The Dialogue™

INFORM ENGAGE IDEATE

Principles for Enabling Responsible AI Innovations in India: An Ecosystem Approach

Authors: Kamesh Shekar[1], Jameela Sahiba[2], Bhavya Birla[3], Garima Saxena[4]

[1] Author is a Programme Manager - Privacy and Data Governance at The Dialogue. Corresponding Author's email: Kamesh@thedialogue.co
[2] Author is a Senior Programme Manager - Emerging Technologies at The Dialogue.
[3] Author is a Research Associate at The Dialogue
[4] Garima is a Research Associate at The Dialogue

List of Expert Reviews

The authors would like to thank the following experts for their expert comments and peer review of the paper. All errors and omissions remain those of the authors

Mr Reggie Townsend
Vice President, Data Ethics Practice (DEP) – SAS Institute

As Vice President of the Data Ethics Practice at SAS, Reggie leads a global effort to build a better, more equitable future through fair, sustainable applications of data, artificial intelligence, and associated technologies. Also, he is a member of the National Artificial Intelligence Advisory Committee that advises the US President and the National Artificial Intelligence Initiative Office on matters related to the National Artificial Intelligence Initiative.

Ms Mitisha Gaur,
Former Artificial Intelligence Researcher, The Italian Data Protection Authority

Mitisha is an engineer turned lawyer with a long-standing and specific interest in the cross-section between law and technology, which led her to apply to the LeADS Program. During her time being associated with the LeADS Project, Mitisha has also been a part of interdisciplinary working groups which are exploring the relationships between Privacy and Intellectual Property and studying data portability and data quality perspectives. Also, she served as secondment with the Department of Artificial Intelligence at the Italian Data Protection Authority, Rome.

Dr Archana G. Gulati
Retired Indian Civil Service Officer

Dr Gulati is a telecom and competition policy expert with more than 30 years of experience, Dr. Gulati is a retired Indian Civil Services. She has served as Senior Deputy Director-General with the Department of Telecommunications and as Advisor and Head of the Combination (M&A) Division of the Competition Commission of India. She is Senior Advisor, Competition Law with a law firm (Trilegal).

Ms Ingrid Soares
Associate Lawyer, Mattos Filho

Ms Soares is an expert in AI governance focused on the private sector.

Table of Contents

Executive Summary

With the rapid proliferation of artificial intelligence (AI) across various domains, discussions surrounding responsible AI have become ubiquitous. These versatile technologies are transforming the nature of our work, interactions, and lifestyles. We are on the brink of witnessing a transformational shift comparable to the impact of the printing press, which revolutionised the world six centuries ago. As a result, several countries and industry bodies are actively engaged in formulating frameworks for algorithmic decision-making that prioritise ethics and the fundamental principles and values associated with responsible AI. Establishing a reliable approach is crucial for fostering "responsible competitiveness" in the realm of AI. This approach is the bedrock on which all individuals and entities involved with AI systems can have confidence that their design, development, and utilisation adhere to legal, ethical, and resilient standards.

To begin with, let us first define responsible AI. Reliability in Artificial intelligence is a by-product of an AI model being trustworthy, safe, fair etc. Its primary objective is to prioritise human agency and well-being while mitigating the potential risks and negative consequences for all involved stakeholders. The ultimate goal of having a reliable AI can be achieved by adopting a trustworthy, purposeful, comprehensive, and responsive approach toward AI development and deployment.

Numerous AI ethics guidelines have been published globally, amounting to multiple sets of guidelines[5]. However, most of the existing literature on the risk management of AI at the development level focuses on uni-stakeholders, i.e., AI developers.[6] However, in this paper, we attempt to suggest an effective governance structure for AI that would be multistakeholder involving AI developers, AI deployers and impact populations, where we discuss their duty towards making AI trustworthy and safe.

Firstly, we will differentiate between harms and impacts emerging at different stages of the AI lifecycle. The objective of doing this is to develop a map of harms and impacts caused by different stakeholders at different stages of the AI lifecycle. In addition, the objective is to declutter and distribute the impact and harm caused by AI, which emerges at different stages so that appropriate steps can be taken.

[5] Ethics guidelines for trustworthy AI | Shaping Europe's digital future. (2019, April 8). Shaping Europe's digital future. Retrieved August 16, 2023, from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[6] While most of the risk management literature talks only about AI developers, some of the key frameworks and policy instructment like NIST Risk Management Framework (NIST 2023) and EU AI Act discusses the role of the multistakeholders within the AI ecosystem. For instance, NIST 2023 highlights that different AI actors have different responsibilities and awareness depending on their roles in the lifecycle.

Followed by mapping the harms and impact to tackle the same, this paper suggests principles to be followed by AI developers, AI deployers and impact populations at the different stages of the AI lifecycle. Mapped critical principles for AI development and deployment advised by the frameworks developed by various governments, intergovernmental organisations, academia, civil society etc., in India and globally.

By not stopping at just mapping the principles, this paper suggests an indicative operational strategy that translates good practice and governance principles into action points. While some of the principles mapped could be universally applied to AI developers, AI Deployers and the Impact population, we realise the fundamental difference when translated into operational action points. For instance, accountability as a principle for AI developers may mean having better internal processes and board-level supervision. However, the same for AI deployers may mean that their processes are open and accountable to impact populations. Therefore, the paper provides an indicative operationalisation strategy to bring out these differences.

1. Introduction

The internet is advancing at an exponential pace; where within a short period, we have seen a transformation of two-dimensional Web 2.0 to technological developments like Artificial Intelligence, which senses the ethos to offer responses to our queries, which is almost near to human reply. Artificial Intelligence is one of the driving forces of change that will shape the Internet in the coming days.[7]

Therefore, making Artificial Intelligence trustworthy will contribute to making the Internet trustworthy.[8] For instance, as the Internet evolved, the face of Web 2.0 has been the intermediaries like social media platforms, search engines, etc.,[9] which has brought to the floor the importance of the safe harbour and online safety debates; similarly, with the evolution to Web 3.0, increasingly we see that Artificial Intelligence is becoming the face of Internet. Therefore, to exert individuals' trust in the internet, tackling concerns emerging with Artificial Intelligence is important.

Taking a step back, it is essential to acknowledge the positive fundamental changes brought by driving forces like artificial intelligence, which would complement human intelligence to solve bigger challenges (refer to box 1). However, Artificial intelligence also creates a sense of uncertainty amongst individuals regarding the future of the Internet, which has got elevated with the recent explosion of innovations around generative AI solutions.

Healthcare: There have been various instances where some healthcare facilities have been improved through AI integration. For instance, AI integrated with ultrasound technology has proven to enhance the measure of the baby's fetal position when it's exiting the womb; this has helped the AI deployers, i.e., healthcare providers, with additional information to make informed decisions that keep both mother and baby healthy. Besides, in simple day-to-day life, many menstrual cycle tracking apps use AI technology to predict women's ovulation period to enhance their health outcomes.

[7] Thomas, M. (2022, August 9). *The future of AI: How artificial intelligence will change the world*. Built-In. Retrieved June 20, 2023, from https://builtin.com/artificial-intelligence/artificial-intelligence-future

[8] While typical use-cases of AI technologies is beyond traditional experience of using internet, however as rightly identified by the Internet Society's Global Internet Report 2017, Artificial Intelligence is one of the driving forces of change that will shape the Internet in the coming days.

[9] O'Neill, S. (2022, January 7). *What's The Difference Between Web 1.0, Web 2.0, And Web 3.0?* MarTech Alliance. Retrieved August 16, 2023, from https://www.lxahub.com/stories/whats-the-difference-between-web-1.0-web-2.0-and-web-3.0

Climate Change: While Satellites can identify hyperspectral images of the Earth, there are limitations between how space images are captured and how Earth operates. Therefore, to bridge this gap, AI integrations to satellites have proven to help climatologists detect events like fast-moving weather patterns and forest fires and prioritise these images. Besides, AI solutions are also helpful for gathering, completing, and interpreting large, complex data sets on emissions, climate impact, and more to predict climate change, which helps us strategise measures to control the same.

Education: In the past decade or so, we've seen some major strides in AI, particularly in the field of natural language processing (NLP) and conversational AI, which has a key impact on the education sector. These technologies have played an important role in taking education to the last mile in terms of enabling conversational/interactive learning and also translating content to multiple languages. Besides, AI solutions have also evolved to a level where they can gauge students' learning styles and pre-existing knowledge to deliver customised support and instruction.

Finance: AI technologies are extensively used in finance services for intermediation, including banks, NBFCs, underwriters, and credit-lending institutes, to make informed decisions when adopting AI systems in areas such as fraud detection, algorithmic trading, credit-lending, and robo-advisory. By integrating trustworthy AI, finance service providers foster a more reliable and responsible financial landscape, benefiting both the industry and its customers.

Agriculture: AI technology is extensively used for environmental sustainability and social impact. AI technologies are specifically used to address challenges specific to agriculture, such as precision farming, soil testing and crop health monitoring. It will provide players within the agriculture sector with the ability to design systems that optimise resource efficiency, reduce environmental impact, and prioritise ethical considerations. Additionally, AI technologies offer responsible applications promoting sustainable and equitable agricultural practices.

To tackle this uncertainty around Artificial Intelligence (which is driving the future of the Internet), various regulatory developments have cropped up worldwide to enhance AI risk management and trustworthiness. While there are various positives and negatives with how other jurisdictions are trying to tackle the issues related to Artificial Intelligence, this paper will discuss why India must consider laying out enabling principles at the ecosystem level to support home-grown AI innovations to serve worldwide. Artificial Intelligence reflects society like a mirror; therefore, through this paper, we emphasise that everyone within the AI ecosystem, including AI developers, AI deployers and the impact population, has a stake in making the ecosystem trustworthy.

This paper will effectively contribute toward the discussion on developing an effective governance structure for AI to enhance its opportunities while mitigating its impact and harms. There are various kinds of literature on the risk management of AI at the development level

focusing on uni-stakeholder, i.e., AI developers.[10] However, the approach to this paper for establishing an effective governance structure for AI would involve multi-stakeholders, including AI developers, AI deployers and impact population, where we map principles for different stakeholders within the AI ecosystem to make it trustworthy and safe.

When Liquified Petroleum Gas (LPG) Cylinders are made for domestic consumption, the manufacturers would have taken most precautions to make the cylinders absolutely safe for domestic consumption; however, if the individuals as users manhandle the LPG Cylinders definitely, the chance of the same causing negative impact is high. Similarly, though AI developers take high-risk management measures, still if AI deployers misuse and impact the population unaware, the fall through the cracks happens. Therefore, the question is, can AI developers be held accountable for AI deployers' behaviour? Can just holding AI developers accountable for their actions enough to tackle the implications of AI solutions? This paper will answer these questions by proposing a Principle-based Multistakeholder Approach as an ecosystem-level intervention.

Chapter 2 of the paper will discuss various global developments in regulating Artificial Intelligence and operationalising key principles to set the context. Following this, in Chapter 3, we will list the five critical implications of AI solutions and try to map out the extent to which AI developers, AI deployers, and the impact population contribute towards manifesting the same. In addition, in Chapter 3, we propose a principle-based multistakeholder approach where we map the principles to be followed by stakeholders, namely AI developers, AI deployers and impact population at appropriate stages. Mapping the principles in the previous chapter. Chapter 3 also discusses indicative operationalisation strategies for AI developers, AI deployers, and the impact population to imbibe the mapped principles. Finally, Chapter 4 discusses the government's role in implementing the principle-based multistakeholder approach.

2. Status-quo of AI Regulations

Regulatory developments have cropped up worldwide to enhance AI risk management and trustworthiness. Against this backdrop, this chapter will discuss various global developments in regulating Artificial Intelligence and operationalising key principles. While various developments are happening around regulating AI worldwide, this chapter discusses some of the critical frameworks that have emerged at the lateral and multilateral levels across the globe.

Box 2: Concerns with AI regulations

[10] Rogers, J. (2023, January 11). *Artificial intelligence risk & governance. AI & Analytics for Business*. Retrieved June 20, 2023, from https://aiab.wharton.upenn.edu/research/artificial-intelligence-risk-governance/

Artificial intelligence governance is fragmented worldwide, primarily because it is rooted in two issues at the heart of the governance of all emerging technologies: The pacing problem and the Collingridge dilemma.[11] Firstly, the pacing problem refers to the act of catching up done by legislatures worldwide, given the rapid advancements in emerging technologies and the countries' slow-paced formulation of laws and regulations. Secondly, David Collingridge proposed the Collingridge dilemma to highlight that we can successfully regulate a given technology when it's still young and unpopular and thus probably still hiding its unanticipated and undesirable consequences, or we can wait and see what those consequences are but then risk losing control over its regulation.

Beyond the pacing problem, Artificial Intelligence is hard to regulate as definitions need continual updating with emerging technologies. A very good example of how fast technology outpaces definitions can be observed in the definitions made by the OECD in 2019. The OECD definition from 2019 did not include 'content generation' within its ambit and, thereby, would not apply to the currently booming generative AI industry. This was corrected in a way under the EU AI Act that includes systems that generate "content" in addition to "predictions, recommendations, or decisions."[12]

Definitional challenges seem to manifest in two distinct trade-offs as well. Whether to define AI technically or through a Human-centric approach and ensure that the scope of the definition is optimal and congruent to the regulatory aims. Human-centric approaches define AI in relation to Human activities. For instance, in the U.S. Department of Defense AI Strategy paper, the definition of AI is "the ability of machines to perform tasks that normally require human intelligence"[13], which is a contrast to the approach taken by the OECD where they define AI as a "machine-based system" that produces "predictions, recommendations, or decisions." Both approaches lead to different outcomes. Moreover, the AI definition is also increasingly evolving in the tangent where where "autonomy" has become an integral element of the definition. For instance, the AI definition within the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF) which is adapted from OECD Recommendation on AI:2019; ISO/IEC 22989:2022 is "an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy".

While the Human-centric approach views AI in socio-economic contexts and accommodates the rapidly changing nature of the technology itself, the latter enables legal precision and enables regulatory

[11] Srinivasan, K. R. (2023, May 2). *Two reasons AI is hard to regulate: The pacing problem and the Collingridge dilemma*. The Hindu: Breaking News, India News, Sports News and Live Updates. Retrieved June 20, 2023, from https://www.thehindu.com/sci-tech/science/ai-regulation-pacing-problem-collingridge-dilemma/article66802967.ece

[12] Murdick, D., Dunham, J., & Melot, J. (2020, June). *AI Definitions Affect Policymaking*. Center for Security and Emerging Technology. https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Definitions-Affect-Policymaking.pdf

[13] US Department of Defense. (2018). *Summary Of The 2018 Department Of Defense Artificial Intelligence Strategy: Harnessing AI to Advance Our Security and Prosperity.* U.S. Department of Defense. https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF

harmonisation as definitions founded upon technical capabilities remain constant across use cases and jurisdictions.[14]

2.1. India

At the soft touch/prescriptive level, through a series of discussion papers under the National Strategy on Artificial Intelligence (NSAI), NITI Aayog put forth various AI principles for the responsible use of emerging technologies. These papers aim to establish broad principles for the design, development, and deployment of AI systems in India.[15] It strives to make India's workforce ready for the future of work through skilling and also recommends various actions like setting up centres of excellence in the AI ecosystem, recommending that the government establish an attractive intellectual property regime for AI, and also introducing subjects teaching emerging technologies in schools. It also flags issues that could arise with the increased usage of AI systems, like algorithmic bias, privacy concerns, and ethical challenges, and requires governments to undertake research to address these challenges. NITI Aayog, in its updates to the AI Strategy document since its first report in June 2018, also provided for certain artificial intelligence principles to ensure the safe and responsible use of AI systems.

Furthermore, existing regulations and upcoming digital laws will apply to AI technologies and their developers. For instance, the Digital Personal Data Protection Act 2023 (DPDPA 2023) will apply to AI developers who develop and facilitate AI technologies.[16] AI developers will collect and use massive amounts of data to train their algorithms to enhance the AI solution so that they might be classified as data fiduciaries. This implies that AI developers may comply with the key principles of privacy and data protection like purpose limitation, data minimisation, consensual processing, contextual integrity, etc., as enshrined in DPDPB 2022. Besides, as contoured during the Digital India Act (DIA) consultation, the government is also considering having provisions within DIA that would define and regulate high-risk AI systems.[17]

---

[14] O'Shaughnessy, M. (2022, October 6). *One of the biggest problems in regulating AI is agreeing on a definition.* Carnegie Endowment for International Peace. Retrieved June 20, 2023, from https://carnegieendowment.org/2022/10/06/one-of-biggest-problems-in-regulating-ai-is-agreeing-on-definition-pub-88100

[15] NITI Aayog. (2022, November). *RESPONSIBLE AI #AIFORALL Adopting the Framework: A Use Case Approach on Facial Recognition Technology.* | NITI Aayog. https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf

[16] Bordoloi, P. (2023, April 15). *India backs off on AI regulation. But why?* Analytics India Magazine. Retrieved June 20, 2023, from https://analyticsindiamag.com/india-backs-off-on-ai-regulation-but-why/; Barik, S. (2023, August 12). MoS IT on concerns around Digital Personal Data Protection Act: There will be checks & balances to ensure personal data is not misused. The Indian Express. Retrieved August 16, 2023, from https://indianexpress.com/article/business/economy/concerns-around-contentious-provisions-of-data-protection-law-mos-it-8889933/

[17] Bordoloi, P. (2023, April 15). *India backs off on AI regulation. But why?* Analytics India Magazine. Retrieved June 20, 2023, from https://analyticsindiamag.com/india-backs-off-on-ai-regulation-but-why/

2.2. OECD AI Principles

The Committee on Digital Economy Policy (CDEP) tabled a proposal for the OECD Council of Ministers to regulate Artificial Intelligence in May 2019. 42 nations adopted the proposal, and it has since become a foundational document for other countries to build their national-level regulatory frameworks in congruence with the principled foundation set by the OECD.[18] The Recommendation identified five complementary values-based principles for the responsible stewardship of trustworthy AI and calls on AI actors to promote and implement them: (a) inclusive growth, sustainable development and well-being, (b) human-centred values and fairness, (c) transparency and explainability, (d) robustness, security and safety, and (e) accountability.

The OECD's principle-based regulatory framework has been incorporated in national frameworks of the EU, Brazil, India and the United States of America, amongst others, in varying proportions. The commonality between all frameworks can be seen in their reliance on human-centred values, fairness, transparency and accountability applicable to AI service providers within their respective jurisdictions.

2.3. European Union

The European Commission's regulatory framework on Artificial Intelligence aims to achieve several specific objectives. Firstly, it aims to ensure the safety of AI systems in the Union market and their compliance with existing laws on fundamental rights and Union values. Secondly, it seeks to provide legal certainty to encourage investment and innovation in AI. Thirdly, it aims to enhance governance and enforcement of laws related to fundamental rights and safety requirements for AI systems. Lastly, it aims to promote a single market for lawful, safe, and trustworthy AI applications and prevent market fragmentation.

Recently (on 14th June 2023), the European Parliament approved the proposed AI Act, which is now passed to the European Council for their approval. The regulation uses a risk-based approach, where systems are classified as having low or minimal risk, limited risk, high risk, or unacceptable risk. The Low-risk systems include spam filters or AI-enabled video games and comprise most of the systems currently being used on the market. The High-risk systems are the ones which can have a significant impact on the life and liberties of a user who is a natural person and are subject to specific requirements such as adequate disclosure, human oversight, transparency thresholds etc. High-risk systems include those used in Biometrics, Critical infrastructure, Education and vocational training, Employment, workers management, access to self-employment, Access to and enjoyment of essential private and public services and benefits,

[18] OECD Council on Artificial Intelligence. (2019, May). *OECD Principles on Artificial Intelligence.* OECD. https://www.oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm

and Health and life insurance. And Finally, Systems with unacceptable risk are those that manipulate behaviour in a way that may result in physical or psychological harm, exploit the vulnerabilities of groups, or are used for social scoring by governments and private actors. The risk-based approach allows regulators to slot new application areas into existing risk categories as AI's use cases evolve, balancing flexibility, supporting innovation and ensuring regulatory certainty.

