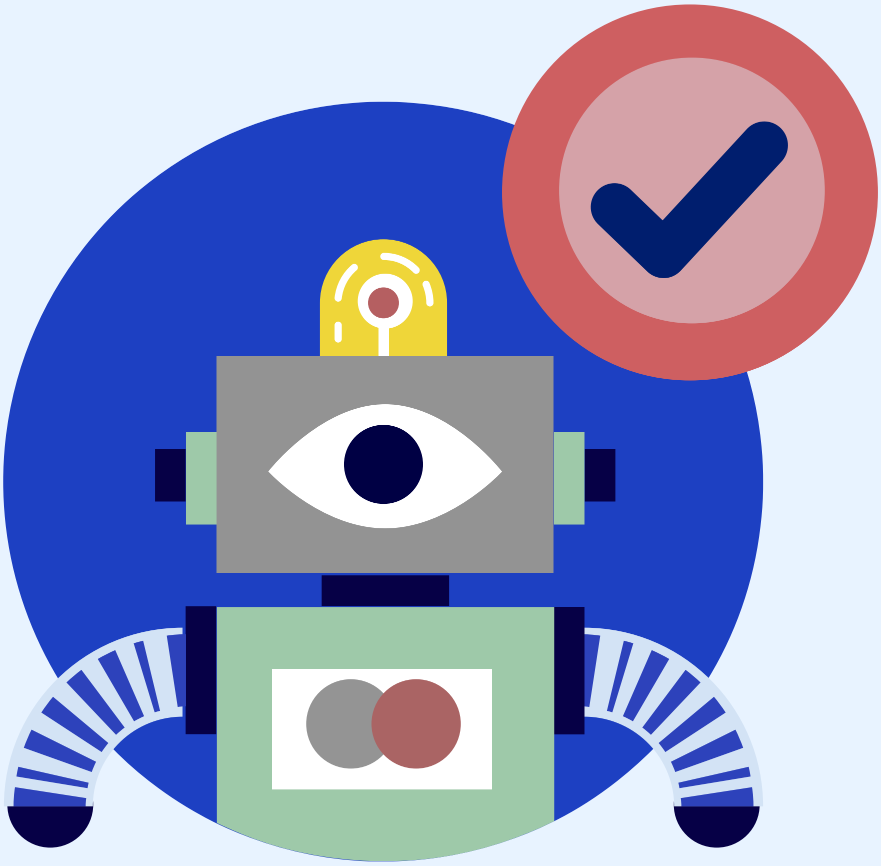
India

The Dialogue™ INFORM ENGAGE IDEATE | Open Loop | Meta

Framework for Incorporating National AI Principles & Prototyping the Principle of Human-Centric AI in India



Authored by

Elonnai Hickok
Principal Investigator & Lead

Nishant Shah
Research & Project Consultancy

Vincent Zhong
Research Support & Project Management

Policy Prototyping for AI Project

ArtEZ University of the Arts

Open Loop India Program

In collaboration with
The Dialogue (India)

Introduction

In recent years there has been a proliferation of AI principles defined by companies, governments, academia, and civil society. AI principles can be understood as a normative framework of high-level values articulated to guide the design, development, and deployment of AI systems.[1] Though there are common themes across principles, there are differences in the interpretation and understanding of these principles and the values they are comprised of across contexts, organizations, and AI systems.[2] While it is difficult to account for the range of values and ethical frameworks that might exist in one context and across contexts, groups of countries have come together around different sets of principles[3] and there are efforts to develop a universal consensus on fundamental aspects of AI.[4] At the same time, there are concerns that frameworks for AI are disproportionately built around Western ethical frameworks and do not account for the context and values that are central in a given culture.[5] It is also difficult for a single set of principles to account for the range of values that may exist in one context. There is also a recognized gap between the articulation of a principle and its implementation, and guidance on how principles can be incorporated consistently into the AI system lifecycle is still emerging.[6]

Experts have emphasized the importance of developing AI systems that bring together multiple voices and account for the intricacies of the context they are deployed in.[7] Yet, little guidance exists on how AI companies can comprehensively and systematically account for context across geographies and different types of solutions. As countries continue to grow domestic AI ecosystems and companies deploy solutions in global contexts, it is essential that AI systems are developed in a way that reflects and takes into account individual and community experiences, local context, and historical understandings. Doing so is an important first step towards developing trustworthy AI systems that account for local realities and can give insight into potential impacts and areas of risk of AI systems.

Motivated by the above, ArtEZ University is collaborating with Open Loop to launch a policy prototyping project to engage and guide companies on the implementation of AI principles across a company and into the AI system lifecycle in a way that emphasizes stakeholder engagement and the incorporation of context.

The policy prototyping project consists of

1. A **framework** that consists of high-level focus areas that can be used by companies to implement AI principles at the organizational level and across the development lifecycle of the AI system. The framework also provides guidance on the development of strategies for stakeholder engagement and the incorporation of context.
2. **Operational guidance** for the principle of 'human-centric' AI that provides examples as to how the principle of human-centric AI can be implemented as per the framework.[8] The guidance is informed by the principle of Protection and Reinforcement of Positive Human Values from the Responsible AI principles defined in India[9] as well as research around the concept of 'human-centric' AI and 'human-centered' machine learning.
3. **Prototyping tasks** that companies can go through to implement the framework and operational guidance with a focus on developing a strategy for stakeholder engagement and incorporating context when contextualizing the principle.

Purpose, Objectives, and Rationale

Through this project we are testing a framework and operational guidance (which together form our “policy prototype”) intended to help AI companies implement AI principles across their company and in the AI system lifecycle. In doing so, the framework includes guidance companies can use to develop a strategy for stakeholder engagement and incorporating context when implementing a principle(s) across a company and the AI system lifecycle. The policy prototype further provides operational guidance with examples on the implementation of the principle of ‘human-centric’ AI.

A framework for implementing AI principles across the company and AI system lifecycle that emphasizes stakeholder engagement and the incorporation of context is central to the concept of 'human-centric' AI and will:

- Guide companies in undertaking meaningful stakeholder engagement throughout the implementation of a principle and the AI system lifecycle.
- Work towards ensuring that AI systems account for the context they are deployed in and the lived experiences of impacted individuals.
- Help companies identify areas of risk and the pressure points across the AI system life cycle that can be improved or addressed as per identified values.
- Raise awareness of the socio-technical impact of AI systems.

The focus on developing a framework for implementing AI principles was chosen because, while there are numerous sets of AI principles that have been developed by companies, industry, and civil society, guidance related to the implementation of AI principles is still emerging. The framework incorporates a strategy for stakeholder engagement and incorporating context as doing so is an important step towards ensuring systems account for local context and the lived experiences of end-users and is also central to a human rights approach,[10] but understanding how to do so consistently with socio-technical systems such as AI is complex. Furthermore, challenges inherent in stakeholder engagement such as power and information asymmetry can be exacerbated in the context of AI, where the technology itself is often a black box.

The principle of 'human-centric' AI was chosen based on

1. The need to bring together existing policy frameworks that promote or provide guidance on the value and/or principle of 'human-centric' AI with technical research that explores the development of human-centered machine learning and synthesize these across the AI system lifecycle.

2. The fluid nature of 'human-centricity' and the criticality of stakeholder engagement and context for informing company implementation of the principle.

3. A lack of shared practices regarding the comprehensive and consistent implementation of the principle of 'human-centric' AI across organizational processes and the AI system lifecycle.

4. The ability to use other relevant existing frameworks to analyze an AI system through the lenses of a particular principle. For example, the OECD Classification framework for AI systems includes the dimension of people and planet which can be relevant to the strategy for stakeholder engagement in this framework.[11]

India was chosen as focus geography as it has adhered to the G20 AI Principles (which in turn reflect the OECD AI Principles).[12] Over the past years, India has taken significant steps to grow a domestic AI ecosystem and has published National AI Principles and taken steps to develop implementation of guidance around the same.

In February 2021, NITI Aayog published the Responsible AI: Approach Document for India Part I - Principles for Responsible AI which articulated a set of seven "Responsible AI Principles" to guide the development of AI ecosystems in India. This included:

1. **Safety and Reliability**
2. **Equality**
3. **Inclusivity and Non-discrimination**
4. **Privacy and Security**
5. **Transparency**
6. **Accountability**
7. **Protection, and Reinforcement of Positive Human Values.**[13]

In August 2021, NITI Aayog published the Responsible AI Approach Document for India Part 2- Operationalizing Principles for Responsible AI.[14] The report provides guidance for government actors, the private sector, and research institutes towards the development of Responsible AI. Aspects of the guidance relevant to this policy prototype guidance include:

- The government should take a risk-based approach to regulating AI in India that is proportional to the likelihood of harm. When assessing the harm, the socio-technical system as a whole must be considered as well as all components of the AI system including development and implementation.
- Until regulatory guidelines are in place, the principles for responsible AI should guide AI development and the development of AI systems should be done in collaboration with multiple stakeholders to identify and address risks.
- The development of guidelines and benchmarks for individual use-cases or specific technologies must be based on the social, economic, political, and cultural realities of the nation while maintaining an international outlook.
- An independent Council for Ethics and Technology (CET) should be developed and be responsible for overseeing, managing, and updating the Responsible AI principles as well as creating guidelines for model review mechanisms that will evaluate the efficacy of AI systems. The CET would also work to harmonize different frameworks relevant to AI across sectors, States, and government departments in India. The CET should be multi-disciplinary and provide a forum for all stakeholders to have a representation.
- It is challenging to provide guidance for the implementation of the Responsible AI principles that is applicable across companies and AI systems. Thus, there should be a focus on developing governance mechanisms that enable the development of reliable, predictable and trustworthy applications. Such governance mechanisms need to begin with stakeholder awareness and education on both capabilities of AI and the risks.
- When developing and implementing such governance mechanisms, there is a need to incorporate multiple stakeholder perspectives from a range of disciplines and backgrounds.
- Responsible AI considerations need to be integrated and embedded into and across the AI system lifecycle. This process needs to be ongoing and iterative.

