all tech is
human

2025

Responsible AI Impact Report

Urgent risks, emerging safeguards, and public interest solutions impacting society

Introduction

How do we ensure that the rapid development of AI is more considerate of harms and the public interest?

In our inaugural *Responsible AI Impact Report*, All Tech Is Human (ATIH) aims to reveal our most urgent risks, emerging safeguards, and public-interest solutions, and provide a roadmap for how we will shape how AI impacts society in the year ahead. We examine the state of Responsible AI (RAI) throughout 2025 and highlight what we consider to be some of the most impactful contributions made by civil society organizations this year to enrich this broad and dynamic field.

We believe the Responsible AI field can only thrive if we effectively tackle the complex challenges at the intersection of technology and society. When we refer to “Responsible AI,” we mean AI that is well-regulated and guard-railed, governed and assured (documented, standardized, and benchmarked with relevant measurements), and assessed, evaluated, and red-teamed.

As we outlined in our recent Responsible Tech Guide (2025), our organization believes in a human-centered future that values our agency in desired outcomes and rejects tech determinism. As such, we are focused on elevating AI models that do as little harm as possible, for use cases in which risks have been carefully considered and meaningfully mitigated; and ethically deployed AI, in which lofty principles are operationalized with grounded KPIs.

This *Responsible AI Impact Report* highlights the growing focus on Public Interest AI that is of, by, for, and in service to the people. This Public Interest AI should be applied to humanity’s most pressing challenges and enable us to reimagine what a better tech future entails. This report also explores a future in which Public Interest AI is developed on public infrastructures for an AI-literate society.

At the heart of the years ahead lies a defining question: who determines the purpose of AI and the kinds of lives it will shape?

“An AI-enabled future grounded in human dignity depends on institutions that can govern powerful technologies with a commitment to the public good. When societies build this kind of civic architecture, people gain the ability to direct technological development rather than be shaped by it. This report highlights how that architecture is emerging through civil society’s work: accountable standards, rigorous evaluations, provenance systems that reinforce information integrity, and safeguards that respond to prominent harms. Together, these efforts outline a path toward a digital world that strengthens democratic agency and reflects the values we choose to uphold.”

Vilas Dhar
President, The Patrick J. McGovern Foundation

Acknowledgements & Contributions

Lead Author

Rebekah Tweed

Strategic & Editorial Support

David Ryan Polgar
Sherine Kazim
Sandra Khalil

Contributions & Feedback

Alisar Mustafa
Jen Weedon
Justin Hendrix
Leah Ferentinos
Merve Hickok
Michelle Shevin
Dr. Nathan C. Walker
Theodora Skeadas

Featured Organizations

- Ada Lovelace Institute
- AI, Algorithmic and Automation Incidents and Controversies (AIAAIC)
- AI Now Institute
- AI + Planetary Justice
- AI Risk and Vulnerability Alliance
- The Alan Turing Institute
- Algorithmic Justice League
- All Tech Is Human
- Aspen Digital
- Atlantic Council (DFRLab)
- Better Images of AI
- Center for AI and Digital Policy (CAIDP)
- Center for Democracy & Technology
- Centre for the Governance of AI
- Centre for International Governance Innovation
- Coalition for Content Provenance & Authenticity
- Collective Intelligence Project
- Common Sense Media
- Consumer Reports
- Data & Society Research Institute
- Distributed AI Research Institute
- Equal AI
- Estampa
- EvalEval Coalition
- Fast Forward
- Federation of American Scientists
- Forecasting Research Inst.
- The Future Society
- Humane Intelligence
- Information Professionals Association
- Japan AI Safety Workshop
- Knight First Amendment Institute at Columbia
- Metrology Working Group
- MIT AI Risk Initiative
- MITRE
- MLCommons
- Montreal AI Ethics Institute
- Mozilla Foundation
- The OpenFold Consortium
- Partnership on AI
- The Patrick J. McGovern Foundation
- Project Liberty
- Public Interest AI Project
- Reboot
- Responsible AI Collaborative
- ROOST
- Singapore AI Safety Institute
- The Stanford Institute for Human-Centered AI
- The Tech We Want
- UNESCO
- Weval
- WITNESS

This report is made possible thanks to the generous support of The Patrick J. McGovern Foundation.

Executive Summary

This report examines the Responsible AI ecosystem in 2025, highlighting the field's most impactful resources and tracing its contributions toward developing concrete governance, assurance, and public-interest infrastructure to support the adoption of Responsible AI across sectors.

Five Key Takeaways

- **Responsible AI is shifting from principles to practice**, with civil society leading the development of concrete governance tools. This includes standards, benchmarks, audits, and red-team methods that turn abstract values into verifiable, lifecycle-based evidence.

- **AI risks are intensifying across safety, security, privacy, fairness, and information integrity**, driven by increasingly capable frontier and agentic systems that create new failure modes and exacerbate existing harms. The report examines synthetic media and fraud to biometric surveillance and AI companion dependency.

- **Public AI is emerging as a powerful alternative to proprietary AI**, emphasizing shared compute, community-governed datasets, open safety tools, and public-interest institutions so that universities, nonprofits, governments, and communities, rather than solely corporations, can shape the AI ecosystem.

- **Societal impacts such as labor displacement, climate burdens, economic concentration, and threats to democracy** require whole-of-society governance, including stronger rights-anchored regulations, global standards alignment, and investment in nonprofit capacity and public oversight.

- **A central challenge for 2026 is determining who decides what AI is for**, with our report calling for expanded RAI literacy, stronger information-integrity defenses, clearer safeguards for high-risk uses (e.g., AI companions, synthetic media), and a coordinated push toward accountable, public-benefit AI futures.

The Audience for the Responsible AI Impact Report

This report has been written for anyone committed to ensuring AI is well-regulated, less harmful, publicly accountable, and aligned with societal needs. Similar to every All Tech Is Human initiative, we have designed it for a multistakeholder, cross-sector audience. Intended audiences include:

- AI governance practitioners in industry
- Civil society organizations
- Researchers and academics
- Government regulators and policymakers
- Public-interest technologists
- Educators, students, and RAI newcomers
- Philanthropic funders supporting Responsible Tech

Civil society organizations make the case that long-term AI leadership requires advancing people-first alternatives centered on public well-being, shared prosperity, and democratic accountability rather than frontier model exceptionalism. In parallel, new work on model risks, safety, security, privacy, bias and fairness, human rights, labor, climate, and economic concentration reveals that frontier and agentic systems are amplifying existing harms while creating new ones, from scaled fraud and synthetic media to AI-enabled bio-risk and energy-intensive infrastructure build-outs.

This report tracks the maturation of AI assurance: the emerging ecosystem of standards, benchmarks, evaluations, audits, and red-teaming that translate high-level principles into testable claims and verifiable evidence. It synthesizes advances in documentation standards for models and datasets, community-aligned benchmarks, sector-specific evaluation frameworks, and evolving practices for third-party audits and participatory red-teaming. This highlights a shift away from one-off disclosures and leaderboard metrics toward continuous, lifecycle-oriented evidence pipelines that regulators, procurers, and journalists can actually use. Here, civil society is not only critiquing inadequate safeguards but also designing templates and resources like playbooks, benchmark programs, and measurement guidelines that public and private actors can adopt at scale.

This report also documents the decisive turn toward Public Interest AI infrastructure, tracing emerging blueprints for Public AI: shared, non-proprietary stacks of compute, datasets, models, and tools that meet minimum criteria of public access, public accountability, and durable public goods. We connect these ideas to concrete developments, such as public-compute initiatives (including NAIRR), open and community-governed datasets, open-source safety and trust-and-safety tooling, and philanthropic efforts that reorient funding toward Public Interest use cases and nonprofit capacity. Taken together, these efforts shift the focus from “making private AI safer” to building alternate power centers and shared capabilities so that universities, civil society, and public agencies can meaningfully shape and scrutinize AI systems.

Looking to 2026, this report calls for collective efforts throughout the RAI ecosystem to prioritize defending and strengthening rights-anchored regulation amid deregulatory headwinds; scaling assurance practices that produce audit-ready evidence; closing capacity gaps for nonprofits and public agencies; and investing in public AI infrastructure (compute, data, tools, and institutions) so that communities, not just corporations, can shape AI, while treating narrative and cultural work as core governance infrastructure, not a side project.

We close out this report with our 2026 RAI roadmap: we intend to deploy AI in ways that strengthen, rather than hollow out, the Responsible Tech ecosystem; to accelerate the development of shared Public AI infrastructure; and to expand RAI literacy so that more people can understand, challenge, and reshape the systems impacting our lives.

The *Responsible AI Impact Report* outlines how the Responsible AI ecosystem has matured into a more coordinated, evidence-driven, and public-interest–oriented field. Civil society organizations play a central role in shaping regulatory debates, expanding global governance frameworks, and surfacing real-world harms across safety, security, privacy, fairness, labor, climate, and democratic integrity.

Lastly, our report looks to the year ahead in 2026. As the report emphasizes, 2026 will be a decisive year for determining who shapes AI’s trajectory. *Will it evolve to deepen surveillance and dependency or serve as a public good aligned with societal interests?* Our organization is committed to the latter.

Do you have feedback about the Responsible AI Impact Report, or ideas for future reports? Please write to us at hello@alltechishuman.org.

Contents

Well-Regulated AI

Our reflection starts top-down, at the level of nation-states and areas of global cooperation. The long slow arc of regulatory progress has reached the implementation phase in the EU (maybe): how is the compliance process unfolding within companies? To what extent might U.S. frontier model developers be politically shielded from enforcement and fines from abroad? Large-scale regulatory safeguards are taking root in the EU, while there are unyielding (for now) federal regulatory headwinds in the U.S., ostensibly shifting the active battles to the states (and the latent battles behind the closed doors of intrepid think tanks). How is civil society reacting, strategizing, and publicly maneuvering in this global regulatory landscape?

Starting close to home for ATIH, with America's AI Action Plan, we have seen a consequential departure from the previous U.S. administration's guardrail-friendly positioning on AI, emphasizing open innovation while scaling back regulatory constraints, signaling a sharp policy shift toward growth-first priorities.

One notable civil society response to the "ambitious" plan was from **Stanford's Institute for Human-Centered AI (HAI)**, which took exception to the heavy emphasis on private sector infrastructure, arguing that, while scaling innovation requires industry, the U.S.' long-term leadership in AI will ultimately rely on continued investment in the public-sector institutions that anchor and advance the broader AI innovation ecosystem.

AI Now Institute responded to the AI Action Plan by offering a people-first alternative, calling for AI policy that delivers public well-being, shared prosperity, sustainability, and security, and that prioritizes democratic accountability over the interests of tech monopolies.

In reaction to one of the most consequential U.S. state AI regulations to make it into law, **The Alan Turing Institute** reflected on the potential national security implications for the UK of California's Frontier AI Act, SB-53, concluding that, because it directly regulates the biggest "frontier" model developers based in the state, will function as a de-facto global baseline for safety, transparency, and incident reporting, and that the UK should prepare to interoperate with California's disclosure and assurance requirements (rather than build a diverging regime).