2.4. United States

While acknowledging the potential that Artificial Intelligence could bring to society, the United States federal government also believes that protecting individuals' rights and tackling the safety concerns that emerge with developments like generative AI is important. On that note, there have been various ongoing efforts within the United States, with a significant one being the 4th May 2023 announcement by the Biden-Harris Administration of new actions that will further promote responsible American innovation in AI and protect people's rights and safety. This is a significant step and has global implications, as most technology companies innovating on generative AI are US-based companies while serving worldwide. Therefore, how the regulations will pan out in the United States will directly impact the development of safe and secure generative AI solutions, which will then be consumed worldwide.

The recent actions add to the federal government's ongoing effort to advance a cohesive and comprehensive approach to AI-related risks and opportunities like the Blueprint for AI Act, etc. The Blueprint for an AI Bill of Rights,[19] a report published by the White House Office of Science and Technology Policy in October 2022, intends to coordinate the efforts of a diverse set of federal agencies around a core set of principles, to address challenges posed by the AI systems to human civil rights. Secondly, the Federal Trade Commission, Consumer Financial Protection Bureau, Equal Employment Opportunity Commission (FTC), and Department of Justice's Civil Rights Division issued a joint statement that they are committed to applying the existing legislation to protect individuals from AI-related concerns.[20] Finally, fulfilling the initial mandate of the US Congress in 2020, the National Institute of Standards and Technology (NIST) recently released a version of the AI Risk Management Framework (AI RMF). The AI RMF offers a collection of risk management practices that organisations can use while designing and using artificial intelligence systems to manage risks and promote responsible development and use of trustworthy AI.[21]

---

[19] Office of Science and Technology Policy. (2022, October). *Blueprint For An AI Bill Of Rights Making Automated Systems Work For The American People.* The White House.
https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

[20] *Joint Statement On Enforcement Efforts Against Discrimination And Bias In Automated Systems.* (2023, April). Federal Trade Commission | Protecting America's Consumers.
https://www.ftc.gov/system/files/ftc_gov/pdf/EEOC-CRT-FTC-CFPB-AI-Joint-Statement%28final%29.pdf

[21] National Institute of Standards and Technology (2022, September 8). *AI risk management framework.* NIST.
https://www.nist.gov/itl/ai-risk-management-framework

2.5. Brazil

The holistic principles of Brazil's regulations on AI can be directly traced to its regulatory landscape. In 2021, the Brazilian House of Representatives approved the Brazilian Artificial Intelligence Act (Bill No. 21/2020), and the Bill proceeded to the Federal Senate for approval. In order to analyse Bill No. 21/2020 and suggest new wording, the Federal Senate established a specific commission composed of AI legal experts. In 2022, the AI commission approved a new draft for the Brazilian AI Act inspired by the EU AI Act and motivated by civil society organisations that demanded more transparency on the AI algorithms and systems and effective mechanisms to ensure accountability of the AI developers.[22] In 2023, the AI commission draft was formally presented by the Brazilian Senate's President as Bill No. 2,338/2023

The current version of the Brazilian AI Act (Bill No. 2,338/2023) establishes specific principles for developing, deploying, and using AI systems in Brazil, including sustainable development and well-being; human participation in the AI lifecycle and effective human oversight of AI; transparency, explainability, and audibility; reliability and robustness of AI systems; traceability of decisions; and prevention, precaution and mitigation of systemic risks derived from intentional or unintentional uses and unforeseen effects of AI systems.

Bill No. 2338/2023 is also centred on three main pillars: safeguarding the rights of the people affected by AI systems, conducting risk-level classification of AI systems, and ex-ante governance measures for organisations involved in designing, developing, deploying, and using such AI systems. It provides for establishing a new regulatory body to enforce the law (Article 18) and takes a risk-based approach by categorising AI systems into different categories. It also introduces a protective system of civil liability for providers or operators of artificial intelligence systems (Chapter V, Article 27,28) and a reporting obligation for significant security incidents (Chapter VII, Article 31). It also requires the AI system organisations to conduct a preliminary algorithmic impact assessment to classify the degree of risk (Chapter III, Article 13) of AI systems as 'Excessive' or 'High'. The Brazilian government has also put forth a National Strategy for Artificial Intelligence intending to stimulate the development and adoption of AI systems to promote scientific development, helping evolve policy designs to solve socio-economic problems, and working for the greater vision of the country.[23]

Therefore, analysis of pathways taken by some of the critical jurisdictions on regulating AI shows that the ounce of tackling concerns about AI is overtly on AI developers. This paper will try to address the gap through discussion at the ecosystem level. This analysis also showcases

[22] Soares, I., Kujawski, F.F., (2023, March). Brazil: AI landscape and what to expect from the upcoming legislation. OneTrust Data Guidance. Retrieved August 15, from https://www.dataguidance.com/opinion/brazil-ai-landscape-and-what-expect-upcoming

[23] *Brazilian National AI Strategy.* The OECD Artificial Intelligence Policy Observatory - OECD.AI. https://oecd.ai/en/dashboards/policy-initiatives/%2F%2Faipo.oecd.org%2F2021-data-policyInitiatives-27104

that there is a lot of effort and literature on risk management of AI at the development level focusing on uni-stakeholders, i.e., AI developers.[24] However, these fall through the cracks as we leave other players undiscussed. Therefore, in the following chapter, we will discuss establishing an effective governance structure for AI involving multistakeholder, i.e., AI developers, AI deployers and impact population, where various principles will be mapped to different stakeholders towards making AI trustworthy and safe.

3. Principle-based Multi-Stakeholder Approach - An Ecosystem-Level Intervention

It is crucial to minimise the impact and harms of Artificial Intelligence to make it a success. As discussed in the previous chapter, countries across the globe are taking steps to regulate AI, such as the recent draft of Brazil's AI Bill, the EU's AI Bill, and the US National Institute of Standards and Technology's AI RMF, NITI Aayog's responsible AI principles. While these regulatory measures are trying to make AI systems trustworthy through risk management, there is less discussion on how we can tackle the adverse implications of AI artificial intelligence at the ecosystem level, involving other stakeholders like AI deployers and the impact population. Besides, in a rapidly changing landscape, regulatory interventions must withstand the test of time. This is the primary reason why draft regulations in development or in the process of becoming a law must be principle-based.[25]

Therefore, through this chapter, we suggest a principle-based multi-stakeholder approach where we discuss various principles across the AI lifecycle bucketed and mapped to respective stakeholders within the AI ecosystem.[26] Firstly, we will differentiate between harms and impacts emerging at different stages of the AI lifecycle. The objective of doing this is to develop a map of harms and impacts caused by different stakeholders at different stages of the AI lifecycle. In addition, the objective is to declutter and distribute the impact and harm caused by AI, which emerges at different stages so that appropriate steps can be taken. Followed by differentiating the harms and impact, to tackle the same, this chapter suggests principles to be followed by identified stakeholders at the different stages of the AI lifecycle. While there are various stakeholders within the AI ecosystem, this chapter covers the three key players, i.e., AI developers, AI deployers, and Impact Population. For the purpose of this chapter, three key stakeholders are defined as the following.[27]

---

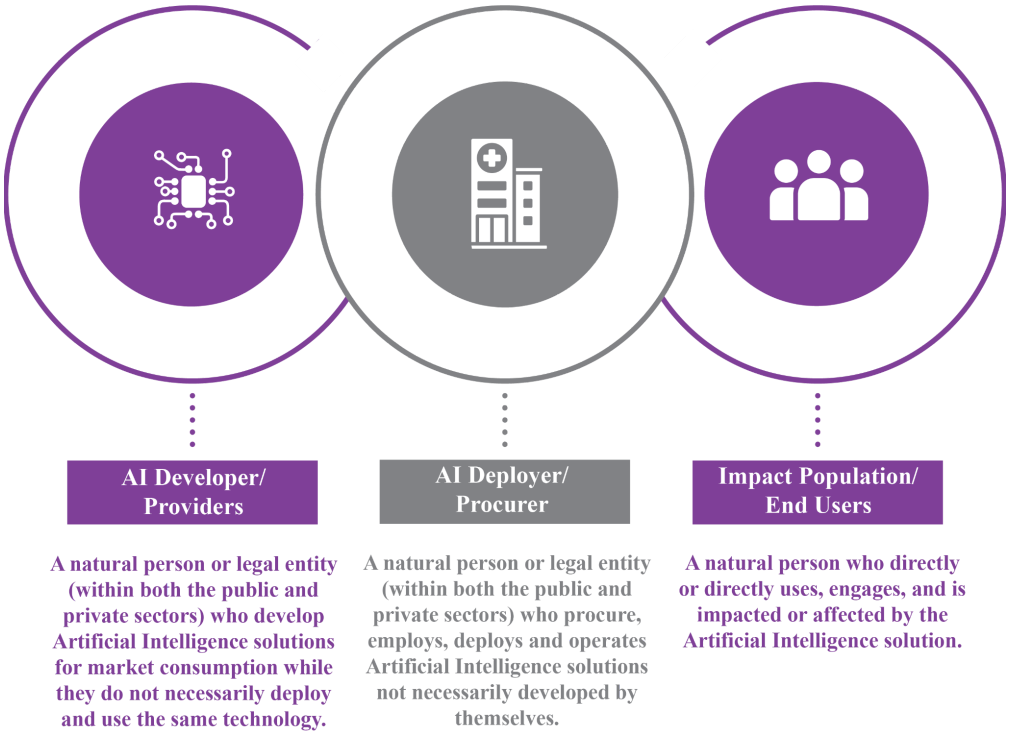
[24] Rogers, J. (2023, January 11). *Artificial intelligence risk & governance. AI & Analytics for Business*. Retrieved June 20, 2023, from https://aiab.wharton.upenn.edu/research/artificial-intelligence-risk-governance/

[25] Maithon, R. (2023, April 11). *India needs a principles-based approach to regulating AI*. Bharat Times. Retrieved June 20, 2023, from https://news.bharattimes.co.in/india-needs-a-principles-based-approach-to-regulating-ai/

[26] The principles should be understood in their cultural, linguistic, geographic, and organizational context, and some themes will be more relevant to a particular context and audience than others. For instance, the definition of transparency or explainability in Brazil may not be the same concept in the US.

[27] The AI developer and AI deployers are not watertight compartments, whereas there are instances where the AI provider/developer could also be an AI operator/user. At such conditions, the entity or natural person must follow the principles bucketed for AI developers and AI deployers at different stages of the AI lifecycle.

Figure 1: Stakeholders



The critical principles mapped for the above-discussed stakeholders in this chapter are advised by the frameworks developed by various governments, intergovernmental organisations, academia, civil society etc., in India and globally. Besides, the principles discussed in this chapter are the key universal and internationally recognised AI design and deployment principles embedded in various responsible AI frameworks across jurisdictions[28], especially India.[29]

3.1. Mapping Harms and Impact across the AI Lifecycle

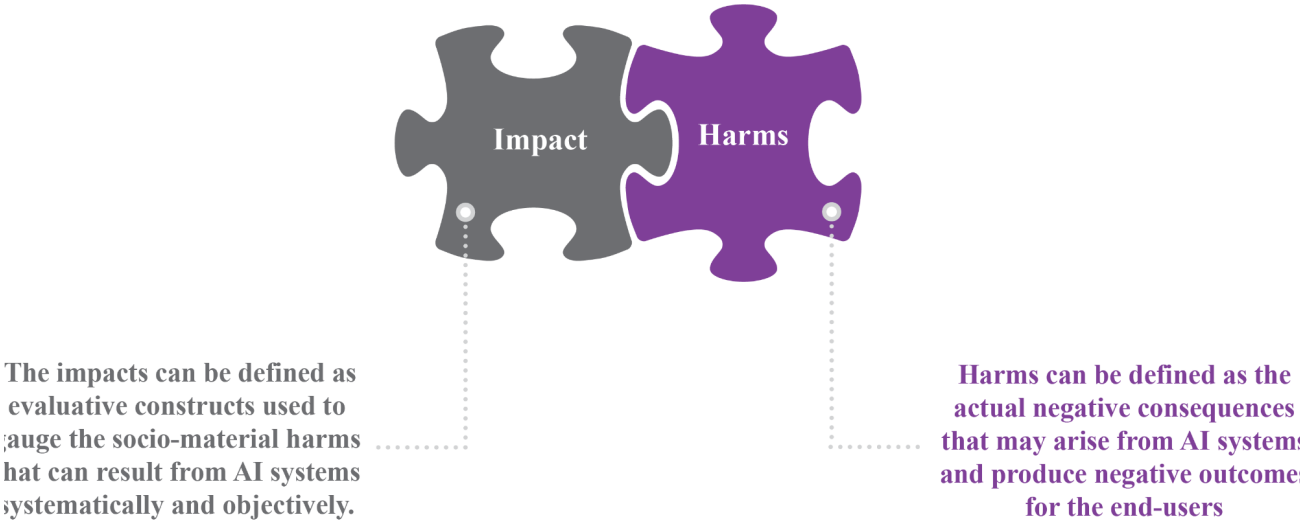
While we interchangeably use the terms such as Impacts and Harms, they are technically not identical. The impacts can be defined as evaluative constructs used to gauge the socio-material harms that can result from AI systems systematically and objectively.[30] These measurable outcomes allow us to understand the consequences of the interaction between AI technologies

[28] Shankar, V., & Casovan, A. (2022, May). *A framework to navigate the emerging regulatory landscape for AI.* The OECD Artificial Intelligence Policy Observatory - OECD.AI.
https://oecd.ai/en/wonk/emerging-regulatory-landscape-ai
[29] NITI Aayog. (2022, November). *RESPONSIBLE AI #AIFORALL Adopting the Framework: A Use Case Approach on Facial Recognition Technology.* | NITI Aayog.
https://www.niti.gov.in/sites/default/files/2022-11/Ai_for_All_2022_02112022_0.pdf
[30] Metcalf J, Moss E, Watkins E, Singh R, and Elish M. (2021, March). *Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts.* ACM Digital Library.
https://dl.acm.org/doi/pdf/10.1145/3442188.3445935

and individuals and society. For instance, the error rates of the AI solution, like the rate of inaccurate information, wrong predictions or disparate errors etc. Defining and measuring impacts allows us to understand the intended and unintended risks, benefits and harms that may arise when the procured AI deployers employ the AI solutions.

However, though the developed AI solutions are working as designed, adverse implications still crop out. This is where the other end of the puzzle, which is less discussed, comes into the picture, i.e., how AI deployers utilise the procured AI solutions for critical functions causing tangible and intangible harms.[31] For instance, as discussed above, the AI solutions might be producing an error or may be designed to capture some biased parameters to produce the suggested outcome; however, real-life harms of such outcomes only translate into action when AI deployers blindly use the same for making real-life decisions. Therefore, this shows that the distinction between harm and impact is rooted in the accountability and responsibility relationship among the stakeholders involved in the AI lifecycle, where both AI developers and AI deployers must follow some key principles to ensure adverse implications of AI solutions are tackled appropriately.[32] Besides, with the evolution of artificial intelligence into Generative AI solutions, real-life harms could also be caused by the impact population. For instance, when an AI solution produces baseless and misleading information, this starts a chain reaction of misinformation, which becomes a wild forest fire as unsuspecting impact populations start sharing the same misleading information within their own network.[33]

Figure 2: Impact Vs Harms



[31] Horowitz, A., & Selbst, A. (2022, June). *The fallacy of AI functionality.* ACM Digital Library. https://dl.acm.org/doi/fullHtml/10.1145/3531146.3533158

[32] Ryan, M. (2020, June 9). *Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications.* Discover Journals, Books & Case Studies | Emerald Insight. Retrieved June 20, 2023, from https://www.emerald.com/insight/content/doi/10.1108/JICES-12-2019-0138/full/html

[33] Discussed in detail below

While there are various forms of adverse implications emerging out of AI solutions, for the purpose of this section, we will be concentrating on five critical implications of AI solutions, i.e., exclusion, false predictions, copyright infringement, privacy infringement, and information disorder. The rationale behind choosing these critical implications is based on the cluster of cases reported on the same, which has been slightly higher.[34] The below illustration showcases how AI developers, AI deployers, and the impact population contribute towards orchestrating the five critical implications. In doing so, the illustration will also showcase at what stages within the AI lifecycle[35] (Refer to Box 3) "impact" and "harm" emerge and how AI developers, AI deployers, and impact populations are associated with the same. While various forms of impact and harm could potentially contribute towards causing the identified adverse implication, for the purpose of this paper, we have mapped some of the predominant causes based on our meta-analytic literature review. Besides, the mapped causes in the form of impact and harm don't exist in water-tight compartments, where some of them could apply universally and could be true for other adverse implications than the one they are mapped to.

Box 3: AI Lifecycle

Plan and Design: This initial stage of the AI life cycle entails early-stage planning and development of the AI systems by data scientists, domain experts and governance experts. The design sub-stage involves articulating the goals and objectives of the systems, stating the underlying assumptions, context and requirements in light of legal and regulatory requirements and ethical considerations, and exploring opportunities for building a prototype. Key players in this stage include C-suite executives, TEVV experts, product managers, compliance experts, auditors, organisational management, etc.

Collect and Process Data: Data stage deals with gathering, validating and cleaning the data and documenting the metadata and characteristics of the dataset. Key players in this stage include Data scientists, data/model/system engineers, AI designers etc.

Build and Use Model: During the model stage, the focus is on creating selection models/algorithms, their calibration, training and interpretation. Various models or algorithms are designed and developed that may be suitable for achieving the intended outcome. Key players in this stage include Modelers, Model Engineers, Data scientists, data/model/system engineers, domain experts, etc.

Verification and Validation: This phase involves executing and tuning models and running tests to assess performance on various factors and metrics. These evaluation metrics are defined based on

[34] European Commission. (2020, March). *The ethics of artificial intelligence: Issues and initiatives.* European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf; *Crime in India 2021 | National Crime Records Bureau*. (n.d.). राष्ट्रीय अपराध रिकॉर्ड ब्यूरो. Retrieved August 18, 2023, from https://ncrb.gov.in/en/Crime-in-India-2021

[35] Advised by OECD and NIST AI lifecycle, however, slightly improvised to fit the model suggested in this paper.

problems and the desired results; frequently used metrics include accuracy, precision, recall, and F1 score.[36] Based on the evaluation results, this stage may also involve developing multiple iterations after identifying the limitations in the previous model and making refinements. This may be done by increasing complexity, revisiting datasets to assess the representativeness of data, considering and evaluating more capable algorithms, and sharing research innovations for the growth of the AI discipline. Key players in this stage include Data Scientists, experts etc.

