KAPOR FOUNDATION

RESPONSIBLE AI

Principles for Advancing a More Equitable Innovation Future

Learn more and
download our principles
kaporfoundation.org

RESPONSIBLE AI PRINCIPLES: EXECUTIVE SUMMARY

Artificial Intelligence (AI) will shape education, labor and the workforce, the economy, civil and human rights, and the environment for decades; and how we design, develop, and deploy AI today will determine who benefits tomorrow. To build a future in which AI benefits society, we must first develop and adopt a set of principles to guide our investments in AI for racial, economic, and environmental justice. Informed by existing research scholarship, frameworks, and definitions advanced by leaders in the field of AI ethics, tech justice, and civil and human rights, the Kapor Foundation has identified the following five core principles to guide our investments:

1. Utilize a sociotechnical framework to identify challenges and meaningful solutions.

We must clearly define the types of societal problems we aim to solve, evaluate whether AI can and should be deployed as a tool to address these challenges, and consider the broader societal dynamics in which the AI tool is situated.

2. Incorporate prosocial design principles and continually assess broader societal impacts.

We must employ prosocial and design justice principles by ensuring that solutions are designed with societal benefit at the forefront, the communities most impacted by AI are centered in design, regular audits of impact are conducted, and the entire lifecycle of AI development can achieve its intended social good.

3. Support AI initiatives that shift power.

We must support new and more inclusive business models, compensation/incentive structures, and investment strategies, while building power across researchers, academic institutions, nonprofits/grassroots organizations, and policy advocates to raise concerns and propose solutions to address harms of AI.

4. Promote critical AI literacy and education across society.

We must expand access to computing education for all students, while advancing critical AI literacies amongst innovators, workers, consumers, and advocates to ensure all are empowered to make decisions about AI's development, adoption/use, and impact.

5. Build collective mechanisms for governance and accountability.

We must build the capacity of a broad coalition of journalists, research scholars, whistleblowers, policymakers, and advocacy groups to shape the future of technological innovation through responsible governance, regulation, and accountability mechanisms.

RESPONSIBLE AI:
PRINCIPLES FOR ADVANCING A MORE EQUITABLE INNOVATION FUTURE

Artificial Intelligence (AI) will shape education, labor and the workforce, the economy, civil and human rights, and the environment for decades; and how we design, develop, and deploy AI today will determine who benefits tomorrow. To build a future in which AI benefits society, we must first develop and adopt a set of principles to guide our investments in AI for racial, economic, and environmental justice.

Recent advancements in AI have permeated all aspects of our lives, from reshaping how we communicate, create, work, and access information to uncovering new potential for breakthroughs in areas including education, medicine, transportation, and the environment. These advancements have driven optimism about the use of AI to tackle some of society's most complex challenges. Yet, despite optimism about AI's potential to transform society for "good," AI advancement and adoption have not come without significant risks and challenges– especially to the most marginalized communities across the globe. Documented detrimental impacts include: the displacement of entry-level workers (disproportionately impacting Black workers) as a result of the hype around possible automation, negative impact of AI chatbots on mental health, discriminatory facial recognition algorithms leading to wrongful arrests, the adverse impact of data centers on public health and climate, deployment of automated decision-making systems in government, and the vast underrepresentation by race and gender in the tech/AI sector and subsequent widening economic inequality. Furthermore, AI has been deployed in ways that cause irreparable harm to vulnerable communities, through the utilization of AI for surveillance of immigrants, protesters, and government targets, and its usage across the federal government to arbitrarily eliminate billions in scientific and social safety net funding. These harms are likely to only be exacerbated by strong opposition to state and federal regulation, and the rolling back of responsible AI commitments, voluntary guidelines, and business practices supporting ethical AI in the private sector to pursue profits at all costs.