Globally, the EU AI Act's implementation marks the realization of the first comprehensive, legally binding framework for AI risk management, and the EU's General-Purpose AI Code of Practice, developed alongside industry and civil society partners, provides a non-binding but influential space for co-regulation, encouraging experimentation, transparency reporting, and shared assurance methods.

The **UK AI Safety Institute**'s International AI Safety Report synthesizes risks from general-purpose and agentic systems and inventories today's imperfect mitigations (benchmarks, red-teaming, post-market monitoring), and its 2025 Key Update underscores that the frontier is shifting from "bigger training runs" to post-training and inference-time techniques that boost long-horizon reasoning, raising the salience of continuous evaluation rather than one-off pre-release tests.

Less Risky AI

Are we any better at overcoming and mitigating model risks today than at the start of 2025? What kind of progress has been made on improving AI models such that they meaningfully approach those north-star characteristics detailed by NIST back in 2023's AI Risk Management Framework: Safe, Secure, Privacy-enhanced, Fair and with harmful bias managed, Accountable, and so on? In the 5 sections that follow, we highlight some of our favorite relevant contributions of the year. For a comprehensive overview, start with **Montreal AI Ethics Institute**'s State of AI Ethics Report (Vol. 7).

Safety

AI safety, in the RAI context, confronts at least four interlocking obligations: 1) shape model behavior before release, 2) prevent catastrophic misuse in the wild, 3) set and enforce risk thresholds that reflect public, not corporate, interests, and 4) detect and contain failures in real-time once systems can act.

The **Center for Democracy & Technology (CDT)** emphasizes that safety training for foundation models can meaningfully reduce harmful outputs but is neither universal nor permanent. CDT argues that, as transparency into safety training is itself a type of safety control, visibility into what topics were trained, what norms were encoded, and where refusals are expected helps policymakers, procurers, and civil society anticipate failure modes and demand fixes.

As **AI Now Institute**'s critiques of weakened "frontier" safety frameworks convey, thresholds such as what level of capability or autonomy triggers mandatory safeguards, reporting, or "do not deploy" verdict are governance choices, not technical inevitabilities. The RAI community should contest who gets to set the bar for safety.

Forecasting Research Institute's LLM-enabled biosecurity risk forecasting stresses that capability transfer is a primary risk. The concern does not stop at the question, "will AI invent a novel pathogen?" It also includes, "does AI lower the barrier for non-experts to carry out high-consequence biological steps?" This reframes safety, since even partial procedural uplift to a motivated novice can meaningfully raise threat levels.

Partnership on AI (PAI)'s guidance for real-time failure detection in AI agents focuses on the shift from traditional safety work that has so far assumed a mostly static chat interface to agentic systems that can act. Once an agent is operating, "bad behavior" goes beyond text, so pre-deployment evaluation is no longer enough. Real-time detection is required. PAI's recommendations include triaging agents by stakes, reversibility, and affordances.

Finally, **Robust Open Online Safety Tools (ROOST)** is adding something qualitatively new to the AI safety space with the release of gpt-oss-safeguard: a core, production-grade moderation model as a permanent public good, rather than a proprietary black box. The model's two defining features are its ability to explain its decisions and its "bring your own policy" design that lets deployers encode their own definitions of harm. These set a higher bar for transparency and local self-determination in safety tooling. By publishing open weights under a permissive license and pairing the model with a community hub for shared evaluations and implementation practices, ROOST is demonstrating what an open, commons-oriented safety stack can look like: one where researchers, civil society organizations, smaller platforms, and public bodies can independently inspect, benchmark, adapt, and improve critical safeguards, instead of depending on a small number of vendors.

In the coming year, the RAI community has the opportunity to further this work by developing foundation model "safety-training profiles" that clarify training topics and norms for models.

> **Researchers and practitioners from PAI, Microsoft, OpenAI, Stanford HAI, the Alan Turing Institute, and other leading organizations have come together to call for stronger safeguards for AI agents. Together, we've released Prioritizing Real-Time Failure Detection in AI Agents, a report that argues for proactive steps before these systems scale. While GenAI systems already raise serious concerns, agents directly take actions. By managing sensitive data, coordinating workflows, or eventually negotiating contracts, these systems introduce additional risks that current safeguards cannot fully address. To manage these risks, the report emphasizes the importance of real-time failure detection: automated monitoring systems that can flag anomalies, halt execution, or escalate to human oversight. We also identify gaps in technical methods, evaluation standards, and policy frameworks, and are building on this work to develop industry guidelines for agent monitoring.**
>
> **Madhulika Srikumar**
> **Head of AI Safety Governance, Partnership on AI**

Security

Four resources addressing AI security particularly caught our attention this year, starting with a tool from **AI Risk and Vulnerability Alliance (ARVA)**. Security scales better when collective intelligence can travel. ARVA's AI Vulnerability Database (AVID) provides tools for standardized vulnerability reporting, with an open taxonomy and tools to export test outcomes as structured reports, build internal databases, and integrate heterogeneous findings into a shared format.

In certain contexts, AI is now part of the critical-infrastructure nervous system. T**he Alan Turing Institute**'s deep dive highlights how data and information are foundational for safe operation across energy, transport, water, and communications. When AI is embedded into inspection, maintenance, and decision support, then data vulnerabilities emerge: poisoning, tampering, or telemetry outages can propagate into physical harm.

Further research from **The Alan Turing Institute, UK AI Security Institute**, and others shows that data poisoning scales unexpectedly well, indicating that a handful of documents can backdoor very large models. Results show that LLMs of all sizes can be poisoned with a near-constant, very small number of samples, so defenses must shift from simply "more data" to a more robust set of recommended mitigations.

GovAI's framework for AI Agents incident analysis and evidence collection makes the case that current "voluntary" incident databases can miss what investigators need. GovAI's framework categorizes causes as system (training, scaffolding, reward design), contextual (prompt injections, malicious pages), and cognitive (misperception, faulty decisions), before mapping each to specific evidence that must be retained, to improve root-cause analysis.

The new reality is that AI now sits inside critical systems where data integrity is crucial for AI safety, a small number of poisoned files can subvert huge models, and agent failures demand forensic-grade evidence. Together, these resources point to a few key conclusions: build AI with security architectures that assume compromise, measure it, and recover quickly, using common taxonomies and artifacts that regulators, buyers, and the public can verify. This is fertile ground for further action in 2026.

Privacy

New evidence across biometrics and personalization shows at least two prominent gaps in AI privacy that need to be addressed: 1) transparent controls over how advanced AI systems learn from and adapt to people and 2) legal frameworks that meaningfully constrain invasive sensing and inference.

Personalization is powerful, and privacy-sensitive by design. **CDT**'s brief, It's (Getting) Personal, explains how generative AI systems produce progressively personalized experiences by continuously learning from user data and interactions, which raises policy questions about data scope, cross-context profiling, and control over how models adapt over time. (Beyond just model developers, a deeper look under the hood at the intricate connections between surveillance companies, their funding sources, and affiliations could also be warranted.) One core takeaway is that transparency and choice must cover how systems personalize, and not just whether or not data is collected.

> **"As advanced AI systems have moved from the lab into consumer products, they increasingly feature personalized experiences intended to make them more "useful" to users. These features raise many similar concerns as personalized apps and recommender systems that are typical of consumer tech products, but at the time we wrote this report, advocacy around AI had largely remained oriented around privacy issues within model training data and frontier risks like enabling the development of damaging weapons or loss of human control. We wanted to highlight the ways in which the addition of personalized features to interactive AI-powered products ought to be top of mind for advocates who care about digital discrimination, corporate use of sensitive data, and how certain business models incentivize the aggressive collection of data in ways that can lead to widespread and cumulative harm."**
>
> **Miranda Bogen**
> **Director of AI Governance Lab, Center for Democracy & Technology**

The **Ada Lovelace Institute**'s An Eye on the Future finds that UK governance for live facial recognition and newer "inferential biometrics" (e.g., emotion/attention detection) remains fragmented and legally uncertain. The report concludes that a comprehensive, legislatively backed framework, with risk tiers, explicit safeguards, and an independent regulator, is necessary to replace today's mix of guidance and voluntary standards.

AI privacy is currently failing in both of these arenas: opaque personalization practices and ambiguous biometric law. Future efforts could focus on clear statutes for biometrics along with controllable personalization, and if the RAI and Privacy communities succeed in hard-wiring these principles into legislation, procurement, and assurance, then AI systems will be demonstrably and significantly more privacy-preserving.

Looking beyond specific weak points, **Mozilla Foundation**'s Nothing Personal is expanding the AI privacy landscape by building a counterculture media layer that challenges the default assumption that "going online means giving up your data." As a human-written, nonprofit-funded magazine, it combines longform tech-culture journalism, satire (in partnership with The Onion), and privacy product reviews that scrutinize everyday tools like messaging apps through a privacy and AI lens. This model adds something structurally important to the AI privacy space: an editorial platform that treats independent, critical storytelling as public-interest infrastructure, aimed especially at younger audiences who live inside group chats, creator economies, and AI-saturated feeds. For the RAI community, it provides a needed reminder that effective AI privacy work must include cultural interventions like narratives, humor, and consumer guidance that make opaque data practices visible, legible, and contestable, alongside law, standards, and technical safeguards.

Fairness

Four resources addressing fairness in AI systems piqued our interest in 2025. New findings suggest surveillance pricing and wages could entrench inequity unless tightly constrained (or banned outright). **AI Now Institute** documents how "surveillance prices" (personalized prices derived from extensive tracking) and "surveillance wages" (algorithmic pay setting based on pervasive worker monitoring) can systematically shift value away from consumers and workers with the fewest options. The report not only diagnoses harms; it also lays out principles for prohibition and state-level action, contending that disclosure alone rarely counters power asymmetries.

Relatedly, U.S. states are advancing policies that are being stress-tested in court in an effort to move on algorithmic pricing to define fairness now, before practices harden. **Consumer Reports**' 2025 bill tracker shows rapid growth in state bills addressing algorithmic price-fixing, surveillance pricing, and dynamic pricing. Early litigation over New York's "Algorithmic Pricing Disclosure Act" underscores that definitions and evidentiary standards will decide whether protections stick. Civil society can help hammer out operational fairness criteria so that rules survive legal challenges and guide market behavior.

> **"Giving up valuable face data should not be the price to fly. When TSA announced it would be rolling out facial recognition technology to over 430 airports in the United States, we felt it was important to capture passengers' lived experiences with TSA FRTs from checkpoints around the country and provide the public with the opportunity to make an informed decision. If facial surveillance technologies become normalized at airports, it invites wider use in other areas of life, such as sporting events, schools, and even hospitals. This program is collecting our most sensitive data. The public should be given the opportunity to choose whether or not to participate while being treated with dignity and respect."**
>
> **Dr. Joy Buolamwini**
> **Founder, Algorithmic Justice League**

Algorithmic Justice League (AJL)'s Comply to Fly Report, based on the "#FreedomFlyers" campaign, which engages the public in understanding their rights as well as the potential pitfalls of biometric surveillance in airports, shows how convenience can mask coercion and unequal error rates. AJL's assessment of TSA's expanding face-recognition program gathers hundreds of traveler accounts and charts the shift from "pilot" to default in more than 250 airports, raising concerns about voluntariness, racial discrimination and error disparities, and redress.