States in India, such as Tamil Nadu[15] and Telangana[16] have also developed AI frameworks to guide and evaluate AI systems deployed at a state level.

Definitions

The policy prototyping project uses the following definitions and understandings:

Artificial Intelligence System:

Recognizing that there are multiple definitions of AI, for the purpose of this project we draw upon the definition put forward by the OECD "An AI system is a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy."[17]

AI Principles and Values:

AI Principles can be understood as a normative framework of high-level values articulated to guide the use and development of AI. As highlighted in the Interim Report on AI Governance released by Japan - principles are one of the most high-level components of a governance framework for AI and can be understood as technology-neutral goals to be ultimately protected.[18] The report Principled Artificial Intelligence finds that though there are common themes across sets of principles that have emerged, there are also significant differences in the interpretation of a principle. Thus, principles need to be understood in the cultural, linguistic, geographic, and organizational context they are implemented within.[19]

AI System Life Cycle:

Recognizing that there are multiple stages in the AI system lifecycle, we use the OECD definition of AI lifecycle and focus on the following six stages 1. Planning and design 2. Data collection & processing 3. Model building and interpretation 4. Verification and validation 5. Deployment 6. Operation and Monitoring.[20]

Policy Prototyping:

"Policy prototyping is a methodology to test the efficacy of a policy by first implementing it in a controlled environment. Policy prototyping applies a user-centered design and user research approach, which is commonplace in product and service design, to the development of law and policy."[21]

AI Company:

For this prototype project we refer to AI companies as organizations that develop AI systems. Within a company, there are a number of roles we see as relevant to implementing a principle(s) across an organization and within the AI system lifecycle.

Definitions

Stakeholder

In the context of the OECD AI Principles, "stakeholders" are defined as “all organizations and individuals involved in, or affected by, AI systems, directly or indirectly.”[22] Recognizing that a stakeholder may fall within multiple groups, for the purpose of this policy prototyping project, we have identified four categories of stakeholders that are important for AI companies to engage with:

- **Experts:** Stakeholders that have relevant expertise or perspective in the a.) definition, conceptualization and contextual implementation of the AI principle in question b.) the design, development, use, and impact of the AI system that is being developed. This can include but is not limited to, practitioners, civil society, academics, and sectoral experts.
- **Accountability Organizations:** Stakeholders that can play a role in defining best practices and holding companies accountable. This can include multi-stakeholder organizations, international organizations, policymakers, industry bodies, standards bodies, auditing and consulting firms, sustainability business organizations, investors, and foundations.
- **AI system operators:** Stakeholders who procure and operate AI systems. This can include governments, companies, and individuals or groups of consumers.
- **End-users and impacted populations:** Stakeholders that directly or indirectly use, engage with, and are affected by AI systems. This can include individuals, groups, and communities with marginalized groups that may require fairness/human rights/human-centered AI considerations.

Stakeholder Engagement

This framework draws upon the definition of ‘Meaningful Stakeholder Engagement’ put forward by the OECD in the context of the extractive industries: Meaningful stakeholder engagement refers to ongoing engagement with stakeholders that is two-way, conducted in good faith and responsive. The OECD clarifies the following characteristics of the engagement: two-way meaning parties may freely express opinions, there is sharing of decision-making power, and stakeholders also lead engagement activities, good faith meaning engagement is done with genuine intention to understand the perspectives put forward, and preparedness to address adverse impacts raised, responsive engagement meaning that there is follow-through on outcomes from the engagement and that stakeholders opinions are taken into account when decisions are made, and ongoing engagement meaning that the stakeholder engagement continues throughout the lifecycle of the project and is not a one-off initiative.[23] The framework also draws upon best practices from the OECD Best Practice Principles on Stakeholder Engagement in Regulatory Policy to inform the stakeholder engagement strategy.[24]

Definitions

Human-Centered AI

This framework attempts to bring together the different ways ‘Human-Centered’ AI has been used at both a policy and technical level. At a policy level, the term ‘Human-Centered AI’ has been put forward as an objective in different governance frameworks for AI. For example, the EU has put forward a strategy for building trust in ‘human-centric’ AI[25] and the G20 has committed to endeavoring to develop an enabling environment for ‘human-centric AI’.[26] Distinct principles have also been defined around ‘human-centric’ AI. For example, the OECD has defined a principle of Human-Centered Values and Fairness,[27] India has defined a principle of Protection and Reinforcement of Positive Human Values’[28] and Japan has defined The Human-Centric Principle.[29] ‘Human-Centered Machine Learning’ has been used to describe a broader approach to AI development that seeks to place the human at the center of AI development and has been defined as “developing adaptable and usable Machine Learning systems for human needs while keeping the human/user at the center of the entire product/service development cycle.”[30] The policy prototype we will be testing through this project uses the terms ‘human-centered AI’ and ‘human-centric’ AI interchangeably.

Digital Rights

This paper uses the term 'digital rights’ to reflect the ongoing discourse over how individual rights, including human rights, should be supported and respected in the use of data-driven technologies, including AI.[31]

Context

Recognizing that there are multiple definitions, understandings, and dimensions of context, for the purpose of this policy prototype project we focus on the following dimensions:

- Environmental: Aspects related to the infrastructure, physical properties, and restrictions in a context. Examples include the availability of infrastructure necessary for the development and functioning of the AI system, levels of digital literacy, levels of mobile penetration, and digital divides.[32]
- Cultural: Aspects of a society that shape how people live - their behaviors, beliefs, expectations, and attitudes.[33] Examples include knowledge and stories, language, value systems, religions, traditions and rituals, techniques and skills, tools and objects, art, food & drink, and social organization.
- Legal & Political: Aspects related to the legal and political environment. Examples include political actors and their power relationships, relevant legal and policy frameworks, government practices, and the strength of democratic institutions.
- Historical: Histories of social, economic, cultural, and political influences. Examples include histories around technological development and use as well as histories of discrimination and bias.
- Economic: Aspects related to the structure of economic life in a context. Examples include resource availability, skilled labor force, markets, and relevant government policies.

Methodology and Scope

This policy prototype project consists of a framework for the implementation of AI principles across a company and in the AI system lifecycle. It provides operational guidance with examples related to the principle of 'human-centric' AI. The framework is built around three parts:

1. **Organizational measures**
2. **Contextualizing the Principle**
3. **Incorporation into the AI System Lifecycle.**

The framework also includes a strategy for stakeholder engagement and incorporating context.

The operational guidance provides resources and examples that companies can use and draw from. These have been further categorized based on:

- **Type of Actor:** The type of actor that developed the resource including industry, civil society, academia, media, standards association, government, international organization, and consulting firm.
- **Type of Resource:** The type of resource or example including report & article, policy submission, guidance & best practice, framework, assessment & certification, principles, standards, policy & regulation, conference, governance & organizational structure.

The framework and operational guidance were developed through a literature review of primary and secondary sources.

While the intention of this project is to guide companies in the implementation of the AI principles with a focus on stakeholder engagement and incorporating context. The limitations of this framework are inherent in the definition and scope of the framework, 'human-centric' AI, stakeholder engagement, and context as well as the sample size of literature.

The prototype task is centered around Part 2 of the framework - Contextualizing the Principle: Defining the principle, Digital Rights, and Impact & Risk with a focus on having companies build a strategy for stakeholder engagement and incorporating context. This includes three stages: 1. Planning and preparation 2. Engaging Stakeholders and Incorporating Input and Context and 3. Evaluation and Learnings. The prototype task also asks a series of reflection questions to companies on the prototype task process.

The policy prototype project engages companies providing B2C services in India from the following sectors: Agriculture, Education, Finance and Healthcare.These sectors were selected based on priority areas for AI development identified by the government, the potential impact that the AI system can directly have on an individual, the role that context plays in ensuring the accuracy of the AI system, and the particular type of AI systems developed and deployed in each respective country.

The policy prototype project is comprised of the following phases:

PHASE 1

Background Research:

Development of a report on the role of Values and National AI principles in AI development. The report was supported by Open Loop and developed independently by ArtEZ University through interviews and a literature review.

PHASE 2

Development of the Framework and Operational Guidance:

In collaboration with Open Loop, develop the framework and operational guidance with examples for the principle of 'human-centric' AI and strategy for stakeholder engagement and incorporating context.

PHASE 3

Prototyping

Engage with companies to prototype the framework and operational guidance as well as learn from their experience in applying the framework and operational guidance to understand how clear, operational, and useful it was in enabling the implementation of AI principles into the AI system life cycle in a way that accounts for context and stakeholder input.

PHASE 4

Final Report

Develop a final report with policy recommendations that draw upon the learnings and inputs from the prototyping exercise.

A Framework and Operational Guidance for Implementing AI Principles

The framework for the implementation of AI principles across an organization and in the AI system life cycle consists of three main parts with focus areas. These include:

The principle of 'human-centric' AI was chosen based on

① **Organizational measures:** Focused on the policies, processes, and resources in place to implement the principle across an AI company with the focus areas of governance, capacity & resources, and responsiveness. This is meant to be an iterative and evolving process.