Concerningly, we are witnessing the power to shape AI slowly becoming consolidated and concentrated into the hands of a select few. The US government and US-based technology companies and venture capital (VC) firms have largely driven the rapid advancement and investment in AI–and subsequent narratives about its potential for drastic societal transformation. The most recent federal budget contains significant investments in AI infrastructure, manufacturing, and energy and billions invested in AI for national defense. In the private sector, the US continues to outpace other countries in AI investments at a rate nearly 12 times greater than China, the second highest in total capital invested. Significant capital investments have come from US-based tech companies; Meta, Microsoft, Amazon, and Alphabet collectively committed $325B towards AI infrastructure in 2025. Moreover, 46% of North American VC funding was invested in AI startups in 2024–and this percentage jumped to 70% in the first quarter of 2025.

This VC funding is coming from an increasingly small number of firms: nine firms raised over half of the total capital raised by VCs in 2024, with just four firms accounting for over one-third of the total capital. This consolidation makes it more challenging for smaller firms to take risks in offering more context-specific solutions, explore pro-social use cases for AI, and expand opportunities for a wider swath of companies to enter the market. Big Tech companies and VC firms have also used capital to play an outsized role in influencing government, with tech spending millions on lobbying efforts at the state and federal levels to advance their policy agendas–and oppose efforts for regulation. Finally, the economic gains from AI advancement have been far from evenly distributed: tech wages are 130% higher than the national median wage, while AI is projected to exacerbate racial wealth gaps and disproportionately displace low-wage and entry-level workers. This concentration of power and wealth within the hands of a few tech companies, VC firms, and governments excludes communities, civil rights organizations, and civil society from contributing to AI's design, deployment, decision-making, and regulation.

For AI to have a positive impact on society, we will need to intentionally identify specific opportunities for its transformational use, while addressing potential risks and harms to both individuals and society. We must resist the assumption that free market forces will ensure that technology will be used for ethical and prosocial purposes. Instead, we believe it is critical to establish principles defining responsible AI that can be adopted by funders across sectors, including philanthropy, venture capital, and public policy.

> **As a Foundation committed to reimagining and rebuilding a more equitable technology ecosystem, we outline the following core principles that shape our approach to investing in a future in which AI benefits everyone. Building upon the existing research scholarship, frameworks, and definitions advanced by leaders in the field of AI ethics, tech justice, and civil and human rights, we have identified five core principles to guide our investments.**

01

UTILIZE A SOCIOTECHNICAL FRAMEWORK TO IDENTIFY CHALLENGES AND MEANINGFUL SOLUTIONS

Technology is not neutral, but is designed with implicit and explicit values and frameworks that guide every decision, from its conception to its deployment and subsequent impacts. Importantly, AI is a sociotechnical system–a system in which technology and its impact on society are intertwined. Overlooking the sociotechnical nature of AI can perpetuate biases and harms and affect the scoping of solutions to some of society's largest challenges by ignoring opportunities or deploying technology in ways that are not suited to address the challenges at hand.

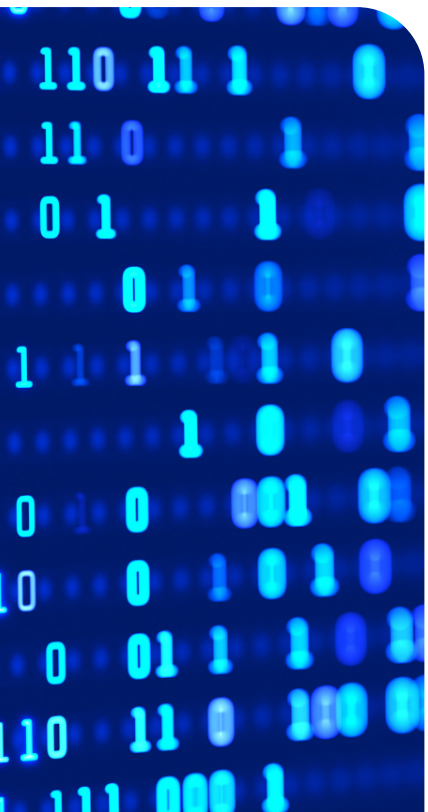
A sociotechnical framework is critical to designing and deploying technology for positive social impact and to understanding that the responsibility for fair, equitable, and just AI lives both within and outside of the technology itself. We must first begin by clearly defining the types of societal problems we aim to solve, evaluating whether AI can and should be deployed as a tool to address these challenges, and considering the broader societal dynamics in which the AI tool is situated. For example, AI has been shown to improve the accuracy and efficiency of medical imaging in early disease detection, to identify mechanisms to improve cancer research and treatment, to improve the detection and classification of oil spills, to increase precision in forecasting wildfire paths to inform communities and first responders, and to expand tracking carbon emissions for environmental monitoring. These are both problems that AI is well-suited to address and are solutions that would benefit society, and the field would benefit from more of this visioning and articulation at the earliest stage.