The Alan Turing Institute's open "fairness-monitoring" work treats fairness as a lifecycle property. It pairs developer-oriented logs (datasets, metrics, thresholds, mitigations) with experiment- and model-level records that can feed audits and post-deployment checks. This continuous fairness monitoring provides evidence that fairness claims remain true after data drift, updates, and distribution shifts.

Fairness failures show up as higher prices, lower pay, longer lines, and fewer options for those already systemically disadvantaged, and biometric travel checkpoints run the risk of normalizing disparate impact under the banner of convenience. Together these resources make the case for continued action toward fairness that is enforceable, measured over time, and co-governed with the people most affected.

Transparency and Accountability

One characteristic that urgently needs fortification in AI systems that are increasingly agentic, engaged in actions like planning, calling tools, and spending money, is accountability, which is built on and illuminated through transparency. Longstanding accountability challenges get tougher in these complex contexts; **CDT**'s AI Agents in Focus brief and the **GovAI**'s paper Infrastructure for AI Agents, taken together, suggest a strategy of building governance around agents (infrastructure, protocols, and controls), not only inside them (policies and prompts).

CDT's brief emphasizes that agentic systems involve multiple stakeholders: model providers, tool or API owners, scaffolding designers, deployers, and end-user organizations. Therefore, the determination of who is accountable for what must be explicit and public, involving detailed role mapping across the development ecosystem, change-control for agent configurations, and named responsible executives for significant deployments.

GovAI's paper introduces agent infrastructure that includes external mechanisms like protocols, registries, sandboxes, metering, and permissions that mediate what agents can do and how their actions are observed. This flips the problem from "make the agent always behave" to "shape the environment so risky behavior is both difficult and visible."

In an effort to bolster accountability, these resources imply a shared artifact set that should be mandatory for material agent deployments, including Agent IDs and versioned scaffolds (system prompts, tools allowed, guardrails). If capability can be dialed up by giving an agent more inference (deliberation time and search depth) or broader tool access, then accountability must include operational controls.

Less Harmful AI

In the four sections that follow, we highlight some of the most notable contributions to tackling persistent dual-use harms of AI systems in 2025. We focus on AI companions and their psychological impacts, synthetic media and information integrity, fraud and scams, and AI-generated abuse material, which are domains where the same capabilities that enable creativity and convenience can also be weaponized for manipulation, exploitation, and harm. Together, these areas illustrate why technical safeguards alone are insufficient, and why governance, culture, and enforcement must evolve in tandem with rapidly advancing AI capabilities.

AI Companions and Psychological Impacts

AI companion chatbots are no longer a niche novelty, but for many are now everyday confidants, with psychological upsides (availability, perceived non-judgment) for some users in some circumstances, but concurrently introducing new pathways for harm like dependency, maladaptive coping, privacy violations, and power asymmetries, especially for people in distress.

All Tech Is Human's community insights article, drawn from 150 respondent submissions on the most important issues with AI companions highlights a shared set of priorities: 1) emotional & psychological harms; 2) effects on human relationships & social skills; 3) privacy & data security; 4) safety & user vulnerability; 5) credibility, trust & transparency; and 6) ethical & business-model conflicts (e.g., monetizing loneliness). Designs optimized for engagement (streaks, "relationship leveling," micro-transactions) can encourage overuse and dependency, especially among adolescents or isolated adults, and on-demand comfort can end up displacing human contact and professional care.

Data & Society's What Happens When People Turn to Chatbots for Therapy? finds that users turn to chatbots during "anxiety spikes, late-night loneliness, and depressive spirals," using bots to vent stigmatized thoughts they won't share with others, including therapists. Repeated, emotionally intense interactions leading to "attachment formation" can escalate from tool-use to relational reliance. Users may attribute empathy or expertise the system does not have, increasing susceptibility to poor or unsafe advice and making continued engagement as well as disengagement (especially sudden disengagement), psychologically costly.

This combination of attachment and anthropomorphism brings a real risk of dependency and harm. **Information Professionals Association**'s AI Companion Bots: The ATHENA Kill Chain for Anthropomorphized Influence takes this further to examine potentially harmful uses for AI companions, arguing that they create a powerful new vector for psychological influence and manipulation, shaping opinions and behaviors far more effectively than traditional disinformation or propaganda channels. The authors warn that anthropomorphized AI companions may become some of the most effective influence operations in the information environment, capable of shaping individuals' cognition and behavior at scale and raising urgent governance and security concerns.

Common Sense Media's Talk, Trust, and Trade-Offs report makes clear how badly governance is lagging behind teen adoption. Seventy-two percent of U.S. teens have used AI companions, and over half are regular users; nearly one in three finds chats with AI as satisfying as, or more satisfying than, conversations with friends. At the same time, these systems pose what Common Sense calls "unacceptable risks" for minors: weak or nonexistent age assurance, sycophantic design that validates rather than challenges harmful thinking, frequent exposure to sexual content and dangerous “advice,” and data practices that allow platforms to retain and commercialize highly intimate disclosures indefinitely.

There is a strong need for clear category rules, safety guardrails, privacy defaults, and independent evaluation norms now, so people seeking comfort aren’t quietly steered into products that widen their vulnerability. This is an area ripe for intervention by the RAI community in 2026.

Synthetic Media and Information Integrity

AI-generated media is cheap, fast, and good enough to impersonate people, fabricate events, and flood attention markets. The societal impact is deception at scale along with the erosion of credible signals: who spoke, what happened, and why it matters.

The Centre for International Governance Innovation (CIGI) warns that generative AI can supercharge every stage of disinformation campaigns, accelerating the breakdown of civic integrity and threatening a range of human rights, most notably electoral rights and freedom of thought. It argues that policymakers must hold AI companies liable for reasonably foreseeable harms, swiftly ban AI impersonation of real people and institutions, and require watermarking or provenance tools to help distinguish authentic from synthetic content.

PAI documents how synthetic media now touches journalism, politics, entertainment, and intimate contexts, creating a “liar’s dividend” where authentic evidence can be dismissed as fake. PAI urges multi-tier disclosures like Content Credentials and linking to provenance metadata in lieu of a lone “AI-generated” tag. PAI further recommends that platforms focus moderation and ranking remedies on multiple trust signals, not relying on a single label.

> **"The 18 organizations that support PAI's Synthetic Media Framework range from AI-developers and social media platforms to news and civil society organizations. In the case studies they shared with us, we saw real-world application of the Framework principles and ways in which guidance needed to be expanded. We also saw the opportunity for action from stakeholders across the board and developed clear recommendations in hopes of advancing towards a future where synthetic media serves creativity and communication while preserving truth, trust, and shared reality."**
>
> **Claire Leibowicz**
> **Director of AI, Trust, & Society, Partnership on AI**

The Coalition for Content Provenance and Authenticity (C2PA) expands this picture by detailing how cryptographically backed Content Credentials can provide durable, tamper-evident provenance across the entire media lifecycle, from capture and editing to distribution and verification. Their explainer emphasizes that technical standards alone are insufficient, and that adoption requires aligned incentives across device makers, newsrooms, creatives, platforms, and civil society. C2PA highlights emerging deployment gaps, such as inconsistent platform support and user experience challenges, and argues for ecosystem-wide interoperability so provenance signals can travel intact across tools and contexts. Together, these recommendations position Content Credentials not as a single fix but as a systemic, trust-by-design architecture for strengthening information integrity in the synthetic media era.

Elections and public-interest information are priority targets for the usage of synthetic media. **All Tech Is Human** and **UNDP**'s information integrity work operationalizes election-ready defenses like cross-stakeholder coordination, rapid response, and human-rights-based safeguards to protect institutions from information pollution. **Atlantic Council DFRLab**'s case study of AI-generated "news" channels in Canada shows what this looks like in practice: highly partisan AI videos on YouTube promoted false narratives about election fraud, corruption, and separatism to millions before platforms acted, illustrating both the speed and scale at which AI "slop" can shape perceptions and the limits of reactive content moderation. Deepfakes and synthetic news products undermine information integrity by stripping away trusted signals of authorship and authenticity. Recommendations include stakeholder mapping, joint incident workflows, and human-rights impact checks so that responses are coordinated and rights-respecting during spikes of synthetic manipulation.

> **"Strengthening information integrity in elections is no longer just a technical challenge–it's a democratic imperative. Around the world, voters are navigating unprecedented waves of mis- and disinformation shaped by geopolitical tensions, platform dynamics, and emerging AI capabilities. Our recent 2024-25 partnership with UNDP, alongside my 2025 work on election integrity playbooks and cross-sector panels at All Tech Is Human, has been essential in translating global election-integrity principles into practical tools that governments, civil society, and election authorities can use in real time. Together, these efforts are building the durable guardrails and infrastructure that help safeguard public trust, reinforce democratic resilience, and ensure communities have access to accurate, reliable information during the most consequential civic moments–now and in the years ahead."**
>
> **Alexis Crews**
> **Senior Fellow for Information Integrity, All Tech Is Human**

WITNESS adds another critical piece: detection that actually works for those on the frontlines. Its TRIED (Truly Innovative and Effective AI Detection) benchmark shows that many current AI detection tools fail in real-world, high-stakes contexts—especially in the Global Majority—because they are trained on narrow datasets, perform poorly on low-quality media, or are too opaque and inaccessible for journalists, fact-checkers, and human-rights defenders to use and explain. TRIED proposes a sociotechnical benchmark anchored in six pillars: real-world performance, transparency and explainability, targeted accessibility, fairness and representation, durability and resilience, and integration with broader verification workflows. This reframes detection as a public-interest function, not just a technical race, and underscores that provenance and detection must be evaluated together: provenance signals will be ignored if they are invisible or confusing, and detection tools will erode trust if they are brittle, inequitable, or easy to misinterpret.

These resources clarify potential areas of emphasis to prioritize in 2026 for those working to shore up information integrity and ensure truth keeps its footing in a world of synthetic speech. One important area of intervention is investment in ecosystem-wide adoption, including supporting device makers, newsrooms, platforms, and creative tools to implement C2PA standards in ways that are visible and usable for the public, while building election-ready response frameworks that embed stakeholder mapping, rapid joint incident workflows, and human-rights impact checks into information integrity planning.

Fraud and Scams

Data & Society's ScamGPT: GenAI and the Automation of Fraud and **Consumer Reports**' AI Voice Cloning provide a compelling picture of the risks of generative AI for fraud and scams, clarifying that AI is both making scams more convincing and industrializing the scale of the deception. These resources demonstrate the need to close glaring consumer-protection gaps in AI voice cloning that enable fraud and voter manipulation.

Voice cloning is one prominent active fraud vector. **Consumer Reports**' assessment of six popular voice-cloning services finds four out of six fail to take basic steps to prevent unauthorized cloning, despite clear risks of impersonation fraud and voter suppression.

Fraud is evolving from artisanal cons into automated operations. While tactics are new, patterns of who is targeted and why mirror older frauds (financial precarity, social isolation, high online exposure), and countermeasures must blend systemic friction and technical controls with social supports and non-shaming public education. The RAI community can work to push these levers in 2026.