② **Contextualizing the Principle:** Focused on contextualizing the principle to a specific AI system with the focus areas of defining the principle, digital rights, and impact and risk. This is meant to be sequential with the contextualization of the principle being necessary before implementation into the AI System Lifecycle.

③ **Incorporation into the AI system lifecycle:** Focused on incorporating the principle into each phase of the AI system lifecycle with the focus areas of planning & design, data collection & processing, model building & interpretation, verification & validation, deployment, and operation & monitoring.

Accompanying each part of the framework is operational guidance with examples as to how the principle of 'human-centric' AI can be implemented.

The framework also includes guidance for the development of a strategy for stakeholder engagement and the incorporation of context when implementing a principle across an AI company and in the AI system lifecycle.

The strategy for stakeholder engagement and incorporations of context is meant to be modular and can be applied to all or any part of the framework including a specific focus area.

Part 1: Organizational measures to consider

For the holistic implementation of a principle(s) across an AI company and the AI system lifecycle, companies could show their commitment to implementing the principle(s) from the outset.[34] This commitment can be manifested at different levels and in the capacity, resources, processes, policies, and mechanisms companies put in place. Within a company, board members, executive management, legal and policy teams, and human resources will play an active role in implementing this part of the framework.

Focus Area	Framework	Operational Guidance for 'Human-Centric' AI
Governance	Aspects of corporate governance to oversee the implementation of the principle across a company can include • Board-level commitment to the principle(s) • Governance structures responsible for incorporating the principle(s). • Alignment of company goals and business priorities with the principle (s). • Internal Communication and transparency regarding the principle (s) and steps to implement the same. • Public Commitment to the principle(s) in reports. • Participation in Multi-Stakeholder Forums: Participating in multi-stakeholder forums committed to values • Benchmarks to track progress towards goals.	Board-level commitment: Resources that highlight the importance of board-level commitment and oversight of AI include • The World Economic Forum has created the Oversight Toolkit for Boards of Directors • The article Why your Board Needs a Plan for AI Oversight highlights why the risks and potential benefits of AI require greater oversight and fluency from boards. • The article Board Responsibility for Artificial Intelligence Oversight highlights the position boards are in to avoid the harms and litigation risks that may arise from the societal impacts of AI. • One of the indicators in the Corporate Accountability Index: Algorithmic Systems by Ranking Digital Rights assesses if senior executives and/or the board of directors' review impact assessments of AI systems. Governance structures: 'Human-centric' examples of AI governance can include creating multi-disciplinary and multi-stakeholder governance structures. For example, • The Fujitsu Group has established a multidisciplinary AI Ethics and Governance Office to ensure the safe and secure deployment of AI. • IBM has created a multidisciplinary AI Ethics Board to cultivate a culture of ethical, responsible, and trustworthy AI. • The Japanese Expert Group on How AI Principles Should be Implemented recommends the creation of AI Management Systems including undertaking a gap analysis between AI governance goals and the current state to understand and address gaps. Alignment with Business Goals: • The Japanese Expert Group on How AI Principles Should be Implemented recommends the setting of AI governance goals based on the potential impact of an AI system. • The article The Key to Growing Human-Centered Business cautions against the prioritization of objectives like profit and efficiency. The article Human-Centric Business emphasizes the importance of empathy, balance, and clarity in organizational approaches. The article Responsible Innovation in Tech: Learnings from Other Industries recommends the integration of Responsible AI into ESG and CSR policies and targets. Internal Communication and Data Sharing: • The Japanese Expert Group on How AI Principles Should be Implemented provides the example of holding cross-sectional meetings across an organization where sensitive AI systems are developed or the company has no previous experience. The Guidance also emphasizes the importance of sharing best practices across organizations.

Focus Area	Framework	Operational Guidance for ‘Human-Centric’ AI
Governance (cont.)	Aspects of corporate governance to oversee the implementation of the principle across a company can include • Board-level commitment to the principle(s) • Governance structures responsible for incorporating the principle(s). • Alignment of company goals and business priorities with the principle (s). • Internal Communication and transparency regarding the principle (s) and steps to implement the same. • Public Commitment to the principle(s) in reports. • Participation in Multi-Stakeholder Forums: Participating in multi-stakeholder forums committed to values • Benchmarks to track progress towards goals.	Public Commitment: This can include working with stakeholders to define AI principles or committing to principles that have been developed through multi-stakeholder processes. For example • The Toronto Declaration: Protecting the right to equality non-discrimination in machine learning is a set of principles developed by civil society through a multi-stakeholder process focused on protecting the rights to equality and non-discrimination in machine learning systems. • The Framework for promoting workforce well-being in the AI-augmented workplace was developed by the Partnership on AI through a multi-stakeholder process outlining a set of best-practices focused on promoting well-being during the implementation of AI in a workplace. Participation in Multi-Stakeholder forums: Participating in multi-stakeholder forums committed to values relevant to ‘human-centric’ AI such as human rights. For example • The Partnership on AI is a multi-stakeholder initiative that brings together members from civil society, industry, and academia to create resources and best practices for advancing positive outcomes for people and society. • Data & Trust Alliance resides within the Center for Global Enterprise and brings together leading businesses and institutions across multiple industries to learn, develop, and adopt responsible data and AI practices. • Global Network Initiative is a multi-stakeholder initiative that brings together industry, investors, media organizations, civil society, and academia guided by a set of principles for protecting privacy and freedom of expression online grounded in international human rights standards. Company members undergo annual independent assessments to determine their progress in implementing the GNI principles. Benchmarks: IEEE has developed benchmarks around safety, accountability, responsibility, and transparency.
Capacity and Resources	Capacity and resources to support the implementation of the principle can include: • Positions: Dedicated positions responsible for implementing the principle(s) including executive and management. • Expertise and capacity: Building internal expertise and capacity through methods like continuous employee training, collaboration with stakeholders, and attending relevant conferences. • Resources: Dedicated budgets and resources that are sufficient to support the commitment to implement a principle(s).	Positions: For example, Toyota Research Institute has created a Human-Centric AI team that focuses on forms of human-AI collaboration. The team partners with Universities and experts to undertake research. Expertise and Capacity: Examples of expertise and capacities and resources related to ‘human-centric’ AI can include building out diverse teams and engagement with external experts to build expertise related to 'human-centric' AI such as human rights, linguistics, ethics, accessibility, anthropology, political science, human-centered computing, and humanists. The report After the Offer: The Role of Attrition in AI’s ‘Diversity Problem’ budgets and resources by the Partnership on AI Provides recommendations on ensuring diversity in AI jobs. Capacity can also be built through employee training and participation in conferences. Examples of conferences related to ‘human-centric’ AI include • Human-Centered AI workshop at NeurPIS 2021 • Intelligence Augmentation: AI Empowering People to Solve Global Challenges • International Conference on Artificial Intelligence in HCI • ICML Workshop on Human in the Loop Learning • INTERACT Workshop on Humans in the Loop - Bringing AI & HCI Together

Focus Area	Framework	Operational Guidance for ‘Human-Centric’ AI
Responsiveness	Mechanisms that can enable companies to respond to internal and external challenges and concerns can include • Organizational assessments, internal and external, to understand how to improve and align company processes, practices, and policies. • Risk mitigation strategies and management frameworks to address risks related to AI. • Grievance and redress mechanism to provide communities and stakeholders a mechanism to exercise their voice and communicate concerns and harms related to the principle(s).	‘Human-centric’ mechanisms that can be used by companies to respond to internal and external challenges include Organizational assessments: **Ethics:** • The Markkula Center for Applied Ethics provides a self-assessment for ethical business practices **Responsible AI:** • PwC offers a Responsible AI Diagnostic Survey • Equal AI provides a certification program for responsible AI. **AI Governance:** • The World Economic Forum has created an Assessment Guide to accompany their Model AI Governance Framework. • The Responsible AI Institute provides an accredited certification program that assesses fairness, bias, and explainability of responsibly built AI systems. **AI Readiness and Preparedness:** • Microsoft has developed an AI Readiness Assessment that companies can take to determine how ready their business is for AI. • Intellico has developed an AI Readiness Assessment to help companies determine if they are investing in the right input and gaining value. • Intel has developed an AI Readiness Model that judges a company's ability to gain value from AI. **Risk mitigation strategies:** Companies can collaborate with stakeholders and external bodies to develop and implement risk mitigation strategies. These can include • Insurance: Brookings has recommended insurance to address risks associated with AI. (Civil society, Report) • Increased oversight in areas where there are regulatory gaps: Brookings has recommended that increased oversight efforts by health systems, hospitals, professional organizations, and insurers may be necessary to ensure the quality of AI systems that fall outside the FDA’s authority. • Risk management frameworks for AI: NIST is developing an AI Risk Management Framework that can be used by companies. BSA has developed Confronting Bias: BSA’s Framework to Build Trust in AI framework companies can use for AI bias risk management. • Accountability Frameworks: The GOA has developed an Artificial Intelligence: Accountability Framework for Federal Agencies and Other Entities • System of escalation and business continuity plan: The Japanese Expert Group on How AI Principles Should be Implemented provides the example of companies having in place a system for escalation of AI incidents and a Business Continuity Plan. **Grievance and redress mechanisms:** A grievance and redress mechanism is a way for communities and stakeholders to exercise their voice and communicate concerns and harms. Companies can work with stakeholders and communities to ensure grievance mechanisms are culturally appropriate and effective. Resources on establishing grievance mechanisms include • Transparency International’s guide Complaint Mechanism: Reference for Good Practice (Civil society, Best practice) • The UNDP guide Supplemental Guidance: Grievance Redress Mechanism. (Civil society, Best practice) • BSR’s guide on Access to Remedy (Civil society, Best practice) • World Bank Group and UNCTAD guide to Grievance Redress Mechanisms (Civil society, Best practice)

Part 2: Contextualizing the Principle

As part of the implementation of a principle(s), companies will need to contextualize the principle to the solution. This includes a process of understanding and identifying values within a principle that will be prioritized across the AI system lifecycle, resolving tensions between principles that might exist and the impact associated with specific tradeoffs, understanding what digital rights are associated with a principle and how these can be incorporated in the AI system lifecycle, and using the principle to guide an understanding and assessment of the potential impact (positive and negative) of the AI system.