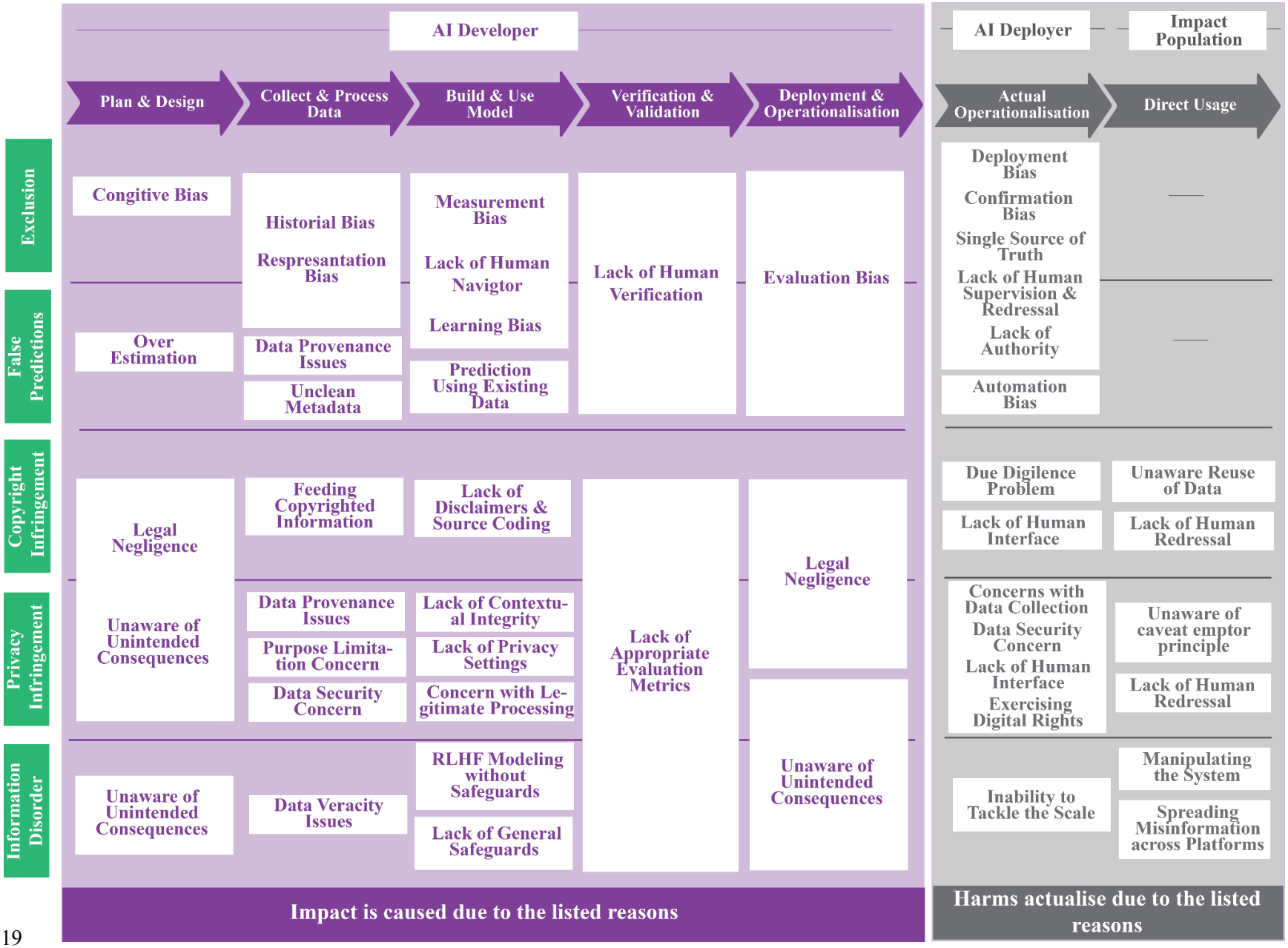
Deployment and Operationalisation: In this phase, the AI system is put into actual production and events such as piloting, compatibility assessment, regulatory compliance, organisational set-up and user experience evaluation are conducted. Followed by this, the AI system is actively used through operationalisation. Key players in this stage include Developers, System Engineers, Procurement experts etc.

Actual Operationalisation: Once the AI developers operationalise the AI solutions, in this stage, AI deployers procure the AI solution from the AI developer (if both are not the same entity). Post-procuring, AI deployers put AI solutions to actual operationalisation by incorporating them with their critical functions and using the outputs for decision-making, service delivery, etc. Key players in this stage include Hospitals, Schools, Law Enforcement, Employers, Banks etc.

Direct Usage: This phase is not universal depending upon the nature of the AI solution, where the impacted population uses the deployed AI solution as part of their day-to-day file. A key player in this stage is the end-user, i.e., individuals like us.

[36] Hodgson, J. (2022, November). *The 5 Stages of Machine Learning Validation. Towards Data Science*. Retrieved June 20, 2023, from https://towardsdatascience.com/the-5-stages-of-machine-learning-validation-162193f8e5db

Figure 3: Mapping Impact and Harms Across AI Lifecycle



3.1.1. Exclusion

One of the main concerns around Artificial Intelligence is producing biased outputs, which could ultimately lead to the exclusion of impact populations traditionally excluded in real life. For instance, alternate credit lending platforms, which use the data points like education attainment, employment history, social media data etc., for underwriting and pricing loans, have been reported to discriminate against individuals based on historical biases.[37] Where individuals who attended colleges/universities dedicated to historically vulnerable populations have been quoted a higher interest rate and were denied credit.[38]

India is a diverse and complex country with various historic dispositions like patriarchy, caste discrimination etc. Against this backdrop, one of the main concerns around AI is producing biased outputs. While AI solutions are not harmful, they replicate biases due to the biases present in its training data set and the way the algorithms are designed. Therefore, it is concerning when there is less clarity on the integrity, quality, and diversity of the data used for training the algorithms of these AI solutions. Besides, as these AI solutions are mostly predictive tools, they might unintendedly replicate the historic disposition causing discrimination and disproportionate harm to the vulnerable population. Moreover, the potential danger caused by AI is not just at the development stage but also at the deployment level, where harm could be caused by AI deployers who may abuse and misuse the technology, as discussed in the below table.

Table 1: Potential Causes for Exclusion

Stage	Cause	Description
AI Developers		
Plan & Design	Cognitive Bias	The human brain simply processes information by prioritising preferred outcomes due to cognitive biases.[39] However, in this scenario, cognitive biases could bring out exclusionary implications. For instance, hypothetically, if the individuals involved in the process of ideating an AI solution are exposed to patriarchal socialisation would think of the exclusion of women as an outcome.
Collect & Process Data	Historical Bias	Exclusion could happen even if the dataset is appropriately measured and sampled because of historical bias, where data carries biases as it is. This could also be attributed to one of the cleanliness

[37] Klein, A. (2022, March 9). *Reducing bias in AI-based financial services*. Brookings. Retrieved June 20, 2023, from https://www.brookings.edu/research/reducing-bias-in-ai-based-financial-services/
[38] Klein, A. (2022, March 8). *Credit denial in the age of AI.* Brookings. Retrieved June 20, 2023, from https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/
[39] Gillis, A. (2022, June 22). *What is cognitive bias?* SearchEnterpriseAI. Retrieved June 20, 2023, from https://www.techtarget.com/searchenterpriseai/definition/cognitive-bias

		issues, which impacts the quality of the data available. For instance, research shows that Natural Language Processing (NLP) models capture the biases associated towards women and vulnerable populations where specific keywords trigger gendered responses. Adding more information on the women and vulnerable populations wouldn't help in such conditions, as more data with historical biases would only add to the exclusionary outputs.
	Representation Bias	This is one of the critical concerns when the development sample of AI solutions is underrepresented with the data of a certain population group could ultimately lead towards the exclusion of the individuals who belong to that population.[40] Representation bias could creep in different forms where the target development sample lacks data of (a) the complete use population while using long a small representative data, (b) an underrepresented population within the use population like women, low-income households etc., (c) diverse ethnic groups within the underrepresented population.[41] For instance, the dataset can be counted to have gender diversity by having data on men, women, LGBTQ+ etc.; however, if the inferences of such data are not diverse, they might produce exclusionary outcomes. For instance, if the dataset has images of women from India only wearing ethnic wear, the inference derived from such a dataset would imply that almost all Indian women wear ethnic wear, excluding other women who don't wear ethnic clothes.
Build & Use Model	Measurement Bias	The label and parameters modelled within the system could bring out exclusion due to the measurement bias, where some proxy labels chosen to approximate some construct could bring our exclusion. For instance, it was reported by a research study that one of the school dropout predictive models had used race directly as a predictor and was also shown to have large racial disparities.[42]
	Lack of Human Navigator	Human navigators are the individuals or organisations who aid impact populations navigating the system, especially in critical sectors like healthcare.[43] However, the lack of modelling human

[40] Lee, N. T., Resnick, P., & Barton, G. (2022, March 8). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms.* Brookings. Retrieved June 20, 2023, from https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

[41] Mehrabi, N., Morstatter, F., & Saxena, N. (2022, January). *A Survey on Bias and Fairness in Machine Learning*. arXiv.org e-Print archive. Retrieved June 20, 2023, from https://arxiv.org/pdf/1908.09635.pdf

[42] Trinidad, J. (2022, March 24). *Spatial analysis of high school dropout: The role of race, poverty, and outliers in New York City*. Retrieved June 20, 2023, from https://doi.org/10.31235/osf.io/9nwst

[43] Natale-Pereira, A., Enard, K., Nevarez, L., & Jones, L. (2014, August). *The role of patient navigators in eliminating health disparities.* PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4121958/

		navigators as a feature within the AI solution could make it difficult for the unaware impact population to navigate the exclusion and seek mitigation.
	Learning Bias	When individuals prioritise one objective at the cost of damaging another as modelling choice brings out disparity and exclusion. For instance, research has shown that using differential privacy tools to enhance privacy ultimately reduces the influence that underrepresented populations have on data samples which ultimately causes exclusion.
Verification & Validation	Lack of human verification	If human verification and validation are not featured at this stage where interpretation of model output takes place, falling through the crack that had happened in the previous stages, as discussed above, might go unnoticed.[44] This implies that the AI solution would move to operationalisation, possibly producing disparity results leading to exclusion.
Deployment and Operationalisation	Evaluation Bias	As the AI solution's pilot, assessment, and monitoring commences at this stage, having less representative and historically biased datasets as a benchmark for evaluation could cause a fall through the crack where exclusionary outcomes will not go undetected.[45]
AI Deployers		
Actual Operationalisation	Deployment Bias	While the AI solution might be developed with all the precautions, deployment bias could bring exclusion. This happens when the AI deployers employ the AI solutions for a different purpose than what it was created for based on the human decision-makers' decision,[46] which is also called the framing trap.[47] For instance, while some prediction technologies in legal enforcement are developed for recidivism, it was noted that such AI solutions are used to determine the length of the sentence.[48]
	Confirmation	When the AI solution produces exclusionary outputs, the

[44] Xu, T. (2021, July 19). *AI makes decisions we don't understand. That's a problem.* Built In. Retrieved June 20, 2023, from https://builtin.com/artificial-intelligence/ai-right-explanation
[45] Reagan, M. (2021, April 2). *Understanding bias and fairness in AI systems*. Medium. Retrieved June 20, 2023, from https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3
[46] Alexisbcook. (2023, April 20). *Identifying bias in AI*. Kaggle: Your Machine Learning and Data Science Community. Retrieved June 20, 2023, from https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai
[47] Weerts, H. (2021, May). *An Introduction to Algorithmic Fairness*. arXiv.org e-Print archive. Retrieved June 20, 2023, from https://arxiv.org/pdf/2105.05595.pdf
[48] Hillman, N. (2019, January). *The use of artificial intelligence in gauging the risk of recidivism*. American Bar Association. Retrieved June 20, 2023, from https://www.americanbar.org/groups/judicial/publications/judges_journal/2019/winter/the-use-artificial-intelligence-gauging-risk-recidivism/

	Bias	confirmation bias of the AI deployers, i.e., confirming their existing belief, might blind them from noticing the error, causing exclusion. For instance, it was reported that one of the school dropout predictive models had used race directly as a predictor and was also shown to have large racial disparities. However, exclusion happened when this model was used in schools to decide which student is deemed to study maths and science, majorly, it was black students in the United States who had to face the brunt.[49]
	Single Source of Truth	When users consider AI solution-based outputs as a single source of truth without any alternative could bring out adverse implications, including exclusion. Besides, using outputs of AI solutions as a single source of truth can also cause a fall through the cracks due to a lack of cross-checking mechanisms, increasing the chances of false negatives and false positives. For instance, the AI tools used by federal agencies for predicting recidivism for individuals have been reported to bring out the disparity in prediction. The tools have brought out false positives in terms of overpredicting the risk of recidivism amongst vulnerable groups and false negatives amongst groups that are not vulnerable.[50]
	Lack of Human Supervision & Redressal	When outputs produced by the AI solutions are blindly incorporated without human supervision, it will make misidentification go unnoticed, which might cause exclusion. Besides, the lack of a human-based grievance redressal mechanism to report exclusionary problems could impact (a) the impact population in voicing their concerns and (b) the feedback loop of AI solutions, where the problem might go unnoticed.
	Lack of Authority	It was reported that a patient was denied pain medication because hospital software confused her medical history with her dog's. Though she tried to rectify it, doctors were afraid to override the systems.[51] Here it is also about the freedom of humans to take decisions of their own though AI solutions have failed. If the system doesn't provide protection and incentive, players will not take hard decisions, as they ultimately want to save themselves from unwanted consequences.

[49] Herold, B. (2022, April 14). *Why schools need to talk about racial bias in AI-powered technologies.* Education Week. Retrieved June 20, 2023, from https://www.edweek.org/leadership/why-schools-need-to-talk-about-racial-bias-in-ai-powered-technologies/2022/04

[50] Dressel, J., & Farid, H. (2018). *The accuracy, fairness, and limits of predicting recidivism*. Science Advances, 4(1). https://doi.org/10.1126/sciadv.aao5580

[51] Szalavitz, M. (2021, August 11). *The pain was unbearable. so why did doctors turn her away?*. Wired. Retrieved June 20, 2023, from https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/

3.1.2. False Predictions

Using an AI-based predictive tool can replicate bias due to the biases in its training set. For instance, AI technologies used for law enforcement purposes have been reported to bring out historical biases where for instance, systems have primarily assigned police parol to the neighbourhoods where discriminated populations reside. The incorrect crime predictions also feed into the system, creating a vicious cycle.[52] Similarly, the utilisation of AI in hiring tools used by companies and recruitment firms has been observed to increasingly discriminate against women. For instance, a company using AI solutions to hire a candidate for a particular role based on human-assigned ratings is reported to predict women as less suitable candidates than men, though the work profiles and qualifications of female candidates were at par with the male candidates. This false prediction scenario may be fed through historical bias against data recording the career growth trajectories of women across corporate settings.[53]

As discussed in Section 3.1.1 in the Indian context, the presence of the historically biased disposition against certain groups could aggravate adverse implications of the AI systems, like false predictions. While false predictions are one half of the story creating impact, the second half is when the AI deployers use those false predictions daily for determining eligibility, profiling etc., causing entry barriers, discrimination etc. There are similarities in causes discussed in Table 1 that also contribute towards causing false prediction at different stages of the AI lifecycle. However, below table 2 discusses some specific causes that predominantly led to false predictions.

Table 2: Potential Causes for False Predictions

Stage	Cause	Description
AI Developers		
Plan & Design	Over Estimation	As a human tendency, we borrow innovations and ideas from different scenarios and streams into our work field subject to those innovations' success rate. However, in this process, we might overestimate the capacity of such innovation and not consider the incompatibility of the same within specific sectors. For instance, while AI-based predictive technologies are extensively used in

[52] Sachoulidou, A. (2023, February 22). *Going beyond the "common suspects": To be presumed innocent in the era of algorithms, big data and artificial intelligence - artificial intelligence and law*. SpringerLink. Retrieved June 20, 2023, from https://link.springer.com/article/10.1007/s10506-023-09347-w

[53] Goodman, R. (2023, February 27). *Why Amazon's automated hiring tool discriminated against women* | ACLU. American Civil Liberties Union. Retrieved June 20, 2023, from https://www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against

		weather predictions and meteorology, it is not necessarily true that similar technology would work completely prejudice-free when used for predicting recidivism.[54]
Collect & Process Data	Data Provenance Issue	Lack of considering the genesis of the data and resulting dataset could lead to false predictions bringing out and amplifying discrimination, biases etc., as AI-based predictive tools predominantly produce predictions based on the data fed into the system. For instance, when the AI-based prediction technology for law enforcement is fed with police and crime datasets, it is important to be aware of the genesis of this data, as research proves that police and crime datasets often carry historical prejudice which may target racial or religious minorities.[55] Unaware of data provenance questions the AI solution's robustness in tackling the unintended consequences.
	Unclean Metadata	AI technologies used for law enforcement purposes have been reported to bring out historical biases where, for instance, systems have mostly assigned police parol to the neighbourhoods where vulnerable populations reside. However, when these incorrect crime predictions also feed into the system as metadata, which creates a vicious cycle.[56]

[54] Rieland, R. (2018, March 5). *Artificial intelligence is now used to predict crime. But is it biased?* Smithsonian Magazine. Retrieved June 20, 2023, from https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/

[55] Verma, P. (2022, July 15). *The never-ending quest to predict crime using AI*. The Washington Post. Retrieved June 20, 2023, from https://www.washingtonpost.com/technology/2022/07/15/predictive-policing-algorithms-fail/

[56] Sachoulidou, A. (2023). Going beyond the "common suspects": To be presumed innocent in the era of algorithms, big data and artificial intelligence. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-023-09347-w

Build & Use Model	Prediction Using Existing Data	The premise of generative AI systems, i.e., current data about the world is enough to understand the world in future, is concerning where it leads to errors. These AI Solutions will be prone to reproduce the same mistakes and patterns in future, causing real-life implications for humans. For instance, in the case of Facial Recognition Technologies, using past datasets to predict future outcomes is concerning, such as potentially resulting in the over-policing of certain communities. This may also impact the allocation of resources to law enforcement agencies (LEAs).[57] Besides, if any individual has exercised their right to be forgotten in the recent past, this information wouldn't be captured by the system, which adds to the inconsistency of the nature of using past data.[58]
AI Deployers		
Actual Operationalisation	Automation Bias	AI Deployers must make AI Users[59] aware of the ability of an AI model to churn out false predictions and may treat the computational results of an AI model as accurate, which may lead to them blindly relying on the results of such an AI model. This is harmful on account of the AI user succumbing to automation bias, especially in cases where the AI user is operating high-risk AI systems causing catastrophic impact.

3.1.3. Copyright Infringement

A problem that could have legal repercussions enforced through monetary claims is that of an AI system infringing intellectual property rights. Since some of the AI innovations, like generative AI technologies, are trained on a wide variety of language models, which include data such as books, articles, and journals, the output to be produced might have the risk of infringing on copyright texts leading to a violation of one's intellectual property rights. For instance, the outcome of generative AI solutions doesn't necessarily show original sources that it has used for deriving an answer; this might cause an infringement of intellectual property. Besides, there is less clarity on the compensation mechanism for using the original work produced through human creativity. Some of the causes for copyright infringement are as follows.

[57] Gentzel, M. (2021). Biased face recognition technology used by government: A problem for liberal democracy. *Philosophy & Technology*, *34*(4), 1639-1663. https://doi.org/10.1007/s13347-021-00478-z

[58] Shekar, K., & Rizvi, K. (2023, February 9). *Regulation of generative AI like ChatGPT and bard mustn't hinder their growth*. Moneycontrol. Retrieved June 20, 2023, from https://www.moneycontrol.com/news/opinion/regulation-artificial-intelligence-chatgpt-bard-hinder-growth-10039461.html

[59] AI users may be employees of the AI deployers who are using AI systems. For instance, a government body using an AI system is the deployer and the AI user is their employee who is using the AI system.