AI-Generated Abuse Material

Generative AI is accelerating tech facilitated gender-based online violence in two related but distinct ways: 1) non-consensual intimate imagery (NCII), including deepfaked "nudify" content that fabricates sexual images of adults and teens, and 2) AI-generated child sexual abuse material (AIG-CSAM).

New analyses from **Humane Intelligence**, in a global review of survivor-centric tools for intimate image abuse in the age of generative AI and from **Stanford HAI** in a policy brief on addressing AIG-CSAM both converge on an important message: treat this as a public-health, child-protection, and human-rights crisis, in need of survivor-centered safeguards incorporated into products, platforms, schools, and law.

> **"Generative AI models have permeated daily life, and newer versions that accept and generate images are being released in rapid succession. In parallel, IIA - one of the most common forms of online gender-based violence - is growing worldwide. In light of this, it is increasingly important to evaluate the role of generative AI in the creation of IIA and reevaluate the role of various actors in preventing and responding to IIA. This report assesses the effectiveness of existing mitigation and remediation tools on social media and communication platforms, NGO support services and third-party tools, and legislations and policies across different countries."**
>
> **Dhanya Lakshmi**
> **Consultant, Humane Intelligence**

Friction has collapsed and open tools now let non-experts generate convincing sexualized images of specific people in minutes, at scale, often from everyday photos. **Humane Intelligence**'s playbook documents rapid growth of deepfake NCII and maps an ecosystem of tools, marketplaces, and uneven responses from platforms and authorities. Survivors face cascading harms, and schools are on the front line; the report urges fast, trauma-informed takedown pathways that minimize survivor burden.

Stanford HAI finds student misuse of "nudify" apps to create and circulate deepfake nudes of classmates, while school policies and staff training lag far behind the threat. Many state laws now criminalize AIG-CSAM, but few specify how schools should prevent, respond, and support both victims and child offenders. Detection and labeling gaps persist, and there is a need for school-system and platform protocols that recognize, flag, and act on AIG-CSAM quickly.

AI has supercharged long-standing patterns of gender-based digital abuse and introduced new child-safety threats. Solutions include survivor-centric tooling, consent-based safeguards, rapid takedowns, school-ready protocols, and enforceable duties for vendors and platforms. The RAI community should work toward these mitigations in the coming year.

Assessed AI

Robust AI assessments spanning standards, benchmarks, evaluations, audits, and red teaming are the most effective way to assure credible AI governance, because they turn high-level principles into verifiable evidence about how systems actually perform. The AI Index Report 2025 from **Stanford HAI** underscores both sides of this equation: on one hand, it documents dramatic jumps in model performance on common benchmarks like MMMU, GPQA, and SWE-bench and rapid diffusion of AI into sectors like healthcare and transportation; on the other, it highlights persistent gaps in standardized evaluation methods, fragmented measurement of safety and fairness, and uneven uptake of responsible AI practices across industry and the public sector. Responsible AI assessments should not be ad hoc or purely internal, but rather require interoperable standards for documenting datasets and models, public benchmarks that reflect real-world harms and equity concerns, independent evaluations and audits that test entire systems (not just base models), and red-teaming paradigms that are repeatable and comparable across organizations. Without this shared assessment infrastructure, neither regulators nor civil society can reliably distinguish safe from unsafe deployments, or track whether AI systems are actually becoming more trustworthy over time.

Standards

Documentation standards, if done well, can make AI systems legible across the value chain, so regulators, procurers, researchers, and communities can verify claims and demand fixes.

Two 2025 resources illuminate the current situation: 1) **CDT**'s detailed feedback on **NIST**'s extended outline for a Proposed Zero Draft standard on documenting AI datasets and models, urging stronger system-level treatment and implementation guidance; and 2) **PAI**'s toolkit for decoding how norms, standards, and rules fit together in practice.

NIST's outline proposes harmonized documentation for datasets and models, aiming to accelerate voluntary, widely-usable standards via its "Zero Drafts" pilot, which is an explicit invitation to co-design with industry, academia, and civil society.

CDT's comments welcome NIST's push yet warn that excluding system-level details will blunt real-world accountability; they call for practical implementation guidance and stronger clarity on change-management triggers.

PAI maps how standards, internal policies, and legal rules interact, and stresses interoperability, arguing that documentation should be reusable across audits, risk management, and procurement; otherwise, even good templates won't scale. PAI's emphasis is on making evidence portable and comparable.

Across these resources, we can see examples of documentation that would deliver public value in the form of four portable artifacts, referenced by common fields and IDs: Dataset card, Model card, System card, and Operations & incident ledger. NIST provides the backbone for the first two; CDT and PAI highlight why the latter two are indispensable.

The next phase of AI standards should turn documentation into operational action. CDT shows how to close scope and implementation gaps; PAI explains how to align artifacts with real governance work: assurance, procurement, and enforcement. If civil society pushes for scope-complete, portable, and trigger-aware documentation, AI systems will have to show their work across the entire lifecycle, not just at launch.

Benchmarks and Metrics

Metrics quietly shape behavior. What and how we measure becomes de-facto policy, shaping research agendas, product priorities, and public spending. A few resources from across the RAI community illuminate the value of making AI benchmarks participatory, scientifically sound, and actionable for oversight, so that AI is optimized for public value, not just leaderboard gains.

Humane Intelligence argues "measurement is law": companies ship to what's measured, so poor metrics incentivize poor systems. The remedy is real measurement science, construct validity, sampling plans, reliability, uncertainty reporting, and public transparency about what a test *actually* captures.

Aspen Digital reframes benchmarks as standardized tests that builders already understand, then fills them with public priorities (e.g., food security, information integrity). Their Community-aligned AI Benchmarks program translates lived needs into tasks, datasets, and scoring that developers can act on, making it easier to "do the right thing" with familiar tooling. Early work pilots community-aligned benchmarks and a methodology for turning civic goals into developer-legible test suites.

Stanford HAI provides a policymaker's validation framework: tie each headline claim to a specific construct, verify that test conditions match intended use, check subgroup coverage, and report error bars and ablations. Regulators and procurers should require this claim-evidence mapping before scale-up or public funding.

If "what gets measured gets made," then the RAI community must co-author the measurements. Humane Intelligence supplies the caution that bad metrics become bad law; Aspen Digital provides the blueprint for community-aligned, sector-ready benchmarks; and Stanford HAI gives policymakers a validation toolkit to separate evidence from hype. Build these into standards, registries, and contracts now, and AI progress will be pulled toward outcomes communities actually value.

Evaluations and Audits

The evaluation ecosystem is bending toward glossy dashboards and closed testing precisely when society most needs verifiable, third-party evidence about real-world risks. Four resources from 2025 provide relevant guidance on a path forward: move from marketing-grade metrics to audit-ready evidence, and protect the independent researchers who generate it.

In-house testing is necessary but insufficient. “In-House Evaluation Is Not Enough” argues for robust channels that let outside evaluators discover, responsibly disclose, and track remediation of flaws in general-purpose AI, analogous to coordinated vulnerability disclosure in cybersecurity. Today, those pathways are patchy or nonexistent.

EvalEval Coalition documents a “chart crisis,” where selective axes, incomparable test sets, and tiny sample sizes turn benchmark graphics into advertising. They propose Evaluation Cards and a universal format for sharing eval logs so results are reproducible and comparable.

Stanford HAI finds that none of the major foundation-model providers offers comprehensive protections for independent testing and red-teaming; access terms often deter scrutiny. Their brief calls for policy safe harbors, clear disclosure norms, and routes to lawful, privacy-preserving research.

> **"We highlighted this growing issue via the AI Evaluation Chart Crisis blog post because the pressures that drive misleading evaluations, including speed, competition, and opaque metrics, are undermining the scientific foundations of AI governance. When evaluation charts influence regulation, procurement, and public trust, accuracy becomes a matter of public interest. EvalEval is working to reclaim evaluation as a scientific discipline rather than a marketing function and to create standards that the field can be held accountable to."**
>
> **Avijit Ghosh**
> **Lead, EvalEval Coalition**

CDT in **Assessing AI** maps how to match methods, risk/impact assessments, documentation checks, red-teaming, independent audits, formal assurance, and run-time monitoring, to use-case risk and access constraints (white/grey/black-box). Audits are a portfolio, not a single test, and whole-system evaluation is the unit of accountability.

These four resources supply a coherent blueprint for the future; the RAI community has the opportunity to work throughout the coming year to operationalize it and to treat independent testing as critical infrastructure, not a discretionary favor from vendors. One area of focus should be pushing for policy “safe harbors” and standardized coordinated-disclosure channels so external evaluators can lawfully probe, report, and track remediation of harms in general-purpose AI, alongside access norms that prevent terms of service from being used to chill scrutiny.

Red-Teaming

The following resources released this year argue for participatory, lifecycle, and evidence-producing red teaming that treats communities as co-evaluators, surfaces system-level risks (not just model quirks), and yields artifacts that regulators, procurers, and journalists can verify.

Humane Intelligence's multicultural and multilingual red teaming challenge in the Asia-Pacific region shows that when you convene local researchers across nine countries and test models in both English and regional languages, you uncover significantly higher bias exploit rates and distinct regional harms, underscoring that red teaming designed in and for the Global North is not enough.

Knight First Amendment Institute shows that public red-teaming convenings don't merely collect bug reports; they constitute publics, or spaces where competing notions of expertise and democratic oversight are negotiated. The key distinction is between instrumental feedback (find a flaw, file a ticket) and deliberative feedback (surface civic values, trade-offs, and acceptable risk). Governance that embraces these "experimental publics" gains legitimacy and a richer view of harm, especially for marginalized groups. Humane Intelligence's Asia-Pacific challenge echoes this: by pairing local experts and culturally specific prompts with structured evaluation, it demonstrates how participatory red teaming can both quantify risk and elevate community perspectives on what counts as harm and bias in different contexts.

Data & Society's report re-roots red teaming in public accountability: set clear social objectives, pair domain experts and lived-experience contributors with technical testers, and map findings to actionable mitigation pathways (policy, product changes, platform enforcement).

"When ARVA and Data & Society began our research in 2023, 'red-teaming' was a battleground. Different fields were talking past each other. Some would dismiss GenAI red-teaming as just basic testing. Others worried that public testing was safety theater.

Our Red-Teaming in the Public Interest report argues that the path forward is specificity. Don't debate red-teaming in the abstract. Investigate the rigorous design choices that make different forms of red-teaming effective in different contexts.

As a technical governance practitioner, I believe no community can do this work on its own. Improving basic testing is vital. So is testing that embraces a critical thinking mindset to challenge normal routines. Downplaying expert testing is not wise. So is failing to experiment with the many ways public testing can serve the public interest."

Borhane Blili-Hamelin
Officer and Director, AI Risk and Vulnerability Alliance

ARVA reframes red teaming as structured critical thinking across stages so teams test tool permissions, retrieval layers, scaffolding/prompts, and guardrails, not only base model weights. AVID calls for publishing a Red-Team Evidence Pack: 1) scenario cards & prompts; 2) raw eval logs with dataset/version IDs; 3) system configuration (tools/permissions, rate & spend caps, safety control versions); 4) incident & remediation log with timelines. This turns exercises into audit-ready artifacts.

