoned business executive of technology companies, currently serves as CEO of Archimedes Controls Corp., a Silicon Valley based technology innovator, designer and manufacturer of AI-powered industrial IoT solutions for industrial controls and automation, cloud and edge data centers, agriculture and environmental and transportation marketplaces.

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