Table 3: Potential Causes for Copyright Infringement

Stage	Cause	Description
AI Developers		
Plan & Design/ Deployment & Operationalisation	Legal Negligence	While AI provides state-of-the-art solutions, this doesn't mean the existing regulations will not apply to AI technologies and their developers. Individuals and businesses still enjoy Intellectual Protection rights (IPR) protections in India under the Patents Act 1970, Trademarks Act 1999 and the Copyright Act 1957.[60] Through IPR, individuals get attribution for their work and flexibility in framing the buyer contract in the physical world. These legislations do apply to AI solutions, though there might be less clarity; it is the responsibility of the AI developers to ensure the solution developed doesn't infringe on existing intellectual property laws and copyrights.
	Unaware of Unintended Consequence	When AI developers would consider certain principles to make AI solutions responsible and ethical, pragmatically when implemented into actionable strategies, some key principles conflict with each other, causing unintended consequences. For instance, while we suggest data quality through more representative and diverse datasets, the unintended repercussion would be infringing intellectual property rights as the diverse dataset might have copyrighted content. [61]
Collect & Process Data	Feeding Copyrighted Information	As direct as it can get if the dataset used for modelling an AI solution as copyrighted information without a contract or formal intimation would cause copyright infringement. For instance, the AI industry has witnessed some interesting developments in the past six months with the release of large language models (LLMs), such as Stable Diffusion, GPT-3, and DALL-E. However, language models do have books, articles, and journals. Therefore, the output to be produced might have the risk of infringing on copyright texts

[60] Rastogi, V., & Bhardwaj, N. (2023, March 23). *Intellectual property rights in India: Laws and Procedures*. India Briefing. Retrieved June 20, 2023, from https://www.india-briefing.com/news/intellectual-property-rights-india-laws-procedures-registration-14312.html/

[61] Adams, S. (2020, January 16). *Comments on the USPTO's Intellectual Property Protection for Artificial Intelligence Innovation*. Center for Democracy and Technology. Retrieved June 20, 2023, from https://cdt.org/insights/comments-on-the-usptos-intellectual-property-protection-for-artificial-intelligence-innovation/

		leading to a violation of one's intellectual property rights.[62] The only exception here could be AI solutions used for activities carried out by research organisations and institutions, journalists, museums, archives and libraries, which do not necessarily constitute a copyright infringement.[63]
Build & Use Model	Lack of Disclaimers & Source Coding	While it is completely fine to use copyrighted content, however, if the AI solution could give out copyrighted content as a response, it is important to model in a display of disclaimers and source information. Otherwise, it would open the possibility for copyright infringement by the users.
Verification & Validation	Inappropriate Evaluation Metrics	Falling through the cracks does happen in previous stages due to the above-discussed reasons. However, if metrics used for verifying and validating the outputs don't have parameters to check for copyright infringement and misuses to emerge at the actual operationalisation and usage stage could let this unintended consequence go unnoticed.
AI Deployers		
Actual Operationalisation	Due Diligence Problem	When an AI deployer doesn't do her due diligence while procuring the AI solutions in terms of checking (a) if the technology is trained using copyright information, (b) modelled with some disclaimers etc., could contribute toward copy infringement when AI solutions are actually operationalised.
	Lack of human interface	If the AI deployer constantly receives feedback from the impact population that there is a copyright infringement, this information has to be funnelled to AI developers. However, if the AI deployers are not provided with a human interface by the AI developer, this information may not reach AI developers on time and appropriately.
Impact Population		
Direct Usage	Unaware Reuse of Data	As discussed above, when the display of citations, disclaimers or source information is not modelled within the AI solution, unaware users might reuse such information without providing attributions, causing copyright infringement

[62] Appel, G., Neelbauer, J., & Schweidel, D. A. (2023, April 7). *Generative AI Has an Intellectual Property Problem*. Harvard Business Review. Retrieved June 20, 2023, from https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

[63] Guadamuz, A. (2017, October). *Artificial intelligence and copyright*. WIPO - World Intellectual Property Organization. Retrieved June 20, 2023, from https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html

	Lack of Human Redressal	Lack of humans in the loop for grievance redressal at the AI deployer end could prevent (a) the impact population from safeguarding their IPR appropriately and (b) the feedback loop of AI solutions, where the problem might go unnoticed.

3.1.4. Privacy Infringement

The AI solutions are trained using a massive amount of data to provide a human-like response. However, there is less clarity on the amount of personal information used by the AI developers as part of the training set and data protection measures taken to secure the same. Besides, there are also data security concerns where it is likely that AI solutions could expose confidential information of individuals causing identity theft, fraud etc. While the dataset has both personal and non-personal data of individuals, however below causations led to privacy infringement.

Table 4: Potential Causes for Privacy Predictions

Stage	**Cause**	**Description**
AI Developers		
Plan & Design/ Deployment & Operationalisation	Legal negligence	It is suggested that the dataset used for training AI solutions includes personal and non-personal data. While personal data warrants protection and security where India's DPDPA 2023 will apply to AI developers, non-personal data unlocks value benefiting individuals, businesses, and communities.
	Unaware of Unintended Consequences	While a massive amount of data is used to enhance AI solutions, one of the unintended consequences would be that AI solutions expose confidential information of individuals causing identity theft, fraud, etc. For instance, recently, it was reported that the Snap AI chatbot had revealed the location of individuals while it had been programmed to say that it doesn't hold such personal information.[64]
Collect & Process Data	Data Provenance Issues	Where the data is sourced from brings out privacy concerns, especially when AI developers aggregate data from multiple public sources[65]. For instance, it is suggested that the dataset used for training Generative AI has billions of words and images scraped

[64] Mahapatra, T. (2023, April 24). *Snapchat's My AI chatbot faces criticism over user privacy and accuracy concerns*. Hindustan Times. Retrieved June 20, 2023, from https://www.hindustantimes.com/technology/snapchats-my-ai-chatbot-faces-criticism-over-user-privacy-and-accuracy-concerns-101682323903867.html

[65] Rafter, D. (2021, January 18). *How data brokers find and sell your personal info*. Norton US. Retrieved June 20, 2023, from https://us.norton.com/blog/privacy/how-data-brokers-find-and-sell-your-personal-info

		from publicly available information from places like websites, articles, blog posts, etc. When the General Court of the European Union held that personal views or opinions of individuals (technically information which is available in articles, blog posts etc.) couldn't be presumed to be personal information;[66] On the other hand, India under its recently enacted DPDPA 2023 takes different course, where the obligations of the bill will not apply to the personal data which has been made or caused to be made available public by the Data Principal themselves to whom such personal data relates. Other publicly available personal information, i.e., not made public by the data principal, can only be processed after obtaining consent from data principles at the commencement of its processing. However, privacy concerns still remained as the publicly available information could also lead or reveal some personal information.
	Purpose Limitation	While data is being used for training the AI solutions, however, if the AI developers don't follow purpose limitation and use training data for purposes beyond what it was aggregated for could contribute to privacy infringement.
	Data Security Concerns	Less clarity on the safeguards equipped at the data storage level brings out data security concerns.
Build & Use Model	Lack of Contextual Integrity	The modelling of an AI solution using information out of context could lead to infringing contextual integrity and privacy, i.e., breaching social relations, which are controlled by the information flow, and cause inappropriateness, i.e., exposing inappropriate information about individuals in a particular social and political setting.
	Lack of Privacy Settings	When privacy settings without deceptive design[67] are not modelled within AI solutions, it doesn't provide a choice to the individuals in terms of protecting their privacy. For instance, using Reinforcement Learning with Human Feedback (RLHF), where every prompt and interaction on the platform is recorded without providing an option for the opt-out to the individual or options like incognito/private tabs.

[66] Quathem, K. V. (2023, April 28). *EU general court clarifies when Pseudonymized data is considered personal data*. Covington. Retrieved June 20, 2023, from https://www.insideprivacy.com/eu-data-protection/eu-general-court-clarifies-when-pseudonymized-data-is-considered-personal-data/

[67] Jarovsky, L. (2022, June 7). *Deceptive patterns in data protection (and what UX designers can do about them)*. Medium. Retrieved June 20, 2023, from https://uxdesign.cc/dark-patterns-in-data-protection-13fdb0c5231d

	Concern with Legitimate Processing	Consent is the bedrock on which not only the EU-GDPR, even India's DPDPA 2023 sits,[68] where it mandates that personal data shall be collected and processed only after obtaining consent from data principles at the commencement of its processing.[69] However, the consent-based approach doesn't consider the complex data processing mechanism for new AI evolution like Generative AI. Besides, this could also cause a fall through the cracks as the determining legitimacy of consent is nebulous in Generative AI operations. However, there must be different means through which individuals' agency over their personal data used for training the algorithms across the data lifecycle is protected and ensured.[70]
Verification and Validation	Lack of Appropriate Evaluation Metrics	The lack of necessary evaluation measures to secure the utilisation of personal data keeping purpose limitations, contextual integrity, and appropriateness intact, could contribute towards privacy infringement and cause legal obligation.
AI Deployers		
Actual Operationalisation	Concerns with Data Collection	Some AI solutions could lead the AI deployers to collect data beyond the purpose for which the solution was developed, causing privacy concerns. For instance, the facial recognition systems installed in the streets and other public spaces for tracking crime also bring in the visuals of every individual using these public spaces who are not part of any illegitimate activities. While it is an essential measure for tackling crime, it could disproportionately hamper the privacy of individuals.[71]
	Data Security Concerns	Easy access to coding tools as part of the information generated by generative AI solutions without safeguards and restrictions could make it easier for cyber attackers to hack, even for non-tech-savvy individuals who lack technical skills.
	Lack of human interface	Similar to copyright infringement, it is important for AI deployers to funnel feedback from the impact population that there is a privacy infringement to AI developers. However, if the AI deployers are not

[68] Shekar, K. (2023, August 4). *Comparative Analysis of India's Digital Personal Data Protection Bill, 2022 and 2023*. The Dialogue. Retrieved August 16, 2023, from https://thedialogue.co/wp-content/uploads/2023/08/Designed-finalDPDPB-2023_Analysis-Paper.pdf
[69] Ibid
[70] Sahiba, J., & Shekar, K. (2023, April 7). *Italy's ChatGPT block: Privacy protection concerns stalk OpenAI and other generative AI developers*. Moneycontrol. Retrieved June 20, 2023, from https://www.moneycontrol.com/news/opinion/italys-chatgpt-block-privacy-protection-concerns-stalk-openai-and-other-generative-ai-developers-10378471.html
[71] Raposo, V. L. (2022). The use of facial recognition technology by law enforcement in Europe: A non-orwellian draft proposal. *European Journal on Criminal Policy and Research*. https://doi.org/10.1007/s10610-022-09512-y

		provided with a human interface by the AI developer, this information may not reach AI developers on time and in an appropriate manner.
	Difficulty in Exercising Digital Rights	While data protection legislations across jurisdictions vest various digital rights on individuals, however, there is less clarity in terms of the applicability of such rights in the context of AI technologies. For instance, there is less clarity regarding how individuals can exercise their right to erasure or correction in the context of Generative AI solutions.
Impact Population		
Direct Usage	Unaware of caveat emptor principle	When individuals are unaware that the AI solution is premised on the caveat emptor principle[72], it could cause privacy infringement. For instance, various Generative AI solutions insist individuals be aware and not share sensitive information during their interaction with Bots,[73] however, if they are unaware of this fact could cause privacy concerns. For instance, recently, Samsung spotted a generative AI solution leaking its confidential information as one of its unaware employees accidentally disclosed sensitive information while interacting with a generative AI solution.[74]
	Lack of Human Redressal	Lack of humans in the loop for grievance redressal at the AI deployers end could prevent (a) the population from safeguarding their privacy by appropriately exercising their digital rights and (b) the feedback loop of AI solutions, where the problem might go unnoticed.

3.1.5. Information Disorder

While quick and easy access to information is useful, lack of understanding about the accuracy of the information received through AI solutions, especially with consumer-facing AI solutions like generative AI, is problematic – especially for high stake information like election-related information, health-related information etc. – given that disinformation and misinformation spread faster than the truth. Therefore, below are some potential causes emerging at different stages of the AI lifecycle contributing to the causation of information disorder.

[72] Corporate Finance Institute. (2020, June 3). *Caveat Emptor (Buyer beware)*. Retrieved June 20, 2023, from https://corporatefinanceinstitute.com/resources/risk-management/caveat-emptor-buyer-beware/
[73] Metz, R. (2023, April 25). *OpenAI Offers New Privacy Options for ChatGPT*. BloomberG. Retrieved June 20, 2023, from https://www.bloomberg.com/news/articles/2023-04-25/openai-offers-new-privacy-options-for-chatgpt
[74] Sharma, D. (2023, May 2). *Samsung restricts use of generative AI tools after employees leak sensitive data using ChatGPT*. India Today. Retrieved June 20, 2023, from https://www.indiatoday.in/technology/news/story/samsung-restricts-use-of-generative-ai-tools-after-employees-leak-sensitive-data-using-chatgpt-2367448-2023-05-02

Table 5: Potential Causes for Information Disorder

Stage	Cause	Description
AI Developers		
Plan & Design/ Deployment & Operationalisation	Unaware of Unintended Consequences	While recent evolution of AI solutions like generative AI is developed to assist humans in various sectors. However, an unintended consequence of these technologies manifests in the form of the capability to generate false records or "deep fakes," imitate individuals, and manipulate information to create politically-altered content. The impact caused by AI-generated deep fake videos and synthetic media could blur the lines between false and truth.[75]
Collect & Process Data	Data Veracity Issues	As simple as it can get if the data fed into the system has issues with veracity, it would definitely regulate the outcomes.
Build & Use Model	RLHF Modeling without Safeguards	Modeling an AI solution to use Reinforcement Learning with Human Feedback (RLHF) without proper checks and balances would get easier for individuals to produce baseless and misleading information to distort the feedback system, causing disinformation.
	Lack of General Safeguards	Lack of technical measures modelled to tackle information disorder faster would exacerbate the issue. For instance, it would be difficult for individuals to distinguish between truth and false if there are no technical measures to differentiate, as AI-based responses are foolproof without typos or grammatical errors.
Validation & Verification	Lack of Appropriate Evaluation Metrics	The lack of necessary evaluation measures which help detect (a) if the system could be potentially tricked to cause information disorder (b) the veracity of outcomes would let the unintended consequence pass through unnoticed.
AI Deployers		
Actual Operationalisation	Inability to Tackle the Scale	When AI deployers deploy AI solutions which can cause information disorders, it puts them at the spot where they will face a scale problem. As AI deployers directly interface with individuals, tackling disinformation and misinformation would be difficult, especially if they serve many individuals. An analogy could be drawn to social media platforms, wherewith mounting pressure on

[75] Bateman, J. (2020, July 8). *Deepfakes and synthetic media in the financial system: Assessing threat scenarios.* Carnegie Endowment for International Peace. Retrieved June 20, 2023, from https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237

		the platforms from the government and individuals to tackle narrative harms, they resort to hard content moderation yet face the problem of scale, causing a fall through the cracks.[76]
Impact Population		
Direct Usage	Manipulating the System	As discussed above, if the AI solution is modelled with RLHF without safeguards individuals could feed baseless and misleading information to deceive, causing 'disinformation'.
	Spreading Misinformation across Platforms	When an AI solution produces baseless and misleading information, this starts a chain reaction of misinformation, which becomes a wild forest fire as unsuspecting impact populations start sharing the same misleading information within their network across platforms.[77]

3.2. Mapping Principles for Stakeholders Across the AI Lifecycle

The various stakeholders within the AI ecosystem contribute in their capacities towards operationalising adverse implications, as discussed in Section 3.1. Therefore, to make the AI ecosystem safe, inclusive, and useful, it is essential to have a concerted effort at the ecosystem level where various stakeholders follow different principles at different stages of the AI lifecycle.

Various governments, intergovernmental organisations, academia, and civil society have developed critical principles for developing and deploying AI. Several regions and countries, including the EU, the US, Brazil, India, etc., have also started developing their national AI strategies to present a vision for AI development and governance of AI (refer to Annexure 1). These frameworks propose principles to ensure that AI technologies are developed and used ethically in a rights-respecting and beneficial manner. The Recommendation on Artificial Intelligence (AI) adopted by OECD in 2019 is the first intergovernmental standard on AI to promote the responsible stewardship of trustworthy AI. The Recommendation sets forth a framework for responsible AI for all stakeholders involved in developing, deploying, and using AI and recommendations for national policies and international cooperation. Similarly, companies such as Microsoft, IBM, Google and SAS have also developed their own principles for responsible AI.

While these frameworks discuss principles for the responsible development of AI solutions, if the users misuse it and the impact population is unaware, it falls through the cracks. Therefore, we need a principle-based intervention that maps responsibilities and principles for various

[76] Douek, E. (2021, June 2). *More content moderation is not always better*. WIRED. Retrieved June 20, 2023, from https://www.wired.com/story/more-content-moderation-not-always-better/
[77] Discussed in detail below

stakeholders (refer to Figure 1) within the AI ecosystem. While in the previous section, we did an implication-by-implication causation analysis, in this section, we will discuss the principles at the consolidated level mapped to various stakeholders to be followed at different stages, as illustrated below.

The below-mapped principles are advised by NITI Aayog's National Strategy for Artificial Intelligence[78], OECD AI principles[79], G20 AI Principles[80], Australia's AI Intelligence Ethics Framework and AI Ethics Principles[81], EU Ethics Guidelines for Trustworthy AI[82], EU-US TTC Joint Roadmap for Trustworthy AI and Risk Management[83], NIST's AI Risk Management Framework[84], Germany, Artificial Intelligence Strategy 2018[85], Singapore National AI Strategy 2019[86], USA's National Artificial Intelligence Research and Development Strategic Plan 2023[87], France's AI for Humanity 2017[88], European Union's Artificial Intelligence for Europe 2018[89], European Union's The Artificial Intelligence Act, 2023[90], United Kingdom's A Pro-Innovation Approach to AI Regulation 2023[91], Japan's Social Principles of Human-Centric AI 2019[92], The

[78] NITI Aayog. (June 2018). *National Strategy for Artificial Intelligence #AIforAll. (2018)*. Niti Aayog. https://niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf
[79] OECD. (2019). *Recommendation of the Council on Artificial Intelligence.* OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
[80] G20. (2019). *G20 AI Principles.* https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf
[81] Australian Government. (2019). *Australia's AI Ethics Principles*.. https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles
[82] European Commission. (2019). *Ethics Guidelines for Trustworthy AI*. High-Level Expert Group on Artificial Intelligence, https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419
[83] European Commission, (2022) *TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management*. https://ec.europa.eu/newsroom/dae/redirection/document/92123
[84] National Institute of Standards and Technology. (2023, January). *Artificial Intelligence Risk Management Framework.* NIST Technical Series Publications. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
[85] German Federal Government. (2020, December). *National AI Strategy.* KI Strategie. https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf
[86] Smart Nation Digital Government Office. (2019, November). *National Artificial Intelligence Strategy.* Smart Nation Singapore. https://www.smartnation.gov.sg/files/publications/national-ai-strategy.pdf
[87] National Science and Technology Council. (2023, May). *The National Artificial Intelligence R&D Strategic Plan 2023 Update.* The White House. https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf
[88] Villani, C. (2018, March). *For A Meaningful Artificial Intelligence: French Strategy.* AI for humanity. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf
[89] European Commission. (2018, April). *Artificial Intelligence for Europe.* EUR-Lex — Access to European Union law. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:237:FIN
[90] European Commission. (2021, September). *The Artificial Intelligence Act*. The AI Act. https://artificialintelligenceact.eu/the-act/
[91] Department for Science, Innovation and Technology. (2023, March). *A pro-innovation approach to AI regulation*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf
[92] The Government of Japan. (2019, February). *Social Principles of Human-Centric AI.* https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf

Global Partnership on Artificial Intelligence's AI principles[93], United Nations' Principles for Ethical Use of AI in UN 2022[94], UNESCO Ethics of Artificial Intelligence[95], and other private sector frameworks.[96] In addition, some of the principles mapped are suggested through research, especially ones mapped to AI deployers and impact populations.

Collectively, we believe the mapped principles (refer to Figure 4) will enhance the digital trust of the impact population such that they feel at ease and safe using AI solutions.

