The Dialogue™
INFORM ENGAGE IDEATE

RESEARCH PAPER

TOWARDS TRUSTWORTHY AI

SECTORAL GUIDELINES FOR RESPONSIBLE ADOPTION

Research Paper

Towards Trustworthy AI: Sectoral Guidelines for Responsible Adoption

AUTHORS

Rama Vedashree in association with **THE DIALOGUE**

Jameela Sahiba
Bhoomika Agarwal
Kamesh Shekar

Designer - Shivam Kulshrestha

The Dialogue™ is a public policy think tank with a vision to drive a progressive narrative in India's policy discourse. Founded in 2017, we believe in facilitating well-researched policy debates at various levels to help develop a more informed citizenry, on areas around technology and development issues. The DialogueTM has been ranked as the world's Top 10 think tanks to watch out for, by the Think Tank and Civil Societies Programme (TTCSP), University of Pennsylvania in their 2020 and 2021 rankings.

For more Information
https://thedialogue.co

Suggested Citation
Vedashree, R., Sahiba, J., Agarwal, B. & Shekar, K.(2024, February). Towards Trustworthy AI: Sectoral Guidelines for Responsible Adoption. The Dialogue™.

Catalogue No
TD/ET/RP/0224/01

Publication Date
Feb 8, 2024

Abhishek Singh, IAS
President & CEO
National e-Governance Division (NeGD), Ministry of Electronics and Information Technology (MeitY), Government of India.

Foreword

In the era of rapid technological evolution, the unprecedented growth and significance of artificial intelligence (AI) have become an undeniable hallmark of our global discourse. As we stand on the cusp of ground breaking advancements in the field of AI, we mustcollaboratively emphasize the trustworthy development, deployment, and utilization of this transformative technology.

Over the past several years, India has taken concrete steps to encourage the domestic adoption of AI in a responsible manner and build public trust in the use of this technology, placing the idea of 'AI for All' at its very core. India is pioneering the approach of harnessing the power of AI for social good, applying AI in education, healthcare, agriculture, languages, and other critical sectors. This aligns with the inclusive development philosophy of our Hon'ble Prime Minister, deeply rooted in the ethics of 'SabkaSaath, Sabka Vikas and SabkaPrayas'.

The Government of India has taken significant steps to enhance data protection and promote trustworthy AI practices. India is enhancing accountability with the Digital Personal Data Protection Act while developing the National Data Governance Policy to optimize ethical data accessibility for improving governance and public services. India is also developing Responsible AI frameworks and standards, demonstrating a comprehensive commitment to promoting Trustworthy AI practices.

As one of the fastest growing economies in the world, India is paving the way for fostering global cooperation on Trustworthy AI, emphasizing a balanced and inclusive discourse. As the Lead Chair for the Global Partnership on Artificial Intelligence (GPAI), this commitment is reinforced in the GPAI Ministerial Declaration- signed by 29 member countries including the European Union at the Annual GPAI Summit, hosted in New Delhi from 12th-14th December 2023.

In this context, it gives me great pleasure to see The Dialogue launch the research paper "Towards Trustworthy AI: Sectoral Guidelines for Responsible Adoption" which has provided a thorough exploration of Trustworthy AI principles and delineated actionable operational strategies for critical sectors. Aligned with our national commitments, the research paper also advocates for a coordinated regulatory approach to ensure the ethical and responsible governance of AI technologies.

This research paper is a valuable resource for policymakers, industry stakeholders, and researchers promoting the ethical advancement of artificial intelligence. As we move forward, let our shared commitment to Trustworthy AI principles be the cornerstone for all our endeavours in furthering the AI frontier.

Rama Vedashree
Former CEO, Data Security Council of India (DSCI)

Kazim Rizvi
Founding Director, The Dialogue

Foreword

We are witnessing a transformational shift in the history of humanity driven by the rise of artificial intelligence.

From Analogue to Digital

The post 2000 era witnessed the integration of technology with sectors driving the economy. Over the next two decades, the primary, secondary and tertiary sectors started getting digitised, which brought great economic value to India and the rest of the world. From finance to healthcare, education to entertainment - every sector saw unprecedented growth brought about by adding a layer or two of technology - driven by the rise of smartphone/high powered computers and internet access.

In healthcare alone, the global digital health market is expected to increase to over USD 500 billion by 2025. And in finance, total transaction value in the Digital Payments market alone is projected to reach USD 11.55 trillion in 2024. From the 1990s, when technology used to be just an afterthought, to the 2020s when we couldn't think of surviving the covid-era without the help of technology - we as a society have come a long, long way in the last 30 years. India embraced digitisation like a duck to water. Powered by the Digital India mission and a ~300 B$ Tech Industry, India's tech prowess is a force to reckon with, globally.

Integrating technology in our professional and personal lives has enhanced our efficiency and productivity, given rise to new businesses, reduced entry barriers for budding entrepreneurs, made public services accessible and inclusive - and it brought the world closer than ever. It helped enhance financial inclusion, made payments systems faster and efficient, improved the quality of healthcare delivery and brought G2C, B2C services to our homes.

From Digital to AI

We are now moving towards the post-digitisation world, which will be dominated by intelligent technological systems, which, from the outset, will reduce human effort and further scale-up our productivity. These systems, also known as Artificial Intelligence, are designed to "help" us go about our lives by "understanding" and "supporting" us. Much like what Alfred Pennyworth was to Batman. This second era of technological transformation is set to bring a much larger impact than the previous era, in a span of time probably less than a decade. This is further evident from the fact that the global artificial intelligence market size is projected to reach USD 1.8 trillion by 2030. We consider ourselves lucky to witness two great eras of tech revolution that continues to evolve our society. The AI era promises to bring even more prosperity, but it can also be fraught with some danger. And that is why we need to tread with caution.

Towards trusted intelligent systems

As much as the world is excited with the advent of AI, fears around a more divided society, loss of jobs, machine taking over humans, socio-economic discrimination, misuse and abuse, are not unfounded. In fact, these fears can threaten the very fabric of human consciousness and the future of humanity, so much so that many believe we are at the crossroads of machines potentially taking over men and women. Only time will tell whether that will

happen, but the question is - can we afford ourselves the luxury of hindsight when the pace of technological evolution is faster than what we can collectively grapple? But we cannot, and must not, inhibit technology and pace of innovation. AI is here to stay, and it is set to solve age old problems, augment human efficiency and make our lives easier. Policymakers therefore will have to find a way to ensure that they hit the sweet spot of maximising the opportunity and minimising potential harm. And the way forward is to build systems that can be trusted. Towards this, it would be imperative that the government engage in multi-stakeholder consultations including discussions with industry members, civil society organisations, legal experts, academia and other various stakeholders to manage diverse expectations and facilitate consensus.

The Nine Commandments of Trustworthy AI

With great hope and promise we embarked on this journey to identify and provide nine principles to make AI trustworthy. These principles, captured through various initiatives taken across the globe, are designed to make AI systems more robust, safer, fair, less biased, transparent, reliable and non-discriminatory. Along with outlining the principles, the paper attempts to outline the operational strategy for all three stakeholders across the AI value chain - the developers, the deployers and the users. It goes in depth to answer simple questions, such as - how do we operationalise non-discrimination? How do we make the systems safer? How do we reduce bias? What should AI developers keep in mind when they design the algorithms for the future? Have the deployers ensured they have everything in place to make systems trustworthy?

As policymakers in India are thinking about these questions, we hope that the paper gives pertinent insights. India demonstrated global leadership when it delivered a successful G-20 summit in 2023. And in 2024, India assumed the presidency of GPAI - a global partnership on AI to bring countries together. It is imperative that India should once again demonstrate leadership when it comes to setting the benchmark for responsible and trustworthy AI.

We are at the fork in the road. One way leads to great prosperity supported by state of the art technology. The other leads to fragmentation and division, which could lead to a total collapse. Trustworthy AI is that sweet spot of maximising opportunities and minimising harm. And if we don't pay attention now, Alfred Pennyworth could end up becoming The Joker.

Acknowledgement

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

Dr. Abhijnan Chakraborty, Assistant Professor, IIT Delhi: He is an Assistant Professor in the Department of Computer Science and Engineering at IIT Delhi. Prior to that, he worked at the Max Planck Institute for Software Systems (MPI-SWS) as a postdoctoral researcher.

Ameen Jauhar, Vidhi Centre for Legal Policy: Mr Ameen Jauhar is a senior resident fellow at the Vidhi Centre for Legal Policy and leads its Centre for Applied Law & Tech Research. Ameens focus areas include AI ethics and governance and the use of AI in legal and justice systems.

Arindrajit Basu, Centre for Internet & Society: Mr Arindrajit Basu is a Non-resident Fellow at the Centre for Internet & Society, India, where he focuses on the geopolitics and constitutionality of emerging technologies.

Arvind Sivaramakrishnan, Chief Information Officer, Karkinos Healthcare: Mr Sivaramakrishnan is CIO at Karkinos Healthcare, where he is responsible for driving the company's IT strategy and digital transformation.

Dr. Avik Sarkar, Indian School of Business: Dr. Avik Sarkar is a faculty at the Indian School of Business working in the areas of Data, Emerging Technology and Public Policy. Dr. Sarkar was the former Head of Data Analytics Cell and Officer on Special Duty (OSD) at NITI Aayog.

Beni Chugh, Dvara Research: Ms Beni Chugh manages the research at the Future of Finance Initiative. Her work focuses on identifying systemic stability and consumer protection concerns in digital finance.

Gaurav Agarwal, Co-founder, 1mg: Mr Gaurav Agarwal is the Co-founder & Chief Technology Officer at 1mg. He has a Bachelor's Degree from IIT Delhi.

Joshua Bamford, British High Commission: Mr Joshua is the Head of Tech and Innovation at the British High Commission in India. He was previously Head of the Trade and Development team at the United Kingdom Mission in Geneva and Head of WTO at the Foreign, Commonwealth and Development Office.

Laura Baldwin, British High Commission: Ms Laura is the South Asia Cyber Lead at the British High Commission in India. In the past, she has worked with the UK-Africa Investment Summit.

Mitesh Bidawatka, Head of Data Analytics, Jio Finance Limited: Mr Mitesh has 17 Years of extensive experience in using Data, Analytics & Technology to solve business problems. He has previously worked with companies like TransUnion, Larsen & Toubro Infotech Ltd, Capgemini and Cognizant.

Nikhil Narendran, Partner, Trilegal: Mr Nikhil Narendran is a Partner in Trilegal in the TMT and general corporate practice area. He has advised companies on a range of matters such as in relation to content liability, M&A, product compliances, data protection, information technology laws etc.

Nitendra Rajput, Senior Vice President, Mastercard: Mr Nitendra Rajput works as Senior Vice President and Head of AI Garage Center in India at Mastercard where he defines and leads the data science and machine learning work for all divisions of Mastercard.

Dr Rajeev Sharma, Vice President Medical Affairs, 1mg: Dr. Rajeev completed his degree from Bangalore Medical College and Research Institute.He has expertise in building systems and processes of clinical excellence as well as productising new technologies.

Contents

List of Figures

Executive Summary

The rapid integration of AI across diverse sectors holds immense potential to drive socio-economic progress. However, concerns regarding ethical and responsible implementation remain a critical roadblock. For AI to reach its full potential and benefit society, it requires widespread adoption and acceptance. Nevertheless, the lack of trust can act as a major barrier to achieving this, thereby dampening the innovation and potential of AI. Therefore, building trust in digital solutions powered by AI becomes paramount. Towards this end, the paper maps out principles that are requisite to building trustworthy AI systems. Further, building trustworthy AI systems demands a departure from generic frameworks and necessitates a nuanced approach that embraces the complexities and unique considerations of each domain. This research paper addresses this crucial need by presenting sectoral guidelines for responsible AI adoption, paving the way for trustworthy applications in two key areas: healthcare and finance. Existing ethical frameworks for AI often fall short by presenting broad, one-size-fits-all principles that fail to capture the intricate challenges and nuances unique to each sector. This paper bridges this gap by offering tailored frameworks for these two critical sectors, acknowledging the distinct ethical landscapes and regulatory environments within each. Our endeavour to do a deep-dive in two sectors, will surely encourage similar efforts in other sectors, where AI has become pervasive.

