SOCIAL INNOVATIONS Journal

Vol. 23 (2024): Social Entrepreneurs Leveraging and Shaping Artificial Intelligence

The What, When, and How of Incorporating AI Governance in Social Enterprises

By: Nidhi Sudhan[1]

[1]Co-founder at Citizen Digital Foundation

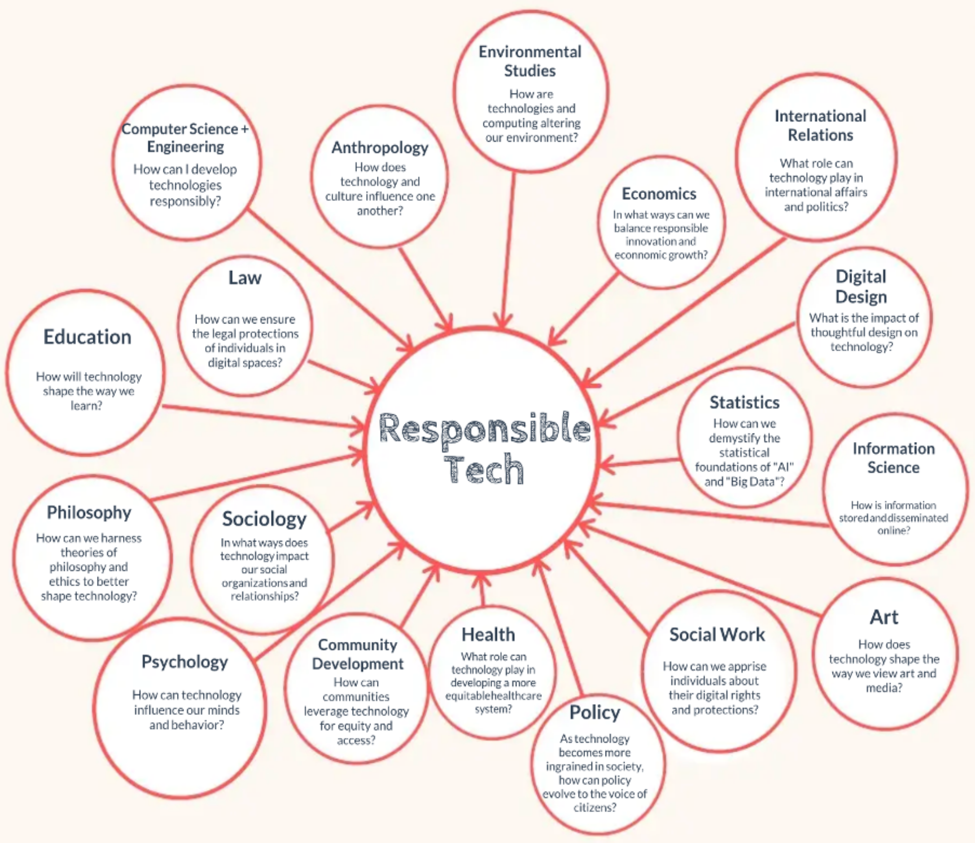
This article has been reviewed and edited by Aditi Pillai.

Keywords: Responsible AI, Artificial Intelligence, generative AI

Abstract

Generative Artificial Intelligence (GenAI) and its corollaries have begun permeating the social fabric, including the social impact ecosystem. As organisations and nation-states increasingly embrace nascent forms of AI, social impact innovators face the formidable challenge of establishing robust foundations for Responsible AI (RAI).

This article delineates actionable steps for the implementation of RAI Frameworks, outlining how these steps can be implemented and the right time to do so. Drawing on insights derived from Citizen Digital Foundation's engagements with business, education, media, and government stakeholders across Southern India, this article provides a practical RAI implementation guide and sheds light on the potential opportunities and real-world challenges encountered by organisations throughout the various stages of AI transformation.


Responsible Tech
Computer Science + Engineering
How can I develop technologies responsibly?
Anthropology
How does technology and culture influence one another?
Environmental Studies
How are technologies and computing altering our environment?
Economics
In what ways can we balance responsible innovation and economic growth?
International Relations
What role can technology play in international affairs and politics?
Law
How can we ensure the legal protections of individuals in digital spaces?
Education
How will technology shape the way we learn?
Digital Design
What is the impact of thoughtful design on technology?
Statistics
How can we demystify the statistical foundations of "AI" and "Big Data"?
Information Science
How is information stored and disseminated online?
Philosophy
How can we harness theories of philosophy and ethics to better shape technology?
Sociology
In what ways does technology impact our social organizations and relationships?
Psychology
How can technology influence our minds and behavior?
Community Development
How can communities leverage technology for equity and access?
Health
What role can technology play in developing a more equitable healthcare system?
Policy
As technology becomes more ingrained in society, how can policy evolve to the voice of citizens?
Social Work
How can we apprise individuals about their digital rights and protections?
Art
How does technology shape the way we view art and media?