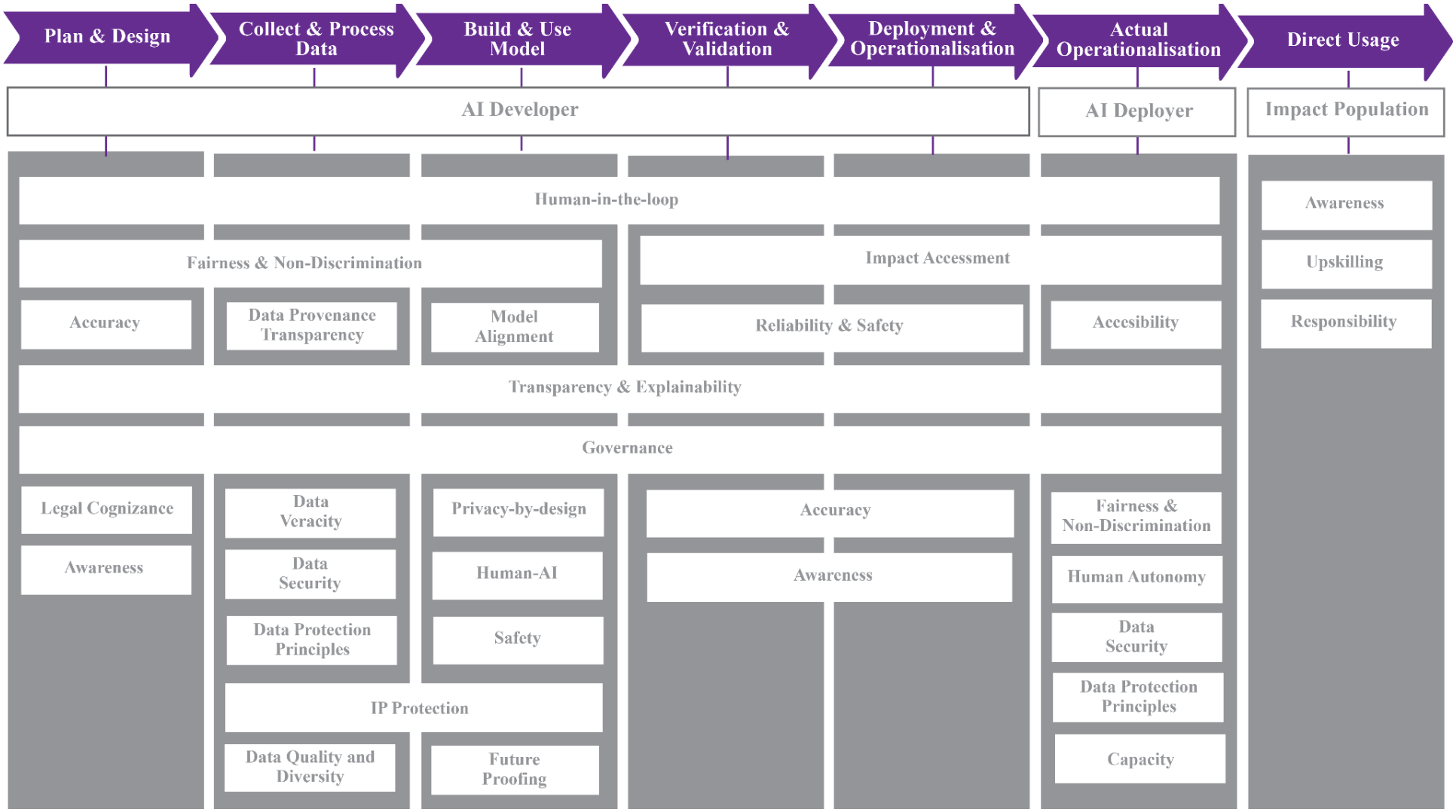
[93] The Global Partnership on Artificial Intelligence's AI principles. (2020, June). *Global Partnership on Artificial Intelligence* - GPAI. https://gpai.ai/about/

[94] UN System Chief Executives Board for Coordination. (2022, September). *Principles for the Ethical Use of Artificial Intelligence in the United Nations System*. United Nations - CEB. https://unsceb.org/sites/default/files/2022-09/Principles%20for%20the%20Ethical%20Use%20of%20AI%20in%20the%20UN%20System_1.pdf

[95] UNESCO. (2023, April 20). *UNESCO adopts first global standard on the ethics of artificial intelligence*. https://www.unesco.org/en/articles/unesco-adopts-first-global-standard-ethics-artificial-intelligence

[96] Schiff J, D., Borenstein, J., & Laas, K. (2021, April 12). *AI ethics in the public, private, and NGO sectors: A review of a global document collection*. Montreal AI Ethics Institute. https://montrealethics.ai/ai-ethics-in-the-public-private-and-ngo-sectors-a-review-of-a-global-document-collection/

Figure 4: Mapping Principles for Stakeholders Across the AI Lifecycle



3.3. Operationalisation of Principles by Various Stakeholders

To ensure the realisation of responsible AI, it is crucial to translate the principles discussed in the above chapter into tangible requirements. While there is a broad consensus regarding the core principles of responsible/ethical AI, there remains a lack of consensus on applying and implementing these principles within organisations effectively. The results of a recent survey conducted by IBM[97] shed light on this issue. The findings indicate that despite a strong recognition of the importance of advancing ethical AI, there exists a gap between the intentions of business leaders and their actual implementation of meaningful actions. Approximately 80% of CEOs are willing to integrate AI ethics into their companies' business practices. However, the survey reveals that less than a quarter of these organisations have successfully operationalised these principles. Moreover, less than 20% of respondents reported that their company's actions align consistently with its AI ethics principles.

Besides, most of the AI principles' operationalisation frameworks have been at the level of risk management with less attention to the responsibilities, which lie at the level of AI deployers and Impact Population. Therefore moving from the uni-stakeholder approach, in this section, we will provide stakeholder-by-stakeholder strategies and means to operationalise the principles discussed in the previous section and their outcomes. While every principle would require/worth a separate research study in terms of operationalisation; however, the purpose of this paper is to map the principles and levers for operationalisation to a limited extent such that future research can be initiated on the same. We believe responsible AI can be effectively achieved by establishing concrete requirements that address the needs and responsibilities of AI developers, AI deployers, and the Impact Population.

3.3.1. AI Developers

The role of the AI developers, as mapped across the paper, is predominant at the development stage, from ideation to deploying the AI solutions. AI developers' role is significant beyond the development stage as they directly/indirectly interface with the AI deployers who procure the AI solutions. Besides, one of the significant ways AI developers can contribute towards making Responsible AI is by tackling the potential impact that the technology could cause when deployed by the AI deployers or directly used by the Impact Population. Therefore, AI developers must operationalise the mapped principles (refer to Figure 4) using some of the following suggested strategies to realise the same at different stages.

[97] IBM Corporation. (2022, April). *AI ethics in action An enterprise guide to progressing trustworthy AI*. IBM - United States. https://www.ibm.com/downloads/cas/4DPJK92W

3.3.1.1. Plan & Design Stage

In this section, we will discuss various principles to be followed by the players, such as C-suite executives, Test & Evaluation, Validation & Verification experts, product managers, compliance experts, auditors, organisational management, etc. may follow to ideate AI solutions which are responsible and safe. In this stage, developers and technologists must focus on understanding their AI systems' potential consequences and implementing appropriate measures to mitigate risks through operationalising the following principles using the suggested strategies.

- **Human-in-the-loop:** As human-in-the-loop could mean many things, here we list this principle to indicate the importance of involving the impact population and other relevant stakeholders as part of this stage using various approaches. One way to operationalise this principle is through adopting a participatory approach, where the impact population is consulted during the ideation.[98] Another way is to use stakeholder engagement tools, as defined by OECD. According to OECD, meaningful stakeholder engagement is to conduct a two-way, ongoing engagement with the stakeholders in good faith and with responsiveness.[99] Therefore, using this tool is important for AI developers to meaningfully engage with stakeholders like the impact population, domain experts, AI deployers, lawyers etc., to bring multiple voices together and account for cultural and contextual intricacies. Besides, during the ideation, it is important to assess whether these technologies truly benefit the impact population, especially the vulnerable population within it, like gender-based minorities, low-income households etc.

  Therefore, AI developers need to conduct landscaping to determine the utility of AI-based interventions for the impact population, emphasising the last mile and the effectiveness of these technologies in resolving the key challenges they face in their daily lives. Adopting Field Scanning, which is interchangeably used for Landscape Scanning — A methodology used in philanthropy to identify gaps,[100] AI developers could find the pain points and needs within the field. Also, understand the opportunities, emerging trends, gaps, and threats of using Artificial Intelligence. AI developers could adopt the Community-based participatory research (CBPR) approach[101] where they may partner with various community members and organisations as part of the process at the different stages of the Field Scanning exercise.

[98] Wolfewicz A. (2023, February). *Human-in-the-Loop in machine learning: What is it and how does it work?* Levity | No-code AI workflow automation platform. https://levity.ai/blog/human-in-the-loop
[99] OECD Secretariat. (2015, April). *Due Diligence Guidance for Meaningful Stakeholder Engagement in the Extractives Sector.* OECD. https://www.oecd.org/daf/inv/mne/OECD-Guidance-Extractives-Sector-Stakeholder-Engagement.pdf
[100] *Analyzing the Landscape: Community Organizing and Health Equity | Published in Journal of Participatory Research Methods*. (2020, June 29). Journal of Participatory Research Methods. Retrieved June 20, 2023, from https://jprm.scholasticahq.com/article/13196-analyzing-the-landscape-community-organizing-and-health-equity
[101] Prabhakaran, V., & Martin Jr, D. (2020, December). *Participatory machine learning using community-based system dynamics*. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7762892/

- **Fairness & Discrimination:** At this stage, the players involved in the ideation process need to be aware[102] of the negative consequences of cognitive bias, which would ultimately lead towards developing a solution that might discriminate and may bring unfair outcomes. While landscaping could help collect information from the field, the players in this stage need to ensure to pick data points which could explicitly showcase the traits of discrimination, inequality, unfair outcomes, etc., which in the Indian context include information about individuals who belong to low-income households, caste and religious minorities, gender minorities, children etc. The inferences collected through stakeholders in the form of lived experinces must be considered while developing the technology. Besides seeking a second opinion from a community organisation, civil society members, field workers etc. in case the individuals from the community might not be aware of their best posible interest, could help during the ideation stage, where they could evaluate if (a) the proposed idea could adversely impact the population and bring out historical biases and discrimination, (b) all the relevant data points collected through landscaping, stakeholder engagement is considered while ideating, (c) the solution can be made better considering different strata of individuals within the impact population.

  Finally, it could also help AI developers devise a fairness index which considers (a) Pertenance: the relevance of the AI solutions ideated for the impact population, (b) Diversity: the level at which ideated AI solutions can serve different strata of individuals within the impact population, especially vulnerable community, (c) Equity: Compatibility of ideated AI solutions to operate within unequally distributed scenarios with power parity concerns within the impact population and (d) Risks: Mapping potential discriminatory and unfair risks what the impact population may face. Once AI developers have a concrete idea, it is natural not to explore alternatives; however, at this stage, they need to run the idea through the fairness index to break the cognitive biases and find alternatives if necessary.[103]

- **Accuracy:** AI developers at the plan and design stage need to thoroughly understand the AI system's business and society requirements. This helps establish practical accuracy goals that align with the specific application and context of the AI system. These goals should consider the intended use, potential risks, and impact on end-users and society.

  AI developers need to put efforts towards accurately predicting potential harms that ideated technology could cause to society such that appropriate impact management measures can be instrumentalised. A constant effort would be needed towards keeping AI

[102] Fallmann, D. (2021, June 14). *Council post: Human cognitive bias and its role in AI.* Forbes. Retrieved June 20, 2023, from https://www.forbes.com/sites/forbestechcouncil/2021/06/14/human-cognitive-bias-and-its-role-in-ai/
[103] *Human-centric AI (India)*. (n.d.). Open Loop. Retrieved June 20, 2023, from https://openloop.org/programs/open-loop-india-program/

impact as close as possible to actual AI harms such that adverse implications, to an extent, can be prevented. To achieve the same, inferences from the stakeholder engagement, especially with experts and community representatives, would be helpful where they could pinpoint the potential harms that the ideated technology could bring on the society, something which impacts the population may or may not be able to voice or understand timely. However, to keep the levels of positive paternalism lower, incorporating learnings from the landscaping would be helpful, especially the part where impact populations might voice their concerns and risks associated with deploying the ideated AI solution.

In addition to using the inferences from landscaping, AI developers should also conduct in-depth analysis and requirements gathering to define the target accuracy levels and performance metrics that best align with the domain-specific applications. Developers should define thresholds and acceptable error rates for the AI system along with an accuracy matrix. These thresholds would determine the point at which the system's performance would be considered acceptable or unacceptable. This will help ensure that the system meets the predefined standards and minimises the risk of unintentional outputs.

- **Transparency and Explainability:** It is important to have a functional organisational procedure for documentation[104] of the ideation process, starting from landscaping to developing an elevator pitch. Documentation would bring out transparency in the process without disclosing much information on the process itself, which is similar to the metadata of the process. In addition to making the ideation process easily explainable, documentation can also help the AI developers refine the process, assessing the alignment of development and deployment goals with that of AI deployers, impact populations, etc. This documentation will help us understand the thought process behind ideating a given AI solution and induce accountability.

- **Governance:** While the ideation stage would involve various players, especially at the executive level, however, it is essential to have external governance/supervision over the process of ideation such that there is (a) separation of power and (b) for checkpoints to assess if the AI developers are moving on the right track. For instance, having domain experts and community members as observers could act as a robust governance structure for the ideation process. Besides, board-level and public-level commitment toward AI principles could act as an appropriate governance structure, where the AI developers (a) could use various platforms, mechanisms and forums to showcase their public

[104] Perifanis N-A, Kitsios F. (2023, February 2). *Investigating the influence of artificial intelligence on business value in the Digital Era of strategy: A literature review*. MDPI. https://www.mdpi.com/2078-2489/14/2/85

commitments to core principles[105] and (b) could also have measurable milestones to check the progress on their commitment.[106]

- **Legal Cognizance:** Any ideation of the AI innovation must be cognizant of the fact that some of the existing regulations and rights protection at the domestic level would apply to them. For instance, the upcoming Digital Personal Data Protection Bill 2022 (DPDPB 2022) will apply to AI developers who develop and facilitate AI technologies. As AI developers will collect and use massive amounts of data to train their algorithm to enhance the AI solution, they might be classified as data fiduciaries. This implies that AI developers may comply with the fundamental principles of privacy and data protection and the provisions enshrined in DPDPB 2022.

- **Awareness:** When AI technology is ideated, it is essential to ensure that (a) unintended consequences are mapped such that the solution might not cause an adverse impact as a byproduct, (b) trade-offs are confronted, (b) both positive and negative externalities which makes a third party benefit or lose is weeded out. AI developers could develop a selection criterion to run past various outcomes and possibilities of ideated AI technologies to operationalise the same.

  Developers can prioritise and integrate awareness into the anticipated system's design by identifying these considerations. In addition, as part of the landscaping study, an impact assessment should be conducted to evaluate the potential effects of the AI system on various stakeholders, including individuals, communities, and society as a whole. These assessments help identify potential risks and unintended consequences, allowing developers to take proactive measures to mitigate them. By considering the broader impact of the system, developers can foster awareness of its potential effects and make informed decisions.

3.3.1.2 Collect and Process Data

In this section, we will explore the crucial stage of collecting and processing data for AI development. Data forms the foundation of AI systems, and its quality and handling significantly impact the outcomes and implications of the technology. During this stage, players such as Data scientists, data/model/system engineers etc., must carefully consider the principles and strategies to ensure responsible and ethical data practices by seeking diverse datasets representing different perspectives, demographics, and societal contexts. Adhering to these principles and employing

[105] Chui, M., & Manyika, J. (2018, November). *Applying artificial intelligence for social good.* McKinsey & Company. https://www.mckinsey.com/featured-insights/artificial-intelligence/applying-artificial-intelligence-for-social-good
[106] *Responsible AI toolkit* | TensorFlow. (n.d.). TensorFlow. https://www.tensorflow.org/responsible_ai

the suggested strategies can enhance the reliability, fairness, and privacy of the data used in AI systems.

- **Data Quality and Diversity:** In the data collection phase, prioritising data quality and diversity is crucial for building reliable and unbiased AI models. It is crucial to evaluate the data sources to ensure they are diverse and representative of the population or domain of interest. Demographic, geographic, and socioeconomic diversity should be considered to minimise biases and comprehensively understand the target problem. To improve data quality, AI developers should undertake data cleaning and pre-processing techniques, such as removing outliers[107] and handling missing values[108]. This would help to improve data quality and ensure that the subsequent analysis and modelling are based on reliable and accurate data. Additionally, techniques like data augmentation[109] can increase data diversity and enhance the generalisability of AI models, ensuring they perform well across various scenarios. Besides, the data quality must also be determined by analysing if the dataset has historical biases which reinforce stereotypes.

  It is also essential to have a mechanism to crosscheck and evaluate the data's integrity and cleanliness, as state and non-state actors would use this for real-life interventions. For instance, mechanising periodic audits for both data collection methods and data could help to cross-check. Besides, comparing the data with an alternative database can also help determine gaps and mistakes in data points within the coordinated dataset.

- **Fairness and Non-Discrimination:** To operationalise this, AI developers should implement strategies that ensure fairness throughout the data collection and processing stages[110]. This includes carefully selecting diverse and representative datasets, debias sampling[111], conducting bias assessments on the data, identifying potential sources of bias, and taking appropriate measures to mitigate them. Developers should also evaluate the performance of their AI systems across different demographic groups to identify and address any disparities or discriminatory outcomes. Regular monitoring and evaluation of the data collection and processing procedure are crucial to ensure ongoing fairness and non-discrimination in AI systems.

---

[107] Singh, H. (2020, May 24). *Data Preprocessing*. Medium. Retrieved June 20, 2023, from https://towardsdatascience.com/data-preprocessing-e2b0bed4c7fb
[108] Singh, H. (2020, May 24). *Data Preprocessing*. Medium. Retrieved June 20, 2023, from https://towardsdatascience.com/data-preprocessing-e2b0bed4c7fb
[109] *Data augmentation for machine learning.* (2023, March). Akkio. Retrieved June 20, 2023, from https://www.akkio.com/data-augmentation-for-machine-learning
[110] N. Mehrabi et al. (2022, January 25) "*A survey on bias and fairness in machine learning*". ACM Computing Surveys (CSUR) (2021) https://arxiv.org/abs/1908.09635
[111] Acharya, S. (2019, March 18). *Tackling bias in machine learning.* Medium. Retrieved June 20, 2023, from https://blog.insightdatascience.com/tackling-discrimination-in-machine-learning-5c95fde95e95

- **Data Provenance Transparency:** Operationalising the principle of data provenance transparency involves ensuring clear visibility and traceability of the origin, history, and handling of the data used in AI systems. AI developers should implement practices that promote transparency and accountability in data collection and processing to achieve this. This includes documenting a trail of how the data was prepared for use in AI models. Metadata about the data, such as its quality, completeness, and any limitations, should be documented to provide insights into the reliability and suitability of the data for the intended AI application. By tracking data lineage at a high resolution, technologists gain insights into how data is processed, enabling a better understanding and control of AI system behaviour[112]. Blockchain technology has emerged as a promising solution to ensure the integrity and immutability of data provenance in the field of AI[113]. By leveraging blockchain, the tamper-proof nature of data provenance can be effectively certified[114].

  Besides, derived data metadata should be viewed as high-risk data as this may cause a feedback loop and compound the harm. For instance, as we move forward, the internet may get filled with data created by generative AI, where generative AI learns from its content, causing a closed loop and a lack of creativity. Therefore, this creates an infinite regression where the homogenisation of content may occur.

- **Transparency and Explainability:** To operationalise the principle of transparency and explainability at the stage of collecting and processing data in the AI lifecycle, it is important to ensure that the processes and methodologies used to collect and process data are clear, understandable, and well-documented. AI Developers should document the data collection methods[115], such as the sources, sampling techniques, any potential biases or limitations associated with the data, and data processing techniques[116], including data cleaning, filtering, and feature selection processes. Furthermore, developers should offer explanations of the decision-making processes and the factors influencing the outcomes, thereby enhancing the understandability of the AI system.