Beyond identifying potential pitfalls and ethical vulnerabilities, the paper delves deeper into specific challenges within each sector and proposes concrete recommendations for mitigating them. For example, the healthcare framework tackles bias in medical diagnoses and patient data privacy concerns, while the finance framework focuses on algorithmic transparency in automated financial decisions and robust cybersecurity measures. These actionable recommendations encompass practical suggestions for data management, algorithm audits, stakeholder engagement, and risk mitigation strategies, empowering diverse actors within each sector to implement AI responsibly. We were fortunate to be guided by experienced domain practitioners while developing the same.

This research paper fulfils two objectives: 1) systematically mapping the global landscape of AI frameworks with the intent of delineating globally recognised principles fundamental to building trustworthy AI, and 2) formulating practical guidelines for responsible AI adoption in high-impact sectors like healthcare and finance. The paper can be categorised into three key parts:

Part I: Navigating the AI Landscape

We conducted an extensive review of prominent AI policy frameworks, encompassing both national (NITI Aayog's National Strategy for AI, UK's A Pro-Innovation Approach to AI Regulation) and international initiatives (OECD AI Principles, G20 AI Principles, EU Ethics Guidelines for Trustworthy AI). Through a comparative analysis, we identified a core set of nine fundamental principles consistently underpinning trustworthy AI: i) transparency and explainability, ii) accountability, iii)

fairness and non-discrimination, iv) reliability and safety/robustness, v) human autonomy and determination, vi) privacy and data protection, vii) social and environmental sustainability, viii) governance and oversight, and ix) contestability. This typology of principles serves as a critical roadmap for navigating the complex ethical considerations in AI development and deployment.

Part II: Operationalizing Trustworthy AI in Action

We shift our focus towards the practical application of these principles in two strategically selected sectors: healthcare and finance. These domains grapple with sensitive data, algorithmic decision-making impacting lives and livelihoods, necessitating a nuanced approach to responsible AI adoption. We propose a comprehensive set of technical and non-technical approaches for operationalizing trustworthiness within these sectors. This includes leveraging explainable AI techniques, bias detection algorithms, fostering user education and awareness, and implementing ethical design principles throughout the AI development lifecycle. Recognizing the vital role of governance, we advocate for robust frameworks and oversight mechanisms tailored to each sector's specific needs and vulnerabilities.

Part III: Building a Trustworthy AI Ecosystem

We acknowledge that genuine progress towards trustworthy AI demands a concerted effort beyond individual actors. Therefore, we identify key drivers for responsible adoption at various levels:

- Domestically: Aligning with existing legal frameworks while fostering adaptability to the rapidly evolving AI landscape.
- Internationally: Promoting harmonisation of global AI regulations to establish a unified ethical foundation transcending national borders.
- Public-Private Partnerships: Harnessing the power of collaborative efforts between governments, industry leaders, academics, and civil society to incentivize ethical practices and develop effective safeguards.

By combining a rigorous understanding of ethical principles with sector-specific operationalization strategies and a multi-stakeholder approach to implementation, we seek to offer a roadmap for responsible and ethical AI integration across diverse domains. Achieving trustworthy AI necessitates a shift from generic principles to context-sensitive frameworks that acknowledge the intricate nuances of each sector. This paper's proposed sectoral frameworks would contribute significantly to this crucial endeavour, paving the way for responsible and ethical AI integration across diverse realms. By embracing this nuanced approach, we can collectively build a future where AI serves not as a source of uncertainty and apprehension, but as a force for good, driving progress and solidifying trust in an increasingly technology-driven world.

Abbreviations

Abbreviations	Expanded form
ABDM	Ayushman Bharat Digital Mission
ABHA	Ayushman Bharat Health Account
AI	Artificial Intelligence
API	Application Programming Interface
CDSCO	Central Drugs Standard Control Organization
CSR	Corporate Social Responsibility
CT	Computed Tomography
DCGI	Drugs Controller General of India
DPB	Data Protection Board
DPDPA	Digital Personal Data Protection Act
DSIT	Department for Science Innovation and Technology
EHR	Electronic Health Record
ESG	Environmental, Social and Corporate Governance
EU	European Union
EU-US TTC	European Union-United States Trade and Technology Council
FAQs	Frequently Asked Questions
FDA	Food and Drug Administration
FEAT	Fairness, Ethics, Accountability, and Transparency
FM	Foundational Model

Acronym	Expanded form
FMEA	Failure Mode and Effects Analysis
GDPR	General Data Protection Regulation
GenAI	Generative AI
GPAI	Global Partnership on Artificial Intelligence
Grad-CAM	Gradient-weighted Class Activation Mapping
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
HDMP	Health Data Management Policy
HIPAA	Health Insurance Portability and Accountability
HTTPS	Hyper Text Transfer Protocol Secure
ICMR	Indian Council of Medical Research
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
IVR	Interactive Voice Response
KPI	Key Performance Indicators
LIME	Local Interpretable Model-Agnostic Explanations
MAS	Monetary Authority of Singapore
METI	Ministry of Economy, Trade and Industry
MFA	Multi-factor Authentication
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NDHE	National Digital Health Ecosystem
NHA	National Health Authority

Acronym	Expanded form
NHS	National Health Service
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
OECD	Organization for Economic Cooperation and Development
PII	Personally Identifiable Information
PHR	Personal Health Record
PRA	Probabilistic Risk Assessment
RBAC	Role-based Access Control
RBI	Reserve Bank of India
RMF	Risk Management Framework
SaMD	Software as a Medical Device
SDG	Sustainable Development Goals
SEC	Securities and Exchange Commission
SEBI	Securities Exchange Board of India
SHAP	Shapley Additive Explanations
SOP	Standard Operating Procedures
UN	United Nations
UNESCO	United Nations Educational, Scientific and Cultural Organization
USA	United States of America
UX	User Experience
XAI	Explainable AI
XG Boost	eXtreme Gradient Boosting

1 Introduction

Artificial Intelligence (AI) stands at the cusp of a transformative wave that promises to revolutionize industries, societies, and economies. Its potential to enhance efficiency, deliver innovative solutions, and optimize processes is undeniable. One of the most remarkable attributes of AI is its ability to enhance efficiency. In sectors such as finance, healthcare, education, etc. AI streamlines operations, optimizes resource allocation, and accelerates decision-making processes. Whether it's automating routine tasks, conducting complex data analysis, or personalizing user experiences, AI offers a toolbox of transformative capabilities. These efficiency gains translate to cost savings, improved productivity, and the creation of new opportunities for innovation.

Moreover, AI holds the promise of delivering innovative solutions to long-standing challenges. In healthcare, AI-driven diagnostic tools and drug discovery processes have the potential to revolutionize patient care and disease management. In finance, predictive algorithms enhance risk assessment, fraud detection, and investment strategies. In education, AI-powered personalized learning experiences cater to individual student needs, fostering better learning outcomes. These innovative applications span numerous sectors, offering game-changing possibilities that were once the realm of science fiction. However, this transformative power of AI is not without its complexities. While AI offers incredible opportunities, it concurrently raises a host of intricate challenges. The rapid proliferation of these technologies has spawned concerns around data privacy, user safety, and the potential displacement of jobs.

As AI reshapes the landscape, it brings to the forefront the critical need for responsible AI governance. Sustainable development and light-touch, principle-based regulation are essential to harness the full potential of AI while addressing the challenges it presents. The delicate balance between reaping the benefits of AI and mitigating unintended consequences hinges on building trustworthiness in AI systems. Building this trust hinges on the fundamental concept that trust serves as the bedrock for the flourishing of societies, economies, and sustainable development. Consequently, the realization of AI's boundless potential on a global scale necessitates the establishment of trust in its capabilities. The perception of AI as trustworthy by its users, such as consumers, organizations, and society at large, is contingent upon its development, deployment, and utilization in a manner that goes beyond mere legal compliance and robust functionality. Crucially, trust in AI is solidified when it aligns closely with overarching ethical principles. This means that not only must AI systems adhere to applicable laws and demonstrate resilience, but they must also prioritize and uphold broader ethical considerations to instill confidence among users.[1] Trustworthiness encompasses critical attributes such as

[1] Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence. Electronic Markets, 31(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4

fairness, transparency, accountability, and security, which are vital in addressing concerns related to data privacy, bias, security, and accountability. This fundamental building of trust not only mitigates unintended consequences but also has profound and lasting effects on the long-term acceptance and prosperity of AI technologies. Trustworthy AI systems engender sustainable adoption by instilling confidence among users and stakeholders, promoting ethical AI practices, contributing to positive economic and social impacts, and facilitating regulatory compliance in an evolving landscape.

In this context, the quest for trustworthy AI principles and sectoral guidelines becomes pivotal. These principles and guidelines provide the roadmap for ensuring that AI technologies align with the highest standards of ethics, security, and reliability. As industries navigate this transformative journey, the responsible adoption of AI emerges as the linchpin for realizing the full potential of these powerful technologies while safeguarding against their inherent challenges.

Against this backdrop, this paper embarks on a profound journey, recognizing that trust in AI is fundamental and that trustworthiness is the bedrock of this assurance. Our journey begins by mapping the intricate landscape of trustworthy AI principles. We undertake an exhaustive study of global AI regulations and standards, distilling from them the core principles that underpin trustworthiness. Principles that are universal, transparent, and capable of serving as a compass in the ever-evolving AI terrain.

We then delve into the practical realm, operationalizing these principles across two critical sectors of finance and healthcare. These sectors epitomize a diverse spectrum of industries undergoing transformative changes, thereby providing a comprehensive lens through which to understand the nuanced challenges and opportunities associated with AI integration. Trust and responsibility emerge as central tenets in these sectors due to the sensitive nature of financial transactions and the critical importance of AI enabled decisions in patient care. In finance, where AI algorithms influence transactions, investments, and risk management, establishing and maintaining trust is paramount for user confidence and the overall stability of economic systems. Similarly, in healthcare, where AI holds promises for personalized diagnostics and treatment, fostering trust is essential for ensuring patient confidence and ethical use of sensitive medical data. Therefore, by operationalizing AI principles in finance and healthcare, we aim not only to address the intricacies specific to these sectors but also to derive insights that can inform trustworthy AI practices across diverse industries. This targeted exploration will contribute to the broader discourse on trustworthy AI adoption, offering nuanced perspectives and guidelines tailored to the unique challenges presented by these critical pillars of societal infrastructure. The endeavor aligns with the imperative to lay the foundation for ethical AI integration, ensuring that the transformative power of artificial intelligence aligns seamlessly with societal values, trust, and responsible governance.

Next, the paper explores the mechanics of implementing a principle-based framework for trustworthy AI. This is not a solitary endeavor; it necessitates concerted efforts at multiple

levels. We investigate the levers essential for implementing such a framework at domestic, international, and public-private partnership levels. Domestic coordination is crucial, where regulations must align with existing sectoral regulations and adapt to the ever-evolving AI landscape. International cooperation becomes imperative, where harmonizing AI regulations is essential to establish a consistent and ethical AI adoption across borders. The global community must converge on shared principles and standards that uphold the responsible adoption of AI. In this landscape, public-private partnerships emerge as a pivotal force. By leveraging market mechanisms, these partnerships can promote responsible AI integration, encouraging developers to prioritize consumer protection and safety as a fundamental value proposition. This aligns market forces with the overarching goal of creating trustworthy AI solutions.

In conclusion, our goal is clear: to usher in a future where AI technologies serve as a force for positive change, one that is responsible, ethical, and aligned with the best interests of society and the economy. Through our collective efforts, we aspire to pave the way for the responsible and trustworthy adoption of AI in diverse sectors, starting with two critical sectors such as finance and healthcare.

1.1 Methodology

The methodology deployed for this paper is grounded in a comprehensive approach that combines one-on-one engagements with relevant stakeholders and extensive secondary research. The primary sources of information were derived from direct interactions with key stakeholders, including industry experts, policymakers, and professionals from the health and finance sectors. These one-on-one engagements provided invaluable insights into the practical nuances and challenges associated with the adoption of AI technologies in these critical domains. Complementing the firsthand information, extensive secondary research was conducted to ensure a robust and well-rounded understanding of existing guidelines, regulatory frameworks, and best practices.