Focus Area	Framework	Operational Guidance for ‘Human-Centric’ AI
Defining the Principle	Principle Definition: Defining the principle. Value Identification: Identifying the key values in the principle and which values will be prioritized in the AI system. Aspects to consider when prioritizing the principle include • The type, function, and use of the AI system and who it will be used by. • The context where the solution will be used. • The potential type and scale of impact. Resolving tensions: Identifying if the principle is in tension with other principles, and how to resolve any trade-offs that might be necessary, and the impact of such trade-offs.	Examples of resources that have defined a principle or value related to ‘human-centric’ AI include • The Indian Responsible AI Principles define a principle of Protection and Reinforcement of Positive Human Values. • The Japanese Social Principles of Human Centric-AI have defined a Principle of Human-Centric AI. • European Commission has adopted an approach towards the creation of human-centric AI including ensuring AI works for people and protects fundamental rights. • G20 AI Principles define a principle of Human-centered values and fairness. • OECD AI Principle 1.2 on Human-centred AI, rationale. • G20 Insights Human-centric AI: from principles to actionable and shared policies • Uni Global Union Top 10 Principles for Ethical Artificial Intelligence defines the principle Make AI Serve People and Planet. • The Global Partnership on Artificial Intelligence's working group on Responsible AI has defined their mission to "foster and contribute to the responsible development, use and governance of human-centred AI systems, in congruence with the UN Sustainable Development Goals." • Chinese National New Generation Artificial Intelligence Governance Expert Committee defines a principle of Harmony and friendliness in the ‘Governance Principles’ for responsible AI. • UNESCO's recommendation on the Ethics of Artificial Intelligence recommends the development of human-centric AI. Drawing from the above, values that have been associated with a ‘human-centric’ approach to AI include: • Legal and Procedural: Respect and rule of law, Human rights, Democratic values, Constitutional values, International standards, Social justice, and International labor rights. • Relationship: Trust, Positive community relationships, Enabling human intervention where necessary, Preservation of social harmony, Promotion of positive human values, Oversight, and Human Control. • State of being: Well-being, Freedom, Dignity, Security • Power: Autonomy, Equality, Human determination, Human agency, Non-discrimination and fairness, Accountability, Fairness, Enabling human intervention, data ownership and control. • Prioritize People: Human relevance vs. programmed intelligence. • Voice: Pluralism, supporting human creativity, diversity including culture and gender.

Focus Area	Framework	Operational Guidance for 'Human-Centric' AI
Defining the Principle (cont.)	Principle Definition: Defining the principle. Value Identification: Identifying the key values in the principle and which values will be prioritized in the AI system. Aspects to consider when prioritizing the principle include • The type, function, and use of the AI system and who it will be used by. • The context where the solution will be used. • The potential type and scale of impact. Resolving tensions: Identifying if the principle is in tension with other principles, and how to resolve any trade-offs that might be necessary, and the impact of such trade-offs.	• Technical configurations: Privacy and data governance, security, transparency, explainability, fairness, reliability, and safety. • Sustainability: Sustainable Development Goals, sustainable development, Societal and environmental well-being, and Inclusive growth. Resolving tensions: The Open Loop report on AI Transparency and Explainability - A Policy Prototyping Experiment discusses the tensions and challenges companies encountered when delving into building XAI solutions at the technical (code) level. The article The Role and Limits of Principles in AI Ethics Towards a Focus on Tensions finds that tensions can include • Moral trade-offs between principles • Tensions arising from societal and/or technological constraints • Tensions arising from different understandings and priorities from stakeholder groups.
Digital Rights	Identification of rights: Use of the principle to identify which rights should be supported in the AI system. In doing so the following can be considered • Where and How: Where in the design of the AI system to support the right. • What Information: Information that needs to be communicated to the user to enable the right and what information needs to be communicated to the public for the company to be held accountable for incorporating that right.	Examples of policy processes/instruments exploring ways to protect and uphold digital rights in the context of automation and algorithms have included the GDPR, the EU Digital Services Act, the EU AI Regulation, and the White House initiative to create a Bill of Rights for an Automated Society. Examples of rights that have been explored include • Notice and Transparency: The GDPR requires notice if automation was used in specific decisions, the draft EU AI regulation requires notice to individuals if the AI system interacts with humans, is used to detect emotions, or generates/manipulates content. The presence of an AI system and the proposed EU Digital Services Act requires disclosure of the use of AI in content moderation as well as a qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error of the automated means used, and any safeguards applied. Article 15 requires notice of information on the use made of automated means in taking a decision related to content. Article 24a requires providers of online platforms to publish the main parameters of recommender systems, the criteria that drive the parameters, and the relative importance of the parameters. • Explanation: The article "Good Explanation for Algorithmic Transparency" explores the explanation of the logic used to reach a decision and how that decision may impact an individual. • Choice and consent: Choice and meaningful consent when interacting with an AI system and in what configuration including for the collection and use of personal data. Article 29 of the proposed Digital Services Act requires providers of online platforms to provide users with at least one option for each for their recommender systems which is not based on profiling.

Focus Area	Framework	Operational Guidance for ‘Human-Centric’ AI
Digital Rights (cont.)	Identification of rights: Use of the principle to identify which rights should be supported in the AI system. In doing so the following can be considered - Where and How: Where in the design of the AI system to support the right. - What Information: Information that needs to be communicated to the user to enable the right and what information needs to be communicated to the public for the company to be held accountable for incorporating that right.	- Portability: The GDPR provides individuals the right to move certain data from one provider to another. - Redress: The Council of Europe’s Expert Committee on Human Rights dimensions of automated data processing and different forms of artificial intelligence prepared the report Responsibility and AI that explored access to redress if harmed by an algorithmic system. - Literacy: Access to information and resources about how algorithmic solutions function. - Where, how and what information: Lessons as to where and how rights can be incorporated can be learned from work that has been undertaken on designing privacy into different services. For example, the article ‘AI, big data, and the future of consent’ explores the use of mechanisms like ‘comic contracts’ to facilitate meaningful consent in the context of big data and AI. - What information: Lessons can be learned from the conditions that must be met for consent to be considered meaningful under regulations like the GDPR including that consent must be freely given, specific, informed, unambiguous, and can be revoked. - Transparency reporting: There is a significant body of work around transparency reporting that companies can draw upon. For example, Ranking Digital Rights has developed a methodology that assesses ICT companies transparency in reporting on aspects related to users digital rights. With respect to AI, this includes transparency about algorithmic content curation and ranking systems.
Impact & Risk	Impact and Risk Assessments: Process of assessing the potential risk and impact of an AI system as per the principle(s) and associated values. The outcomes of impact assessments should further inform the goals and design of the AI system as well as decisions as to if the AI system should be developed or the use be limited to specific stakeholders.	Impact assessments: Examples of impact assessments from perspectives relevant to ‘human-centric’ AI include: Human Rights and Rule of Law - Artificial intelligence, human rights, democracy, and the rule of law: a primer. The Council of Europe. This primer aims to provide some background information on the areas of AI innovation, human rights law, technology policy, and compliance mechanisms covered therein. - The UN Guiding Principles on Business and Human Rights provides a framework to guide companies in respecting and protecting human rights throughout the course of their business practices and product development. - The OECD has published guidance on how companies can undertake human rights due diligence in the context of AI. - Article 26 of the proposed Digital Services Act requires providers of very large online platforms to assess systemic risks stemming from the design, including the design and functioning of algorithmic systems, functioning and use made of their services. This includes negative effects on the exercise of fundamental rights, human dignity, protection of personal data, freedom and expression, prohibition of discrimination, the rights of the child, and consumer protection. - The Business & Human Rights Resource Centre has developed a white paper outlining recommendations on how to apply UN Guiding Principles to AI Microsoft has completed a Human Rights Impact Assessment by Article 1 to understand the impact of its AI-related processes and products on human rights. Microsoft has also published a Responsible AI impact Assessment Template and Guide that can be used by companies.