02

INCORPORATE PROSOCIAL DESIGN PRINCIPLES AND CONTINUALLY ASSESS BROADER SOCIETAL IMPACTS

Developing AI tools with the intention to have a positive social impact (e.g., in areas like healthcare, government benefits, and education) alone is insufficient, and an unjust or underdeveloped design and development process could create wide-ranging unintended consequences. Critiques of current approaches to AI design include the environmental cost of training, the exploitation of international data workers, the lack of informed consent and transparency on data collection and harvesting, suppression of voices raising ethical concerns, and the lack of diversity on AI teams. The deployment of AI technologies and their potential impacts on society must also be considered by creators in the process of design and continuously evaluated and addressed. Potential societal impacts of AI deployment to consider include data usage and impacts on climate change and sustainability; the potential for displacement of entry-level and low-wage workers; discriminatory impacts of bias in AI models in health, criminal justice, and employment; and increased economic inequality.

To create a just design process, we must employ prosocial design principles by articulating ways to engineer products towards positive social impact (as opposed to maximizing for engagement or profit) and incentivizing these approaches. We must also leverage design justice principles, ensuring that the communities most impacted by AI are involved and centered in the design process and that solutions are sustainable and benefit communities. To prevent further harm, particularly amongst marginalized communities, we must conduct regular assessments and audits of AI models and tools to assess for bias and unintended harm.



We must also deploy a principled stance on prohibitions of AI uses that defy fundamental beliefs in human and civil rights, including identifying targets for lethal force, mass surveillance, and for reinforcing patterns of injustice (see: Algorithmic Justice League's definition of accountable and equitable AI). In all AI use cases, there must also be articulated processes for appealing decisions, redressing harms, or halting development in totality. Embedding justice into the design and development of AI ensures that environmental, civil, and human rights remain at the forefront, and that the entire lifecycle of AI can achieve its intended social good.

03

SUPPORT AI INITIATIVES THAT SHIFT POWER

Our current innovation ecosystem concentrates power in the hands of a select few, allowing them to exercise an outsized influence in shaping the future of AI–and benefitting economically from its success. US-based tech companies, VC firms, and the US government have invested billions in AI startups, infrastructure, and in adopting pro-AI policies that impact global markets, geopolitics, and civic order. Further, the market dominance of a select few AI industry titans has restricted transparency, competition, and innovation to the detriment of national interests. Broad swaths of society are excluded from the process of designing AI solutions and making decisions about their deployment. Often, the most marginalized communities (including Black, Latine, Native people, differently abled, low-income, LGBTQ+, and immigrants) have little involvement in its creation and are excluded from its economic benefits–while also being the most impacted by irresponsible AI adoption.

Large corporations are profiting from AI models built on the intellectual property of human artists, journalists, authors, and individual citizens without their consent (including the use of copyrighted material), while the creators remain uncredited and uncompensated. Simultaneously, workers—from civil servants to software engineers and artists—are being replaced by AI, while startups are being incentivized to limit employee headcount. The clamor from an elite few to automate rote tasks, improve efficiencies, and increase productivity through AI tools has the potential to fundamentally change the labor workforce and exacerbate economic inequality through the elimination of a multitude of roles across all sectors and industries.