2 Typology of AI Principles

In recent years, the rapid advancement of artificial intelligence (AI) has sparked a global conversation about the ethical implications of this transformative technology. Discussions surrounding regulating this fast-paced technology often revolve around the delicate balance between mitigating potential risks and promoting innovation and adoption. However, it is important to recognise that these two objectives are not mutually exclusive but rather interconnected.

Therefore, through this paper, we aim to transcend the idea of a trade-off between risk mitigation and innovation/adoption, and instead emphasise the importance of a balanced approach that acknowledges the interconnectedness of these objectives. Further, navigating the expansive landscape of AI governance frameworks reveals a proliferation of principle-based approaches. The diversity in these frameworks, while reflective of the multifaceted considerations in AI governance, can be daunting. However, amidst this multitude, a discernible pattern emerges – a convergence of principles that forms a common thread across these frameworks. Despite the apparent variations, foundational similarity exists at the principle level. Leveraging the alignment of these principles provides a strategic pathway to streamline the myriad approaches, fostering a unified and standardized framework. To accomplish the same, in this chapter, we map and analyze trustworthy AI principles by conducting a comprehensive landscape study of regulatory frameworks from around the world. The principles discussed in this section may serve as guiding values, shaping the development and implementation of governance frameworks pertaining to AI.

2.1 Landscape Study

This section embarks on a landscape study of the various ethical AI frameworks that have emerged across the globe. It explores the different approaches taken by governments and organizations worldwide in formulating ethical AI frameworks. By examining the motivations, objectives, and methods employed in these frameworks, we aim to assess the key principles they espouse and their potential impact on AI development and deployment.

Box 1: Purpose of the landscape study

The purpose of this study is twofold.

Firstly, it provides a comprehensive overview of the diverse range of currently published AI ethical frameworks. These frameworks encompass a wide spectrum of principles, guidelines, and values designed to govern the development, deployment, and use of AI systems. By examining and analyzing these frameworks, we gain valuable insights into the common principles, variations, and challenges associated with evolving ethics frameworks for AI.

Secondly, it identifies a typology of trustworthy AI principles specifically tailored to the objectives of this paper. With the abundance of ethical AI frameworks, it becomes crucial to identify the key elements that contribute to the credibility and effectiveness of these principles. By studying the landscape of AI ethical frameworks, we can identify the core principles that consistently emerge across multiple frameworks, thereby enabling the formulation of a typology of trustworthy AI principles.

Thirdly, the findings of this landscape study will help showcase a principle-based congruence at the global level, fostering discussions on a global scale and facilitating consensus building among diverse stakeholders.

2.1.1 NITI Aayog's National Strategy for Artificial Intelligence[2]

The strategy document envisions five key areas where the deployment of AI can prove to be revolutionary. These include: a) healthcare, b) agriculture, c) education, d) smart cities and infrastructure and e) smart mobility and transportation. It identifies challenges like lack of enabling data regime, inadequate research capacity, unclear ethical regulations etc., that might potentially hinder the realisation of the full potential of AI. To tackle these challenges, certain recommendations have been enumerated that would help in realisation of #AIforAll, i.e. developing an AI roadmap which is inclusive and beneficial for all. These recommendations include strengthening and incentivising national and international research capacity through the establishment of research centres, re-skilling of the workforce, accelerating the adoption of AI across the value chain through the creation of a multi-stakeholder marketplace that would help address information asymmetry among small players and developing responsible AI. It identifies three pillars as integral to the development of 'Responsible AI', including ethics, privacy and security. An ethical AI would encompass fairness (bias elimination) and transparency (explainability). Privacy in AI would involve informed user consent in

[2] NITI Aayog. (2018). National Strategy for Artificial Intelligence #AIFORALL. https://niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf

addition to other principles inherent to a data protection framework, including data minimisation, whereas security in AI emphasises on determining the party at fault and attributing accountability.

Based on the national strategic blueprint, a series of measures were enacted. An exemplar instance is the "Responsible AI #AIFORALL" flagship AI initiative, led by NITI Aayog in 2021, which has yielded a compendium of Responsible AI approach documents produced in collaboration with the World Economic Forum Centre for the Fourth Industrial Revolution. The initial segment, titled 'Approach Document for India Part 1, Principles for Responsible AI,' was released in February 2021, constituting an extension of the underlying National Strategy for AI.[3] The fundamental objective of this Approach Document resides in the formulation of comprehensive ethical precepts governing the conception, evolution, and implementation of AI technologies within the context of India.

Subsequently, the 'Approach Document for India: Part 2 - Operationalizing Principles for Responsible AI,' was issued in August 2021. This document operationalizes the seven guiding principles that originated from the antecedent phase of the approach. Furthermore, it elaborates upon the governmental and interdisciplinary frameworks while proffering recommendations directed towards the private sector, research entities, academic institutions, and other pertinent stakeholders.[4]

Recently, the Economic Advisory Council to the Prime Minister released a working paper wherein it has suggested a Complex Adaptive System Framework (CAS) to regulate AI. Under CAS, guardrails and partitions would be established to limit undesirable AI behaviour.[5]

2.1.2 OECD AI Principles[6]

OECD AI Principles are a set of internationally agreed principles that seek to promote human-centric AI. The document is divided into two parts: first, it delineates five key principles that all AI actors are encouraged to adopt for responsible stewardship of trustworthy AI. These principles include: 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 document stresses the complementary nature of these principles. The second part of the legal instrument lays down recommendations for countries to help them implement the above-mentioned principles. The recommendations range from facilitating investment in R&D for fostering innovation in trustworthy AI to framing enabling policies and increased cooperation at international forums.

In 2021, the OECD released a report assessing how governments have implemented policy recommendations from the OECD Principles on Artificial Intelligence.[7] The report highlighted effective practices, and explored emerging trends in AI policy. The main focus

[3] NITI Aayog (2021) Responsible AI #AIFORALL: Approach Document for India Part 1 – Principles for Responsible AI. https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf
[4] NITI Aayog (2021) Responsible AI #AIFORALL: Approach Document for India: Part 2 - Operationalizing Principles for Responsible AI. (2021). https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf
[5] EACPM (2024) A Complex Adaptive System Framework to Regulate Artificial Intelligence. https://eacpm.gov.in/wp-content/uploads/2024/01/EACPM_AI_WP-1.pdf
[6] OECD. (2019). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449). Adopted on May 22, 2019. Amended on November 8, 2023. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
[7] OECD. (2021). State of implementation of the OECD AI Principles: Insights from national AI policies (No. 311). OECD Digital Economy Papers. OECD Publishing, Paris. https://doi.org/10.1787/1cd40c44-en

was on how different countries were putting into action the five recommendations for policymakers outlined in the OECD AI Principles. Recently, in May 2023, Secretary-General Mathias Cormann announced that the OECD is planning to update its guidelines to include rules for generative AI.[8]

2.1.3 G20 AI Principles[9]

Drawing reference from OECD principles, the G20 also adopted identical principles for responsible stewardship of trustworthy AI in June 2019, so as to promote and implement: (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 aim is to foster beneficial outcomes, including augmenting human capabilities, reducing inequalities, and protecting the environment. The principles of transparency and responsible disclosure enable informed decision-making, while robustness, security, and safety mitigate risks. These principles emphasize traceability, risk management, and accountability, addressing concerns such as privacy, digital security, safety, and bias.

During the recent G20 Summit in India, a historic milestone was achieved with the endorsement of the New Delhi Leaders' Declaration, reflecting a collective dedication to actively address the challenges and opportunities posed by Artificial Intelligence (AI). Aligned with the New Delhi Declaration, the G20 nations have chosen to adopt a "pro innovation" regulatory approach, emphasizing the optimization of AI benefits while judiciously managing associated risks. Notably, the Declaration underscores the commitment to the G20 AI Principles of 2019, which provide comprehensive guidelines for the "responsible stewardship" of "Trustworthy AI." This commitment reinforces the importance of ethical AI practices and responsible deployment. Furthermore, the G20 has pledged to facilitate the exchange of information regarding approaches to leverage AI for solutions in the digital economy. In a visionary move, the G20 countries have resolved to champion the use of responsible AI as a strategic tool for advancing Sustainable Development Goals (SDGs), emphasizing the technology's potential to contribute positively to global societal and economic objectives.

2.1.4 Australia's AI Ethics Framework[10]

Australia has established 8 AI Ethics Principles aimed at ensuring the safety, security, and reliability of AI systems. These principles serve several purposes: (a) promoting safer and fairer outcomes, (b) reducing the risk of negative impacts on those affected by AI applications, and (c) encouraging businesses and governments to uphold high ethical standards throughout the AI design, development, and implementation processes.

The key principles include prioritising (a) human, societal, and environmental well-being, (b) respecting human rights and autonomy, (c) ensuring fairness and non-discrimination, (d) protecting privacy and data security, (e) maintaining reliability and safety, (f) promoting transparency and

[8] Taguchi, S., & Tsuji, T. (2023, May 29). OECD pursues guidelines on regulating generative AI: leader. Nikkei Asia. Retrieved January 10, 2024, from https://asia.nikkei.com/Editor-s-Picks/Interview/OECD-pursues-guidelines-on-regulating-generative-AI-leader

[9] G20 Trade Ministers and Digital Economy Ministers. (2019, June 9). Ministerial Statement on Trade and Digital Economy. G20 Ministerial Statement on Trade and Digital Economy. https://wp.oecd.ai/app/uploads/2021/06/G20-AI-Principles.pdf

[10] Department of Industry, Science and Resources. (2022). Australia's AI Ethics Principles. Australia's Artificial Intelligence Ethics Framework | Department of Industry, Science and Resources.
https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

explainability, (g) allowing contestability when significant impacts occur, and (h) establishing accountability for the outcomes of AI systems. These principles emphasise the importance of AI systems benefiting individuals, respecting diversity, and upholding privacy rights and data protection. They also emphasise the need for reliable operation, transparency, responsible disclosure, human oversight and accountability throughout the AI system lifecycle.

Recently, the Australian Government announced its intention to regulate AI more stringently and has further released a consultation paper to ensure AI is used in a responsible and safe manner.[11]

2.1.5 EU Ethics Guidelines for Trustworthy AI[12]

The European Commission constituted a High-Level Expert Group on Artificial Intelligence to develop guidelines for the promotion of trustworthy AI. The Guidelines identify three components of trustworthy AI: lawful, ethical and robust. Using fundamental rights as the basis for developing trustworthy AI, the guidelines devise four ethical principles that should be adhered to during the development, deployment and usage of AI: (i) Respect for human autonomy, (ii) Prevention of harm, (iii) Fairness (iv) Explicability (transparency, openness, explainability). Building on these principles, seven requirements are delineated that can be met through technical and non-technical methods. These include: a) Human agency and oversight, b) Technical robustness and safety, c) Privacy and data governance, d) Transparency, e) Diversity, non-discrimination and fairness, f) Societal and environmental well-being , g) Accountability. The guidelines further provide an assessment list for the actors to ensure that the AI complies with these principles. The guidelines acknowledge the possibility of potential conflicts between principles and emphasises the need for determining trade-offs based on evidence and reason.

In February 2020, the European Commission expanded upon the previously established guidelines through its white paper titled "On Artificial Intelligence: A European Approach to Excellence and Trust."[13] This document introduced forthcoming regulatory measures and outlined the fundamental components of the prospective regulatory framework. One pivotal aspect underscored was the adoption of a risk-based paradigm, advocating the imposition of obligatory legal requisites rooted in ethical principles upon AI systems deemed of high risk. Building on the paper and subsequent consultation, the AI Act was introduced.