---

[112] M. Herschel et al. (2017, October 16)*"A survey on provenance: What for? What form? What from?"* The VLDB Journal (2017). https://link.springer.com/article/10.1007/s00778-017-0486-1

[113] M. AlShamsi et al. (2020, September 1)*"Artificial intelligence and blockchain for transparency in governance"*. Artificial Intelligence for Sustainable Development: Theory, Practice and Future Applications. Springer, 2021 https://link.springer.com/chapter/10.1007/978-3-030-51920-9_11

[114] D. N. Dillenberger et al. (2019, February 20) *"Blockchain analytics and artificial intelligence"*. IBM Journal of Research and Development (2019). https://ieeexplore.ieee.org/document/8645631

[115] Javaid, S. (2022, June 16). *AI/ML data collection in 2023: Guide, challenges & 4 methods.* AIMultiple. Retrieved June 20, 2023, from https://research.aimultiple.com/data-collection/

[116] Baheti, P. (2023, February). *Data Preprocessing in machine learning [Steps & techniques]. V7* - AI Data Platform for Computer Vision. Retrieved June 20, 2023, from https://www.v7labs.com/blog/data-preprocessing-guide

Further, in the cases wherein AI systems are used by government bodies to perform functions which have a direct impact on the life, liberty and freedoms of their citizens, such as prediction of the rate of criminal recidivism, prediction of tax fraud etc., we recommend that the AI systems developed for this use are focused on creating an intrinsically explainable model, instead of a black-box AI model which is later explained through explainable AI (XAI) techniques. This is especially crucial since government functions and decisions not only owe a degree of transparency to the citizens, but the doctrine of the principle of natural justice requires all administrative actions to have a duty to provide a *reasonable explanation* to the persons who are subjected to such administrative decisions. Therefore, without an inherent level of explainability regarding the inner workings of an AI system, it is difficult to rely on the computation of AI systems when carrying out administrative or judicial functions.

- **Governance:** At this stage, operationalising the principle of governance would involve establishing robust policies, frameworks, and processes to ensure responsible and ethical handling of data. AI Developers should adhere to relevant laws, regulations, and industry standards governing data privacy, security, and consent. Developers should document and communicate their data governance practices to stakeholders, including data subjects, regulators, and auditors. They should provide clear information about the purpose of data collection, the types of data being collected, and the rights and choices available to data subjects.

- **Data Veracity:** AI Developers should assess the quality of the collected data by evaluating its completeness, consistency, relevance, and accuracy. This may involve data profiling[117], data cleansing[118], and data normalisation[119] techniques to identify and correct errors, outliers, and inconsistencies. In addition, to evaluate the credibility and reliability of the data sources, AI developers should consider factors such as the data provider's reputation, the methodology used for data collection, and any potential biases or limitations associated with the data source.

- **Data Security:** AI Developers should establish data security measures to safeguard collected data against unauthorised access, breaches, and misuse. This involves implementing appropriate access controls, encryption techniques, and secure storage systems to safeguard the data from unauthorised access, data breaches, or tampering. Developers should use secure communication channels when transferring or sharing data

[117] Nova. (2023, March). *Data Profiling: The Developer's Secret Weapon*. AItech Trend. Retrieved June 20, 2023, from https://aitechtrend.com/data-profiling-the-developers-secret-weapon/
[118] Goel, U. (2023, June 10). ML | *Overview of data cleaning*. GeeksforGeeks. Retrieved June 20, 2023, from https://www.geeksforgeeks.org/data-cleansing-introduction/
[119] Alam, M. (2020, December 14). *Data normalization in machine learning.* Medium. Retrieved June 20, 2023, from https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02

with third parties and establish data-sharing agreements that outline all parties' security requirements and responsibilities. This helps ensure that data is protected during transit and that data recipients adhere to the same security standards. Further, conducting regular security audits and vulnerability assessments to identify and address any potential security weaknesses or vulnerabilities in the data collection and processing systems is crucial. This helps in proactively identifying and mitigating security risks.

- **Data Protection Principles:** AI developers follow principles starting from the stage of data collection to data expunction. They should collect only the necessary data and minimise collecting sensitive or personally identifiable information (PII) whenever possible (Data Minimisation). By reducing the amount of sensitive data collected, developers can lower the potential risks of storing and processing such information. AI Developers should also adhere to applicable privacy regulations and ensure that appropriate privacy protections are in place. The information on the processing mechanism of the data must be simple and documented. The data protection impact assessment and other audit reports must be made public.

  In addition, AI developers should establish guidelines for data retention and disposal to ensure that data is retained only for as long as necessary and securely disposed of when no longer needed. This includes implementing secure data deletion techniques to prevent data recovery.

- **IP Protection:** At this stage, it is essential to identify any intellectual property (IP) assets involved in the data collection and processing process, such as proprietary algorithms, datasets, or trade secrets. This helps implement appropriate measures to protect these assets, including applying copyrights, trademarks, patents, or trade secret protection mechanisms. Using non-disclosure agreements (NDAs) when collaborating with external parties, such as data providers or third-party vendors, is crucial to ensure the confidentiality and protection of sensitive or proprietary information shared during the data collection. NDAs help outline the terms and conditions for handling and sharing confidential information and help safeguard intellectual property rights.

3.3.1.3. Build and Use Model

In this stage, AI developers (i.e., players like Modelers, Model Engineers, Data scientists, data/model/system engineers, domain experts, etc.) face the crucial task of carefully selecting suitable algorithms, building the model architecture, and establishing the specific techniques and methodologies to be employed. This stage is pivotal in achieving essential attributes such as robustness, explainability, fairness, generalisation, and privacy protection in the AI model's

design. The thoughtful consideration of these factors ensures that the algorithm is effective, trustworthy, and aligned with responsible AI principles.

- **Future-proofing:** Operationalising this principle will require AI developers to design systems with scalability and flexibility in mind. This allows for easier integration of new features, algorithms, and data sources as they become available. In addition, AI developers should incorporate mechanisms for continuous learning and adaptation into the AI system. This includes updating models and algorithms based on new data and feedback, enabling the system to improve over time and adapt to changing environments and user needs. Further, staying informed about emerging industry standards, best practices, and regulations related to AI is crucial. This helps ensure that the AI system is designed to be compatible with these standards to avoid future compliance issues or the need for major system modifications. Besides, it is also important to be aware of the limitation of using current data about the world to understand the world in future such that we can appropriately mitigate the error.

- **Model Alignment:** Alignment refers to the practice of fine-tuning AI models to align with human intent and human values. Models that work in line with the human intention are deemed to be aligned.[120] The practice involves training AI models on human feedback under a reinforcement learning model to align the AI model's outputs to human values and not merely to the best computable answer. AI Alignment also empowers users to a) correct the models when they commit mistakes, b) ensure that they align with human values even when they progress beyond human intellectual limitations, and c) enable capacity to be fine-tuned over time as human values aren't permanent.[121] The importance of alignment can be seen in the fact that major AI developers are currently allocating considerable resources to ensure that their AI services provide outputs that are safe and aligned with human values. Alignment takes centre stage for Artificial generalised Intelligence ('AGIs') as the impact population is projected to and even in the present relies upon the accuracy of outputs received from AI services.[122] Since the reliance on AI services is likely to increase with time, alignment of AI models must be given adequate attention and importance from the initial stages as alignment in the status quo is a tedious process requiring considerable human and compute resources as is evidenced by the longstanding ethical study conducted by the self-driving industry and other ancillary industries addressing AI alignment.[123]

---

[120] Christian, B. (2020) *The Alignment Problem*, Norton Publishing. ISBN: 978-0-393-86833-3
[121] Russell,S. (n.d) *The Value Alignment Problem*, Leverhulme Centre for the Future of Intelligence. Retrieved on June 20,2023 from http://lcfi.ac.uk/projects/completed-projects/value-alignment-problem/
[122] Leike,J et al. (2022 August 4) *Our approach to alignment research*, Open AI. Retrieved June 20, 2023 from https://openai.com/blog/our-approach-to-alignment-research
[123] Hansson, S.O., Belin, M.A. & Lundgren,B. (2021 August 12) *Self-Driving Vehicles—an Ethical Overview*, Journal of Philosophy & Technology, Springer. Retrieved on June 20, 2023, from https://link.springer.com/article/10.1007/s13347-021-00464-5

- **Fairness and Non-Discrimination:** In this stage, it is important to continuously monitor and evaluate the AI system's development to identify any instances of unfairness or discrimination. To operationalise the same, bias mitigation techniques are employed. These techniques can be categorised into two primary approaches: debias sampling and debias annotation.

  Debias sampling involves the identification and selection, or annotation, of data points in a manner that mitigates bias. However, it is important to note that merely having a dataset that reflects the user population does not guarantee fairness. Statistical methods and metrics may still favour majority groups, so it becomes necessary to consider task difficulty. For instance, tasks like recognising speech in less-spoken accents can inherently be more challenging due to data scarcity[124]. Therefore, system developers must consider task difficulty when constructing and evaluating fair AI systems. Debias annotation involves choosing the appropriate annotators, particularly when dealing with underrepresented data. For instance, selecting experts who know rarely heard accents is essential when annotating speech recognition data. This ensures that human bias is minimised and prevents the introduction of biased annotations. Careful consideration should be given to selecting experts who can provide accurate and unbiased annotations, especially when dealing with data from underrepresented groups.

  Besides, it has to be on the conscience of AI developers that the model to be developed doesn't elevate or create discrimination or positive and negative externalities.

- **Explainability:** It is difficult to achieve transparency in the context of AI systems because ML models encode correlations between input and output that are learned and not that of what developers have specified, which makes these systems highly opaque by default. Therefore, while it is tough to bring transparency to the statistical and algorithmic portions of the AI systems, instead, AI developers could bring transparency to the development process, datasets, and other connections around the model through the documentation process.[125] Technologists should continuously improve these processes by implementing interpretability techniques and methods. This can involve using model-agnostic approaches like LIME[126] (Local Interpretable Model-agnostic

[124] A. Koenecke et al.(2020, March 23) *"Racial disparities in automated speech recognition"*. Proceedings of the National Academy of Sciences (2020) https://www.pnas.org/doi/10.1073/pnas.1915768117
[125] *ABOUT ML Reference Document*. (2021, September 7). Partnership on AI. Retrieved June 20, 2023, from https://partnershiponai.org/paper/about-ml-reference-document/1/#Section-0
[126] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August 12). *Local interpretable model-agnostic explanations* (LIME): An introduction. O'Reilly Media. https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/

Explanations) or SHAP[127] (Shapley Additive Explanations) to provide insights into the decision-making process. While LIME or SHAP would bring mathematical explainability, it is also crucial to present information clearly and understandably, allowing AI users to interact with the system, inquire about its decision-making process, and access relevant explanations. User-friendly interfaces facilitate transparency and empower users to make informed judgments about the system's outputs.

Further, providing clear information about the AI system's capabilities, limitations, and intended use is pertinent. This helps ensure that stakeholders understand the purpose and objectives of the system and provides them with channels for feedback, complaints, and redress. Regular reporting and communication on system performance and outcomes are essential for transparency and explainability.

Another practical approach to achieve explainability is by integrating an explanation task into the AI model. This method is commonly utilised in tasks such as Natural Language Processing (NLP) based reading comprehension, where supporting sentences are generated to provide a clear rationale[128]. To ensure effective training for the explanation task, it is advantageous to gather explanations or supplementary information that may not be directly tied to the primary task. These explanations can be obtained through direct input from annotators[129] or through automated techniques. By collecting and incorporating such explanatory data, the interpretability of the AI system can be enhanced.

- **Governance:** At this stage, it is essential to create internal governance structures to oversee the development and use of AI systems. This may include establishing an AI ethics committee or a dedicated team responsible for monitoring and enforcing compliance with governance policies. These structures can ensure accountability, provide guidance, and facilitate decision-making processes. Further, implementing robust data governance practices is essential to protect user privacy and ensure compliance with data protection regulations.

- **Privacy by Design:** The principle of Privacy-by-Design can be operationalised by integrating privacy considerations into the build and use state of the AI system. This involves adopting a privacy-centric mindset and placing privacy as a core requirement

[127] Verma, Y. (2022, March 26). *A complete guide to SHAP* - Shapley additive explanations for practitioners. Analytics India Magazine. Retrieved June 20, 2023, from https://analyticsindiamag.com/a-complete-guide-to-shap-shapley-additive-explanations-for-practitioners/

[128] M. Tu et al.(2020 February) "*Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents*". Proceedings of the AAAI Conference on Artificial Intelligence. 2020. https://arxiv.org/abs/1911.00484

[129] S. Wiegreffe et al.(2021, December 7)"*Teach Me to Explain: A Review of Datasets for Explainable NLP*".https://arxiv.org/abs/2102.12060

rather than an afterthought. This includes techniques such as data anonymisation [130], and differential privacy[131] (without causing unintended consequences of underrepresentation), which involves removing or encrypting personally identifiable information (PII) from the data to protect individuals' identities. Further, enabling individuals to exercise their privacy rights effectively is also essential. This includes providing mechanisms for individuals to access, rectify, delete, or restrict the processing of their personal data. Implement processes to respond to privacy-related requests and inquiries promptly and transparently. Privacy protection measures, such as data access controls and secure data storage, should also be implemented to safeguard sensitive information and uphold user privacy rights. Further, industry-standard security practices should be adopted to prevent unauthorised access, data breaches, and other privacy-related incidents.

Besides, AI developers could also plugin Privacy-Enhancing Technologies to enhance privacy quotient. Where on the supply side, PETs aid businesses in adhering to some of the fundamental principles of data protection like data minimisation, proactive data protection, end-to-end security and privacy-by-design, and in turn, aid in compliance.[132] On the demand side, PETs are placed in a unique position where their consumer-facing solutions aid individuals in securing their data from privacy harm like financial loss,[133] discriminatory treatment,[134] exclusion,[135] restrictions on free speech,[136] and enhance user agency. [137] Together, supply-side and demand-side PETs together can aid AI developers in fixing the privacy void at different data lifecycle stages.

---

[130] Yang, S. (2020, December 15). *Data Anonymization with Autoencoders.* Medium. Retrieved June 20, 2023, from https://towardsdatascience.com/data-anonymization-with-autoencoders-75d076bcbea6

[131] Nguyen, A. (2022, January 15). *Understanding differential privacy*. Medium. Retrieved June 20, 2023, from https://towardsdatascience.com/understanding-differential-privacy-85ce191e198a

[132] Ruan, W., Xu, M., Jia, H., Wu, Z., Song, L., & Han, W. (2021). *Privacy Compliance: Can Technology Come to the Rescue?* Retrieved from IEEE Security & Privacy: https://www.computer.org/csdl/magazine/sp/2021/04/09444564/1u3mFH7L9gA

[133] Prasad, S. (2019, October 29). *An Analysis of 'Harm' defined under the draft Personal Data Protection Bill, 2018.* Retrieved January 17, 2022, from https://www.dvara.com/research/blog/2019/10/29/an-analysis-of-harm-defined-under-the-draft-personal-data-protection-bill-2018/

[134] Khan, L. M. (2017, January 3). Yale Law Journal - Amazon's Antitrust Paradox. The Yale Law Journal. Retrieved January 17, 2022, from https://www.yalelawjournal.org/note/amazons-antitrust-paradox

[135] A Taxonomy of Privacy - ORG Wiki. (2013, January 8). ORG Wiki. Retrieved January 17, 2022, from https://wiki.openrightsgroup.org/wiki/A_Taxonomy_of_Privacy#Exclusion;
Falling through the Cracks: Case Studies in Exclusion from Social Protection - Dvara Research. Retrieved January 17, 2022, from https://www.dvara.com/research/social-protection-initiative/falling-through-the-cracks-case-studies-in-exclusion-from-social-protection/

[136] Freedom of Expression & Privacy. (n.d.). The Centre for Internet and Society. Retrieved January 17, 2022, from https://cis-india.org/internet-governance/blog/freedom-of-expression-and-privacy.pdf

[137] *Personal Data and Individual Agency*. (n.d.). IEEE. https://ethicsinaction.ieee.org/wp-content/uploads/ead1e_personal_data.pdf

- **Human-AI:** To operationalise this principle, AI developers need to develop AI systems with a user-centric approach, considering the needs, preferences, and limitations of human users. This can be achieved by involving users in the design process through user research[138], feedback sessions[139], and usability testing[140]. In addition, it is essential to design AI systems with appropriate levels of human oversight. This can be achieved through mechanisms such as a human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach[141]. The choice of the appropriate mechanism depends on the specific application and the level of human intervention required. For example, in the HITL approach, humans can intervene in every decision cycle of the system. However, human intervention at such a granular level may not always be practical or desirable. However, incorporating some of the key features would make human interaction with AI easier and reduce unintended consequences. For instance, a simple addition of a feature, such as a citation for the information generated by the AI, could take us a long way in protecting copyrights. Besides, to operationalise the principle of human-AI, we must be aware of the contextuality of the data, AI use and human behaviours, which differ based on the context and environment.

- **Safety:** To operationalise this principle, AI developers must perform rigorous testing and validation of the AI system to ensure its safety and reliability. AI systems need to be tested under various scenarios and conditions to identify and address any potential safety issues. Real-world data and simulations can be used to evaluate the system's performance and identify potential vulnerabilities. In addition, safety measures and safeguards need to be implemented, including building redundancy, fail-safe mechanisms, and error-handling capabilities. The system needs to be designed to minimise the likelihood of accidents, malfunctions, or harmful behaviours. Further, AI developers need to establish a robust incident response plan to address any safety incidents or failures promptly and define procedures for reporting, investigating, and resolving safety-related issues. Contingency plans to recover from any potential disruptions caused by safety incidents should also be developed.

---

[138] Butler, C. (2017, March 12). *Testing AI concepts in user research*. Medium. Retrieved June 20, 2023, from https://uxdesign.cc/testing-ai-concepts-in-user-research-b742a9a92e55

[139] Barnett, J. (2018, August). *The Future Of Feedback: How AI Fosters A Human Connection At Work.* Forbes. Retrieved June 20, 2023, from https://www.forbes.com/sites/jimbarnett/2018/08/07/the-future-of-feedback-how-ai-fosters-a-human-connection-at-work/?sh=46f564b3f653

[140] Roose, J. (2017, March). *How to conduct usability testing in six steps.* Toptal Design Blog. Retrieved June 20, 2023, from https://www.toptal.com/designers/ux-consultants/how-to-conduct-usability-testing-in-6-steps

[141] European Commission. (2019, April). *Ethics guidelines for trustworthy AI*. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

- **IP Protection:** Staying updated with relevant intellectual property laws, regulations, and best practices is important to ensure compliance. This includes understanding the legal requirements for protecting intellectual property rights, respecting copyright and trademark laws, and adhering to licensing agreements when using third-party data or intellectual property for modelling AI solutions.

3.3.1.4. Verification and Validation

In the verification and validation stage in the AI lifecycle, developers and technologists (Data Scientists, experts etc.) delve deeper into ensuring the responsible and safe operation of AI systems before deployment. Building upon the principles outlined, this stage requires a meticulous focus on comprehending the potential consequences of AI systems and implementing effective risk mitigation measures. By overlaying the deployment context and making informed choices, developers can establish a robust foundation for successfully integrating AI systems while addressing potential risks and ethical concerns.

- **Human-in-the-loop:** Operationalising the principle of human-in-the-loop involves incorporating human involvement and oversight into the system's testing and evaluation processes. This can be achieved by including human reviewers who access the decisions made by the AI system and provide feedback and correction where needed. These reviewers can be domain experts or individuals with relevant knowledge or expertise. They ensure that the system's outputs align with desired outcomes and ethical considerations. Further, it is essential to establish clear criteria for human intervention or override in certain circumstances. This ensures that human judgement can be applied when the AI system's outputs are uncertain, questionable, or have significant implications.

- **Impact Assessment:** The principle of impact assessment involves evaluating the potential effects and consequences of the AI system on various stakeholders and the broader environment. This assessment aims to understand and mitigate any negative impacts and maximise the positive outcomes of the system. To operationalise this, several steps can be taken. Firstly, it is important to identify the key stakeholders who may be affected by the AI system, such as end-users, employees, communities, and society at large. Next, AI developers should define appropriate metrics and indicators to measure the impact of the AI system. These metrics can include aspects such as fairness, privacy, safety, economic implications, and societal well-being. Then the AI developers should conduct thorough testing and evaluation to assess the system's performance against the identified impact criteria. This includes analysing the system's outputs, potential biases, unintended consequences, and any risks associated with its deployment. The feedback derived can help identify any biases, errors or limitations in the system's performance and

inform improvement. Both quantitative[142] and qualitative[143] methods can be employed to collect data and evidence for impact assessment.