Responsible AI solutions must evolve to include more inclusive participation and address data collection strategies, business models/incentive structures, alongside broader community and economic benefit models, and powerbuilding. We must expand who participates in all aspects of AI design, development, and deployment beyond those who identify as technologists, recognizing and including other forms of expertise, including domain-specific experts (e.g., doctors, teachers, social workers) and community-specific experts (e.g., language and culture bearers). We must intentionally seek to include diverse and marginalized communities' perspectives, given their lived expertise and the importance of designing with those who will be using or impacted by the tools to ensure systemic inequalities are not reinforced and exacerbated.

We must move beyond VC as the only investment strategy, encourage alternative financing models that don't rely solely on rapid scale (and encourage extractive and exploitative tactics), and support more inclusive business models that create profit, incorporate shared compensation structures, and re-envision community data ownership and economic distribution models. We must also build power across sectors to raise concerns and address harms of AI innovation, including supporting researchers, academic institutions, unions, and other worker-led organizations, nonprofits/grassroots organizations, and policy advocates.

04

PROMOTE CRITICAL AI LITERACY AND EDUCATION ACROSS SOCIETY

The rapid development and adoption of AI tools has outpaced the critical AI literacy of those using and impacted them. Usage of AI tools, chatbots, and companions has increased exponentially, while the deployment of AI into high-stakes decision-making in such areas as housing, lending, hiring, healthcare, and surveillance has already demonstrated harmful impacts on vulnerable communities. The ability to weigh the benefits and risks of these AI innovations across sectors will require critical AI literacy of those using and impacted by the tools. Moving forward, the field must clearly define what constitutes AI literacy and the types of skills and knowledge that are critical to being consumers and creators of AI tools, contributing to its governance, and participating in a tech-driven society.

We must advance critical AI literacy in K-12 education, which equips all students and teachers with the knowledge and skills to critically interrogate the development, deployment, and impacts of AI, and moves significantly beyond an understanding of how to use AI tools. The use of AI tools has the potential to improve efficiencies, tailor learning experiences, and expand accessibility, but it also has shown great potential for harm, such as hallucinating information, exacerbating biases, and lowering critical thinking skills. We want students and teachers to become equipped to undertake equitable decision-making in AI adoption, including advocating for tools with gap-closing impact, and weighing the option to reject the deployment and usage of AI tools. Importantly, critical AI literacy alone is insufficient, as we must also expand equitable access to computing education, which includes the skills and knowledge required to build AI products and tools and to pursue technical careers.

As AI usage expands, individuals must be empowered to navigate this new world as workers, consumers, and advocates for a future which benefits society. Workers will need to navigate the changes to the labor market, including supporting critical assessments of adoption of AI technologies in their roles and companies, and investing in upskilling efforts that augment workers' existing expertise and ensure access to high-quality jobs in growing sectors. Consumers will need to be empowered with knowledge about not only the opportunities, but also the challenges, risks, and harms of AI, such as data privacy and surveillance issues, bias in training data and outputs, mental health risks associated with companion chatbots, and the risks of chatbot "sycophancy." Advocates (including community groups, grassroots organizations, and policymakers) will need to be equipped with knowledge and resources to examine the impact of AI integration on their communities and constituents and ways to take action as consumers and as advocates at the local, state, and federal levels.

05

BUILD COLLECTIVE MECHANISMS FOR GOVERNANCE AND ACCOUNTABILITY

In the AI innovation space, longstanding debates on the role that guardrails and regulation should play have been strong and polarizing. Opposing views on regulation have emerged, with some fearing that regulation will stifle innovation and global competitiveness, while others argue regulation is necessary to provide protection from harms and existential risk. Yet, these debates have been primarily dominated by the tech industry and VC perspectives and interests, including massive lobbying efforts to oppose regulation.