2.1.6 EU-US TTC Joint Roadmap for Trustworthy AI and Risk Management[14]

The U.S.-EU Joint Statement of the Trade and Technology Council in May 2022 expressed a commitment to developing a Joint Roadmap

[11] Taylor, J. (2023, June 2). Australia is looking to regulate AI – what might they be used for and what could go wrong? | Artificial intelligence (AI) | The Guardian. Retrieved January 10, 2024, from the Guardian website: https://www.theguardian.com/technology/2023/jun/03/australia-is-looking-to-regulate-ai-what-might-they-be-used-for-and-what-could-go-wrong

[12] High-Level Expert Group on Artificial Intelligence set up by the European Commission. (2019, April 8). Ethics Guidelines for Trustworthy AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[13] European Commission. (2020, February 19). White Paper on Artificial Intelligence: A European approach to excellence and trust. Retrieved January 10, 2024, from European Commission website: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[14] European Commission. (2022, December 2). TTC Joint Roadmap for Trustworthy AI and Risk Management. https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management

on evaluation and measurement tools for trustworthy AI and risk management. The purpose of this roadmap was to minimise the negative impacts of AI systems while maximising their positive contributions, aligned with the shared values of democratic societies. The roadmap aims to guide the EU and the United States in developing tools, methodologies, and approaches for AI risk management and trustworthy AI. Both sides recognise the significance of procedures that advance transparency, openness, fair processes, impartiality, and inclusiveness in shaping these standards. By prioritizing these principles, the roadmap seeks to ensure that AI standards support the values of safety, security, fairness, and non-discrimination, while fostering innovation and compatibility in diverse markets. It also seeks to support international standardisation efforts and promote trustworthy AI based on a shared dedication to democratic values and human rights. The roadmap emphasises practical steps to advance trustworthy AI and uphold the OECD Recommendation on AI, reflecting a collective commitment to responsible AI development and deployment.

2.1.7 NIST's AI Risk Management Framework[15]

NIST AI Risk Management Framework (AI RMF 1.0) was released in January 2023 with the intent to offer practical guidance to organisations on identifying and managing risks arising from AI and promote trustworthy development and use of AI systems while allowing organisations the flexibility to operationalise the principles in differing capacities as and when needed. The first section outlines the characteristics of trustworthy AI systems and analyses how businesses might interpret the risks associated with AI. Characteristics of trustworthy AI systems include: (a) valid and reliable, (b) safe, secure and resilient, (c) accountable and transparent, (d) explainable and interpretable, (e) privacy-enhanced, and (f) fair with harmful bias managed. Section two outlines four particular tasks that can be performed to help enterprises manage the risks associated with AI systems. These include 'Govern', 'Map', 'Measure' and 'Manage'.

2.1.8 Germany Artificial Intelligence Strategy 2018[16]

On 15th November 2018, the German Cabinet adopted the AI Strategy, to promote the development and implementation of artificial intelligence in the country. In December 2020, an updated AI Strategy was released in response to new developments in the AI field. This update refines, strengthens, and supplements the measures to support AI in Germany and Europe. As part of the implementation of the AI strategy, the German Government focuses on the AI standardization roadmap. This includes the development of test criteria to evaluate the robustness, safety, security, reliability, integrity, transparency, explainability, interpretability, and non-discrimination of (hybrid) AI systems. The strategy emphasizes the importance of promoting AI development, increasing funding, and ensuring the robustness and ethical standards of AI systems.

[15] U.S. Department of Commerce. (2023, January). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

[16] German Federal Government. (2020, December). Artificial Intelligence Strategy: 2020 Update. https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf

2.1.9 Singapore National AI Strategy 2019[17]

With a vision to be the global leader in developing and deploying scalability by 2030, Singapore adopted its AI Strategy in 2018, reflecting a comprehensive approach to accelerating AI deployment in key sectors while addressing risks and ensuring societal readiness. The Strategy focuses on the effective deployment of AI solutions through collaboration among public, private, and research institutions. As per the strategy, three key aspects guide the effective deployment of AI in Singapore. Firstly, the focus is on leveraging AI to serve human needs and deliver benefits to citizens and businesses. This approach aligns with Singapore's Smart Nation initiative and emphasizes the application of AI technology for practical purposes rather than its development for its own sake. Secondly, the strategy addresses the risks and governance challenges associated with increased AI usage. It emphasizes the need to preserve societal and institutional responsibilities and accountabilities in the face of automation, detection, and prediction facilitated by AI systems. Thirdly, the strategy aims to build an AI-ready population and workforce. At the societal level, digital literacy promotion includes raising awareness of AI to prepare citizens for technological change and engage them in discussions about its implications.

In February 2022, the Monetary Authority of Singapore (MAS) further unveiled five white papers that elucidate evaluation methodologies concerning the principles of Fairness, Ethics, Accountability, and Transparency (FEAT). These papers are designed to provide guidance to financial institutions (FIs) in responsibly employing artificial intelligence (AI).[18] In January 2024, Singapore's Infocomm Media Development Authority (IMDA) announced a public consultation on its draft Model AI Governance Framework for Generative AI showing the potential future policy interventions relating to generative AI.[19]

2.1.10 US's Blueprint for AI Bill of Rights and USA's National Artificial Intelligence Research and Development Strategic Plan 2023[20]

In a significant move to address societal concerns regarding the use of Artificial Intelligence (AI), the White House introduced a Blueprint for an AI Bill of Rights. This comprehensive document delineates five fundamental protections that every individual in America should enjoy when engaging with AI and automated systems. These protections encompass ensuring the safety and effectiveness of AI systems, preventing algorithmic discrimination, safeguarding data privacy, providing clear notice and explanation of AI processes, and advocating for the consideration and development of human alternatives and fallback mechanisms. In tandem with this initiative, NIST released a framework in January 2023, aimed at enhancing the management of risks

[17] Singapore. (2019). National Artificial Intelligence Strategy. Singapore National AI Strategy. https://wp.oecd.ai/app/uploads/2021/12/Singapore_National_Artificial_Intelligence_Strategy_2019.pdf
[18] Joshi, K. (2022, November 10). Singapore Guidelines on Artificial Intelligence: How Singapore Policies Impact the Future of AI. https://xai.arya.ai/article/singapore-guidelines-on-artificial-intelligence-how-singapore-policies-impact-the-future-of-ai
[19] IMDA & AI Verify Foundation (2024) Proposed Model AI Governance Framework For Generative AI: Fostering A Trusted Ecosystem. https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf
[20] Select Committee on Artificial Intelligence of the National Science and Technology Council. (2023, May). National Artificial Intelligence Research and Development Strategic Plan: 2023 Update. https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf

associated with AI for individuals, organizations, and society at large. The synergy between these frameworks not only offers valuable guidance to researchers but also presents avenues for further research and exploration in the evolving landscape of AI governance.

The USA's National Artificial Intelligence Research and Development (R&D) Strategic Plan 2023 builds upon previous strategic plans (2016[21], 2019[22]) and outlines a principled and coordinated approach to advancing AI research. The plan encompasses nine key strategies, emphasizing the importance of long-term investments in fundamental and responsible AI research to drive innovation and maintain the USA's leadership position in AI. The strategies cover a wide range of areas, including the development of effective methods for human-AI collaboration, understanding and addressing the ethical, legal, and societal implications of AI, ensuring the safety and security of AI systems, and developing shared public datasets and environments for AI training and testing. The plan also highlights the need to measure and evaluate AI systems through standards and benchmarks, understand the national AI R&D workforce needs, and expand public-private partnerships to accelerate advances in AI. Additionally, the plan underscores the significance of establishing a principled and coordinated approach to international collaboration in AI research. This strategy prioritizes international partnerships to address global challenges, such as environmental sustainability, healthcare, and manufacturing. By fostering responsible progress in AI R&D and collaborating on international guidelines and standards, the USA aims to drive innovation, promote equity, and address critical societal issues through AI.

More recently, the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (EO)[23] was released by the White House. In the order, the US relies on voluntary disclosures to authorities, emphasizing the development of AI safety standards and prioritizing the safety, security, and trustworthiness of AI systems without explicit prohibitions on applications like social scoring or facial recognition. The EO takes a more proactive approach underscoring the significance of research and talent development and actively promoting AI innovation across diverse sectors, including healthcare and climate change.

2.1.11 France's AI for Humanity 2017[24]

France's AI strategy for humanity places a strong emphasis on leveraging artificial intelligence (AI) to drive growth and create job opportunities within various industries. The strategy focuses on several key areas, including trusted, explainable, and certifiable AI systems. This highlights the importance of building AI solutions that are transparent, trustworthy, and capable of meeting certification standards. One significant aspect of the strategy is the integration of AI into embedded systems, which are autonomous electronic systems used to perform specific tasks. This involves utilizing AI for various purposes such as design, simulation,

[21] National Science and Technology Council. (2016, October). The National Artificial Intelligence Research and Development Strategic Plan. https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf
[22] National Science & Technology Council. (2019, June). The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update. A Report by the Select Committee on Artificial Intelligence. https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf
[23] White House. (2023, October 30). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
[24] AI Watch. (2021, September 1). France AI Strategy Report. https://ai-watch.ec.europa.eu/countries/france/france-ai-strategy-report_en

development, testing, and logistics. Additionally, AI plays a crucial role in maintenance and the implementation of Industry 4.0 principles, which encompass automation and data exchange in manufacturing processes. The strategy also addresses critical concerns regarding the performance, reliability, and robustness of AI systems, particularly in critical applications. This underscores the need to ensure the safe and dependable use of AI in systems where failures or malfunctions could have severe consequences. Furthermore, the strategy promotes open innovation through industry engagement and encourages collaboration, knowledge sharing, and the adoption of open innovation principles. By fostering collaboration among industry stakeholders, the strategy aims to drive innovation, enhance competitiveness, and create a conducive environment for the development and deployment of AI technologies.

2.1.12 European Union's Artificial Intelligence for Europe 2018[25]

The European Union (EU) introduced the Artificial Intelligence for Europe initiative in 2018, with a focus on addressing ethical concerns and promoting fundamental rights in the development and deployment of artificial intelligence (AI) technologies. As a key step towards this goal, the EU aimed to develop AI ethics guidelines taking into account the Charter of Fundamental Rights of the European Union. The aim was to tackle various critical issues, including the future of work, fairness, safety, security, social inclusion, and algorithmic transparency. They sought to examine the impact of AI on fundamental rights, encompassing areas such as privacy, dignity, consumer protection, and non-discrimination.

2.1.13 European Union's Artificial Intelligence Act, 2023[26]

The AI Act, proposed by the European Commission in April 2021, is a law aimed at regulating the development and use of AI systems in the European Union. It focuses on high-risk AI systems used in areas like human resources, banking, and education. In December 2023, the Council of the European Union and the European Parliament reached a provisional agreement on the AI Act. The provisional agreement is now expected to be formally adopted by both the Council and the Parliament in the first half of 2024. Often referred to as the "GDPR for AI," it imposes significant penalties for non-compliance and includes a wide range of mandatory requirements for organizations involved in AI development and deployment.

The AI Act adopts a risk-based approach, categorizing systems as low risk, limited risk, high risk, or unacceptable risk. Low-risk systems, such as spam filters and AI-powered video games, are already prevalent in the market. High-risk systems, which can significantly impact a user's life chances, are subject to specific requirements. These high-risk systems include those used in biometrics, critical infrastructure, education, employment, access to essential services, and health and life insurance. Unacceptable risk systems, outrightly prohibited, include those

[25] European Commission. (2018, April 25). Artificial Intelligence for Europe. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237

[26] European Commission. (2021, April 21). Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021) 206 final; 2021/0106 (COD)). https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

that have a significant potential for manipulation either through subconscious messaging and stimuli, or by exploiting vulnerabilities like socioeconomic status, disability, or age. The Act prescribes certain obligations for high-risk systems in regard to data governance including examining data for potential biases, maintaining transparency, human oversight, robustness etc.