- **Reliability and Safety:** To operationalise reliability, AI developers should conduct comprehensive testing to verify that the AI system consistently produces reliable and consistent results. This includes testing the system's performance across different scenarios, inputs, and datasets to assess its robustness and reliability. Rigorous testing methodologies, such as unit testing[144], integration testing[145], and stress testing[146], can help uncover any potential issues or vulnerabilities. On the other hand, safety considerations involve identifying and addressing risks associated with the AI system's operation. This includes analysing potential safety hazards, such as unintended consequences, biased decision-making, or negative impacts on users. Developers should conduct risk assessments and employ techniques like fault tolerance[147], fail-safe mechanisms[148], and continuous monitoring[149] to minimise risks and ensure the system's safe operation. Further, AI developers should establish clear benchmarks and criteria for evaluating reliability and safety. This may involve setting performance thresholds, defining acceptable error rates, and establishing safety protocols.

- **Transparency and Explainability:** At this stage of the AI lifecycle, transparency and explainability can be achieved through proper documentation of the various steps involved in the verification and validation process. This includes documenting the data used for testing, selecting and evaluating the performance metrics, the methodologies and techniques employed, and the results obtained. By documenting these details, developers

---

[142] European Commission. (2019, April). *Ethics guidelines for trustworthy AI.* Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[143] European Commission. (2019, April). *Ethics guidelines for trustworthy AI*. Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[144] Pykes, K. (2021, December 7). *Testing machine learning systems: Unit tests*. Medium. Retrieved June 20, 2023, from https://medium.com/pykes-technical-notes/testing-machine-learning-systems-unit-tests-38696264ee04

[145] Kukkuru, M. G. (2023). *Testing imperative for AI systems*. Infosys - Consulting | IT Services | Digital Transformation. https://www.infosys.com/insights/ai-automation/testing-imperative-for-ai-systems.html

[146] Chan-lau, J. (September 4, 2019). *Stress-testing applications of machine learning models.* Risk.net. https://www.risk.net/stress-testing-2nd-edition/7084211/stress-testing-applications-of-machine-learning-models

[147] European Commission. (2023, May). *AI act: A step closer to the first rules on artificial intelligence* | European Parliament. https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence

[148] European Commission. (2023, May). *AI act: A step closer to the first rules on artificial intelligence* | European Parliament. https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence

[149] Office of Science and Technology Policy. (2023, May). *The National Artificial Intelligence R&D Strategic Plan.* The White House. https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf

provide insights into how the AI system was tested and validated, making it easier for others to understand and assess its reliability. In addition to documenting the technical aspects, it is also essential to document any ethical considerations, limitations, and assumptions made during the verification and validation process. This provides transparency regarding the ethical framework within which the AI system operates and helps stakeholders understand the system's limitations and potential biases. Furthermore, documentation should include any insights gained from the verification and validation process. This can involve recording observations, key findings, and lessons learned during the testing and evaluation. By sharing these insights, developers contribute to the collective knowledge in the field and facilitate continuous improvement and learning.

- **Governance:** At this stage, AI developers must establish clear governance policies that outline the principles, objectives, and guidelines for verification and validation. These policies should align with ethical standards, legal requirements, and industry best practices. Further, they should develop standardised processes and methodologies for verification and validation to ensure consistency and reliability. These processes should include data collection, preprocessing, model evaluation, and testing guidelines. By following standardised procedures, developers can ensure that the AI system undergoes thorough and reliable verification and validation.

- **Accuracy:** To operationalise this principle, AI developers need to ensure the collection of high-quality and representative data that is relevant to the AI system's intended use. This includes careful designing of data collection processes to minimise biases and errors. Factors such as data source diversity, sample size, and data labelling techniques should be considered to enhance the accuracy of the testing dataset. Based on the feedback received, rigorous data preprocessing techniques should be applied to clean and normalise the data. This includes removing outliers, handling missing values, addressing the class imbalance, and reducing noise. Proper preprocessing helps improve the quality and accuracy of the data, directly impacting the accuracy of the AI system. Further, cross-validation techniques should be employed to assess the model's generalizability. This involves splitting the data into multiple subsets and testing the model on different combinations of these subsets. Evaluating the model's performance on each subset and analysing accuracy, precision, and recall metrics. This approach helps to measure and improve the overall accuracy of the AI system.

- **Awareness:** Similar to the awareness principle discussed in the plan & design section, when AI technology is tested, it is essential to ensure that (a) unintended consequences are tested, (b) trade-offs are confronted, (b) both positive and negative externalities where the ideated AI solution makes a third party benefit or lose is weeded out.

3.3.1.5. Deployment and Operationalisation

The deployment and operationalisation stage is crucial in operationalising AI principles. It entails deploying AI systems onto real products and their interaction with the environment and users. This stage focuses on fine-tuning the AI system to ensure its effectiveness and reliability in real-world scenarios. In this stage, AI Developers and technologists (Developers, System Engineers, Procurement experts etc.) work towards refining the system's performance, addressing any issues that arise, and optimising it for seamless integration into existing processes. The goal is to ensure that the AI system functions effectively and delivers the intended outcomes in real-world applications.

- **Human-in-the-loop:** To operationalise this principle, AI systems should aim to involve human input in making decisions in specific situations or contexts where the system's outputs may have significant consequences. This allows human judgment to be considered and helps prevent potential biases or errors. This can be done by defining predetermined thresholds or triggers that signal when human input is necessary. These thresholds can be based on various factors, such as the level of confidence or uncertainty in the AI system's predictions, the potential impact of the decisions, or the presence of sensitive or high-stakes scenarios. When these thresholds are met, the AI system can prompt human intervention or provide recommendations for human review and decision-making.

- **Impact Assessment:** Conducting a comprehensive analysis of the potential positive and negative impacts of the AI system across different dimensions is critical at this stage. This analysis should consider both immediate and long-term effects, as well as potential indirect consequences. Further, AI developers should establish mechanisms for ongoing monitoring and evaluation of the AI system's impact throughout its operational lifecycle. This allows for the identification of emerging issues, the assessment of the effectiveness of mitigation measures, and the adaptation of strategies as needed.

- **Reliability and safety:** AI developers should implement and adopt error handling mechanisms and fail-safe measures to handle unexpected situations or errors during operation. One approach is incorporating redundancy, where critical components or functions are duplicated to ensure backup functionality in case of failure. Redundancy can be implemented at the hardware or software level, allowing the system to continue functioning even if one component fails. Another approach is through fallback mechanisms that provide an alternative course of action when the primary system encounters errors, offering a fail-safe option. For instance, a fallback mechanism could switch to a safer mode or prompt the human driver to take control in autonomous driving. Further, error correction techniques play a role in rectifying errors or inaccuracies in the system's outputs, improving accuracy. By analysing user feedback or using machine

learning algorithms, error correction techniques help the system learn from mistakes and make necessary adjustments. Besides, at the ex-ante level, AI developers could consider practices such as red teaming[150] where AI solutions is subjected to systematic adversarial attacks to identify the potential harms to constitute mitigation strategies accordingly.

- **Transparency and Explainability:** At this stage, AI developers can implement techniques enabling the system to explain its outputs. This can be done through methods such as generating textual or visual explanations highlighting the factors or features the AI system considers in reaching a decision. These explanations can help users and stakeholders understand the reasoning behind the system's outputs, increasing transparency and fostering trust. Furthermore, AI developers can consider incorporating model interpretability techniques that make the internal workings of the AI system more understandable. Techniques such as feature importance analysis, attention mechanisms, or rule extraction methods can provide insights into which features or factors contribute most significantly to the system's decisions.

- **Governance:** At the deployment and operationalisation stage, developers should identify and assign specific roles and responsibilities to individuals or teams responsible for tasks such as system configuration, monitoring, maintenance, and performance evaluation. This helps create a clear structure and ensures everyone understands their responsibilities and is accountable for their assigned tasks. Defining roles and responsibilities includes clarifying each individual or team's authority and decision-making powers. This helps establish a hierarchy and ensures that the appropriate individuals or teams make decisions about the AI system's deployment and operation with the necessary expertise and knowledge. In addition to assigning roles and responsibilities, it is essential to establish clear lines of accountability. This means that individuals or teams should be accountable for the outcomes and consequences of the AI system's deployment and operation. They should be aware of the potential risks and ethical considerations associated with the system and take responsibility for addressing any issues that may arise.

- **Accuracy:** Fine-tuning the AI model is necessary to optimise its performance and accuracy. This involves adjusting hyperparameters, such as learning rate, regularisation, or network architecture, to enhance the model's ability to generalise and make accurate predictions. In addition, creating a feedback loop between the AI system and users or domain experts can significantly improve accuracy. Developers should collect feedback on the system's predictions or outputs and use this information to identify areas of improvement. User feedback, manual reviews, or continuous learning techniques can be employed to enhance the system's accuracy over time iteratively. Further, AI developers

[150] Introduction to red teaming large language models (LLMs) - Azure OpenAI Service. (2023, July 18). Microsoft Learn. Retrieved August 17, 2023, from https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming

should conduct thorough error analysis to identify the root causes of inaccuracies or mistakes made by the system. By understanding the types of errors and their underlying causes, AI developers can take targeted actions to address them. This may involve improving the training data, updating the model architecture, or implementing error-handling mechanisms to mitigate potential inaccuracies.

- **Awareness:** Establishing a monitoring and evaluation framework allows developers to maintain awareness of the system's performance and identify any deviations or issues. Monitoring can include tracking key performance indicators, conducting regular audits, and leveraging user feedback to assess the system's effectiveness and identify areas for improvement. This ongoing evaluation ensures that developers remain aware of the system's performance and can take timely actions when necessary. Besides, it is important for AI developers to constitute an adequate internal policy which keeps the process of testing the deployment aware of the individual's concerns. Besides, the AI developers may constitute a user guide based on the interference collected from this stage such that the AI deployers are informed and aware of the issues (which technically emerged during testing) while deploying the AI technology in real-world scenarios.

3.3.2. AI Deployers

AI deployers refer to individuals, organisations, or entities that utilise artificial intelligence solutions or systems in their operational processes. These users are the recipients or consumers of AI technology and leverage its capabilities to perform various tasks, make informed decisions, deliver services, or enhance their operations. AI deployers can span across different industries and sectors, such as healthcare, education, finance, manufacturing, law enforcement, and more. They interact with AI systems, either directly or indirectly, to leverage the outputs, insights, or recommendations generated by AI algorithms and models. AI deployers play a critical role in effectively implementing and utilising AI solutions, driving innovation, efficiency, and data-driven decision-making within their respective domains.

3.3.2.1. Actual Operationalisation

After the AI developers have operationalised and made the AI solutions available, AI deployers procure these solutions (if both are not the same entity). Once procured, AI deployers integrate the AI solutions into their operational processes, leveraging the outputs generated by the AI system for decision-making, service delivery, and other critical functions. The active participation of AI deployers in the AI lifecycle is integral to the successful integration and utilisation of AI solutions. By embracing responsible AI practices and operationalising the principles outlined in the AI lifecycle (refer to Figure 4), AI deployers can harness the full potential of AI technology to drive positive outcomes in their respective domains.

- **Human-in-the-loop:** To implement the human-in-the-loop principle at the actual operationalisation stage, AI deployers can review and validate the AI system's outputs before taking action, considering the expertise and judgment of humans in critical situations. They can also assess the context and circumstances surrounding the AI system's recommendations, incorporating ethical, legal, and social considerations. Human judgment can help ensure that the AI system's outputs align with the organisation's or user's desired goals and values. In addition, AI deployers should incorporate mechanisms that allow human operators to override or modify AI decisions when necessary, based on their expertise. This allows users to intervene in situations where they believe the AI system's outputs are inappropriate or require adjustment based on their expertise or domain knowledge. Further, AI deployers should continuously monitor the performance and behaviour of the AI system during its operational use. This includes tracking the accuracy, reliability, and fairness of the system's outputs and detecting any potential biases or errors. Human monitoring and intervention can help identify and rectify issues that may arise during the AI system's actual operationalisation.

- **Impact Assessment:** To operationalise this principle, AI deployers should establish metrics or indicators to assess the impact of the AI system on various aspects, such as efficiency, productivity, cost-effectiveness, user satisfaction, and societal impact. These metrics should align with the organisation's or user's goals and objectives. Next, AI deployers should collect relevant data to accurately measure the AI system's impact. This may involve gathering data on key performance indicators, user feedback, system performance, and any unintended consequences or side effects resulting from the AI system's use. Further, AI deployers should analyse the collected data and evaluate the impact of the AI system. This analysis may involve comparing the system's performance against predefined benchmarks or evaluating its effectiveness in achieving the desired outcomes. It should also include an assessment of any ethical, legal, or social implications arising from the AI system's deployment. Based on the impact assessment findings, AI deployers should identify areas for improvement and take necessary actions to enhance the positive impacts and mitigate any negative effects.

- **Accessibility:** AI deployers should perform accessibility audits on the AI system to identify any barriers or challenges marginalised users face within the impact population. By actively involving marginalised users in the operationalisation process and seeking their input, AI deployers can gain insights into their specific accessibility needs and challenges. This can involve reviewing the system's user interface, interactions, and content to ensure they are accessible. This engagement can help inform the design and implementation of accessibility features that cater to a diverse user base. Further, AI deployers should offer comprehensive training and support to users, focusing on

accessibility features and best practices. This can include providing documentation, tutorials, and resources that guide users in utilising accessibility features effectively.

- **Transparency and Explainability:** AI deployers should prioritise using AI systems that provide transparency and explainability. This involves selecting systems that clearly explain their decision-making processes, allowing users to understand how and why certain decisions are made. Further, AI deployers should establish mechanisms to audit the AI system's performance and ensure accountability. This can involve regularly evaluating the system's outcomes, monitoring for biases or errors, and addressing any issues that arise. Checklists and quantitative testing are widely used approaches for evaluating fairness[151], transparency[152], and reproducibility[153]. In addition to this, in the event of harmful or unintended consequences of the AI system, AI deployers should take appropriate remedial actions and provide redress to affected individuals or groups. This may involve updating the system, compensating for damages, or addressing biases and discrimination promptly and responsibly.

- **Governance:** Conducting periodic audits[154] of the AI system is a crucial aspect of governance for AI deployers. Audits serve as a systematic and thorough evaluation of the AI system's compliance with governance standards, legal requirements, and ethical guidelines. The audit aims to identify any gaps or deviations from established policies and procedures. It helps uncover potential risks, biases, errors, or ethical concerns that may arise from the AI system's deployment and operation. Audits provide a comprehensive and objective assessment of the system's performance, highlighting areas that require improvement or corrective actions. In addition to audits, maintaining an AI registry could enhance governance, as they would capture information in terms of data flows, data processing, risk developed etc., for auditors to understand the AI system better.

- **Fairness and Non-Discrimination:** AI deployers should continuously monitor the performance of the AI system to identify any potential biases or discriminatory patterns in its outputs. This includes analysing the system's decisions and outcomes across different demographic groups to detect any disparities. If biases or discriminatory

[151] M. A. Madaio et al. (2020, April 23) "*Co-designing checklists to understand organizational challenges and opportunities around fairness in AI*". Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. 2020.https://dl.acm.org/doi/abs/10.1145/3313831.3376445
[152] L. Schelenz et al.(2020, 2nd April) "*Applying Transparency in Artificial Intelligence based Personalization Systems"* https://arxiv.org/abs/2004.00935
[153] J. Pineau et al. (2020, 30th December) "*Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program*". Journal of Machine Learning Research (2021). https://arxiv.org/abs/2003.12206
[154] Minkkinen, M., Laine, J. & Mäntymäki, M.(2022 October 4) *Continuous Auditing of Artificial Intelligence: a Conceptualization and Assessment of Tools and Frameworks*. DISO 1, 21 (2022). https://doi.org/10.1007/s44206-022-00022-2

outcomes are identified, AI deployers should take corrective actions promptly. This may involve adjusting the system's algorithms, retraining with updated data, or implementing additional measures to mitigate biases. Further, AI deployers should establish feedback channels for users and stakeholders to report instances where they believe biases or discrimination affect the system's decisions. This feedback loop can provide valuable insights and help identify areas for improvement.

- **Human Autonomy:** To operationalise this principle, AI deployers should clearly define and establish the level of control they want to maintain over the AI system's decisions and actions. This includes identifying critical areas where human input and decision-making should be prioritised. In addition, AI deployers should establish boundaries for the AI system's decision-making authority. This involves identifying scenarios or contexts where human intervention or override capabilities are necessary to ensure the system's outputs align with desired outcomes and values.

- **Data Security:** AI deployers should assess the security practices and protocols of AI vendors before procuring or using their systems. This includes evaluating their data security measures, adherence to industry standards and best practices, and their track record in handling data security incidents. Further, AI deployers should regularly update and patch their AI systems to ensure they have the latest security fixes and protections against known vulnerabilities and security incidents. This includes staying informed about security updates the AI system vendors released and promptly applying them.

- **Data Protection Principles:** As there could be instances where AI deployers would be collecting data from the impact populations while operationalising AI solutions, it is important to follow some key data protection principles. The impact population must consent to the data collection and have adequate notice of how their data will be used and processed. There shall be a limit to the extent of data collection via fair and appropriate means, and the purpose of data collection must be specified at the data collection stage. The data collected must be used only for the stipulated purpose, nothing incompatible with the specified purpose. Besides, in case of a change in purpose, the individuals must be notified for fresh consent. Digital rights like the right to data correction etc., must be vested in the hands of the individuals. Moreover, consumer-facing privacy and data protection policies must be written in layman’s terms. Those documents must enhance the ease of exercising informed consent by making policy simple to understand.

- **Capacity:** AI deployers should invest in training and education programs to enhance their understanding of AI technologies and their potential applications. This can include attending workshops, webinars, or training sessions conducted by AI experts or industry professionals. Building knowledge and skills in AI can help users make informed

decisions, build capacity and effectively utilise AI systems. Further, AI deployers should embrace a culture of continuous adaptation. This involves staying updated on the latest developments in AI, exploring emerging technologies, and being open to incorporating new knowledge and approaches into their practices. Regularly assessing and reassessing AI strategies and adjusting them based on lessons learned can help build capacity and improve the effectiveness of AI systems.

3.3.3 Impact Population

In the context of AI, the term "impact population" refers to the individuals or groups who are directly affected by the deployment and use of AI systems. The impact population includes the end-users, customers, or beneficiaries of AI applications, as well as any stakeholders who may be affected by the outcomes or consequences of the AI system. These individuals or groups may experience the direct impact of AI-generated decisions, services, or products.

3.3.3.1. Direct Usage

During the direct usage stage, the end-users, who are individuals or groups within the impact population, interact with the AI system daily. These end-users could be individuals like us, utilising the AI solution to perform tasks, make decisions, or access services that are facilitated or enhanced by AI technology.

- **Awareness:** To build awareness, impact populations should be cognizant of the privacy implications of AI systems. They should understand the types of data being collected, how it is stored and secured, and the potential risks associated with using and disclosing personal or sensitive information. Being informed about privacy considerations enables individuals to make conscious decisions about the data they share and the level of control they have over their personal information when interacting with AI systems. Moreover, impact populations should recognise the impact of AI technology on fairness and human rights. They should know how AI systems can perpetuate biases or discriminate against certain groups, potentially amplifying existing societal inequalities. Besides, the impact population must also be cognizant of the fact that (a) in many cases, the AI developer and AI deployers follow the caveat emptor principle, and (b) all the outcomes generated through AI are not true. However, to ensure such information and awareness appropriately reaches the impact population, assistance from the private sector (both AI developers and AI deployers) and the Public sector (utilising public resources adequately) is essential. Besides, the awareness activities must capture specific requirements regarding vulnerable groups using AI systems, like children, elderly, disabled persons, gender minorities (women, LGBTQ+) etc.