Focus Area	Framework	Operational Guidance for 'Human-Centric' AI
Impact & Risk (cont.)	Impact and Risk Assessments: Process of assessing the potential risk and impact of an AI system as per the principle(s) and associated values. The outcomes of impact assessments should further inform the goals and design of the AI system as well as decisions as to if the AI system should be developed or the use be limited to specific stakeholders.	- The Ad Hoc Committee on Artificial Intelligence Policy Development Group created a Human Rights, Democracy and Rule of Law Impact Assessment of AI systems. - Aapti Institute's report 'Artificial Intelligence and Potential Impacts on Human Rights in India', commissioned by the United Nations Development Programme under the Business and Human Rights in Asia programme and the European Union explores the impact of AI deployment by businesses in India on human rights of consumers in sectors of healthcare and financial services, and the labor force in sectors of retail and gig economy. Legal, technical, and ethical aspects: - The European AI Alliance has created an AI Impact Assessment & Code of Conduct checklist that can be used by companies to assess legal, technical, and ethical implications of an AI solution. - The AI Pulse has created a method to reproducibly estimate the ethical impact of Artificial Intelligence. Societal impact and risk: - From Principles to Practice An Interdisciplinary framework to operationalise AI ethics by the AI Ethics Impact Group recommends measuring impact through the intensity of potential harm on fundamental rights, number of people affected, and impact on society and the dependence on the decision including if it was fully taken by AI, the ability for the individual to switch systems, and the ability for an individual to access redress. - Open Loop has prototyped an AI Impact Risk Assessment as part of the Automated Decision Impact Assessment policy prototype. Well-being: Including aspects like community, culture, education, economy, environment, health, emotional well-being - IEEE has developed the 7010 standard for assessing well-being implications of artificial intelligence. - The Partnership on AI has developed a framework for Promoting Workforce Well-being in the AI-Integrated Workplace. Specific rights and sectors: - The Ada Lovelace Institute has designed an algorithmic impact assessment for the healthcare sector that assesses the impact of AI on the right to health. - Algorithm Watch has developed the Automated Decision-Making Systems in the Public Sector - An Impact Assessment Tool for Public Authorities to assess the use of automated systems in public administration. and use of AI in the public sector. - The AI Now Institute has developed the Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. - The OECD Framework for the classification of AI systems: a tool for effective AI policies provides a framework for classifying AI tools and assessing potential risk.

Part 3: Incorporation into the AI System Life-Cycle

Companies can undertake a process of implementing the principle at each stage of the AI system life-cycle.

Focus Area	Framework	Operational Guidance for ‘Human-Centric’ AI
Planning and Design	Use of the principle to guide: • Objectives and Intended Use: Defining the concept and objectives of the AI system including underlying assumptions, intended use, context and requirements. This can include: • Design Method: The method and approach to the design of the system. • Design elements: Including user interface and explainability. • Human-AI collaboration: When and how humans are able to interact with the AI system.	Objectives and Intended Use: Companies can work with stakeholders to ensure AI systems have ‘human-centric’ objectives and meet the needs of end-users. For example • Intended use has been defined by Microsoft in the Responsible AI Impact Assessment as “a description of who will use the system, for what task or purpose, and where they are when using the system." • The article “Hello AI”: Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making Research into human-centered AI development” engaged pathologists to understand how non-AI-experts interact with an AI tool. • The article “Human-Centered Artificial Intelligence and Machine Learning” explores AI that understands humans from a sociocultural perspective and AI systems that help humans understand them. • The article “Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy” explores AI that supports human efficacy, mastery, creativity, and responsibility. Design Method: Companies can bring human needs and perspectives into the design of the AI system. For example, The article “Human-centered AI: The role of Human-Centered Design Research in the development of AI” identifies the following user-centered methods: • Human-centered design: Focuses on human needs and sees the human as central to the design of the system. • Social design: Focuses on the designer’s role and responsibilities in design choices. • Participatory design: Emphasizes the democratization of participation in the design of a system and questions of power, democracy, and control. • Inclusive Design: Focuses on the needs and behaviors of diverse groups to make AI systems more accessible and usable. • Interaction Design: Focuses on observation of human behavior, action, and cognitive processes to inform the design of human-machine interactions. • Human-Centered Computing: Focuses on incorporating context through integrating diverse views. • Co-design: A design approach in which community members are treated as equal collaborators in the design process. Design Elements: Companies can bring a human perspective into the design elements of an AI system. For example, Understandability: The article “Questioning the AI: Informing Design Practices for Explainable AI User Experiences” identifies potential questions that end-users may have about an AI system. The article “Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the Global South” sought to understand how users in different contexts understand algorithms to better inform aspects of explainability.

Focus Area	Framework	Operational Guidance for ‘Human-Centric’ AI
Planning and Design (cont.)	Use of the principle to guide: • Objectives and Intended Use: Defining the concept and objectives of the AI system including underlying assumptions, intended use, context and requirements. This can include: • Design Method: The method and approach to the design of the system. • Design elements: Including user interface and explainability. • Human-AI collaboration: When and how humans are able to interact with the AI system.	Explainability and Interpretability: • The report “People-Centric Approaches to AI Explainability” features a draft AI Explainability Framework, which provides guidance on how to design and develop explainability experiences in AI-powered products. • A resource for understanding how AI systems work “Systems Cards” • The article “Explainable machine learning in deployment” explores developing end-user-driven and informed explainability and interpretability goals. • Usability: Future AI identifies the following best practices for developing AI systems that are usable, acceptable, and deployable: • Engaging with end-users to identify requirements • Testing for usability • Developing usability metrics • Ensuring solutions can be integrated technically and into workflows • Providing training material assuming non-AI experts will be using the system, monitoring and assessing for changes that may be needed. Human AI Collaboration: The working paper “Human-Algorithm Ensembles” and the article “Society-in-the-Loop: Programming the Algorithmic Social Contract” explore configurations for human-AI collaboration and decision making: • Human in the loop: Allows humans to give direct feedback into an AI system or collaborate with the AI in different configurations. • Human out of the loop: The AI system functions autonomously. • Human-on-the loop: Provides human oversight of an AI system but does not require human feedback to operate. • Society in the loop. Integration of humans in the governance and regulation of AI including in the metrics that determine performance. Questions companies can consider that have been identified in the article “The Feedback Loop: How Humility in AI Impacts Decision Systems” includes • What should trigger human involvement in the system? • In what configuration should humans be involved? • What should be the response of the AI system when a human is involved? The Partnership on AI has developed considerations that can guide company decisions around human-AI collaborations including • End-user: Level of digital literacy, the vulnerability of the population, ability of end-user, end-user needs, degree of impact on end-users including the rights of the end-user. • Contextual: Historical and political context, social norms, societal understanding of technology, infrastructure, and legal frameworks.

Focus Area	Framework	Operational Guidance for 'Human-Centric' AI
Data Collection and Processing	Use of the principle to guide practices around: • Data Sourcing: Including what data is sourced from where, and how. • Cleaning and Annotation: Including developing annotation instructions, sourcing annotation work, and maintaining the quality of the dataset as well as documenting the metadata and characteristics of the dataset.	Data Sourcing: Human-centered considerations when sourcing data can include: • Accounting for local challenges and nuances that might be associated with data and using locally curated datasets, local open data sets, and collaborating with stakeholders to develop datasets. For example, the article Indigenous dataset: A listing to help AI perform better in India highlights challenges such as differences in fonts, language, and license plate design in creating local datasets such as datasets of license plates. • Ensuring that sourced data is representative, inclusive, and diverse including relevant demographics, religion, gender, language, features, accents abilities, and ensuring data is accurate, complete, and inclusive.[35] For example, the Tamil Nadu government has created a 'Deep Max' scorecard for AI systems that, among other things, assess the system for a diversity of training data in race, gender, religion, language, color, features, food habits, accent, etc. In a submission to NIST, the World Institute on Accessibility has highlighted the importance of developing datasets that account for different abilities. The report for the Data Governance WG at the Global Partnership on AI highlights the need to account for power dynamics that can exist when data is sourced. This includes respecting the privacy and other rights of end-users. Data Annotation and Cleaning: The article Data-Centric AI: AI Models are Only as Good as Their Data Pipeline by Stanford University's Centre for Human-Centered Artificial Intelligence highlights the following best practices when annotating and cleaning data include: • Situating the data within the context that it is being used. • Transparency of how data is cleaned, how sensitive data is handled, and the error rates of an algorithm. • Undertaking multiple data collection processes and simulations. • Development and application data benchmarks. Niti Aayog in the Approach Document for Operationalizing Principles for Responsible AI recognizes the importance of accounting for data source reliability, missing data, duplicate data, correlated variables, and outliers when ensuring the quality of data. Methods of annotation and cleaning that can be relevant to 'human-centric AI' include • Community-driven models of annotation. For example, AI4Dignity is a project that brings together AI developers, fact-checkers, anthropologists, and policy experts in collaborative models of annotation and coding. • Co-creation of AI datasets and AI pipelines. For example, the article Empowering Local Communities Using Artificial Intelligence explores co-creating the design of AI systems, co-curating data sets, and creating localized AI pipelines.