2.1.14 United Kingdom's A Pro-Innovation Approach to AI Regulation 2023[27]

The UK's Department for Science Innovation and Technology (DSIT) released its White Paper on AI regulation, titled "pro-innovation approach to AI regulation," on March 29, 2023. The White Paper proposes a light touch regulation of AI through a sector-specific, principle-based framework, with a focus on fostering agility and promoting innovation. The White Paper delineates five key principles to guide the development and application of AI across different sectors of the economy. These five principles include- (a) ensuring safety, security, and robustness; (b) promoting appropriate transparency and explainability; (c) upholding fairness; (d) emphasizing accountability and governance; and (e) facilitating contestability and redress. UK has been actively working towards governing AI. In 2023, the country organised the AI Safety Summit hosting several countries. Recently, the UK's Information Commissioner Office (ICO) launched a consultation series on how aspects of data protection law should apply to the generative AI models.[28]

2.1.15 Japan's Social Principles of Human-Centric AI 2019[29]

Japan's Social Principles of Human-Centric AI, introduced in 2019, outlines a set of principles that aim to guide the implementation of social frameworks for AI across Japanese society. These principles are applicable to various stakeholders, including national and local governments, as well as multilateral frameworks, in the pursuit of creating an "AI-Ready Society." The principles prioritize human well-being and emphasize education on AI literacy. They stress the importance of privacy protection, secure operations, and fair competition in the AI ecosystem. Ensuring accountability, transparency, and fostering innovation while considering ethical implications are central to Japan's approach in promoting AI development for societal progress. Japan has also been actively shaping its AI landscape through several significant developments and strategies. The Governance Guidelines for Implementation of AI Principles, established by METI, delineates actionable targets for incorporating Social Principles into AI implementation.[30] These guidelines elucidate processes for creating and updating AI governance structures through collaborative efforts with stakeholders within an agile governance framework.

The AI Governance in Japan Ver. 1.1 report, published by METI in July 2021, underscored the complexities of AI innovation that hinder the feasibility of legally binding horizontal requirements. It acknowledges that

[27] 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

[28] ICO (2024). ICO consultation series on generative AI and data protection. https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-series-on-generative-ai-and-data-protection/

[29] Council for Social Principles of Human-centric AI. (n.d.) Human-Centric Artificial Intelligence. https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf

[30] Expert Group on How AI Principles Should be Implemented, AI Governance Guidelines WG. (2022, January 28). Governance Guidelines for Implementation of AI Principles. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20220128_2.pdf

regulations might struggle to keep pace with the rapid and intricate advancements in AI technology and therefore, a rigid and detailed regulatory framework could potentially impede innovation. It recommends excluding specific AI technologies from mandatory regulations and urges careful consideration of the scope of application to avoid unintended consequences. It further highlights the need to tailor regulations based on the specific use of technologies, acknowledging variations in benefits and societal impact.[31] It also advocates for "agile governance," respecting companies' voluntary AI governance efforts and offering nonbinding guidance grounded in multi stakeholder dialogues.[32] The Japan AI Strategy 2022 endeavors to address global challenges. With a focus on enhancing societal resilience, industrial competitiveness, and disaster response, this strategy emphasises objectives like bolstering AI reliability and promoting its government utilisation.[33]

2.1.16 The Global Partnership on Artificial Intelligence's AI principles[34]

The Global Partnership on Artificial Intelligence (GPAI) is a collaborative international initiative aimed at promoting the responsible development and utilization of AI. The GPAI brings together various stakeholders from around the world to guide the advancement of AI technologies in a manner that aligns with human rights, fundamental freedoms, and shared democratic values. The GPAI's AI principles are aligned with the OECD Recommendation on AI, which provides a comprehensive framework for responsible AI development. These principles emphasize the importance of upholding human rights, fundamental freedoms, and democratic values in the context of AI. By adhering to these principles, the GPAI aims to guide the global community in harnessing the potential of AI while mitigating risks and addressing societal concerns. It seeks to create a global ecosystem that prioritizes transparency, accountability, inclusivity, and the well-being of individuals and societies.

2.1.17 UNESCO Ethics of Artificial Intelligence[35]

The UNESCO Ethics of Artificial Intelligence framework encompasses several key principles that aim to guide the responsible development and deployment of AI technologies. These principles address various aspects of AI systems, focusing on ensuring ethical practices and upholding human rights and fundamental freedoms. The key principles include: (a) Proportionality and doing no harm; (b) Safety and Security; (c) Fairness and Non-discrimination; (d) Sustainability; (e) Right to Privacy and Data Protection; (f) Human oversight and determination; (g) Transparency and explainability; (h) Responsibility and accountability; (i) Awareness and literacy; and (j) Multi-stakeholder and adaptive governance and collaboration. The objectives of the UNESCO Ethics of Artificial Intelligence are to establish a universal framework of values, principles, and actions that guide states in

[31] Expert Group on How AI Principles Should be Implemented. (2021, July 9). AI Governance in Japan Ver. 1.1: Report from the Expert Group on How AI Principles Should Be Implemented. https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_8.pdf
[32] Habuka, H. (2023). Japan's approach to AI regulation and its impact on the 2023 G7 presidency. https://www.csis.org/analysis/japans-approach-ai-regulation-and-its-impact-2023-g7-presidency
[33] Chambers and Partners. (n.d.). Artificial Intelligence 2023 - Japan. Global Practice Guides. https://practiceguides.chambers.com/practice-guides/artificial-intelligence-2023/japan/trends-and-developments/O13573
[34] OECD.AI. (n.d.). The Global Partnership on AI (GPAI). https://oecd.ai/en/gpai
[35] UNESCO. (2023, December 1). Ethics of Artificial Intelligence. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

formulating AI-related legislation and policies in accordance with international law. Additionally, it aims to ensure that ethics are embedded in all stages of the AI system life cycle, safeguarding human rights, fundamental freedoms, human dignity, equality (including gender equality), and the interests of present and future generations. The ethics framework also emphasizes the importance of preserving the environment, biodiversity, ecosystems, and respecting cultural diversity. Another objective is to foster multi-stakeholder and multidisciplinary dialogue to address ethical concerns related to AI systems, promoting consensus building. Lastly, it seeks to promote equitable access to AI developments and knowledge, ensuring the sharing of benefits, with a particular focus on the needs and contributions of less developed countries, including least developed countries, landlocked developing countries, and small island developing states.

2.1.18 United Nations' Principles for Ethical Use of AI in UN 2022[36]

The principles, derived from UNESCO's Ethics of AI recommendation, are designed to ensure that the UN employs AI in the best interest of the people it serves. These principles establish a framework for the ethical utilisation of AI by UN organizations across all stages of an AI system's lifecycle. The primary objective is to foster trustworthiness and prioritise human dignity, equality for all individuals, environmental preservation, biodiversity and ecosystems, respect for cultural diversity, and responsible handling of data.

2.2 Mapping Trustworthy AI Principles

Through an extensive analysis of various ethical AI frameworks worldwide, it has become evident that certain principles play a pivotal role in ensuring the development and deployment of trustworthy AI technology. It is also important to acknowledge that not all of them are centered on promoting ethical or trustworthy AI. Certain frameworks outlined above place a greater emphasis on AI regulation and governance, underscoring the importance of adhering to legal and operational standards, in contrast to those that are centred on establishing ethical guidelines to promote trustworthy and responsible AI. However, within this diverse array of frameworks, our synthesis has uncovered a set of core principles that consistently surface across different contexts. These principles underscore their fundamental importance regardless of the framework's primary focus. These recurring principles are instrumental in shaping responsible AI practices. They often revolve around concepts such as transparency, accountability, fairness, and the protection of individual rights and privacy. Regardless of whether a framework's explicit goal is to address ethics, their inclusion of these principles signifies a broader recognition of their pivotal role in guiding AI development and deployment. Hence, in our effort to outline fundamental principles for this paper, we've compiled a checklist of

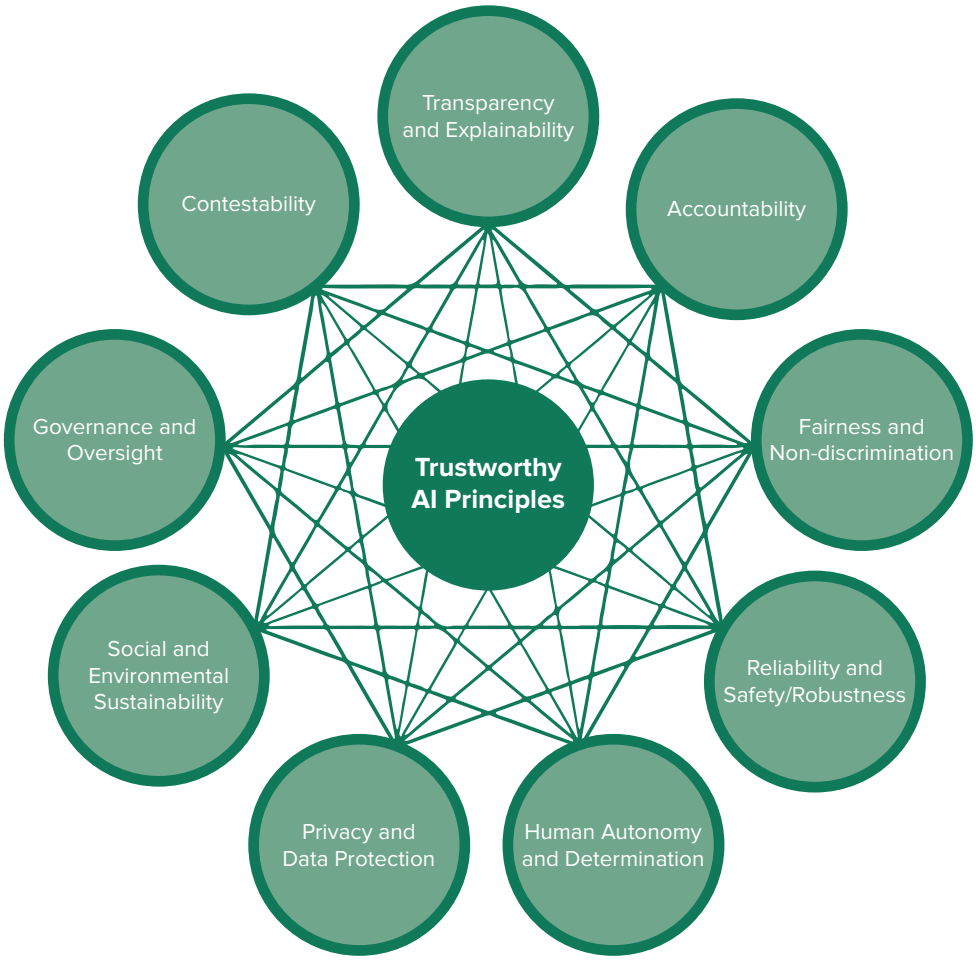
[36] United Nations System. (2022, September 20). High-Level Committee on Programmes (HLCP) Inter-Agency Working Group on Artificial Intelligence: Principles for the Ethical Use of Artificial Intelligence in the United Nations System. https://unsceb.org/sites/default/files/2022-09/Principles%20for%20the%20Ethical%20Use%20of%20AI%20in%20the%20UN%20System_1.pdf

overarching principles found consistently across the various regulations we've examined. It's essential to highlight that some of these regulations align closely with nearly all the principles outlined below. This alignment is expected, given that some of these laws and regulations explicitly prioritize trustworthiness, while others take a broader approach. The broader regulations may not encompass all these principles to the same extent, as their scope is more broad-ended and extends beyond ethical considerations.

Table 1: Trustworthy AI Principles Checklist

Framework	Transparency and Explainability	Accountability	Fairness and Non-discrimination	Reliability and Safety/ Robustness	Human Autonomy and Determination	Privacy and Data Protection	Social and Environmental Sustainability	Governance and Oversight	Contestability
OECD	✓	✓	✓	✓	✓	✓	✓		
NITI Aayog	✓	✓	✓			✓			
G20	✓	✓	✓	✓			✓		
Australia	✓	✓	✓	✓		✓	✓		✓
EU Ethics Guidelines	✓	✓	✓	✓	✓	✓	✓		
EU-US TTC	✓		✓				✓		
NIST	✓	✓	✓	✓	✓	✓	✓	✓	
Germany	✓	✓	✓	✓					
Singapore		✓					✓		
USA	✓		✓	✓		✓			
France	✓			✓					
EU's AI for Europe 2018	✓		✓	✓		✓	✓		
EU's AI Act, 2023	✓	✓	✓	✓	✓	✓			
United Kingdom	✓	✓	✓	✓				✓	✓
Japan	✓	✓	✓	✓		✓			
GPAI	✓	✓	✓	✓	✓	✓	✓		
UNESCO	✓	✓	✓	✓	✓	✓	✓		
UN	✓	✓	✓	✓	✓	✓	✓		

Figure 1: Principles for Trustworthy AI



Further, we categorise the identified trustworthy AI principles within three perspectives: the technical perspective, the user perspective, and the social perspective, for a more comprehensive analysis[37]. The categorization of trustworthy AI principles into technical, user, and social perspectives is a strategic approach designed to address the diverse needs and concerns of stakeholders within the AI ecosystem. These distinct categories provide a unique lens through which principles can be comprehensively analyzed and applied.