- **Upskilling:** AI technologies and trends constantly evolve. Impact populations should stay updated with the latest developments in AI through industry publications, research papers, conferences, and webinars. This continuous learning will enable them to keep pace with advancements and leverage new opportunities AI systems offer. Further, to upskill themselves and foster a collective understanding of AI systems, impact populations should actively engage in open discussions and dialogues while also seeking to educate themselves. This can be achieved through participation in community forums, attending public meetings, or utilising online platforms dedicated to AI discussions. By actively participating in these conversations, individuals can share their unique perspectives, voice their concerns, and highlight their personal experiences related to AI systems. Engaging in such discussions helps to create a space for exchanging knowledge and insights, facilitating a broader understanding of the societal impact of AI. Through these interactions, impact populations can contribute to developing a well-informed community that is cognizant of the opportunities and challenges posed by AI systems, enabling them to make more informed decisions and actively shape the future of AI technologies. However, to facilitate the same and to scale such upskilling activities, assistance from the private sector (both AI developers and AI deployers) and the Public sector (utilising public resources adequately) is essential.

- **Responsibility:** When using AI technology, impact populations handle their data cautiously and ensure that sensitive personal information is not indiscriminately shared with AI systems. By being mindful of the data they input, individuals can protect their privacy and mitigate potential risks associated with the misuse or unauthorised access of personal data.

  Besides, it is important to combat the pre-existing beliefs; as a thumb rule, we as users should introspect whether the information we are about to share complies with our ideology. If it does, we have to take one step backwards and cross-check the integrity of the information by referencing multiple credible sources to cut the chain of misinformation. Even before applying the said thumb rule, we must be aware of our biases and ideologies. Confirmation bias (one of the cognitive biases) is inevitable, but confronting it helps us work our way through it. Besides, to make users aware of their biases, the pedagogical programs should conduct an implicit association test and also use the test results to customise the program accordingly.

4. Implementation of Principle-based Multistakeholder Approach

Coordination of various factors like regulatory landscape, geopolitics etc., is essential for the seamless implementation of the principle-based multistakeholder approach. In this section, we will discuss the government's role in implementing the principle-based multistakeholder

approach by establishing different forms of coordination. While there are various levels at which India could need coordination to adopt a principle-based data multistakeholder approach, in this chapter, we will discuss three essential levels, i.e., Domestic Coordination, International Coordination, and Public-Private Coordination.

4.1. Domestic Regulatory Coordination

The zero step towards implementing the principle-based multistakeholder approach would require domestic stability in terms of regulations. The primary regulatory issue would be recognising this framework as a legitimate lens to establish responsible AI innovations in India. If the regulation and enforcement fall under the ambit of multiple regulators domestically, discussed in this section, recognition of this framework might not be uniform as some might recognise it while others refrain from it. In addition, the existence of different regulators/authorities will pave the way for multifarious interpretation/understanding of the framework, which gives birth to slightly different versions of the principle-based multi-stakeholder approach at the implementation level, causing confusion and conflict. Moreover, this conflict and differences at the implementation level will impact AI innovations, causing compliance uncertainty and regulatory arbitrage. Therefore, consistent recognition and implementation of a principle-based multi-stakeholder approach at domestic regulatory levels are crucial.

Though laying down principle-based intervention that maps responsibilities and principles for various players within the AI ecosystem to support home-grown AI innovations is the way forward. However, concerns related to harmonising various existing/upcoming regulations and coordinating various ministries and sectoral regulators remain unaddressed. Though in the long term, it is ideal to have single consistent AI regulation for India as envisioned by the government[155], in the short term, we would require high-level coordination amongst the regulators and policymakers to recognise and implement the principle-based multistakeholder approach. The regulatory coordination envisioned must happen at two levels, as discussed below.

- **Horizontal Regulation:** Various existing and upcoming digital laws and regulations (horizontal regulatory frameworks) apply to all applications of AI, agonistic to the sectors. For instance, the upcoming Digital Personal Data Protection Bill 2022 (DPDPB 2022), will apply to AI developers who develop and facilitate AI technologies. AI developers will collect and use massive amounts of data to train their algorithms to enhance the AI solution; therefore, they might be classified as data fiduciaries. This implies that AI developers may comply with the key principles of privacy and data protection like purpose limitation, data minimisation, consensual processing, contextual

[155] Mathew, L. (2023, June 10). *Will bring regulations for AI to keep digital citizens safe: Minister*. The Indian Express. Retrieved June 20, 2023, from https://indianexpress.com/article/india/will-bring-regulations-for-ai-to-keep-digital-citizens-safe-minister-8655167/

integrity etc., as enshrined in DPDPB 2022. Besides, as contoured during Digital India Act (DIA) consultation, the government is also considering having provisions within DIA which would define and regulate high-risk AI systems. Moreover, the recent government has also expressed that there will be a separate overarching AI regulation for India.[156] On the other hand, some of the other non-tech regulations like Intellectual Protection rights (IPR) protections in India under the Patents Act 1970, Trademarks Act 1999 and the Copyright Act 1957, The Competition (Amendment) Act, 2023[157], Consumer Protection Act, 2019[158], Consumer Protection (Direct Selling) Rules, 2021[159] etc. also applies to both AI developers and AI deployers.

While the path the government takes through various policy instruments, as discussed above, is different, the end objective of these instruments together could make the AI ecosystem safe and responsible. Therefore, as these upcoming laws and existing legislations separately handle various concerns with AI solutions, we believe more effort is needed to establish coordination between various policy instruments such that different building blocks work in tandem to tackle harm posed by the technologies. The first step towards it is to have a consensus on the definition of AI solutions such that it clarifies which policy instruments apply to them. Followed by that could harmonise the applicable policy instruments through (a) weeding out the overlapping and conflicting scopes and bringing them to congruence with a proposed principle-based multistakeholder approach while enforced in a coordinated way, (b) extending the non-tech laws to recognise the principle-based multistakeholder approach such that they extend to the AI innovations within the digital realm. A similar set of strategies was proposed in the Report of the Financial Sector Legislative Reforms Commission (FSLRC)[160] to consolidate some of the provisions in financial regulation. For instance, while significant data fiduciaries under the upcoming DPDPB 2022 must appoint a privacy officer, how we align responsibilities between privacy officers and other internal officers who would be looking into other AI issues, including privacy, could be sorted through establishing coordination between different policy instruments.

---

[156] Mathew, L. (2023, June 10). *Will bring regulations for AI to keep digital citizens safe: Minister.* The Indian Express. Retrieved June 20, 2023, from https://indianexpress.com/article/india/will-bring-regulations-for-ai-to-keep-digital-citizens-safe-minister-8655167/

[157] Garg, R. (2023, May 27). *Analysis of competition (Amendment) Act, 2023*. iPleaders. Retrieved June 20, 2023, from https://blog.ipleaders.in/analysis-of-competition-amendment-act-2023/

[158] Mahawar, S. (2022, April 29). *Consumer Protection Act, 2019*. iPleaders. Retrieved June 20, 2023, from https://blog.ipleaders.in/consumer-protection-act-2019-2/

[159] Department of Consumer Affairs. (2021, December). *Consumer Protection (Direct Selling) Rules, 2021*. Ministry of Consumer Affairs Food and Public Distribution | Government of India. https://consumeraffairs.nic.in/sites/default/files/232214.pdf

[160] Mishra, A. R. (2012, October 1). *Committee for single financial sector authority*. mint. Retrieved June 20, 2023, from https://www.livemint.com/Politics/pRD4I0Wcj5T4UEEqpmHwgP/Committee-for-single-financial-sector-authority.html

- **Vertical Regulation:** In vertical regulation, notified use cases are regulated with sector-specific rules. An independent or established regulator regulates the nascent industry in such regulatory frameworks. The vertical regulatory frameworks and due diligence requirements for the financial, health, environmental sectors etc., would apply to sector-specific AI solutions. For instance, if AI-based fintech solutions engage in the activities of a payment aggregator, they would require authorisation from the Reserve Bank of India (RBI) and need to adhere to the technical and security-related recommendations suggested by the RBI.[161] Similarly, certain fintech providers are directly regulated by RBI by licensing them as Non-Banking Financial Companies[162] or Fintech (who may be an AI developer) indirectly regulated through regulated entities like banks, NBFCs etc. (who may be an AI deployer).[163] In the insurance sector, if an AI solution aids in online aggregation where the impact population could compare and choose the appropriate insurance, such technologies could require operationalisation approval from the Insurance Regulatory Development Authority of India.[164]

  Therefore, while the proposed principle-based multistakeholder approach is sector agnostic, the sectoral regulators need to recognise this approach to tailor the principles for various stakeholders to fit the needs and requirements within the respective sector.

4.2. International Regulatory Cooperation

While domestic regulatory coordination is crucial, there are also various other roadblocks to implementing the principle-based multistakeholder approach towards the AI ecosystem, which can't be solved exclusively at the domestic level. A concerted effort is needed between India and other jurisdictions beyond its borders to make AI innovations responsible and safe. In an increasingly interconnected world, international regulatory cooperation has emerged as a crucial pillar of regulatory policy[165]. Various jurisdictions have also emphasised this in the context of AI

[161] Bhalla, T., & Shukla, S. (2022, April 23). *RBI lens on companies seeking payment aggregator licence.* The Economic Times. Retrieved June 20, 2023, from https://economictimes.indiatimes.com/tech/startups/rbi-ups-scrutiny-on-fintechs-as-it-issues-payments-aggregator-licences/articleshow/91013336.cms?from=mdr

[162] Sood, N. (2023, June 8). *RBI releases new FLDG guidelines for banks and fintech lenders*. YourStory.com. Retrieved June 20, 2023, from https://yourstory.com/2023/06/rbi-guidelines-on-default-loss-guarantee-agreement-fldg-fintechs-bank

[163] Sood, N. (2023, June 8). *RBI releases new FLDG guidelines for banks and fintech lenders*. YourStory.com. Retrieved June 20, 2023, from https://yourstory.com/2023/06/rbi-guidelines-on-default-loss-guarantee-agreement-fldg-fintechs-bank

[164] *Rules and Regulations Relating to FinTech Laws in India*. (2023, February 6). Online Legal India. Retrieved June 20, 2023, from https://www.onlinelegalindia.com/blogs/fintech-laws-regulation-in-india

[165] OECD. (2021). *Why does international regulatory cooperation matter and what is it?* OECD iLibrary. Retrieved June 20, 2023, from https://www.oecd-ilibrary.org/sites/62c39d12-en/index.html?itemId=/content/component/62c39d12-en

governance, where they believe concerted international-level regulatory cooperation is the way forward.[166]

Box 4 - Importance of International Cooperation

There are several reasons why international regulatory cooperation is essential. Firstly, it helps to minimise regulatory fragmentation and inconsistencies that can hinder international trade and investment. Regulatory approaches and requirements differ significantly across countries, creating business barriers and complexities and limiting market access. By fostering cooperation and convergence, regulatory systems can be harmonised, reducing unnecessary regulatory burdens and facilitating smoother cross-border activities.

Secondly, international regulatory cooperation enables the exchange of knowledge, expertise, and experiences among regulatory authorities. It allows regulators to learn from each other's successes and challenges, identify emerging trends and risks, and develop more informed and effective regulatory strategies. Through dialogue and collaboration, countries can leverage collective intelligence and resources to develop robust regulatory frameworks that address common concerns such as public health, environmental protection, consumer safety, and financial stability.

Thirdly, international regulatory cooperation promotes regulatory coherence and enhances policy effectiveness. By aligning regulatory approaches and promoting the adoption of best practices, it improves the overall quality of regulations and enhances their efficiency and effectiveness. This reduces duplication, streamlines processes, and facilitates compliance for businesses operating in multiple jurisdictions. It also helps to ensure that regulations are evidence-based, proportionate, and responsive to societal needs and challenges.

Furthermore, international regulatory cooperation contributes to building trust and confidence among nations. By fostering dialogue, transparency, and collaboration, it strengthens relationships between regulatory authorities, promotes understanding, and resolves potential conflicts or disputes cooperatively. This trust-building is crucial for maintaining a stable and predictable global regulatory environment and fostering international cooperation on broader policy objectives, such as sustainable development, innovation, and the protection of public interest.

4.2.1. Principles of International Cooperation

Some of the key principles to be considered by the domestic regulators and governments in enhancing international-level coordination and cooperation are:

[166] Kerry, C. F., Meltzer, J. P., Renda, A., Engler, A., & Fanni, R. (2022, March 9). *Strengthening international cooperation on AI*. Brookings. Retrieved June 20, 2023, from https://www.brookings.edu/research/strengthening-international-cooperation-on-ai/

- **Balanced Discretion:** While the principles allow for domestic-level discretion in implementation, this act has to be balanced where interpretation is not too different from the preamble of the principle-based multistakeholder approach, i.e., building consensus through balancing differences in national constraints and practices while respecting international principles of Artificial Intelligence. Besides, the exemption must be less discretionary. Concertedly, countries must lay down fair procedures and scenarios for exemptions.

- **Trinity Thumb Rule:** While jurisdictions have various economic and national interests to cater to, countries must strive to follow the Trinity thumb rule, i.e., safety, cooperation and growth as part of any actions taken related to AI. These three elements also form the backbone of the principle-based multistakeholder approach. Besides, countries must strive for a positive-sum game and not compromise on one element to achieve the other.

- **Collaborative Formulation:** Governments must actively engage with the private sector businesses[167] and other policy actors while implementing the principle-based multistakeholder approach such that the operationalisation is smooth. Also, jurisdictions should work in tandem with businesses while defining vertical regulations, i.e., sector-specific rules.

- **Recognition of Distributed Accountability Principle:** The government and concerned regulators must acknowledge that different stakeholders in the AI lifecycle have varying responsibilities and liabilities based on the impact and harm they could inflate.

4.2.2. Means to Enable International Cooperation

There are various existing multilateral (both binding and non-binding)/multistakeholder arrangements that India could utilise to introduce a principle-based multistakeholder approach. The arrangements discussed in this section include agreements, strategies, and declarations to which India is currently a signatory, as well as arrangements to which India could potentially consider being a signatory in future for establishing responsible AI innovations.

- **Global Partnership on Artificial Intelligence Summit:** As a chair of the 2023 Global Partnership on Artificial Intelligence Summit, India hints towards initiating a conversation on creating a well-thought-through regulatory environment for AI. Therefore, the principle-based multistakeholder approach could contribute to this effort by initiating a rich multistakeholder and multilateral discussion at the global and

[167] Kerry, C. F., Meltzer, J. P., Renda, A., Engler, A., & Fanni, R. (2022, March 9). *Strengthening international cooperation on AI*. Brookings. Retrieved June 20, 2023, from https://www.brookings.edu/research/strengthening-international-cooperation-on-ai/

especially at the Asia-pacific level on tackling the AI issues at the ecosystem level involving various players beyond AI developers like AI deployers and impact population.

AI regulations are approached differently by India and other countries to cater to their respective domestic concerns and needs. However, our research on the cross-jurisdictional analysis of AI regulations and multilateral frameworks shows that there is potentially a principle-level congruence. We believe this similarity at the principle level could act as a means to initiate a conversation at GPAI to enable a principle-based multistakeholder approach for AI regulation through consensus building.

- **QUAD:** The QUAD members have expressed interest in terms of strengthening cooperation on the responsible development of AI and deploying this technology to transform the economy.[168] However, it has been reported that they face challenges in approaching governance of technological progress and geopolitics.[169] Therefore, leveraging this opportunity, India must introduce a principle-based multistakeholder approach with QUAD nations to enable responsible AI technological development.

- **UNESCO's Global Agreement on the Ethics of Artificial Intelligence:** About 193 member countries of UNESCO, including India, adopted this agreement to define shared values and principles for enabling the responsible development of AI innovations.[170] The principles and values defined in the agreement, like fairness, diversity, inclusivity etc., are similar to that of the proposed principle-based multistakeholder approach. Therefore, through the means of this agreement, India, in collaboration with UNESCO, could consider introducing the approach as the way forward in terms of implementing the principles meaningfully.

- **OECD Development Centre:** As India had joined the OECD development centre,[171] this could be an appropriately open and credible communication channel with the developing world on the principle-based multistakeholder approach for AI regulations as the centre acts as a forum for policy dialogue and comparative research into the emerging issue. In addition, India as a country sets precedence and benchmark for other global south

[168] Chahal, H., Luong, N., Abdulla, S., & Konaev, M. (2023, June 9). *Assessing AI-related Collaboration between the United States, Australia, India, and Japan.* Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/quad-ai/

[169] Chahal, H., Luong, N., Abdulla, S., & Konaev, M. (2023, June 9). *Assessing AI-related Collaboration between the United States, Australia, India, and Japan.* Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/quad-ai/

[170] Choudhary, A. (2021, December 1). *Ethics of AI: 193 members of UNESCO adopt recommendations*. Analytics India Magazine. Retrieved June 20, 2023, from https://analyticsindiamag.com/ethics-of-ai-193-members-of-unesco-adopt-recommendations/

[171] OECD. (2021, February). *India Joins OECD Development Centre.* OECD.org. https://www.oecd.org/newsroom/indiajoinsoecddevelopmentcentre.htm#:~:text=08%2F02%2F2001%20%2D%20The,countries%20and%20the%20developing%20world

countries, especially in south Asian countries, in terms of policy directions; striking a dialogue at the OECD development centre on a principle-based multistakeholder approach for AI regulations is an ideal way forward.

4.2. Establishing Public-Private Collaboration

Implementing the AI regulations is a fresh start for regulators and domestic industries in many jurisdictions, especially in the global south. The range of AI innovations to be tackled will be immensely vast, starting from big tech to MSMEs to government agencies. While a one-size-fits-all approach towards AI regulation might bring in compliance (at a cost) among the horizontally (AI general) and vertically (AI narrow) diverse range of AI developers and AI deployers, it might not bring cooperation. Therefore, governments must operationalise various market mechanisms to build a healthy relationship and cooperation with AI developers and AI deployers with a limited disposal capacity.

The governments could follow normative theories of regulation[172] and institute market mechanisms such as a (a) audit of features for AI developers and AI deployers based on the principles mapped for them and (b) market for principles-based accreditation, enabling a competitive edge for platforms. While an independent auditing agency must perform the audit, a government or authorised entity must perform the accreditation process at a nominal cost based on defined principles. The accreditation process must have a well-laid process and procedure that balances transparency and safeguards to protect intellectual and proprietary information. Besides, the accreditation process must be aspirational such that it pushes the AI developers and AI deployers toward performing better on the user outcome aspect, i.e., securing the impact population from the adverse implications of AI technologies.

5. Conclusion

Humans are the heart of the Internet, and everyone should benefit from the open and trustworthy Internet. However, the Internet is going through a paradigm shift driven by key technological developments like Artificial Intelligence. These technological developments pose challenges to the internet at different levels, like (a) gaps in the regulatory parameters, (b) technological differences, (c) lack of interoperability for networking, (d) safety and security concerns impacting trust etc. These challenges directly implicate how humans perceive the Internet's future, which is currently filled with anxiety and uncertainty, as highlighted by the previous version of the global Internet report.

[172] UNESCO. (2021, November). *Recommendation on the ethics of artificial intelligence.* https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence

Therefore, to transform the status quo, it is important to reinstate trust within disruptive technologies like Artificial Intelligence, which will be the face of the internet in the future. To achieve the same, there is a need for a governance framework which would enhance opportunities afforded by Artificial intelligence by making it trustworthy while minimising harm. Therefore, this is where our paper comes into the picture, adding value to efforts towards making AI development and deployment trustworthy by proposing an ecosystem-level principle-based approach which appropriately maps the harms and impact at the different stages and suggests principles for various stakeholders for tackling the same. Going further, this paper could set the context for future research on how the stakeholders can pragmatically put to action the identified principles and indicated operational strategies at scale.