The scope of the challenge to promote responsible and ethical AI solutions is significant and will require efforts beyond the private sector's embrace of more responsible innovation. It must also include public sector regulation and governance–especially as the federal government continues efforts to restrict any guardrails or regulations, states diverge on AI regulatory policies, tech companies roll back guidelines and commitments to safe and responsible AI, and AI companies and VC lobbying firms continue tactics to oppose regulations. For-profit companies and their financial interests cannot be the only voice in the conversation. Government, civil society, and the public can and should have power in shaping the future of the technologies that will impact their lives.

AI innovations have impacted communities across the globe through the erosion of civil rights and privacy, exploitation of human labor, fueling of hate speech and discrimination, exacerbation of the climate crisis, furthering of inequality, and destabilization of the democratic process. Given the widespread misuse of AI tools, a broad coalition of individuals, organizations, and policymakers is required to shape the future of technological innovation through regulation and accountability mechanisms. We must ensure that grassroots, civil society, and movement organizations have the capacity to understand and examine AI innovation from technical, sociotechnical, and policy perspectives, and to advocate for policies on behalf of their constituents.

To that end, we must support multi-stakeholder coalitions of journalists, research scholars, whistleblowers, policymakers, and advocacy groups to work in concert to document and identify harms of AI, raise awareness among the general public, and advance policies or promote other accountability mechanisms. In one instance, a multitude of researchers, civil rights advocates, and tech workers organized against technology companies providing surveillance technologies to law enforcement; this organizing campaign, in part, drove the companies involved to halt this practice. We must also raise collective consciousness about both responsible and irresponsible examples of innovation–to provide alternative visions, drive support for responsible innovation practices and policies, and further marginalize bad actors. We must use these tactics collectively to sideline harmful practices; endorse ethical adoption, development, and use policies; and promote investment in responsible AI solutions that benefit all of society.

TAKE ACTION

The trajectory of AI innovation and deployment that is currently being driven by Big Tech, leading AI companies, and VC firms–as well as the narrative that AI will inherently transform society for the better–is often overhyped and is far from inevitable. In the face of the pressure toward unregulated growth, resistance to guardrails, and the belief that extractive and exploitative models are required for success in a profit-at-all-costs climate, we are reaching a critical inflection point. This moment will require collective action amongst philanthropy, entrepreneurs, scholars, journalists, organizers, educators, nonprofits, policymakers, and investors to mobilize resources and action towards ensuring a just design, development, and deployment process for future AI innovations, while demanding safeguards to protect communities from harm.

> **Rooted in the Kapor Foundation's mission to build a more equitable technology sector, economy, and society, the Foundation calls for collaboration in integrating these principles into grantmaking strategies and will seek out a coalition of aligned partners who share this vision for a responsible AI future. As the field continues to rapidly evolve, we anticipate the principles to evolve as well, and to incorporate new and innovative ideas as they emerge. We are also in the process of translating these principles into a set of investment policies and criteria to engage the broader investment, venture, and entrepreneurship ecosystem and seek collaboration in investing in a more just and equitable AI future.**

Suggested Citation

Scott, A., Hinton, L., Koshy, S., Gangas, L., & Jones, N. (2025). Responsible AI: Principles for Advancing a More Equitable Innovation Future. Retrieved from: kaporfoundation.org/responsibleai.