[37] Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A. K., & Tang, J. (2021). Trustworthy AI: A Computational Perspective. https://arxiv.org/pdf/2107.06641.pdf

Box 2: Definitions

Technical Perspective:
From a technical standpoint, trustworthy AI principles encompass aspects related to the development, deployment, and functionality of artificial intelligence systems. This perspective emphasizes the need for robust technical architectures, algorithmic transparency, and mechanisms for accountability. Technical principles guide the implementation of AI in a manner that ensures reliability, accuracy, and adherence to ethical standards. By focusing on the technical dimension, these principles aim to mitigate risks associated with system vulnerabilities, biases, and unintended consequences, fostering a foundation of technical trustworthiness.

User Perspective:
The user perspective delves into the individual experiences, rights, and needs of those interacting with AI systems. Trustworthy AI principles within this category aim to empower users by ensuring transparency, explainability, and user-friendly interfaces. Protection of user privacy, informed decision-making, and the provision of recourse mechanisms are integral components. The user perspective acknowledges the importance of establishing trust at the user level, addressing concerns such as algorithmic bias, user autonomy, and the overall user experience. Principles within this realm contribute to building a user-centric approach that fosters confidence and engagement.

Social Perspective:
The social perspective broadens the scope to encompass the societal impact and implications of AI technologies. Trustworthy AI principles within this category consider the ethical and societal ramifications, promoting fairness, inclusivity, and accountability on a larger scale. Principles here address issues like the equitable distribution of benefits, societal fairness, and the avoidance of discriminatory practices. The social perspective recognizes AI's potential influence on societal structures, emphasizing the responsibility of developers and policymakers to consider the broader social context in which AI systems operate. By integrating these principles, the aim is to contribute to the establishment of AI technologies that align with societal values, ethical norms, and foster positive societal impacts.

This endeavor acknowledges the complex interplay between technical functionalities, user needs, and societal impact, creating a cohesive and balanced approach to AI development and regulation. This framework serves to enhance trustworthiness by systematically covering three crucial dimensions: the intricacies of AI development and deployment, user trustworthiness, and the broader societal adoption of technologies. By delineating these perspectives, the principles become more tailored and applicable, offering specific guidelines for technical robustness, user empowerment and protection, and societal impact. This multifaceted approach ensures that the principles are

comprehensive, fostering a well-rounded trustworthiness approach that resonates across the intricate technical landscape, individual user experiences, and the broader societal context of AI implementation.

As AI continues to evolve and permeate various aspects of our lives, a principled and multi-dimensional approach is vital to ensure that AI technologies align with human values and contribute positively to our society. Each perspective illustrated below offers a unique lens to evaluate and address different aspects of AI development and deployment. The sum total of the different perspectives will help provide regulators with a comprehensive and balanced framework to achieve trustworthy AI. It will enable them to address complex challenges, foster innovation responsibly, and ensure that AI technologies serve the best interests of individuals and society as a whole.

Table 2: Principles for Trustworthy AI within different perspectives

Perspective	Principles
Technical	Transparency and Explainability; Reliability and Safety/Robustness
User	Human Autonomy and Determination; Privacy and Data Protection; Contestability
Social	Accountability; Fairness and Non-discrimination; Governance and Oversight; Social and Environmental Sustainability

From a technical perspective, trustworthy AI is expected to show the properties of transparency, robustness, and explainability. Transparency and Explainability enable us to understand and interpret the decisions made by AI systems, providing insight into the reasoning behind their outputs. Gaining these insights into the internal mechanisms of AI models and algorithms promotes transparency at the process level, allowing for the identification of potential biases, errors, or unethical behaviors.[38]. On the other hand, reliability and Safety/Robustness ensure that AI systems perform to an extent consistently and accurately, even in dynamic and complex real-world environments. Robust AI systems are resilient to uncertainties, variations, and adversarial attacks, thereby reducing the risk of erroneous or harmful outcomes.

[38] Burt, A. (2019, December 13). The AI transparency paradox. Harvard Business Review. https://hbr.org/2019/12/the-ai-transparency-paradox

Box 3: Example

In Natural Language Processing (NLP) applications, like sentiment analysis or chatbots, transparency and explainability are essential to detect biases in language models. Techniques like LIME (Local Interpretable Model-Agnostic Explanations)[39] can be applied to explain the model's predictions for specific instances, revealing potential biases and enabling developers to fine-tune the model to avoid discriminatory outcomes. Further, in medical diagnosis, AI models are used to examine medical images, such as X-rays or MRI scans, to detect diseases[40]. By incorporating explainability techniques like Grad-CAM (Gradient-weighted Class Activation Mapping)[41], the model can highlight regions of the image that contributed to its decision, providing insights into the reasoning behind the diagnosis. This transparency will help doctors understand and trust the model's predictions, making it easier to identify potential biases or errors.

From a user perspective, trustworthy AI should possess the properties of human autonomy, privacy, and contestability. Human autonomy and determination emphasise the critical role of human involvement in decision-making processes, ensuring that ultimate responsibility and accountability rest with human agents rather than solely relying on automated systems. Privacy and Data Protection principles safeguard individuals' personal information, promoting trust and preserving their autonomy and rights in the context of AI-driven technologies. Finally, Contestability principles foster a culture of openness, allowing users to challenge decisions made by AI systems and seek redress in cases of unfair or biased outcomes.

Box 4: Example

In healthcare, AI-driven decision support systems can assist doctors in diagnosis and treatment recommendations[42]. However, human autonomy remains crucial in the final decision-making process. AI systems can present doctors with evidence-based suggestions, but the ultimate responsibility of choosing the appropriate treatment option lies with the healthcare professional, ensuring that patients' well-being is prioritized. Further, AI-based loan approval systems use complex algorithms to assess applicants' creditworthiness[43]. Contestability principles enable applicants to seek explanations for their loan rejections and challenge the decisions made by AI systems. By providing transparent explanations for loan approvals or denials, financial institutions can build trust with their customers and ensure fair and unbiased decision-making.

[39] Ribeiro, M. T. (2016, February 16). "Why should I trust you?": Explaining the predictions of any classifier. arXiv.org. https://arxiv.org/abs/1602.04938
[40] IBM. (n.d.). Artificial Intelligence in Medicine. https://www.ibm.com/topics/artificial-intelligence-medicine
[41] MathWorks United Kingdom. (n.d.). Grad-CAM reveals the why behind deep learning decisions. MATLAB & Simulink. https://uk.mathworks.com/help/deeplearning/ug/gradcam-explains-why.html
[42] IBM. (n.d.). Artificial Intelligence in Medicine. https://www.ibm.com/topics/artificial-intelligence-medicine
[43] Corestrat. (2023, July 27). Exploring Automated Loan Approval Systems: AI's Impact on Borrowers and Lenders. https://corestrat.ai/blog/exploring-automated-loan-approval-systems-ais-impact-on-borrowers-and-lenders/

From a social perspective, trustworthy AI should be accountable, fair/non-discriminatory, law-abiding, and environmentally friendly. Accountability ensures that individuals and organisations are held responsible for their actions and decisions in the AI ecosystem, providing recourse in case of harm or misuse. Fairness and Non-discrimination principles emphasise the importance of eliminating biases and ensuring equal treatment and opportunities for all individuals, regardless of their characteristics or background. Governance and Oversight mechanisms play a key role in establishing regulations, standards, and policies to ensure that AI systems are developed, used, and governed in a manner that aligns with societal values, ethical considerations, and legal frameworks. Finally, the Social and Environmental Sustainability principle emphasises the need to consider the broader societal and environmental impact of AI technologies, ensuring that they contribute positively to the well-being of individuals and the planet.

Box 5: Example

AI-based systems are used to optimise the allocation of healthcare resources[44], such as organ transplants or hospital beds. Fairness and non-discrimination principles play a vital role in ensuring that these resources are allocated without any bias towards certain groups or demographics, ensuring equitable access to healthcare services. Further, in the development and deployment of autonomous vehicles, accountability is crucial[45]. When an accident occurs involving an autonomous vehicle, the responsibility for the incident needs to be identified and addressed. Properly implementing accountability principles ensures that the relevant parties, such as manufacturers or developers, are held accountable for any errors or malfunctions in the AI system. AI is also being used in smart city initiatives to optimise resource usage, reduce energy consumption, and improve transportation systems[46]. Social and environmental sustainability principles guide the development and deployment of AI systems in smart cities, ensuring that these technologies contribute positively to the well-being of citizens and the environment.

2.3 Mapping Synergies and Conflicts

It is important to recognise that the principles outlined above in the three perspectives of technical, user, and social are not independent of each other. While these principles generally complement and reinforce each other, there are instances where conflicts may arise. The next section will delve deeper into this.

[44] Deloitte Insights. (n.d.). Smart use of artificial intelligence in health care. https://www2.deloitte.com/us/en/insights/industry/health-care/artificial-intelligence-in-health-care.html

[45] Omeiza, D., Web, H., Jirotka, M., & Kunze, L. (2021). Towards Accountability: Providing Intelligible Explanations in Autonomous Driving. In 2021 IEEE Intelligent Vehicles Symposium (IV) (pp. 231–237). IEEE Press. https://doi.org/10.1109/IV48863.2021.9575917

[46] Herath, H. M. K. K. M. B., & Mittal, M. (2022, April). Adoption of artificial intelligence in smart cities: A comprehensive review. International Journal of Information Management Data Insights, 2(1), 100076. https://doi.org/10.1016/j.jjimei.2022.100076

2.3.1 Synergies

Trustworthy AI principles are not siloed but rather work in concert, complementing each other and building synergies to achieve trustworthy AI. The interplay between these principles strengthens the overall integrity and reliability of AI technologies. By recognizing their interdependence and fostering synergistic relationships, it becomes possible to create a comprehensive and balanced framework that promotes trustworthy AI practices.

- **Transparency and Explainability with Privacy and Data Protection:** The technical perspective, which emphasises transparency and explainability, is closely tied to the user perspective of privacy and data protection. Consider an AI-driven financial advisory platform where transparency and explainability play a pivotal role. The platform, in adherence to technical principles, provides clear insights into the data it analyzes, ensuring users understand how their financial information contributes to personalized recommendations. Simultaneously, from the user standpoint, the emphasis on privacy and data protection ensures that sensitive financial data is securely managed, with strict protocols in place to prevent unauthorized access and user confidentiality is maintained. This harmonization of technical transparency and user privacy principles establishes a foundation for user trust and confidence in AI-driven financial services.

- **Human Autonomy with Accountability:** The principles of human autonomy and determination, as viewed through the lens of user perspective, exhibit a compelling synergy with accountability from the social perspective. In essence, the correlation underscores the idea that when individuals are endowed with the capacity to autonomously make decisions, the process of holding them accountable becomes inherently more transparent and streamlined. This interconnection emphasizes the profound significance of empowering users with the autonomy to make decisions that align with their preferences and values. Essentially, the more users are entrusted with the ability to exercise self-determination, the more seamlessly accountability mechanisms can operate within the societal framework. This intricate relationship illuminates the interplay between user-centric principles and the overarching social structure of accountability, underscoring the pivotal role that individual autonomy plays in fostering a more transparent and responsible society.