Responsible Tech Ecosystem. Responsible Tech Guide 2023, All Tech Is Human.

Introduction

The emergence of Generative AI has reinvigorated the discourse around Responsible AI (RAI). Driven by economic incentives and the multi-polar trap, organisations and nation-states have been hastily and often unscrupulously deploying and adopting nascent forms of AI. It is becoming evident that a fear of opening up to AI's true potential for maximum social impact could lead to widening disparities, and a failure to coordinate in the AI arms race could result in unprecedented tragedies and the collapse of our socioeconomic fabric.

Social impact organisations and innovators who choose to leverage AI must prepare to see their intent, value, and impact diminish or backfire if they don't fortify AI adoption with accountability practices across the design, development, and deployment stages of AI, to mitigate adverse effects. Without RAI frameworks, efforts meant to tackle complex challenges using new technologies will end up solving problems narrowly or, worse, redirecting them elsewhere. At the same time, scarcity of knowledge and resources, lack of sufficient precedence, and being weighed down by ethical responsibility in the face of operational pressures may justifiably act as deterrents. We aim to address some of the challenges we have come across during our interactions with business, education, media, and government stakeholders in southern India.

Make AI governance your focus

Many decision-makers find it overwhelming and cumbersome to initiate RAI within organisations and don't know where to start.

Focusing on 'AI governance' instead of 'Responsible' or 'Ethical' AI helps take some weight off the shoulders. Being ethical – even performatively, as we see in token CSR or DE&I initiatives – calls for a consistent higher order of behaviour, and it's only human to find that stressful, especially when it goes against the popular order of things. Managements are more familiar with 'governance' - a collaborative journey using a set of normative frameworks that help achieve long-term impact and value creation, small milestones at a time. Such framing could go a long way in getting started on RAI.

Carefully assess your organisation's AI opportunity and maturity

Social enterprises working with AI often have the intent but are rarely equipped operationally to move from 'intent' to 'compliance' when it comes to AI governance, according to Simon Zadek's issue maturity scale. Depending on how mature a region is on AI adoption, this might further impede AI governance resourcefulness.

Outlining opportunities that organisations stand to gain through the adoption of AI governance can help provide the impetus to initiate crucial first steps in time. For instance, it can help:

1. Fortify your work against upcoming and future AI regulations.
2. Distinguish the organisation as research-minded and futuristic; build trust among stakeholders, and enhance reputation and impact.
3. Mitigate future economic risks: loss of data, infrastructure, manpower, and negative externalities.

4. Attract and retain fresh talent skilled in emerging technologies, especially since the younger the talent, the more they seek value alignment at work.

Leverage like-minded networks to bring down costs and speed up your learning curve

Despite the opportunities, we come across practical hurdles in implementing AI governance. High costs, operational priorities, lack of knowledge or resources to foresee scenarios, inability to visualise negative externalities, and prioritisation of short-term goals, come in the way of enterprises considering AI governance among their top priorities. On the other hand, AI's *outward promise* of efficiency and cost-effectiveness are most attractive to social sector organisations and innovators who are consistently working with limited resources and goals of delivering large-scale social impact.

Collaborating with 'good-tech' civil society organisations, RAI communities, and forums, exploring open-source AI, and conducting AI maturity/risk assessments using support infrastructure from trade & industry bodies, could help bring down initial costs, allocate resources, and set goals as you test the water.

Visualising negative externalities is aided by several AI governance frameworks (discussed later) that help set team and data hygiene, plan red-teaming exercises, account for redressal systems and explainability practices like provenance and AI model cards, structure external reviews and audits with subject matter experts and affected communities, etc. Precedent reports, and government-advised approaches help bolster arguments that convince investors and donors of why these measures are indispensable for the scale and efficacy of impact, as well as sustenance.

These are complex conversations, the stewardship of which requires managing polarities, where this helpful framework could be used.

For cradle-to-cradle, just digital futures that reward all stakeholders equitably we need a paradigm shift that maximises social impact. This can look like joining local, representative efforts to create platform cooperatives instead of newer models driven primarily by fast growth and short-term profits. In the same vein, enterprises that aim to deliver value instead of landing unreal valuations are now banding together as 'zebras' in a counter-narrative to 'unicorns.'