Focus Area	Framework	Operational Guidance for ‘Human-Centric’ AI
Model Building and Interpretation	Use of the principle to guide: • Model Selection: Creation or selection of the model or algorithm. • Model Training: Training and/or interpretation of the model or algorithm.	Model Selection: The article De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research highlights an open-source approach to AI development to address divides that exist in access to the infrastructure needed to develop AI. Model Training: The article Reliance on Metrics is a Fundamental Challenge for AI recommends • Defining a broad range of metrics and parameters to enable a comprehensive picture. • Integrating metrics with qualitative data. • Engaging with a range of stakeholders to define metrics. The article Aligning AI to Human AI Means Picking the Right Metrics recommends re-evaluating metrics on a regular basis to ensure they are relevant through processes like double-loop learning. The article Towards a More Transparent AI recommends transparency around metrics and parameters used and if/how they changed as the solution is deployed and evolves. Metrics relevant to the concept of ‘human-centric” AI have included those relating to well-being. For example, IEEE P7010 standard defines well-being metrics for autonomous and intelligent systems. Feedback loops can also support ‘human-centric’ goals. For example, the Article Artificial Intelligence and Community Well-being: A Proposal for an Emerging Area of Research explores how feedback loops can lead to community-driven AI development and ensure that gains from AI do not increase divides and inequality. The article Personalized Ranking with Diversity explores how to incorporate diversity into personalized ranking objectives using implicit user feedback.
Verification and Validation	Use of the principle to • Model Evaluation: Evaluate the AI system based on various dimensions and considerations and refine the model.	Model Evaluation: Aspects related to the concept of ‘human-centric’ AI that models can be evaluated for include bias, accuracy, and fairness. The OECD Catalogue of Tools for Trustworthy AI includes tools for assessing fairness, accountability, and transparency of AI systems. Part of this can include moving beyond statistical fairness and developing context-specific fairness frameworks. For example, the article Re-imagining Algorithmic Fairness in India and Beyond explores a framework for fairness in India which considers proxies and harms like caste, gender, religion, ability, class, gender identity & sexual orientation, and ethnicity. The article Human-Centered Approaches to Fair and Responsible AI recommends understanding human perceptions of fairness within the context where the solution is being deployed as well as understanding how algorithmic decisions are used by humans to further inform contextual understandings of fairness.

Focus Area	Framework	Operational Guidance for 'Human-Centric' AI
Deployment	Use of the principle to guide: • Use: The use of the AI system. • Regulatory Compliance: Ensuring the AI system is in compliance with regulatory requirements. • Organizational Change: If applicable, managing organizational changes that have resulted from the use of the AI system. • Evaluating User Experience: Including ensuring end-users have the capacity and literacy needed to understand and use the AI system.	Use: There are a number of uses that have been attributed to 'human-centric' AI. For example, the Japanese Council for Social Principles of 'Human-centric' AI in the Social Principles of Human-Centric AI includes examples like expanding individual abilities and creativity, enabling individuals to pursue their own well-being, achieving sustainable development goals, and addressing social issues such as declining birth rate, aging populations, and increased fiscal spending. Regulatory Compliance: This can include, but is not limited to, regulation directly related to AI, consumer protection, privacy, anti-discrimination, human rights as well as sectoral regulation. Organizational Change: The Partnership on AI's "Framework for Promoting Workforce Well-being in the AI- Integrated Workplace" provides a conceptual framework and a set of tools to guide employers, workers, and other stakeholders towards promoting workforce well-being throughout the process of introducing AI into the workplace. Evaluating user experience: The principle of 'human-centric' AI in the Japanese Principles of Human-Centric AI emphasizes digital literacy as a means to help protect against an over-dependence on AI, and misuse of AI including to manipulate others. The principle also emphasizes accessible interfaces so as to avoid creating a digital divide between those with skills and information and those without.
Operation and Monitoring	Use of the principle to guide: • Monitoring: Monitoring of the AI system for effectiveness and reasonably foreseeable negative/positive and intended/unintended impacts against objectives and ethical considerations. • Improvement: Improving and adapting the AI system including retiring an AI system if necessary. • Communicating Product Changes: Incorporating and communicating policy and product changes.	Monitoring: Monitoring AI systems based on 'human-centric' dimensions. For example, the article 'Survey of Human-Centered Evaluations in Human-Centered Machine Learning' surveys literature that has explored monitoring the following dimensions of an AI system • Quality and accuracy of the model • Perceived quality based on user observations • Transparency • Interpretability • Trustworthiness • Effectiveness The article also surveys the use of tools to carry out evaluations and bring in qualitative information including • User evaluations, surveys, and questionnaires • Participant self-reports and Likert scales • Interviews • Expert reviews • Case studies and use cases • Crowdsourced experiments • Empirical feedback • Long term collaboration with experts. • User feedback including implicit and explicit

Focus Area	Framework	Operational Guidance for 'Human-Centric' AI
Operation and Monitoring (cont.)	Use of the principle to guide: • Use: The use of the AI system. • Regulatory Compliance: Ensuring the AI system is in compliance with regulatory requirements. • Organizational Change: If applicable, managing organizational changes that have resulted from the use of the AI system. • Evaluating User Experience: Including ensuring end-users have the capacity and literacy needed to understand and use the AI system.	The article The Flaw of Policies Requiring Human Oversight of Government Algorithms has noted the limitations of relying on human review alone to bring meaningful oversight into AI systems and has stressed the importance of also assessing human interventions in AI systems. The article The False Comfort of Human Oversight as an Antidote to A.I Harm highlights aspects to consider when shaping oversight mechanisms including • Information needed for meaningful oversight to be provided. • Potential influence of the AI system on the decisions taken by the human oversight mechanism or bias that might exist within the human oversight mechanism. • The role that factors external to the AI system may play in its functioning. The Expert Group on How AI Principles should be implemented in the Governance Guidelines for Implementation of AI Principles recommends that when an 'AI Incident' occurs, companies provide end-users with an explanation, identify the extent of the impact, take steps to prevent further spread, clarify legal responsibilities, and consider relief measures. Companies can also share information with repositories focused on documenting 'AI incidents'. The Partnership on AI has developed an AI Incidents Database that documents intelligent systems causing safety, fairness, and other real-world problems and can be used by companies. Improvement: The impact assessments outlined in the section 'Impact & Risk' can also be used as tools to monitor the impact of an AI system, and identify areas for improvement including taking decisions to retire an AI system. Communicating Product Changes: Article 12 of the proposed Digital Services Act requires intermediaries to inform recipients of any significant changes to the terms and conditions including the use of algorithmic decision-making in content moderation.

Strategy for Stakeholder Engagement and Incorporating Context

Stakeholder engagement and incorporating context is critical to the process of a company defining and implementing a principle(s) across the company and AI system lifecycle. It is also fundamental to the implementation of the principle of 'human-centric' AI as it allows for the incorporation of domain expertise, context, and the lived realities of end-users. It is also a step towards ensuring that AI systems reflect and meet a specified need and that potential and real impacts are identified.[36]

This framework draws upon the definition of 'Meaningful Stakeholder Engagement' put forward by the OECD in the context of the extractive industries: Meaningful stakeholder engagement refers to ongoing engagement with stakeholders that is two-way, conducted in good faith and responsive. The OECD clarifies the following characteristics of the engagement: two-way meaning parties may freely express opinions, there is sharing of decision-making power, and stakeholders also lead engagement activities, good faith meaning engagement is done with genuine intention to understand the perspectives put forward, and preparedness to address adverse impacts raised, responsive engagement meaning that there is follow-through on outcomes from the engagement and that stakeholders opinions are taken into account when decisions are made, and ongoing engagement meaning that the stakeholder engagement continues throughout the lifecycle of the project and is not a one-off initiative.[37] It also draws upon best practices from the OECD Best Practice Principles on Stakeholder Engagement in Regulatory Policy. These include a range of instruments

1. Defining a policy for an open and balanced public consultation including oversight mechanisms to ensure quality of the process.
2. Engaging stakeholders at each stage of the governance lifecycle through appropriate methods and in a way that is proportionate to the impact and significance of the regulation.
3. Ensuring engagement takes place with sufficient time to incorporate stakeholder inputs.
4. Providing an explanation of how stakeholder input has been assessed and incorporated to stakeholders as well as steps taken to ensure the balancing of different interests.
5. Emphasizing engagement with the least represented.
6. Providing stakeholders with sufficient time to provide input and incorporate the same as well as sufficient, specific, and accessible information to inform stakeholder input.
7. Using appropriate consultation tools.
8. Regularly evaluate the stakeholder engagement policy and process.[38] It is also important that principles of equity, accountability, transparency, and participation are applied to stakeholder engagement.

Companies should engage with stakeholders to the extent possible when implementing a principle across a company and throughout the AI system lifecycle. Similarly, different aspects of context may be relevant to various stages of a company implementing a principle across a company and in the AI system lifecycle. Thus, at each stage, we recommend companies identify which contextual aspects are relevant, the method to identify these contextual aspects, and how each aspect will be incorporated into the decision-making process of a company.

The below steps for forming a strategy for stakeholder engagement and incorporating context can be applied to each part of the framework as well as a specific focus area.

Stage 1- Planning and Preparation

As the first stage of engaging stakeholders and incorporating context in the implementation of a principle across a company and the AI system lifecycle, companies will need to develop a strategy to do so. This includes compiling background information, mapping stakeholders, developing an engagement strategy, and mapping context.

1.0 Background Information

To inform decisions around which stakeholders will be important to engage with, what contextual aspects to account for, and how, companies will need to document relevant information about the AI system, the context(s) the AI system will be deployed in, the principle, and specific information as it pertains to the framework.

Focus Area	Description
Purpose and Objectives	The purpose and objectives of the AI system.
Intended use	The intended use of the AI system.
Context	The geographical context(s) where the AI system will be deployed including relevant languages.
Framework	The part of the framework that stakeholder engagement will inform.
Principle(s)	The principle(s) that is being implemented across the company and AI system lifecycle.

1.1 Stakeholder Mapping

Companies will need to undertake a process of identifying the stakeholders they will seek to engage with when implementing the part(s) of the framework. In doing so, companies will need to identify the relevance of the stakeholder, ensure diversity and equity, and identify/ account for potential safety concerns related to stakeholder participation.