  A use case that exemplifies the interplay between human autonomy, determination, and accountability is in the context of AI-driven medical diagnosis systems[47]. This use case involves a medical institution deploying an AI system to aid doctors in diagnosing medical conditions. Here, the principle of human autonomy ensures that doctors retain the final decision-making authority, with the AI system offering support and

[47] Amann, J., Blasimme, A., Vayena, E., et al. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310. https://doi.org/10.1186/s12911-020-01332-6

recommendations. Simultaneously, in cases of errors, the accountability principle holds the institution responsible, ensuring corrective measures. This symbiotic relationship showcases how user-centric AI, aligned with accountability, enhances decision-making, improves patient care, and fosters trust between users and AI technologies.

- **Fairness and Non-Discrimination with Reliability and Safety/Robustness:** The principles of fairness and non-discrimination from the social perspective are closely related to reliability and safety/robustness from the technical perspective. Fairness requires that AI systems are designed to avoid biases and ensure equal treatment for all individuals. Reliability and safety/robustness ensure that AI systems consistently produce accurate and unbiased results, reducing the risk of discriminatory outcomes. Analysing the interplay between fairness and reliability/safety/robustness reveals that these principles complement each other. A fair AI system needs to be reliable and robust to ensure that it consistently delivers unbiased outcomes, while a reliable and robust AI system helps enhance fairness by reducing the risk of discriminatory results. By integrating these principles into the design and development of AI systems, we can move towards building AI technologies that not only perform accurately but also promote fairness and ethical decision-making, contributing to a more equitable and inclusive society.

- **Governance and oversight with Reliability:** Governance and oversight, a social perspective principle, serves as a bridge between the technical and social dimensions. Effective governance and oversight frameworks ensure that AI systems are developed and deployed in a reliable manner that upholds ethical standards, protects user rights, and aligns with societal values[48]. The existence of such mechanisms not only strengthens the ethical foundation of AI but also instills confidence in users, making the technology more reliable for widespread use. A use case illustrating the synergy between the two principles would be in the development and implementation of AI-based medical diagnostic systems. Healthcare institutions can adopt governance and oversight frameworks that define the data sources and types of data that can be used in these systems. By adhering to strong governance and oversight practices, the medical AI system can be trusted by both healthcare professionals and patients. It will help build confidence in the system's reliability and accuracy, promote responsible AI usage, and ultimately lead to better healthcare outcomes.

- **Transparency & Privacy Vs Contestability:** Transparency and explainability mechanisms ensure that the decision-making processes of AI systems are clear and understandable. In contexts where decisions may impact individuals significantly, these principles foster user trust and confidence. Contestability complements these by emphasizing the need for mechanisms

[48] Ibm. (2021, April 26). AI governance: Ensuring your AI is transparent, compliant, and trustworthy. IBM. https://www.ibm.com/analytics/common/smartpapers/ai-governance-smartpaper/

that allow users to challenge AI outcomes perceived as unfair, biased, or harmful. By creating avenues for redress and accountability, contestability ensures that users have the means to seek explanations for AI decisions. This dual approach not only enhances user understanding but also empowers them to contest decisions that may have negative consequences, contributing to a more accountable and trustworthy AI ecosystem.

2.3.2 Conflicts

While the principles identified in the technical, user, and social perspectives of trustworthy AI generally complement and reinforce each other, it is important to acknowledge that conflicts between these principles can arise. This is primarily because implementing these principles involves navigating complex trade-offs and balancing competing interests.

- **Transparency and Explainability vs. Privacy and Data Protection:** The principle of transparency and explainability emphasises the need for AI systems to provide clear insights into their decision-making processes. However, this can conflict with the principle of privacy and data protection, as revealing certain information may compromise individuals' privacy. For eg., let's consider a use case where a financial institution deploys an AI-driven credit scoring system to assess individuals' creditworthiness for loan approvals[49]. The AI system uses various data points, including financial history, employment records, and spending habits, to predict credit risk and determine loan eligibility. The tension emerges in how much information the financial institution can provide to individuals regarding the AI-driven credit scoring system's decision-making process without compromising the privacy of sensitive financial and personal data.

  If the AI system provides a highly detailed breakdown of its decision-making process, including the specific factors and data points used, it may inadvertently expose individual borrowers' sensitive information. For instance, if the system uses a small number of borrowers from a particular demographic to train its model, the explainability process might reveal patterns that can be traced back to these individuals, violating their privacy.

- **Robustness vs. Privacy and Data Protection:** To ensure robust AI systems, comprehensive data collection and analysis are often necessary. However, this can clash with privacy and data protection principles, as extensive data usage raises concerns about unauthorised access or misuse of personal information. Finding ways to balance the need for robustness and safety with individuals' privacy rights is crucial in navigating this conflict. Let's consider a use case in the healthcare industry where a hospital deploys an AI system to analyze patient data and predict medical conditions to provide timely and accurate diagnoses.[50] The AI system requires comprehensive data collection and analysis of patients' medical history, symptoms, test results,

[49] Hicham Sadok, Fadi Sakka & Mohammed El Hadi El Maknouzi (2022) Artificial intelligence and bank credit analysis: A review. https://doi.org/10.1080/23322039.2021.2023262

[50] Mills, T. (2022, February 16). AI for Health and Hope: How machine learning is being used in hospitals. Forbes. https://www.forbes.com/sites/forbestechcouncil/2022/02/16/ai-for-health-and-hope-how-machine-learning-is-being-used-in-hospitals/?sh=1ee7761255be

and treatment outcomes to build accurate predictive models. The conflict arises when balancing the need for comprehensive data collection to ensure a reliable and safe AI system while respecting individuals' privacy rights. Comprehensive data collection may require access to a vast amount of patient information, which could be seen as intrusive and raise concerns about data security and privacy. Patients may be reluctant to share their medical history and health-related data if they feel that their privacy is at risk.

- **Robustness vs. Human Autonomy and Determination:** Achieving robustness in AI systems often involves minimizing human intervention and relying on automated decision-making processes whereas, human autonomy and determination highlight the importance of human involvement in critical decisions. Automation helps create more consistent and standardized outcomes, especially in routine and well-defined tasks. By limiting human involvement, AI systems can exhibit increased efficiency, reliability, and predictability. However, it is essential to strike a careful balance to ensure that the automated processes align with ethical considerations, address potential biases, and provide adequate mechanisms for human oversight in critical and nuanced situations. Striking a balance requires determining the appropriate level of human oversight to ensure safety and reliability while still incorporating human judgment and decision-making. For example, let's consider the deployment of AI-driven diagnostic tools. These tools utilize sophisticated algorithms to analyze medical data and assist in diagnosing diseases. To enhance robustness, these AI systems aim to reduce the need for extensive human intervention and provide swift and accurate diagnoses. However, the conflict arises as human autonomy emphasizes the crucial role of healthcare professionals in decision-making processes, especially in complex or ambiguous cases. Achieving a balance involves incorporating human expertise to validate AI-generated diagnoses, ensuring that healthcare providers remain actively engaged in the decision-making process. This collaborative approach safeguards against potential errors, promotes trust in the diagnostic outcomes, and leverages the strengths of both AI technology and human medical expertise.

- **Governance and Oversight vs. Contestability:** Governance and oversight mechanisms are designed to regulate AI systems, ensuring compliance with ethical and legal standards. However, contestability principles advocate for open challenges and scrutiny of AI decisions. Balancing these two principles necessitates the establishment of effective governance frameworks that promote transparency and accountability while accommodating contestability without impeding progress or stifling innovation. Let's consider a scenario in the healthcare sector where a hospital employs AI algorithms to assist in diagnostic processes. Governance and oversight mechanisms would be crucial to ensure that the AI adheres to ethical standards, patient privacy regulations,

and medical guidelines. This involves establishing protocols for data security, accuracy, and overall compliance with healthcare regulations. On the other hand, contestability principles in healthcare might involve allowing medical professionals to challenge or question the AI's diagnostic recommendations. For instance, if an AI suggests a specific treatment plan for a patient, contestability would empower healthcare practitioners to review and contest the decision based on their medical expertise and the unique circumstances of the patient. Balancing these two principles requires a governance framework that ensures regulatory compliance and patient safety while providing room for healthcare professionals to contest AI recommendations when necessary.

- **Contestability vs. Reliability:** The principle of contestability in AI emphasizes the importance of providing individuals or stakeholders with the ability to challenge AI decisions when significant impacts are involved. This ensures that AI systems remain accountable and that users have recourse in case of unfair or biased outcomes. However, this principle can create tensions with the need for certainty and stability in AI decision-making processes, especially in critical domains like healthcare, finance, and autonomous vehicles. In these contexts, unpredictability or frequent challenges to AI decisions can lead to inefficiencies, delays, and potential risks to safety and security.

  To address this conflict, developers and policymakers must strike a delicate balance between contestability and reliability. They can implement mechanisms that allow for challenges but establish thresholds or criteria for when contestation is permissible. For instance, in the financial industry, if a bank uses AI algorithms to make loan approval decisions based on applicants' creditworthiness[51], the principle of contestability can be incorporated to provide applicants with clear explanations of the factors considered in the decision. If an applicant disagrees with the decision or believes there may be errors, they have the option to request a manual review. The bank sets guidelines for contestability, such as allowing challenges for loan applications that fall within a specific credit score range or have approval ratings on the borderline. This approach strikes a balance between allowing contestation and maintaining a certain level of certainty and efficiency in the loan approval process.

[51] Hicham Sadok, Fadi Sakka & Mohammed El Hadi El Maknouzi (2022) Artificial intelligence and bank credit analysis: A review. https://doi.org/10.1080/23322039.2021.2023262

3 Operationalisation of Trustworthy AI Principles

As various stakeholders strive to embrace AI's potential, there arises a pressing need to develop a comprehensive operational strategy that translates identified principles into actionable steps. This chapter introduces a distinctive approach to operationalizing trustworthy AI principles. It attempts to not only address the theoretical aspects but also delves into the pragmatic realm, offering valuable insights and methodologies for stakeholders, applicable across the two sectors of Health and Finance, to navigate the implementation of trustworthy AI in a nuanced and effective manner.

Our methodology, drawing from an array of ethical guidelines and best practices, endeavours to go beyond mere theoretical discussions. We delve into practical implementation, at both technical and non-technical levels. The operationalization process, as explained in this chapter, focuses on three key participants: AI developers, AI deployers and AI users. Through this approach, we aspire to demonstrate the relevance of our strategy and encourage its adoption across sectors, ultimately fostering a responsible and ethical AI ecosystem for the betterment of society as a whole. For purposes of this paper, we have limited our scope of stakeholders in the AI lifecycle to AI developers, deployers and end-users.

Figure 2: Stakeholders in AI ecosystem

AI Developer/ Providers

A natural person or legal entity (within both the public and private sectors) who develop Artificial Intellignece solutions for market consumption while they do not necessarily deploy and use the same technology.

AI Deployer/ Procurer

A natural person or legal entity (within both the public and perivate sectors) who procure, employs, deploys and operates Artificial Intelligence solutions not necessarily developed by themselves.

Impact Population/ End Users

A natural person who directly and indirectly uses, engages, and is impacted or affected by the Artificial Intelligence solution.

3.1 Principles for Operationalisation

This section will provide a comprehensive explanation of the identified principles, offering users a clear and in-depth understanding by demystifying the principles. These principles serve as a comprehensive blueprint for ensuring the trustworthiness of AI and data-driven technologies, regardless of the specific domain in which they are deployed. This broad mapping serves as a starting point, allowing for sector-specific adaptations and operationalization while maintaining a common principles based framework.