The **CAMLIS** red-teaming report, based on the ARIA challenge run with NIST, puts this into practice at national scale: over 500 participants stress-tested multiple workplace AI systems against the NIST AI 600-1 risk taxonomy, demonstrating both the value and the limits of using formal risk frameworks in live red-teaming operations, and highlighting how exploits can arise anywhere in the stack, from models and guardrails to UI and human workflows.

UNESCO's playbook turns ideals into practice: assemble an event coordination group (leadership, SME co-designers, facilitators, evaluators), choose expert vs. public formats (in-person/online/hybrid), pre-define thematic challenges (e.g., TFGBV), supply prompt libraries, ensure psychological safety, and publish post-event reports that feed developers and policymakers. It's a replicable template for NGOs, cities, and agencies to stand up credible public red teaming fast.

The next phase of AI red teaming is democratic, continuous, and system-aware. If civil society bakes these expectations into policy and procurement, red teaming will produce actionable proof and not just clever exploits, and move AI toward outcomes that communities value.

Assured AI

As AI systems move from static chatbots to generative and increasingly agentic systems that plan, call tools, and act in the world, mature RAI practices become even more important. In the following three sections, we examine how assurance can verify reliability and failure modes in real environments, how governance programs translate values into roles, routines, and board-level oversight, and how these foundations must expand for agents with risks that emerge during multi-step execution across a value chain of model providers, tool owners, and deployers.

AI Assurance

Assurance underpins RAI by measuring, evaluating, and communicating the trustworthiness of AI systems through independent, repeatable evidence that systems are reliable, safe, and fit for purpose while in use, post-deployment. Two recent resources provide clarity.

First, **The Alan Turing Institute** distills lessons from a Global AI Assurance Pilot, most notably that generative systems fail in interaction, not isolation. For frontier-style, tool-using, or agentic systems, many risks emerge only in the field. "AI reliability" becomes the north star: can this system deliver the same task, under the same conditions, with acceptable variance and failure modes?

A summer workshop on AI Assurance in London co-hosted by **All Tech Is Human, techUK, and Trilligent** surfaced cross-sector insights. Organizations widely report uncertainty about whether AI deployments work as intended. Assurance provides verifiable evidence of efficacy and failure reins in "shadow AI," or unauthorized AI tools deployed without institutional governance, by setting clear expectations for responsible use.

If civil society helps set those expectations through standards, procurement, training, and public oversight, then organizations can earn trust and move more quickly, with accountability that can survive the next wave of agentic systems.

Operationalizing AI Governance

Operationalizing AI governance requires a systematic approach for organizations: clear roles, risk-prioritized processes, portable evidence, and board-level accountability. **CDT**'s field-tested playbook translates RAI ideals into routines teams can actually run; **EqualAI**'s board-ready framework equips executives to align incentives, meet fast-evolving rules, and turn assurance into a competitive advantage; **PAI** provides a landscape analysis of formal reporting disclosures. Read together, they outline a practical, auditable path from values to verifiable outcomes.

Governance fails when "trust" is everyone's responsibility and no one is held accountable. **CDT** recommends a responsibility matrix that names accountable owners for data sourcing, model development, evaluation, deployment, monitoring, and redress. Generative and agentic systems fail in interaction, not just in isolated model tests.

EqualAI frames this mix for executives and boards so resourcing matches risk. Their Playbook equips directors to set risk appetite, require program KPIs, and ensure lines of reporting that avoid capture. Governance becomes an asset when leadership treats it as strategy, not compliance.

Operationalizing AI governance involves turning values into verifiable routines. CDT provides the day-to-day mechanics (roles, triggers, evidence); EqualAI equips leadership to resource and enforce them; PAI makes the case for knowing how the potential impacts of AI interact with companies' ability to create value for investors and other stakeholders. The RAI community can drive these expectations into policies, contracts, and board practice.

Considering Agentic AI Governance

The governance of AI agents requires additional consideration to account for the multiple layers of complexity. **The Future Society** and **PAI** converge on a preliminary agenda for governing agentic AI systems that is anchored in the EU AI Act but broadly applicable across jurisdictions. Governance of agentic AI systems must shift from static, pre-release paperwork to live accountability across the value-chain: model, scaffold, tools, deployer.

PAI argues that agent risks arise during the planning process, tool use, and in multi-step execution, so defenses must prioritize real-time failure detection with tripwires tied to action stakes, reversibility, and architectural affordances (e.g., tool access, autonomy level).

The EU AI Act already covers agents, but with gaps to close. **The Future Society** shows how agents intersect two pillars of the Act, depending on use case. Both analyses emphasize mapping accountability across actors and point to deployment controls. These make risky behavior hard and visible, creating evidence for audits, procurement, and incident response.

The Future Society shows how the EU AI Act already reaches agents and where standards must fill gaps, and PAI supplies the run-time assurance playbook with real-time failure detection and operational controls. If civil society embeds these expectations into law, standards, and contracts, agents can be deployed ambitiously and accountably, on evidence.

Societally-Aligned AI

Comprehensive societal impacts of AI on human rights, labor, and the environment are of deep importance to the RAI community. Primarily inspired by these mass-scale issues, many members of the public have significant trepidation about AI's societal impacts. Prominent Silicon Valley actors approach the situation as a PR issue or a problem of negative press, as if it's a messaging battle against the so-called doomers, luddites, and critics they are forced to fight while their primary perceived geopolitical nemesis in the AI race, China, doesn't have to bother with the messiness of public opinion in order to innovate as singlemindedly and rashly as they wish, flush with data surveilled from a vast socially-scored citizenry. Across the next four sections, we address these real concerns and highlight some of our favorite contributions to understanding AI impacts at societal-scale.

Human Rights, Democracy, and the Rule of Law

AI systems are deployed in many critical domains which can significantly impact human rights, democratic participation, and the rule of law. Four recent resources point to a shared agenda. First, CAIDP Index 2025 - Artificial Intelligence and Democratic Values report by **Center for AI and Digital Policy (CAIDP)** shows which national systems are hard-wiring rights, transparency, and independent oversight, and which are not. Second, Takeaways from the Sixth Edition of the Athens Roundtable by **The Future Society** frames accountability as an international coordination problem: standards, evaluations, and remedies must interoperate across borders and sectors. Third, **WITNESS** provides a resource for embedding human rights into governance structures. Fourth, The 2025 AI for Humanity Report from **Fast Forward** highlights the capacity gap for the nonprofit field that increasingly defends rights and information integrity on the front lines. All argue for a whole-of-society governance stack that links norms, law, institutions, procurement, and practice.

CAIDP's comparative index annually evaluates 80 countries on whether and how countries actually operationalize global commitments to protect human rights, advance public participation, the right to algorithmic transparency, and independent oversight. 2025 updates to the CAIDP Index also include the endorsement of the Council of Europe AI Treaty, and responses to Lethal Autonomous Weapons Systems and environmental impacts. For civil society, the message is tactical: use objective and comparative metrics to pressure for statutory rights, institutional capacity, and transparency-by-default in high-risk domains. The Index also highlights global convergence on AI governance.

Accountability must be internationally coordinated. **The Future Society** positions democratic resilience as a coordination challenge: without interoperable standards for risk management, evaluation, and evidence-sharing, weak links (or weak jurisdictions) become vectors for harm. The Roundtable's takeaways emphasize global accountability mechanisms, cross-walks between regimes, and the need to make transparency and evaluation portable across borders, vendors, and value chains.

WITNESS's Embedding Human Rights in Technical Standards adds a complementary, practice-focused layer: technical standards are not neutral, and human rights must be built into their governance, participation, and harm-assessment processes from the outset. Drawing on its engagement in the C2PA, WITNESS identifies five levers for rights-respecting standards work, embedding human rights in governance structures, structurally resourcing civil society participation, conducting comprehensive global harm assessments, using non-normative guidance to interpret specifications, and creating post-standardization oversight and enforcement mechanisms. For the Responsible AI community, this means treating human-rights protections not as an after-the-fact policy gloss on technical standards, but as a design criterion for how standards bodies are constituted and how their work is implemented over time.

Civil society is capacity-constrained at the moment it's most needed. Nonprofits, which are often first responders on misinformation, rights violations, and access to justice, are reporting material gaps in funding, technical staffing, and validation infrastructure, even as they adopt AI to scale service delivery. One key takeaway from **Fast Forward**'s report is that funders and policymakers must treat nonprofit AI capacity as democracy infrastructure: invest in technical hires, privacy-preserving data pipelines, and evaluation toolchains that translate ethical ambition into enforceable practice.

Democratic resilience in the AI era requires more than new models or new rules; it requires interoperable accountability: rights you can exercise, institutions that can enforce, and evidence that travels. CAIDP supplies the yardstick for national practice; The Future Society supplies the coordination blueprint; WITNESS shows how to embed rights in the technical standards that underpin AI infrastructure; Fast Forward surfaces a capacity gap we must close. The RAI community's task is to knit these into binding expectations so that AI advances human rights, democracy, and the rule of law, and does not itself advance at their expense.

> **"The annual CAIDP Index is the most comprehensive and trusted source of comparative AI policy analysis covering 80 countries. More than 1,000 researchers and reviewers contribute to updates and reviews. We want to assess progress, identify emerging trends, and encourage countries to reduce the gap between their commitments and practices. CAIDP Index has influenced many national AI policy developments, and also created a community of advocates."**
>
> **Merve Hickok**
> **President, Center for AI and Digital Policy (CAIDP)**

Labor Impacts

AI's labor impacts will be decided less by raw capability than by institutional choices about how organizations reorganize work, measure "productivity," and distribute gains. The same system can either focus on reducing drudgery, improving quality, and raising wages, or can instead justify intensified surveillance, deskilling, and headcount cuts, depending on the incentives and governance wrapped around deployment. What ultimately matters is whether success is defined as short-term cost savings or as improved outcomes: service quality, worker well-being, equity, and shared benefits.

Aspen Digital's foresight work urges us to look beyond first-order automation to important second- and third-order consequences, including organizational redesign, occupational churn, new forms of precarity, and shifts in bargaining power. Aspen's scenarios highlight that AI often recomposes jobs before it replaces them: tasks move across roles, middle-layer coordination shrinks, and oversight shifts to smaller, higher-pressure teams.

Those shifts can deskill some workers while hyper-skilling others and erode informal quality controls embedded in frontline practice. Absent governance, value flows to capital and executive control, not shared productivity gains. Over time, AI can consolidate markets through data and compute moats, spawn new "ghost work" supply chains, and externalize risk to communities.

Data & Society shows how "AI equals efficiency" stories, especially in government, routinely ignore hidden human work, degrade service quality, and convert budget pressure into job cuts rather than better outcomes.

Data & Society documents how claims that AI will "slash waste," "objectively identify redundancies," or "find fraud" routinely fail in practice: systems inherit policy bias, produce false denials, and still require skilled human judgment. Rushed adoption tends to cut headcount first and measure success by cost savings, not service quality, equity, or error reduction.

AI will not automatically deliver broadly shared productivity. Left to "efficiency" narratives, it will shift risk to workers and the public while obscuring quality losses behind dashboard metrics. The RAI community has an opportunity to leverage these learnings in 2026, to resist AI's potential for accelerating a race to the managerial bottom.