ACKNOWLEDGEMENTS

We are grateful to our external partners for their deep technical expertise and thoughtful feedback to develop and shape these principles rooted in the core values of equity and justice:

Alex Hanna, Director of Research, DAIR Institute
Elaine O. Nsoesie, Faculty Affiliate and Associate Professor, Boston University School of Public Health, DAIR Institute
Tina M. Park, Research Associate, DAIR Institute

Caroline Siegel Singh, Manager for AI and Emerging Technologies, FAS
Clara Langevin, AI Policy Specialist, FAS
Karinna Gerhardt, JD Candidate, New York University School of Law, FAS
Oliver Stephenson, Associate Director for AI and Emerging Technologies, FAS

Govind Shivkumar, Director of Investments, Omidyar Network

Shana V. White, Director, CS Equity Initiatives, Kapor Foundation

Akina Younge, Director of Movement Collaborations at UCLA Center on Resilience & Digital Justice (CRDJ)
Deirdre K. Mulligan, Professor, School of Information at University of California, Berkeley
Nicole Ozer, Executive Director, Center for Constitutional Democracy at UC Law San Francisco

We are also indebted to the researchers, scholars, and leaders who have contributed their knowledge and ideas to growing the field of "responsible AI" and centering equity, human rights, and justice in technical, social, and political discussions on the future of innovation. This scholarship was central to the development of the Kapor Foundation principles. To learn more from our field scan, please see the Suggested Readings section.

SUGGESTED READINGS

Business and Venture Capital

- Business Roundtable Roadmap for Responsible Artificial Intelligence (Business Roundtable, 2022)
- PwC's Responsible AI: AI you can trust (PwC)
- Responsible AI for Startups (Responsible Innovation Labs)
- Responsible AI Playbook for Investors (World Economic Forum, 2024)
- Supporting Tech for Justice-Impacted Communities: Strategies to Supercharge Justice Tech Investing (Markovich et al., 2022)

Government

- Artificial Intelligence Risk Management Framework (AI RMF 1.0) (National Institute of Standards and Technology, 2023)
- Blueprint for an AI Bill of Rights (White House, 2022)

Institutes, Non-Profits, and Researchers

- 2023 Landscape: Confronting Tech Power (Kak and Myers West, 2023)
- A Matrix for Selecting Responsible AI Frameworks (Narayanan and Schoeberl, 2023)
- AI for Whom? Shedding Critical Light on AI for Social Good (Moorosi et al., 2023)
- AI Index Chapter 3: Responsible AI (Reuel, 2025)
- Design Justice Network Principles (Design Justice Network, 2018)
- Don't ask if artificial intelligence is good or fair, ask how it shifts power (Kalluri, 2020)
- Explainer: A Sociotechnical Approach to AI Policy (Chen and Metcalf, 2024)
- Responsible AI in Philanthropy (Project Evident and TAG)
- Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (Fjeld et al., 2020)
- Principles for Accountable Algorithms and a Social Impact Statement for Algorithms (FAT/ML)
- Radical AI Principles (Radical AI Network)
- TechEquity AI Policy Principles (TechEquity, 2024)
- The AI Con (Bender and Hanna, 2025)
- The Algorithmic Justice League's 101 Overview (The Algorithmic Justice League, 2020)

International Frameworks

- Global Index on Responsible AI: Our multidimensional framework (Global Index on Responsible AI)
- EU guidelines on ethics in artificial intelligence: Context and implementation (Madiega, 2019)
- Montreal Declaration for a Responsible Development of Artificial Intelligence (University of Montreal, 2018)
- OECD AI Principles overview (OECD, 2024)
- Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2022)

Philanthropy

- A Guiding Framework for Vetting Technology Vendors Operating in the Public Sector (Ford Foundation, 2023)
- AI & Machine Learning (Center for Democracy & Technology, 2019)
- Bending Generative AI's Trajectory Toward a Responsible Technology Future (Omidyar Network, 2025)
- Data Capitalism and Algorithmic Racism (Milner and Traub, 2021)
- National Funders Commit $25M to Center and Accelerate Responsible, Equitable and Ethical AI at Inaugural Joint
- California Summit on Generative AI (Kapor Foundation, 2024)
- Philanthropies launch new initiative to ensure AI advances the public interest (Ford Foundation, 2023)
- Responsible AI and Tech Justice: A Guide for K-12 Education (Kapor Foundation, 2024)
- Responsible Tech Guide (All Tech is Human, 2024)