Focus Area	Framework
Stakeholders	Identification of the stakeholders that will be engaged. Categories of stakeholders highlighted in this framework that could be useful for companies to engage with when implementing a principle(s) across a company and the AI system lifecycle include • Experts: Stakeholders that have relevant expertise or perspective in the a.) definition, conceptualization and contextual implementation of the AI principle in question b.) the design, development, use, and impact of the AI system that is being developed. This can include but is not limited to, practitioners, civil society, academics, and sectoral experts.
Relevance	The relevance of the stakeholder to the AI system, the context, the framework, and/or the principle. • Accountability Organizations: Stakeholders that can play a role in defining best practices and holding companies accountable. This can include multi-stakeholder organizations, international organizations, policymakers, industry bodies, standards bodies, auditing and consulting firms, sustainability business organizations, investors, and foundations. • AI system operators: Stakeholders who procure and operate AI systems. This can include governments, companies, and individuals or groups of consumers. • End-users and impacted populations: Stakeholders that directly or indirectly use, engage with, and are/or are affected by AI systems. This can include individuals, groups, and communities. This includes marginalized groups that may require fairness/human rights/human-centred AI considerations. • Other: Other stakeholders that may be relevant to the AI system, the context, the framework, or the principle.
Diversity	The diversity of stakeholders engaged with including representation of different backgrounds, demographics, abilities, genders, geographies, and language with an emphasis on including vulnerable and marginalized communities.[39]
Equity	The sharing of decision-making power between companies and stakeholders. The ability of stakeholders to access and equally participate independently in the engagement process. Potential power dynamics between stakeholders and stakeholders and companies should be identified and addressed. This can include reaching out to stakeholders whose voices are underrepresented.
Trust & Safety	The potential risks that stakeholders may face by participating in the engagement process. Accepting anonymous contributions and holding meetings under Chatham House Rule.

1. 2 Engagement Strategy

After identifying the relevant stakeholders and ensuring diversity, accounting for power dynamics, companies will need to develop an engagement strategy. This includes identifying the purpose, timing and scale, and method for engagement. It also includes identifying the questions that will be asked and the information that will be shared with stakeholders to facilitate the engagement.

Focus Area	Framework
Purpose	The purpose for the stakeholder engagement. This can include, but is not limited to: • Bringing needed expertise and perspective to the AI system, the context, the part(s) of the framework, and/or the principle. • Identifying and articulating impact (positive and negative). • Holding companies accountable and working to strengthen processes.
Timing and Scale	Criteria to guide decisions around if, when, how frequently, and the scale of stakeholder engagement can include: • Human Impact: Companies should seek to collaborate with stakeholders if a solution impacts individuals directly or indirectly. Companies can draw upon a number of resources including frameworks related to the risk and potential harm of AI[40] and incident databases[41] to identify if their system has a human impact (positive and negative). • Proportionate engagement: Stakeholder engagement should be proportionate to the scale and scope of potential impact as well as the size of the organization. This includes how many stakeholders are engaged and how frequently. • Timing: Stakeholder engagement should take place early in the process to ensure that stakeholders have enough time to share inputs and the inputs incorporated into multiple versions if needed.
Method	Companies should seek to collaborate with stakeholders through an inclusive and accessible approach, method, and format: • Approach: Companies should ensure that the engagement is open, participatory, and transparent. • Format: There are a range of methods companies can use for stakeholder engagement. Some of these include interviews (one on one and semi-structured), surveys, focus groups, user groups, consultations, questionnaires, 'citizen juries' town meetings, listening tours, conferences, nominal group technique, Delphi technique, modeling, concept mapping, and scoping studies.[42] When determining the format for engagement, companies should pursue meaningful engagement that recognizes and places stakeholders as experts. To build processes that are two-way, companies should ensure that stakeholders co-lead or lead the engagement. Lessons can also be learned from approaches like Innovative Citizen Participation,[43] Citizen's Councils,[44] and Deliberative Democracy.[45] • Accessibility and Inclusion: Companies should seek to build inclusive engagement processes. This includes ensuring that the format for engagement is inclusive of relevant languages and accessible formats.

Questions

The information that is sought through the stakeholder engagement and the questions that will be asked.

Information to Inform Engagement

The knowledge and information stakeholders need about an AI system to engage meaningfully. Companies should ensure this information is presented in relevant languages and accessible/understandable formats. This information can include, but is not limited to[46]

Overview of AI System

- Purpose and goals of the AI system including how it will meet the needs of end-users.
- Key functions and features of the AI system including the languages supported and aspects related to usability, explainability, fairness, and security.
- Information about the accuracy rates of the algorithm.
- Data requirements of the AI system, sources of training data, and the types of data the AI system will collect and use.
- The Human-AI configuration including
  - Level of control end-users will have over the AI system.
  - Extent to which the AI system will take outputs autonomously, semi-autonomously, partially support human decision-making, or fully support human decision-making.
  - Information about the AI system will be shared with end-users.
  - Interaction between humans and the AI system (will the system display human behavior etc.)

Use of the AI system

- Intended use of the AI system.
- Expected end-users and impacted populations that directly or indirectly use, engage with, and are/or are affected by AI systems. This can include individuals, groups, and communities. This can also include marginalized groups that may require fairness/human rights/human-centred AI considerations.
- Expected AI system operators who procure and operate AI systems. This can include governments, companies, and individuals or groups of consumers.
- The complexity of the conditions that the AI system will be deployed in.
- The impact that failure of the AI system could have.
- Known limitations of the AI system.
- Potential ways the AI system could be misused and safeguards in place to prevent this misuse.
- The geographies where the AI system will be deployed including information about the legal, political, and cultural context.

1.3 Mapping Context

Companies will need to go through a process of mapping contextual elements that will be relevant to the AI system, the principle, and the part of the framework that is being implemented. They will also need to identify the method used to identify the contextual elements.

Focus Area	Framework
Contextual Aspects	Which contextual aspects will be relevant is dependent on the type of AI system, the intended use, and the intended AI operators and end-users/impacted populations. Contextual aspects that companies can consider accounting for, are not limited to, but can include: • Environmental: Aspects related to the infrastructure, physical properties, and restrictions in a context. Examples include the availability of infrastructure necessary for the development and functioning of the AI system, levels of digital literacy, levels of mobile penetration, and digital divides. [47] • Cultural: Aspects of a society that shape how people live - their behaviors, beliefs, expectations, and attitudes.[48] Examples include knowledge and stories, language, value systems, religions, traditions and rituals, techniques and skills, tools and objects, art, food & drink, and social organization. • Legal & Political: Aspects related to the legal and political environment. Examples include political actors and their power relationships, relevant legal and policy frameworks, government practices, and strength of democratic institutions. • Historical: Histories of social, economic, cultural, and political influences. Examples include histories around technological development and use as well as histories of discrimination and bias. • Economic: Aspects related to the structure of economic life in a context. Examples include resource availability, skilled labor force, markets, and relevant government policies.
Methods	Methods: Research has emphasized the importance of interdisciplinary, transdisciplinary, and systems approach when trying to account for context.[49] Examples of research methods that have been identified as relevant for learning about context include:[50] • Ethnographic research: Observation and interviews to understand human behavior. [51] • Contextual inquiry: Semi-structured interviews with interviewees situated in the relevant context. • Expert interviews: Interviews with sectoral or domain experts. • Surveys: Collection of comparative data from a pool of participants. • Action research: Ongoing collaboration with stakeholders to understand the interdependencies between technologies and communities.[52] Where relevant, methods specific to different parts of the AI system lifecycle have also been identified in the operational guidance of the framework.

Stage 2- Engaging Stakeholders and Incorporating Input and Context

2.1 Engaging Stakeholders

After developing a stakeholder engagement strategy, companies would need to engage with stakeholders. This includes documenting the inputs received from stakeholders, recognizing stakeholder contributions, and sharing information about the stakeholder engagement with the public.

Focus Area	Description
Documenting Engagement	The process and format that input can be provided including the language the engagement is held in, if anonymous inputs are accepted, and if workshops will be held under the Chatham House rule.
Recognizing Contributions	The process and format for recognizing the contributions of stakeholders engaged with. This can include compensation and attribution.
Transparency	As a form of accountability, the process of sharing with the public high-level information about the stakeholder engagement including the purpose and objectives, scale and scope, and the type of stakeholders that were engaged.

2.2 Evaluating and Incorporating Stakeholder Input and Context

After engaging with stakeholders and mapping context, companies will need to undertake a process of evaluating and incorporating stakeholder input and contextual elements.

Focus Area	Description
Evaluating Stakeholder Input	The process of evaluating the information received through stakeholder engagement. When doing so questions to consider include: • How can stakeholders be brought into the decision-making process? • What is the criteria that will be used to evaluate input? • Are there differences in terms of voice, access, and independence of inputs received? • Is there conflicting input? How can both perspectives be represented?
Incorporating Stakeholder Input and Context	The process of determining where and how inputs from stakeholders and contextual aspects will be incorporated into the framework. This can include incorporation into an organizational process, policy, product/AI system lifecycle.

Sharing Changes	As a form of accountability, the process of sharing back with stakeholders, and where appropriate, the public: • The criteria used to evaluate stakeholder input. • How conflicting stakeholder input was represented. • The stakeholder input that was received and the input that was incorporated. • The contextual aspects that were incorporated. • How and where the input and contextual aspects were incorporated such as informing decisions related to a policy, process, or product.