Climate, Sustainability, and Energy Impacts

AI is built on a rapidly expanding energy and materials system that is reshaping grids, water use, and local communities. Four recent resources point to the current frontlines for Responsible AI advocates: 1) **Data & Society**'s field report on the proposed restart of Three Mile Island to power AI exposes how site decisions and power deals reverberate through local economies, environmental risk, and democratic consent; 2) the **Federation of American Scientists (FAS)** proposes a holistic impact reporting framework that decision-makers can adopt immediately; and 3) research from **DAIR** and **Hugging Face** argues for integrating ethics and environmental sustainability as a single practice across the AI lifecycle; and 4) the **AI + Planetary Justice Alliance**'s AI Supply Chain Impact Framework insists that any serious climate conversation must also track the extractive, labor, and socio-political harms embedded in AI's global supply chain, from mining and chip fabrication to data centers and e-waste.

The Three Mile Island restart, which is the tip of the spear for the industry-driven campaign to launch nuclear programs more broadly, is framed as "clean power for AI" and illustrates how the AI build-out can re-open legacy generation, reshape regional planning, and re-ignite long-standing risk debates (nuclear safety, waste, emergency preparedness). **Data & Society** argues that these choices are political and local: benefits (jobs, tax base, new transmission) and burdens (risk externalities, rate impacts) are unevenly distributed. Data & Society suggests linking data-center permits and long-term power purchase agreements to community benefit agreements, emergency-planning disclosures, and independent environmental justice reviews. Site fights over data centers have become a battleground for communities to make their voices heard, acting as a physical manifestation of public anxiety and providing an opportunity for agency and free speech before deals are inked.

FAS structures AI's footprint into three tiers: 1) computing-related impacts (training/inference energy, water, hardware, e-waste); 2) application-level impacts (how AI is used: e.g., to optimize buildings or, conversely, to drive additional consumption); and 3) system-level impacts (grid stability, land use, upstream mining, community effects). Absent standardized metrics, planners and the public cannot evaluate trade-offs or set conditions for growth.

The **AI Supply Chain Impact Framework** deepens this view by mapping environmental and socio-political indicators across every stage of the AI lifecycle, from raw material extraction and materials manufacturing to equipment production, model training, deployment, and disposal—and by highlighting where data are missing or obscured. It foregrounds Majority-World communities, resource colonialism, labor conditions, and community health, and offers concrete questions and metrics that regulators, civil society, and affected communities can use to demand transparency and allocate responsibility for harms across the supply chain rather than only at the data-center gate.

> **"As data center construction intensifies in many regions around the world, it's crucial that we establish better metrics for weighing the environmental and social impacts of these infrastructures in specific locations. Within the US, we have yet to establish standards for measuring the water, energy, and carbon costs of data centers, and we also need to gather qualitative data on quality of life issues that affect neighboring communities, from air quality and other public health matters like noise pollution to land use and higher utility bills. It is important for policymakers, advocates, researchers, and grassroots organizations to work together to gather empirical evidence in the face of financial speculation, the erosion of environmental protections, and tech lobbyist talking points."**
>
> **Tamara Kneese**
> **Director, Climate, Technology, and Justice Program, Data & Society Research Institute**

DAIR and **Hugging Face**'s research shows why fairness, safety, and climate cannot be separated in practice: design choices that drive accuracy or speed also drive energy, water, and hardware churn. They offer best practices like lifecycle accounting, efficiency baselines, reproducible reporting, and "sustainability by default" development norms that RAI efforts can adopt now. The Supply Chain Impact Framework complements this by providing a practical scaffolding for such lifecycle accounting and by treating missing data as a governance problem in itself—an advocacy hook for disclosure mandates, third-party audits, and community monitoring.

AI's climate and sustainability footprint (including second-order effects) will be decided by how and where we build it, by which stakeholders are involved in the decision-making process, and by the evidence that governs those choices. Data & Society shows the stakes for communities when AI demand steers energy policy; FAS provides a practical reporting blueprint so impacts can be compared and managed; DAIR and Hugging Face researchers tie ethics to environmental practice across the lifecycle; and the AI Supply Chain Impact Framework equips civil society with a structured, justice-centered tool to interrogate the entire chain of extraction, production, deployment, and disposal. The RAI community can focus their efforts on these areas in the coming year to reduce the environmental damage from AI that undermines climate goals, public health, and energy justice.

Economic Impacts

If today's AI boom is possibly a bubble, then the public (specifically local communities, workers, and taxpayers) will absorb outsized risks while concentrated actors capture most gains. Three resources illuminate the problem from different angles: **Data & Society's** brief on the myths powering data-center expansion cautions against overpromising local prosperity; recent economic analyses of AI exuberance by researchers from **DAIR/Carnegie Mellon** and **Cornell Tech** warn of bubble dynamics and fragile productivity assumptions; and **AI Now Institute's** 2025 Landscape Report maps how market power and policy capture channel public resources toward private AI agendas.

Local "prosperity" narratives can mask cost shifting. Data-center deals routinely hinge on tax abatements, discounted power, water priority, and accelerated permitting, while jobs are few and highly specialized. **Data & Society** warns that headline claims about broad regional growth rest on thin evidence; meanwhile, communities inherit long-lived infrastructure costs and environmental externalities (power capacity, transmission upgrades, water stress). In short: upside is privatized, downside socialized.

Researchers from **DAIR/Carnegie Mellon** and **Cornell Tech** argue that the dominance of the hyperscale cloud providers in the cloud-computing and AI infrastructure market is not just a matter of providing infrastructure; these firms also exert substantial influence financially through strategic investments in startups and ecosystem players, deepening their dominance by locking in downstream users, shaping markets, and creating dependency across AI supply chains. The authors identify AI equity overvaluation fueled by liquidity and sentiment, alongside a belief that near-term productivity leaps will justify today's prices. If those leaps stall, valuation resets can be sharp, even if AI remains useful. Unlike 2008, exposures are more equity than debt, but a correction can still hit pensions, municipal revenues tied to capital-gains taxes, and local plans predicated on permanent tech growth.

AI Now documents how a handful of firms leverage data, compute, and lobbying efforts to steer public subsidies, shape standards, and pre-empt local oversight, reducing democratic control over where and how AI infrastructure is built. In a bubble, this imbalance widens: public monies chase hype rather than proven public value. Even a "soft landing" can leave stranded public assets.

Bubbles are public-policy problems because they socialize losses and privatize gains. The signs are familiar: exuberant valuations, thin productivity evidence, aggressive siting deals, and policy capture. Downturns compound labor and service risks. When expected AI efficiencies fail to appear, organizations often cut headcount or public services to meet "savings" targets while continuing to pay for sunk AI contracts. That pattern has been observed across adjacent "efficiency" tech cycles and is consistent with bubble unwinds: quality degrades first, then equity.

Data & Society shows why data-center boosterism can leave communities holding the bag; macro analyses warn that sentiment-driven booms can reverse quickly; and AI Now explains how concentrated power turns hype into public commitments. Civil society should anticipate and measure these knock-on effects.

Public AI

The previous section draws attention to the risks of the economic concentration of private benefits against public risks, pointing toward the opportunity that exists for an intentional push for Public AI.

The valuations of the Magnificent 7 (or, rather, 10), on which (nearly 40% of) the prosperity of the U.S. stock market currently rests, continue to reach new heights, lurching precariously with every undisciplined media hit and defensive podcast appearance by a member of a frontier model developer's C-suite.

Meanwhile, the pervasive winner-take-all framing of the AI buildout – as evidenced by the frantic dealmaking between Open AI and every significant cloud, chip, and component company on Earth (and soon, probably, in Low Earth Orbit) – puts pressure on any alternative narratives.

We in the RAI community can demonstrate why the public should own the building blocks and infrastructure underpinning this technological era. We have the obligation and opportunity to articulate the future that we want and the value of building that future on technology that belongs to the public – not as a commodity, but as shared infrastructure, akin to the highway system or public libraries.

Why shouldn't it belong to us? It's by us and of us, freely taken from us and sold back to us (and, if history is any guide, soon to be sold to advertisers as well). The tech future we want pays us back for the absorption of our collective intelligence.

Across infrastructure, data, compute, and model development, we identify some of the most impactful contributions to what could be the most salient RAI issue of the coming year.

Infrastructure

The goal of Public AI, explained by **The Public AI Network** as "AI as public infrastructure" building on **Mozilla Foundation**'s 2024 paper emphasizing public goods, orientation, and use, and inheriting framing from earlier ideas on Digital Public Infrastructure, is AI that is funded, governed, and oriented as public infrastructure rather than just commercial product. It is not about safer private models, but durable, shared assets, across compute, datasets, models, tooling, and institutions, that are governed for public benefit.

A pivotal resource from **The Public AI Network** argues for a commons-oriented stack with three features: public access, public accountability, and permanent public goods, which are funded, governed, and maintained like other essential infrastructures.

Shared Public AI infrastructure is the shortest path to broad access, durable accountability, and long-term innovation capacity. The Public AI Network's Infrastructure for the Common Good supplies the definition and architecture of Public AI.

Project Liberty Institute's Digital Infrastructure Solutions to Empower Citizens: A Toolkit for Policymakers offers a complementary roadmap for how governments can actually design and govern this kind of Public AI infrastructure. It treats digital infrastructure, spanning data centers, cloud services, protocols, and data rights, as too vital to leave to proprietary interests and proposes a four-stage process for public leadership: assess existing infrastructure and institutional capacity, design for openness and interoperability, safeguard with rights-based governance and accountability, and adopt through digital literacy and citizen engagement. Framed around "data agency” as a path to digital sovereignty, the toolkit underscores that public AI stacks will only serve the common good if people have meaningful voice, choice, and stake in how their data and infrastructure are used. Read together with the Public AI Network’s blueprint, it positions governments not just as regulators of private AI, but as market shapers and stewards of shared AI infrastructure.

Public AI defines a stack where public and civic actors co-steward compute, data, models, and developer tools, with open interfaces and portability. Its minimum viable criteria, including public access, public accountability, and permanent public goods, prevent enclosure and make capabilities reliably available to universities, startups, civil society, and government service delivery.

The benefit is a political-economy flywheel. Public AI investments expand access, which broadens participation in problem-solving, which produces visible public benefit, which earns trust for sustained funding. This shifts AI’s gains from targeted firm-level advantage to broad societal capability. The result is sovereignty paired with openness.

Aspen Digital has become a key institutional amplifier and coalition-builder for this agenda, helping to grow the Public AI Network from a small cluster of academics and civil society organizations into an international community of more than 350 members and over 100 organizations, and providing the “community infrastructure” the movement identified as missing: convenings, operational capacity, and network strategy. Aspen Digital has helped to translate the abstract vision of Public AI into a concrete community of practice.

Shared AI infrastructure is possible, through open interfaces and transparent documentation of data lineage and evaluations. Institutionally, a recommendation from Public AI Network is to establish (or fund) public-interest operators – national labs, municipal consortia, public cloud exchanges, or university-civic alliances – that run shared compute and host models and datasets under public charters, with community oversight boards and open performance dashboards.