Refrain from anthropomorphising AI

One of the most problematic issues is emerging AI's capability to produce convincingly real text, image, video, and sound outputs even when far from accurate. This, layered with the human inclination to anthropomorphise technologies, results in end-products, even with the best intent, often taking on a human face, name, tone, or persona. Such products are designed to harvest implicit trust for efficacy but make it difficult for end users to distinguish where the human act ends and tech's limitations begin. This masking of tech's limitations using human traits could, for instance, have dire consequences on AI-enabled helplines for children, women, or other vulnerable communities or AI-enabled applications that people use for mental health support, companionship, or astrological advice.

Our innate trust in the empirical accuracy of algorithms and faith in their efficiency to reduce human errors unknowingly leads to the propagation of discriminatory patterns even without intent, as detailed by Cathy O'Neil in her book 'Weapons of Math Destruction.' This leads to the exacerbation of existing social inequities in AI-based solutions for law enforcement, employment, financial services, social welfare, information dissemination, labour welfare, gender injustices, etc.

Refraining from anthropomorphising AI-based products, however tempting, is key. Maintaining the self-disclosure of AI personas is crucial in building trust with your stakeholders. Consciously building representative design teams and databases can prevent AI models from relying on discriminatory datasets and producing false positives and negatives.

Pay close attention to the provenance of your data and ownership rights

Enterprises adopting AI often tend to offset copyright challenges faced by foundational models to the companies building them without comprehending how these may have ripple effects on our products or, in the power tussle accompanying the AI boom, even be passed on to users. Since many of the chips are still in the air on these issues, what we can do at this point is build best practices that absorb the shockwaves when (not if) they happen.

Some of these best practices include building provenance records into everything you create. The work by the Content Authenticity Initiative (CAI) and MIT Center for Constructive Communication shines a light on how this can be achieved through upcoming tools and databases like Content Credentials and Data Provenance Explorer. Additionally, an organisation should strive to follow existing copyright laws and record and credit sources where possible. Establishing Consent, Credit, Control, and Compensation mechanisms when sourcing content from artists is imperative.

How to go about it?

It may not be possible to score a hundred on RAI from day one. It's not a target to achieve and move on, but rather a journey. We advocate some of the existing and continuously evolving resources that guide organisations in their AI governance journey.

The AI Blindspot Discovery framework developed by the Assembly Program at Berkman Klein Center is a comprehensive framework that proposes several steps to conduct a thorough analysis of the status quo and initiate systems and processes that help establish AI governance at the Planning, Building, Deploying, and Monitoring stages of an AI model or system.

The Responsible AI Resource Kit by NASSCOM is a sector-agnostic resource with multi-stakeholder inputs consisting of RAI principles, RAI maturity assessment, governance framework, and an expansive guide for design architects. Conducting annual RAI maturity assessments using either of these frameworks would help evaluate progress or regression in each process, allowing for necessary steps to be taken toward course correction. A Plan-Do-Study-Act (Deming's cycle) process at each testing phase would additionally enhance the quality of the product and processes, allowing for a dynamic application of the learnings.

When is the right time?

A much-debated point, the right time for each organisation to initiate AI governance can be assessed using the Collingridge Dilemma. As Collingridge himself put in his book The Social

Control of Technology, “When change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult, and time-consuming." We believe an AI risk or RAI maturity assessment right at the outset could act as a compass to initiate an AI governance system and would turn out to be well worth it in the long run. Much like fire, vulnerabilities are best assessed as part of architecture design to allow the planning of exit pathways and installation of prevention mechanisms during construction.

People’s expertise and experience drive their work in their respective fields, and there’s added commitment in the social sector. It’s natural to sometimes find civil society organisations, government bodies, start-ups, tech, medicine, and journalism communities get fiercely territorial of the work they do. The Responsible Tech Ecosystem (Pg 14, Responsible Tech Guide, ATIH 2023) highlights the need for us to open up to parallel conversations in allied fields or risk being siloed in our approaches to problem-solving. Isolated efforts can only be incremental and true systems change calls for interdisciplinary collaboration.

The consequences of a bicycle accident are different from that of a train accident. Responsibility and accountability are proportional to the vehicles’ capabilities. Similarly, as much as there is a race in every sphere to harness AI, it’s not a basic race; it’s more like a lemon and spoon race. Wherein there is no point in coming first if you end up dropping the lemon halfway.

References

All Tech is Human. 2023. “Responsible Tech Guide” https://www.scribd.com/document/476272088/Responsible-Tech-Guide-by-All-Tech-Is-Human

Assembly Program - Berkman Klein Center: https://www.berkmankleinassembly.org/

BCG Gamma. “Are You Overestimating Your Responsible AI Maturity?” March 2021. https://web-assets.bcg.com/b5/4b/8386b5cf409e835bba50306c39d2/slideshow-final-website-version-2021-rev.pdf

Braungart, Michael, and McDonough, William. 2002. *Cradle To Cradle: Remaking the Way We Make Things.* North Point Press.