Stage 3- Evaluating the Process of Engaging Stakeholders and Incorporating Context

3.1 Evaluation and Learnings

When evaluating the process of engaging stakeholders and incorporating context, companies should identify successes, challenges, and areas for improvement in key dimensions of the engagement. They should also learn about the experience of stakeholders.

Focus Area	Description
Composition and Size of Engagement	Evaluating if the composition of stakeholder engagement was diverse in terms of engaging with stakeholders from different backgrounds, demographics, abilities, genders, geographies, and language with an emphasis on including vulnerable and marginalized communities. A process of assessing if the engagement was proportionate to the scale of potential impact of the AI system, if and the size of the organization.
Logistics and Efficient Use of Resources	Evaluating the logistics of the stakeholder engagement and resources used.
Quality of Inputs and Changes Made Based on Inputs	Evaluating the quality and comprehensiveness of inputs received through the stakeholder engagement including if the questions asked were comprehensive and the contextual information relevant as well as evaluating how inputs were incorporated including thoroughness and transparency and equity of the process.

Stakeholder Experience	Evaluating the experience of stakeholders in the engagement process. Questions companies can pose to stakeholders to learn more about their experience can include: • Was the engagement process accessible? • Was the question/information sought relevant? • Did you have sufficient information about the consultation to engage meaningfully? • Did you have sufficient information about the AI system and its intended use to engage meaningfully? • Do you feel your input was adequately accounted for and incorporated? • Do you feel that you have decision-making power? • Do you feel that you were able to express your opinions freely? • Is there something you would change about the engagement process?

References

[1] As noted by The Alan Turing Institute in the report Understanding Artificial Intelligence Ethics and Safety "In the context of practical ethics, the word 'normativity' means that a given concept, value, or belief puts a moral demand on one's practices, i.e. that such a concept, value, or belief indicates what one 'should' or 'ought to' do in circumstances where that concept, value, or belief applies." For more information see: https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf

[2] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3403010

[3] For example, in 2019 the G20 published AI principles based on the OECD recommendations on AI. For more information see: https://oecd-innovation-blog.com/2020/07/24/g20-artificial-intelligence-ai-principles-oecd-report/

[4] For example, UNESCO has finalized an ethical framework for AI that can be applied across contexts. The framework was developed through international negotiations and consultations. For more information see: https://www.unesco.org/en/articles/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence

[5] For example, in the article On becoming human: An African notion of justice and equity in Machine Learning, Sabelo Mhlambi explores understandings of justice and equity in the Ubuntu culture and notes "A homogeneity is assumed when discussing ethics in technology and posits western ethics as the default framework in addressing the harmful effects of technology." For more information see: https://sabelo.mhlambi.com/ubuntåu/

[6] For example, in 2020 the Partnership on AI launched an interactive project to identify gaps in the implementation of ethical AI principles. For more information see:https://partnershiponai.org/pai-launches-interactive-project-to-put-ethical-ai-principles-into-practice/

[7] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3403010 https://deliverypdf.ssrn.com/delivery.php?ID=952112098007007071098076029025103024031086037020053013069080084065096030112124099087055031099031027019034092124089124093014003033047002081054083083100116123107029039076022117009090127085107003004080094027067119028114103084109125127079064078009085116&EXT=pdf&INDEX=TRUE

[8] This policy prototype includes reference to publicly available materials and toolkits produced by different public and commercial entities, in order to support the Framework understanding and use. Such reference does not imply recommendation or endorsement by the Open Loop consortium. These examples do not endorse the accuracy of any information stated by these companies or organisations.

[9] Responsible AI principles - Principle of protection and reinforcement of positive human values - AI should promote positive human values and not disturb in any way social harmony in community relationships." For more information see: https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf

[10] Principle 18 of the UN Guiding Principles states "In order to gauge human rights risks, business enterprises should identify and assess any actual or potential adverse human rights impacts with which they may be involved either through their own activities or as a result of their business relationships. This process should: (b) Involve meaningful consultation with potentially affected groups and other relevant stakeholders, as appropriate to the size of the business enterprise and the nature and context of the operation." For more information see: https://www.ungpreporting.org/reporting-framework/management-of-salient-human-rights-issues/stakeholder-engagement

[11] https://wp.oecd.ai/app/uploads/2022/02/Classification-2-pager-1.pdf

[12] https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm

[13] https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf

[14] https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf

[15] https://elcot.in/sites/default/files/AIPolicy2020.pdf

[16] https://it.telangana.gov.in/wp-content/uploads/2020/07/Govt-of-Telangana-Artificial-Intelligence-Framework-2020.pdf

[17] https://oecd.ai/en/ai-principles

References

[18] https://www.meti.go.jp/press/2020/01/20210115003/20210115003-3.pdf

[19] https://cyber.harvard.edu/publication/2020/principled-ai

[20] https://www.oecd-ilibrary.org/sites/8b303b6f-en/index.html?itemId=/content/component/8b303b6f-en#:~:text=The%20AI%20system%20lifecycle&text=The%20design%2C%20data%20and%20models,operation%20and%20monitoring%20(Figure%201.5

[21] AI_Impact_Assessment_A_Policy_Prototyping_Experiment.pdf

[22] https://oecd.ai/en/ai-principles

[23] https://mneguidelines.oecd.org/OECD-Guidance-Extractives-Sector-Stakeholder-Engagement.pdf

[24] https://www.oecd-ilibrary.org/sites/39416960-en/index.html?itemId=/content/component/39416960-en

[25] https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence

[26] https://www.mofa.go.jp/files/000486596.pdf

[27] https://oecd.ai/en/dashboards/ai-principles/P6

[28] https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf

[29] https://ai.bsa.org/wp-content/uploads/2019/09/humancentricai.pdf

[30] For example, UNESCO defines culture as "A set of distinctive spiritual, material, intellectual, and emotional features of society or a social group, and that it encompasses, in addition to art and literature, lifestyles, ways of living together, value systems, traditions and beliefs." For more information see: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8038476/#B161-sensors-21-02514

[31] This discourse is reflected in policy initiatives like the White House initiative to define a Bill of Rights for AI: https://www.whitehouse.gov/ostp/news-updates/2021/10/22/icymi-wired-opinion-americans-need-a-bill-of-rights-for-an-ai-powered-world/, the proposed EU AI regulation:https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206, and the proposed EU Digital Services Act: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en

[32] In implementation guidance, Niti Aayog has noted that the diversity, scale, digital divide, lack of awareness, and inequality in India can serve as vectors of harm for AI. It has also recognized that factors like digital divides can impact the completeness and representation found in datasets. For more information see: https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf

[33] https://cit4vet.erasmus.site/module-1-the-concept-of-nculture/5/

[34] Frameworks focused on the implementation of human rights across a company have emphasized the importance of company commitment starting with board-level commitment. For example see: https://rankingdigitalrights.org/index2020/

[35] https://gpai.ai/projects/data-governance/role-of-data-in-ai.pdf

[36] https://partnershiponai.org/methodsforinclusion/

[37] https://oecd.ai/en/wonk/classification. Researchers have outlined harms emerging from faulty inputs, faulty conclusions, and a failure to test. https://yjolt.org/sites/default/files/23_yale_j.l._tech._special_issue_1.pdf

[38] For example, The Partnership on AI has developed an incident database for AI incidents. https://partnershiponai.org/workstream/ai-incidents-database/. The Center for Security and Emerging Technology documents created a searchable repository of AI incidents and codes them according to safety, fairness, industry, geography, timing, and cost. https://incidentdatabase.ai/taxonomy/cset?lang=en.

[39] https://www.ncbi.nlm.nih.gov/books/NBK62556/

[40] https://www.oecd.org/governance/innovative-citizen-participation/

References

[41] For example, in the UK the National Institute for Health and Care Excellence has established a Citizens Council that provides input into overarching moral and ethical issues that NICE should take into consideration. This has included defining societal values to be considered when evaluating trade-offs equity and efficiency. For more information see: https://pubmed.ncbi.nlm.nih.gov/28230944/

[42] Lessons can be learned from approaches such as deliberative democracy which emphasizes structured consultations with those affected by collective decisions through norms to guide engagement, good information, and equal access and participation guided by the principles of deliberation and sortition. For more information see: https://news.stanford.edu/2021/02/04/deliberative-democracy-depolarize-america/

[43] This section draws upon categories assessed in the Microsoft Responsible AI Impact Assessment Template. For more information see: https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf

[44] In implementation guidance, NITI Aayog has noted that the diversity, scale, digital divide, lack of awareness, and inequality in India can serve as vectors of harm for AI. It has also recognized that factors like digital divides can impact the completeness and representation found in datasets. For more information see: https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf

[45] https://cit4vet.erasmus.site/module-1-the-concept-of-culture/5/

[46] https://www.happycounts.org/uploads/2/4/4/6/24468989/artificialintelligenceandcommunitywellbeing_1.pdf

[47] For a list of human-centered methods and design tools see:https://www.vic.gov.au/methods-human-centred-design-tools-and-references

[48] https://www.nature.com/articles/s42256-021-00323-0

[49] https://www.happycounts.org/uploads/2/4/4/6/24468989artificialintelligenceandcommunitywellbeing_1.pdf

[50]For a list of human-centered methods and design tools see:https://www.vic.gov.au/methods-human-centred-design-tools-and-references

[51]https://www.nature.com/articles/s42256-021-00323-0

[52]https://www.happycounts.org/uploads/2/4/4/6/24468989/artificialintelligenceandcommunitywellbeing_1.pdf