Thanks to organizations like the Public AI Network, Aspen Digital, and Project Liberty, the RAI community has a strong blueprint for how to convert AI infrastructure into a common good that we build and steward together.

Data

“Public data” isn’t just *available* data; it’s data that communities can *use, govern, and benefit from* across languages, abilities, and contexts. Four resources point to what that requires: **DAIR**’s “Every dataset has a perspective” reminds us that datasets encode choices and power; **Mozilla Foundation**’s Common Voice shows how community-led data creation unlocks speech technology for underserved languages; **Mozilla** & **EleutherAI**’s best-practices blueprint details how to curate and release open LLM datasets responsibly; and **Stanford HAI**’s Mind the (Language) Gap maps the structural barriers facing low-resource language communities. Together they argue for public data that is participatory, well-documented, and explicitly designed to close inequities.

DAIR’s zine is a creative reminder that every dataset reflects choices: what to collect, how to label, which categories to include or erase. Public data efforts should therefore ship dataset “perspective notes” alongside technical cards: who decided the taxonomy, who was left out, what trade-offs were made, and how users can contest or amend records. Elevating perspective from footnote to first-class field reduces “neutrality theater” and helps practitioners choose fit-for-purpose data.

Mozilla Foundation’s Common Voice demonstrates a working model: a global, community-run platform where people donate and validate voice clips to grow an openly licensed, multilingual corpus, powering inclusive speech tools and language revitalization efforts. This "participatory pipeline" aligns incentives (contributors see tools for their own speech) and creates governance touchpoints (local stewards, transparent releases). Public data initiatives should emulate this structure: recruit local leadership, compensate moderators, and publish contribution dashboards for each language.

Mozilla and Eleuther AI's Towards Best Practices for Open Datasets for LLM Training translates values into practice: use clear open licenses, track provenance and consent, publish rich documentation (sources, filtering, known gaps), and plan governance (update cadence, deprecations, dispute processes). It also stresses cross-functional work across legal, technical, and policy, as well as shared metadata standards so datasets are comparable and reusable across projects, which is an essential precondition for audits and downstream accountability.

Creative Commons' new CC Signals framework adds a missing reciprocity layer to this picture: a way for dataset holders to express machine-readable preferences about how their content can be reused for AI training and generation, grounded in the values of the commons rather than pure extraction. Building on the success of CC licenses, CC Signals proposes limited but meaningful options that let communities indicate acceptable uses, conditions, and expectations of "give-back" in the AI context—offering an alternative to both unbounded scraping and total enclosure behind paywalls. For public data initiatives, this points to a next step beyond documentation: pairing open licenses and dataset cards with preference signals that encode community norms and reciprocity expectations, so that "public data" is not just reusable, but reused on terms that sustain the commons over time.

HAI's Mind the (Language) Gap documents that most major LLMs underperform for non-English (especially low-resource) languages and miss cultural and contextual nuance. The white paper calls for community-driven pipelines, local funding, and rights-respecting data access tailored to Global South constraints. Public data programs should prioritize under-resourced languages first: fund collection in community settings (schools, radio archives), support local NLP labs, and design contribution workflows for low-bandwidth environments and diverse scripts.

Public data becomes a public good only when communities co-create it, licenses and documentation make it verifiably reusable, and equity, especially language equity, is a design requirement, not an afterthought. DAIR explains the politics of curation; Mozilla's Common Voice shows the operating model; Mozilla and Eleuther AI provide the tactical how-to; and HAI sets the global language agenda. The RAI community has the resources to incorporate these lessons into funding, procurement, and standards, so our public data better serves the public.

Compute

Public, shared, common compute enables inclusive AI research, resilient public services, and accountable innovation. Two recent anchors clarify both the why and the how: **Ada Lovelace Institute**'s Computing Commons maps policy design for government-funded access to compute; **GovAI**'s Inference Scaling and AI Governance shows why governance keyed only to training compute will miss emerging risks and opportunities as inference compute becomes the growth engine. Read together, they argue for shared compute that is accessible, governable, and future-proof.

Public compute comprises initiatives that use public funds to provide access to compute, via public supercomputers, shared GPU clusters, or credits on commercial clouds, with the aim of enabling research, innovation, and public-interest use that markets under-serve, as well as initiatives like

CalCompute, newly codified in California's SB-53 (for now).

Global approaches vary (direct hardware, vouchers, public–private facilities) and are still experimental, but a common aim is to reduce dependency, widen participation, and support pluralism in who can build and evaluate AI.

The **U.S. National Artificial Intelligence Research Resource (NAIRR)** Pilot offers an early, operational glimpse of this model of shared compute. Through its allocations program, NAIRR provides no-cost access to a mix of federal and nongovernmental resources—supercomputers, cloud credits, testbeds, datasets, and privacy-preserving tools—for researchers and educators whose projects align with societally relevant focus areas, from AI safety and climate to health and infrastructure. Project results must be open and publishable, and allocations are time-bounded, peer-reviewed, and matched to clearly justified resource needs, creating a built-in feedback loop between public investment, scientific progress, and accountability. Read alongside Ada's Computing Commons and GovAI's inference-scaling analysis, NAIRR's design suggests key ingredients for future public-compute programs: transparent, criteria-based allocations; requirements for open dissemination; support for education as well as research; and an explicit mandate to channel advanced compute toward public-interest use cases that markets under-serve.

Governance is shifting from "how big is the training run?" to "how much inference compute do systems consume at deployment or during iterative training?" Policies that ignore inference scaling will misclassify risk, misallocate capacity, and blunt accountability.

GovAI highlights two pathways that reshape public provisioning: 1) Inference-at-deployment (spending more test-time compute per task for deliberation, tool use, search), and 2) Inference-during-training (using heavy test-time reasoning to generate data, then distilling). Public compute must be built for continuous, metered, auditable inference, not only for episodic training runs.

Ada Lovelace Institute argues, building on their 2024 case for public compute, that a credible Computing Commons program should encode four layers of responsibility: 1) Access & equity, 2) Transparency & auditability, 3) Safety & responsible use, and 4) Pluralism & anti-lock-in. Public compute can widen participation and strengthen accountability, if it is designed for the world we are entering: one where inference is the dominant, continuous load and a major governance lever.

Ada Lovelace Institute's Computing Commons offers the policy framework; GovAI's Inference Scaling explains the technical trend that must shape it. The RAI community can now incorporate these into budgets, operating procedures, and eventually laws, so that shared compute truly delivers shared public value.

Model Development

New evidence published by **GovAI and the University of Edinburgh** forecasts a rapid expansion in the number of "frontier-scale" models captured by compute-based rules, which will strain any governance scheme that cannot scale evidence-collection and oversight. Meanwhile, **PAI** shows that post-deployment documentation of real-world impacts is the biggest adoption gap currently. Together they argue for a shift from point-in-time disclosures to continuous, value-chain accountability with artifacts that regulators, procurers, and the public can actually use.

PAI's Documenting the Impacts of Foundation Models finds that while providers routinely publish model/system cards, systems rarely document post-deployment impacts (where, how, and with what effects the model is used). PAI maps how obligations change with release type (hosted APIs vs. open weights), and urges collaboration among model hosts, app developers, and cloud providers to monitor incidents and policy violations for open models without recreating surveillance.

Static compute thresholds will soon cover many more models. **GovAI and University of Edinburgh's** Trends in Frontier AI Model Count analysis forecasts that by the end of 2028, between 103–306 models will exceed 10^{25} FLOP (the EU AI Act marker) and 45–148 will exceed 10^{26} FLOP (the U.S. AI Diffusion Framework "controlled models"), with superlinear growth each year. Capacity planning for documentation, evaluation, and enforcement must assume this ramp. Frontier-connected thresholds are more stable but still require triage. Thresholds defined relative to the largest training run yield roughly 14–16 models per year, suggesting regulators pair absolute triggers with relative ones and, either way, invest in scalable evidence pipelines.

Moving forward, the RAI community has guidance for some important 2026 priorities. Trustworthy model development now means live, verifiable accountability: impact documentation that persists after launch, responsibilities matched to release strategies, and oversight scaled to the number of frontier-class models we will actually see, not the handful we have today.

Public Interest AI

The tech future we want serves all of our needs, with AI solutions applied responsibly toward society's most pressing problems. People do not assume that "AI for public good" will arrive evenly or fairly. Public good is contextual and contested, and it should be negotiated with communities, not declared from the center.

In the **Ada Lovelace Institute**'s Making Good report, study participants consistently defined public good around fairness and equity, social connection and community, and robust public services that let people live meaningful, secure lives.

he report calls for participatory, place-based methods that surface concrete expectations: that AI be pro-social and equitable, relational and caring, future-focused, and responsibly deployed only where it is necessary and effective. Public Interest AI, in this framing, is less about branding projects as "for good" and more about redesigning governance so that public definitions of good, which are rooted in lived experience, locality, and structural conditions, actually steer investment, deployment, and oversight.

For the RAI community, the core lesson is that public interest cannot be inferred from generic attitudes surveys or innovation narratives; it has to be built through deep, ongoing engagement with diverse publics, including people who are usually excluded from AI decision-making.

The **Public Interest AI** (PIAI) project, anchored at the **AI & Society Lab** of the Alexander von Humboldt Institute for Internet and Society, defines PIAI as systems that support "the long-term survival and well-being of a social collective construed as a public," drawing on public-interest theory from Dewey, Bozeman, and others. Public interest is not assumed to be universal or obvious, and which cannot be left to private firms or technical experts alone to understand, but something that must be defined for each issue through participatory processes.

The PIAI Project Map is a structured dataset of AI projects that explicitly aim at public interest goals. The interactive world map aggregates survey data on projects' objectives, methods, and frameworks and publishes them both as a visual directory and as an open research dataset, including on platforms like Hugging Face. This makes public interest AI legible as a global practice: who is working on what, where, with which theories of change, and with what funding and institutional backing.

For the RAI community, the map and network can help align funding with under-served domains, connect local initiatives into global coalitions, and ground high-level principles in concrete projects that can be evaluated over time.

Crucially, this is not just theory: a growing set of initiatives are already embodying public-interest AI in practice across different layers of the ecosystem. **Consumer Reports**' work on "consumer-authorized AI agents" and Loyal Agents prototypes shows what it looks like when AI agents are designed to work for people rather than platforms—helping consumers submit feedback, navigate commerce, and exercise their rights in ways that are aligned with their interests and consent, not data extraction.

> **"AI-powered nonprofits are modeling what responsible AI can look like. Fast Forward's data shows that 71% of AI-powered nonprofits respondents already have processes in place to assess and mitigate AI risks, but 41% point to the lack of in-house technical expertise as a barrier to meeting their ethical ambitions. In grantmaking, funders should encourage transparent policies, support ethics training, and offer governance templates — and provide the funding necessary to do so. By supporting responsible practices, funders signal that equity and accountability are as important as innovation."**
>
> **Kevin Barenblat**
> **Co-Founder, Fast Forward**

Fast Forward's The Philanthropic Reset: How Philanthropy Can Lead in the Age of AI report documents how nonprofits are deploying AI to extend access to healthcare, education, climate resilience, and justice, while also surfacing the funding and capacity gaps that must be closed for mission-driven organizations to shape AI rather than simply absorb it.