Calderon, Ania, Taber, Dan, Qu, Hong and Wen, Jeff. “AI Blindspots”. https://aiblindspot.media.mit.edu/

Collingridge, David. 1980. *The Social Control of Technology*. Cambridge University Press.

Content Authenticity Initiatives https://contentauthenticity.org/

Content Credentials https://contentcredentials.org/

Data Provenance Explorer https://www.dataprovenance-explorer.org/

Deloitte. 2023 Gen Z and Millennial Survey. https://www.deloitte.com/global/en/issues/work/content/genzmillennialsurvey.html
Karkera, Kiran. “Why is Provenance Important for AI?” July 10, 2020. https://kaal-daari.medium.com/an-example-of-art-provenance-records-for-the-curious-d3a5e4a1dd77

Genus, Audley, and Stirling, Andrew. 2018. “Collingridge and the Dilemma of Control.” Research Policy. February 2018. https://www.sciencedirect.com/science/article/pii/S0048733317301622?via%3Dihub

Google. “Model Cards” https://modelcards.withgoogle.com/about

Gurteen, David. “Multipolar Traps: Acting against our collective interests” In *Multipolar Traps.* https://conversational-leadership.net/multipolar-trap/

Hays, Kali. 2023. “Google, OpenAI, and Microsoft are Blaming Users When Generative-AI Models Show Copyrighted Material”. *Business Insider*. Nov 7, 2023. https://www.businessinsider.com/google-openai-microsoft-users-responsible-ai-copyrighted-material-2023-11?IR=T

Hendrycks, Dan, Mazeika, Mantas, and Woodside, Thomas. “An Overview of Catastrophic AI Risks”. Center for AI Safety. 9 Oct. 2023.

https://arxiv.org/pdf/2306.12001.pdf

IBM. Global AI Adoption Index 2022. May 2022.
https://www.ibm.com/downloads/cas/GVAGA3JP

Lu, Yingying. 2023. "AI Will Increase Inequality and Raise Tough Questions About Humanity, Economists Warn". *The Conversation*. April 27, 2023.
https://theconversation.com/ai-will-increase-inequality-and-raise-tough-questions-about-humanity-economists-warn-203056

MIT Center for Constructive Communication. "Data Provenance for AI"
https://www.ccc.mit.edu/project/data-provenance-for-ai/

NASSCOM. "Responsible AI Resource Kit". https://indiaai.gov.in/responsible-ai/homepage

NITI Ayog. "RESPONSIBLE AI. Approach Document for India: Part 1 – Principles for Responsible AI" February 2021. https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf

NITI Ayog. "RESPONSIBLE AI. Approach Document for India: Part 2 - Operationalizing Principles for Responsible AI." August 2021. https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf
Center for Creative Leadership. "Are You Facing a Problem or a Polarity?" November 18, 2022. https://www.ccl.org/articles/leading-effectively-articles/are-you-facing-a-problem-or-a-polarity/

Olay. "AI-shu Chatbot"
https://ai-shu.in/

O'Neill, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown.

Platform Cooperativism Consortium:
https://platform.coop/

Saujani, Reshma. 2023. "We Don't Have to Choose Between Ethical AI and Innovative AI" *Time.* Dec. 5, 2023. https://time.com/6342280/ai-paid-leave-social-good/
Center for Humane Technology. "The A.I. Dilemma". March 9, 2023.
https://www.youtube.com/watch?v=xoVJKj8lcNQ

Tech Stewardship. "How Can We Ensure Tech is Beneficial for All?" 2023.
https://techstewardship.com/the-change/

The Authors Guild. 2023. "More than 15,000 Authors Sign Authors Guild Letter Calling on AI Industry Leaders to Protect Writers." July 18, 2023.
https://authorsguild.org/news/thousands-sign-authors-guild-letter-calling-on-ai-industry-leaders-to-protect-writers/

The Consilience Project. "Challenges to Making Sense of the 21st Century". March 30, 2021.
https://consilienceproject.org/challenges-to-making-sense-of-the-21st-century/

The W. Edwards Deming Institute. "Deming Cycle"
https://deming.org/explore/pdsa/

Tveit, Alex, Abbott, Mark, and Lajoie, Jason. "Tech Stewardship as a foundation for Multi-Stakeholder Collaboration (MSC) to enable STI4SDGs"
https://sdgs.un.org/sites/default/files/2023-05/B47%20-%20Tveit%20-%20Tech%20Stewardship.pdf

Zebras Unite Cooperative:
https://zebrasunite.coop/

Zadek, Simon. "The Path to Corporate Responsibility" *Harvard Business Review.* Dec. 2004.
https://hbr.org/2004/12/the-path-to-corporate-responsibility