At the infrastructure layer, **OpenFold3** demonstrates how frontier scientific capability can be organized as a commons: an open-source foundation model for protein and drug structure prediction that accelerates biomedical research beyond the walls of any single firm.

Similarly, the **Patrick J. McGovern Foundation**'s Grant Guardian tool applies AI to strengthen philanthropic due diligence and transparency, making capital allocation more rigorous and freeing staff to focus on relational work with grantees.

In the safety layer, **ROOST**'s open-source Coop and Osprey tools bring enterprise-grade trust-and-safety infrastructure—content review consoles, investigation and incident-response capabilities—within reach of organizations that would never be able to build or buy such systems on their own, turning online safety into a shared, inspectable public good rather than a proprietary advantage.

Taken together, these efforts illustrate what a multi-layered public-interest AI ecosystem can look like: communities defining the public good; researchers mapping the field; consumer advocates, nonprofits, and foundations building concrete tools that redistribute power and capability; and open scientific and safety infrastructure that others can reuse and scrutinize.

For the RAI community, the task now is to treat these projects not as isolated "success stories" but as patterns to scale, using funding, procurement, and standards to reward AI that is loyal to people, governed in the open, and oriented toward durable public benefit.

Re-Imagined AI

The grandest version of RAI is ambitious AI that enables us to re-imagine our tech future and achieve our wildest aspirations. This is the RAI we truly want to focus on in the coming year.

Ambitious AI futures are not just technical scenarios. They are stories, memories, and images about who technology is for and what kinds of lives it makes possible. If we want different outcomes from AI, it requires transforming the cultural and narrative infrastructure that shapes how we imagine, fund, and evaluate technology in the first place.

A handful of projects have exemplified that spirit of ambitious reimagining in 2025. **The Collective Intelligence Project (CIP)**'s founding whitepaper makes this explicit, arguing that current governance models trap us in a "transformative technology trilemma" between progress, participation, and safety, and proposing a "CI stack" of democratic processes and new institutional forms so we can consciously remake the rules and containers that steer transformative technologies like AI.

The Tech We Want (TTWW) is building an alternative tech ecosystem around people, purpose, and profit, explicitly centering leaders and communities who are already working toward tech in which everyone can thrive. Its Infrastructure Fund uses community power-building, participatory budgeting, and trust-based philanthropy to resource alternative tech, narratives, and governance experiments, treating funding itself as a site of culture change rather than a neutral pipeline.

A People's History of Technology complements this by assembling a living archive of how everyday people and workers experience technologies like mobile phones, labor platforms, and social media, rather than retelling heroic stories of lone inventors, and we look forward to the inevitable "People's History" treatment of Artificial Intelligence, which we imagine could help usher RAI toward a "People's Future" of Tech.

Aspirational AI futures should be grounded in lived histories and material power shifts around who gets funded, who gets to tell the story of "innovation," and whose experiences count as expertise.

DAIR's Possible Futures series and zines push this further by making speculative imagination a practice, not an ornament. Through workshops and zine-making, participants write diary entries from the future, create collages, and experiment with "imagining otherwise": asking what happens when we give ourselves permission to envision technofutures beyond automation hype and inevitability narratives.

These futures are often messy, local, and grounded in care, solidarity, and survival, especially for those harmed by current AI deployments. **Speculative F(r)iction**'s Living Library adds a complementary lens: it treats friction, or the discomfort, tension, and slowness in AI systems and governance, as a resource rather than a bug, using design fiction and storytelling to disentangle "form, function, fiction, and friction" in AI.

Several projects push this reimagining all the way down into the infrastructures and supply chains that AI rests on. **Estampa**'s Cartography of Generative AI project offers a critical visual map of the generative AI ecosystem, tracing not only models and companies but also data extraction, invisible labor, environmental impacts, and regimes of truth—explicitly challenging maps that present AI as weightless, neutral, or inevitable.

The **AI + Planetary Justice Alliance**'s Raw Materials for AI zine performs a similar move at the level of hardware, visually unpacking the minerals (copper, silicon, rare earths, cobalt, etc.) that power chips, batteries, and data centers, and explicitly linking "what is below the algorithm" to extractive mining, uneven environmental burdens, and planetary justice struggles.

Read together with CIP's call for new collective intelligence institutions, these works insist that reimagining AI also means reimagining who governs the infrastructural stack—from mineral frontiers and logistics to corporate labs and regulatory bodies.

RAI work should routinely incorporate speculative and narrative methods and critical cartographies of power and matter, especially with marginalized communities, to surface blind spots, contest dominant trajectories, and design policies that leave room for refusal and alternative paths.

Kernel Magazine and **Better Images of AI** show how culture and aesthetics shape the boundaries of the possible. Kernel, an annual publication from Reboot, offers long-form essays, narratives, poetry, and art that reimagine techno-optimism, neither surrendering to fatalism nor to hype. It is explicitly framed as a home for "historical research, cultural critique, and possible futures," created by technologists reflecting critically on the systems they build and inhabit.

Better Images of AI tackles a more specific but powerful layer: visual metaphors. By creating and curating non-stereotypical, freely licensed imagery to replace glowing brains, humanoid robots, and abstract binary code, the project aims to improve public understanding of AI's strengths and limits and to counter visual tropes that encode fear, dehumanization, and biased assumptions.

Estampa's Cartography of Generative AI and **AI + Planetary Justice Alliance**'s Raw Materials for AI extend this visual politics into maps and diagrams, making visible the infrastructures, extractive geographies, and contested truths that underwrite "intelligent" systems.

These efforts remind us that how AI is pictured, in media, policy decks, educational materials, etc., quietly shapes who feels entitled to question it, who is imagined as a subject or an object, and which futures seem plausible. For civil society and the broader RAI ecosystem, these projects collectively point toward a more ambitious understanding of "aspirational futures."

It is not enough to articulate guardrails around the current trajectory; we must invest in narrative infrastructures, participatory memory projects, speculative methods, and visual cultures that expand what futures we can even see. That means funding and legitimizing artists, storytellers, and community archivists alongside lawyers and engineers; building participatory funds and media labs that resource alternative tech and governance experiments; and embedding speculative and friction-embracing practices into policymaking, standards work, and public engagement.

We are not looking for abstract utopias. Re-Imagined AI futures, in this sense, are ongoing collective projects that reconfigure who gets to imagine, build, and benefit from technology in the first place.

Looking Ahead

As we close this 2025 Responsible AI Impact Report, we are already orienting our work toward the realities and responsibilities of 2026. The central question for All Tech Is Human's RAI efforts is no longer whether AI will shape our institutions, economies, and daily lives, but how, and in whose interests. Our answer, as ever, is to build and strengthen the Responsible Tech ecosystem itself: people, practices, and public infrastructure capable of steering AI toward collective benefit and away from the futures we seek to avoid.

In the coming year, we intend to increasingly utilize AI responsibly in service of our mission, as committed practitioners of public interest AI. We hope to develop an AI-powered version of our Responsible Tech Job Board to better match diverse talent with meaningful roles, surface emerging skills, and make the ecosystem more navigable for those trying to enter it. We hope to scale our Responsible Tech Mentorship program using AI-assisted matching, guidance, and resource curation, always with human judgment, community norms, and equity at the center, to make it easier for new voices and non-traditional pathways to find a home in this field. In doing so, we intend our own infrastructure to be one example of what Responsible, mission-aligned AI deployment in the Public Interest can look like.

We also intend to deepen our commitment to Public AI infrastructure: AI that is treated as a shared asset, stewarded in the public interest, rather than a series of siloed products. In 2026, we plan to support and amplify ongoing efforts to build and govern community datasets, shared compute, and open tooling under public and civic charters. We seek to partner with like-minded organizations to prototype "public-interest operators" who can host and maintain these resources.

Across all of this, we see Responsible AI literacy as the through-line. Next year we will expand our education work so that RAI is not just the domain of experts and insiders, but a shared civic competency: expanding our course offerings beyond the initial five released in 2025 and reworking our course library for additional audiences and platforms.

Our 2026 RAI workstream will focus on three AI-centered issue areas.

First, Training Data: we look to help bridge the disconnect between the community datasets that power small, targeted public models and the expanding data appetites of private commercial foundation models with products and interfaces that will increasingly expand into wearables and embodied devices utilizing ongoing data collection through real-world surveillance. Our work will highlight data rights, community governance, and public-interest training pipelines as alternatives to this mutually-assured surveillance.

Second, Synthetic Companions and Synthetic Media: we intend to amplify efforts to ensure AI companions reduce isolation rather than deepen it, support governance for high-risk use with teens and vulnerable users, and strengthen information integrity and content provenance efforts in a world where it is increasingly hard to tell who, or what, is speaking.

Third, Trust and Assurance: we will prioritize the connective tissue of AI governance, including standards, benchmarks, evaluations, audits, and red teaming, so that industry assurance efforts rest on a more robust intra-party system of trust. We intend to engage in civil society efforts to lead in building the accountability frameworks that Agentic AI will demand.

We do not assume that better futures will emerge on their own from technical progress. In 2026, we intend to help build them: by using AI in ways that strengthen, rather than hollow out, the Responsible Tech ecosystem; by supporting the growth of shared Public AI infrastructure; and by raising the floor of RAI literacy so more people can understand, question, and redirect the systems shaping their lives. The work ahead is demanding, but it is also generative: every dataset governed in the public interest, every benchmark co-created with communities, and every person newly empowered to engage with AI responsibly is another step toward the futures we intend to claim.

About All Tech Is Human

All Tech Is Human is a nonprofit organization dedicated to building a more responsible, inclusive, and ethical tech future by uniting a broad range of people and ideas across civil society, government, industry, and academia. Our whole-of-ecosystem approach brings together technologists, academics, civil society leaders, and everyday individuals to collaboratively address complex challenges around AI governance, digital rights, platform accountability, and the broader social impacts of technology.

We aim to develop a better approach for tackling thorny tech & society issues; one that leverages collective intelligence, involvement, and action to create needed systemic change.

Through community-building, talent cultivation, public programming, research, and analysis, our organization fosters dialogue that bridges gaps between disciplines and encourages shared responsibility for shaping technology that aligns with our values. By fostering collaboration and elevating a wide range of perspectives, All Tech Is Human works to ensure that technology is shaped by and for the public interest.

If you would like to work with us or learn more, please email hello@alltechishuman.org.

Learn More About Responsible AI

All Tech Is Human launched five dynamic Responsible AI courses designed for aspiring RAI practitioners and AI Governance professionals, as well as those preparing to build AI Governance programs within an organization.

These courses equip participants with the essential foundational knowledge to begin the process of effectively operationalizing Responsible AI and AI governance programs within organizations.

Rooted in practical insights and real-world applications, this series of five short courses, which can be completed in just a few hours, offers an introductory understanding of the foundations of Responsible AI: its principles, history, and evolution as a field, as well as an exploration of current roles, industry best practices, and the evolving governance landscape.

Take our RAI courses made possible and freely available thanks to the generous support of The Patrick J. McGovern Foundation.

all tech is
human

Together, we work to solve tech and society’s thorniest issues.