3.1.1 Transparency and Explainability

Transparency and explainability in AI, while often used together, differ significantly in their depth and scope. Transparency provides a broad view into the workings of an AI system, allowing stakeholders to grasp its overall functioning, data inputs, and general decision-making processes. It enables modelers, developers, and auditors to gain insights into the AI's training data, evaluation metrics, and high-level decision boundaries. This transparency is valuable for understanding the system's behavior at a macro level and for ensuring accountability.[52]

On the other hand, Explainable AI (XAI) delves much deeper into the intricacies of AI systems.[53] It not only reveals algorithms' operations but also provides explicit and interpretable explanations for individual decisions or recommendations. XAI aims to make the AI's decision-making logic clear and comprehensible to users and customers. It goes beyond transparency by answering questions like "Why was this decision made?" or "Why is this recommendation being provided?" This level of granular insight empowers users to trust AI systems and helps AI practitioners identify potential biases, errors, or ethical concerns at a micro level. The quest for explainability stems from the need to demystify the black-box nature of AI algorithms and provide meaningful insights to stakeholders[54]. However, demystifying the black-box nature of AI algorithms can be a formidable challenge, and it's often more productive to prioritize clarity in other dimensions of AI systems. While explainability seeks to shed light on the internal workings of complex models, it's important to recognize that some AI algorithms can be highly intricate and challenging to fully unveil. Instead of attempting to completely unveil these intricate black boxes, a more pragmatic approach is to emphasize clarity in different dimensions of AI systems. This includes transparent documentation of data collection processes, explanations of system design choices, detailed process documentation, and clear articulation of decision logic. By focusing on these aspects, we can enhance stakeholders' understanding and trust in AI systems without delving into the complexities of the black box. This approach promotes transparency, facilitates validation, and supports ethical AI practices.

[52] Building Transparency into AI Projects. (2022, June 20). Harvard Business Review. Retrieved August 25, 2023, from https://hbr.org/2022/06/building-transparency-into-ai-projects
[53] IBM. (n.d.). What is explainable AI? https://www.ibm.com/topics/explainable-ai
[54] Vorras, A., & Mitrou, L. (2021). Unboxing the Black Box of Artificial Intelligence: Algorithmic Transparency and/or a Right to Functional Explainability. In T. E. Synodinou, P. Jougleux, C. Markou, & T. Prastitou-Merdi (Eds.), EU Internet Law in the Digital Single Market (pp. 145-159). Springer, Cham. https://doi.org/10.1007/978-3-030-69583-5_10

Box 6: Datasheets for Datasets[55]

Microsoft has introduced the 'Datasheets for Datasets' initiative, aiming to enhance communication between dataset creators and consumers while fostering transparency and accountability within the machine learning community. A datasheet documents the dataset's motivation, composition, collection process, recommended uses, and more.

Microsoft's Aether Data Documentation Template,[56] offers a structured framework to guide developers in creating transparent AI. This template includes key questions covering various aspects, such as:

1. Overview of the dataset.
2. Intended purposes of the AI and potential inappropriate uses.
3. Process followed to collect data, including obtaining consent.
4. Inclusiveness and representativeness of the dataset, specifying demographic groups included.
5. Data quality, detailing steps taken to verify data and addressing inaccuracies.
6. Cleaning and labeling processes applied to the data.
7. Privacy concerns, along with any privacy reviews undertaken.

This initiative encourages a responsible approach to AI development by providing a standardized methodology for documenting key aspects of datasets, promoting ethical and accountable use of machine learning technologies.

3.1.2 Accountability

Accountability is a critical principle that underpins the entire lifecycle of an AI system[57]. It demands that all stakeholders involved in the development and deployment of AI systems take responsibility for ensuring that the technology aligns with human values. This accountability is achieved through careful product design, reliable technical architecture and a thorough assessment of potential impacts. Transparency plays a fundamental role in facilitating the accountability of an AI system by providing the means to understand and justify its decisions and actions. Derived from accountability, the concept of auditability also comes into play, requiring that the justification of an AI system be subject to review, assessment, and auditing[58].

[55] Microsoft Research. (2022, August 25). Data Documentation. https://www.microsoft.com/en-us/research/project/datasheets-for-datasets/overview/
[56] Microsoft Research. (2022, August 25). AETHER Data Documentation Template (Draft 08/25/2022). https://www.microsoft.com/en-us/research/uploads/prod/2022/07/aether-datadoc-082522.pdf
[57] Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. AI & Society. https://doi.org/10.1007/s00146-023-01635-y
[58] Williams, R., Cloete, R., Cobbe, J., Cottrill, C., Edwards, P., Markovic, M., . . . Pang, W. (2022). From transparency to accountability of intelligent systems: Moving beyond aspirations. Data & Policy, 4. https://doi.org/10.1017/dap.2021.37

3.1.3 Fairness and Non-discrimination

The principle of fairness and non-discrimination in AI systems underscores the importance of eliminating biases and ensuring equal treatment across all individuals, irrespective of factors like race, gender, or socioeconomic status. It aims to prevent unjust outcomes in AI decision-making processes. This becomes especially crucial in critical domains like financial risk assessment, recruitment, and face identification, where the utilization of AI systems can lead to systematic disadvantages for certain groups, resulting in negative social impacts and biases. Such biases not only erode trust in AI but also hinder the technology's overall potential to benefit society. Consequently, practitioners must prioritize the fairness of AI systems to avoid perpetuating or exacerbating social bias. A primary goal in achieving fairness in AI systems is mitigating the effects of biases, which can manifest in various forms during the development and application of AI technologies, such as data bias, model bias, and procedural bias[59]. Often, biases result in the unfair treatment of specific groups based on their protected characteristics, such as gender, race, ethnicity, low purchasing power, etc. Two key factors that contribute to bias are group identity (sensitive variables) and the system's response (prediction). Sensitive variables are attributes like race, gender, age, and more, which are often at the heart of discrimination. When these attributes are included in the training data, models can inadvertently learn and perpetuate biases associated with them. The system's response, on the other hand, pertains to the model's output or predictions. Biases can manifest when the model provides different outcomes for different groups, even when the input data should logically result in similar predictions. These disparities are often a consequence of how the model interprets and processes sensitive variables, resulting in discriminatory or unfair outcomes.

Box 7: Google's Model Cards[60]

To enhance transparency in machine learning, Google has introduced Model Cards, a tool that offers a structured framework for reporting on various aspects of ML models, including their origin, usage, and ethical considerations. These cards aim to provide a clear and comprehensive overview of a model's functionality, target audience, maintenance, architecture, and training data. By offering detailed insights into a model's suggested uses and limitations, Model Cards cater to developers, regulators, and end-users alike. The goal of Model Cards is to make transparency accessible to both experts and non-experts. Developers can utilize them to design applications that highlight a model's strengths while informing users about its weaknesses.

[59] Ferrara, E. (2023). Fairness And Bias In Artificial Intelligence: A Brief Survey Of Sources, Impacts, And Mitigation Strategies. https://arxiv.org/ftp/arxiv/papers/2304/2304.07683.pdf
[60] Google Cloud. (n.d.). Google Cloud Model Cards. https://modelcards.withgoogle.com/about

Two of the model cards released by Google include: face detection and object detection. The face detection card provides details on the model's performance across different demographic characteristics, while the object detection card outlines how the model performs on various classes of objects. They also highlight where the model performs well or poorly. Both cards offer insights into performance, limitations, and tradeoffs for the respective models.

Model Cards can play a crucial role in addressing issues such as unfair bias by examining how a model performs across diverse groups of people. For example, they can reveal whether a model consistently performs well or exhibits unintended variations based on factors like skin color or region.

3.1.4 Reliability and Safety/Robustness

Reliability and safety/robustness are fundamental principles in ensuring the trustworthy operation of AI systems.[61] Reliability refers to the ability of an AI algorithm or system to consistently perform accurately under varying conditions and inputs. A reliable AI system should produce consistent and dependable results, instilling confidence in its users and stakeholders. Banking on reliability, robustness goes further ahead and encompasses the ability of an AI system to handle unexpected situations, errors, or erroneous inputs gracefully.[62] A robust AI system can adapt to dynamic and diverse environments and still produce reliable results[63]. It should be resilient to variations in data, changes in input distributions, or the presence of outliers. Different organisations have come up with innovative initiatives to further safety and accuracy. For instance, Model Evaluation on Amazon Bedrock lets developers compare different foundational models ('FMs') for their specific use case based on custom metrics, such as accuracy and safety, allowing developers to select the FM that is best suited for specific use cases.[64]

3.1.5 Human Autonomy and Determination

The principle of "Human Autonomy and Determination" in the context of Trustworthy AI regulation emphasizes the critical role of human involvement in decision-making processes related to AI systems. It recognizes that while AI can be a powerful tool to assist and augment human decision-making, the ultimate responsibility and accountability for those decisions should rest with human

[61] Msteller-Ai. (2023, July 28). Responsible and trusted AI - Cloud Adoption Framework. Retrieved from https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/innovate/best-practices/trusted-ai#:~:text=Reliability%20and%20safety,-For%20AI%20systems&text=An%20organization%20should%20establish%20rigorous,performance%20can%20degrade%20over%20time.

[62] Singh, R. (2020, November 2). Trustworthy AI. Retrieved from https://arxiv.org/abs/2011.02272

[63] For example, A robust AI-driven fraud detection system would be capable of adapting to ever-evolving patterns of fraudulent activities, even in dynamic and diverse financial environments. By continuously learning and adjusting to new tactics employed by fraudsters, the system can consistently deliver reliable outcomes, effectively identifying and preventing fraudulent transactions across a range of scenarios. Aziz, Layla, & Andriansyah, Yuli. (2023). The Role Artificial Intelligence in Modern Banking: An Exploration of AI-Driven Approaches for Enhanced Fraud Prevention, Risk Management, and Regulatory Compliance, 6, 110-132. https://www.researchgate.net/publication/373489510_The_Role_Artificial_Intelligence_in_Modern_Banking_An_Exploration_of_AI-Driven_Approaches_for_Enhanced_Fraud_Prevention_Risk_Management_and_Regulatory_Compliance

[64] Amazon Bedrock Developer Experience. https://aws.amazon.com/bedrock/developer-experience/

agents rather than solely relying on automated systems.[65]

3.1.6 Privacy and Data Protection

Privacy protection is a fundamental aspect of building trustworthiness in AI systems. It involves safeguarding personally identifiable data from unauthorized access or use that could potentially identify individuals or households. The personal data at risk, cover a wide spectrum of information, including but not limited to names, ages, genders, facial images, fingerprints, and other personally identifiable details.

A commitment to privacy protection is essential because it not only respects individuals' rights to privacy but also plays a crucial role in determining the overall trustworthiness of an AI system.[66] When users entrust their data to AI systems, they expect that their personal information will be handled with utmost care and confidentiality. Any compromise in data privacy can lead to breaches of trust and undermine the credibility of the AI system and the organisations behind it.

3.1.7 Social and Environmental Sustainability

The principle of "Social and Environmental Sustainability" in Trustworthy AI regulation emphasizes the responsible development, deployment, and governance of AI systems with consideration for society's well-being and environmental preservation.[67] It involves AI developers and users ensuring that AI technologies contribute positively to social progress and do not perpetuate inequalities or discrimination.[68] Social sustainability prioritizes fairness, inclusivity, and equal treatment for all individuals, while environmental sustainability aims to minimize the negative impact of AI on the environment, encouraging energy-efficient algorithms and hardware.

3.1.8 Governance and Oversight

The principle of "Governance and Oversight" refers to the establishment of effective mechanisms and frameworks for governing the development, deployment, and usage of AI systems. It emphasizes the need to ensure that AI technologies are developed and utilized in a manner that aligns with ethical principles, societal values, and legal regulations. Governance and oversight involve setting up regulatory bodies, industry standards, and policies that guide the responsible use of AI. These mechanisms aim to hold AI developers, organizations, and users accountable for their actions and decisions within the AI ecosystem. They also provide recourse in case of misuse or harm caused by AI systems.

65. De Cremer, D. (2021, August 30). AI should augment human intelligence, not replace it. Retrieved from https://hbr.org/2021/03/ai-should-augment-human-intelligence-not-replace-it
66. Reinhardt, K. (2022). Trust and trustworthiness in AI ethics. AI And Ethics. https://doi.org/10.1007/s43681-022-00200-5
67. Heilinger, J., Kempt, H., & Nagel, S. K. (2023). Beware of sustainable AI! Uses and abuses of a worthy goal. AI And Ethics. https://doi.org/10.1007/s43681-023-00259-8
68. UNESCO. (2021, November 23). Recommendation on the Ethics of Artificial Intelligence. Paris, France. https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence

3.1.9 Contestability

The principle of "Contestability" refers to the ability of individuals or groups to challenge and question the decisions made by AI systems.[69] It emphasizes the importance of creating mechanisms that allow for transparency, accountability, and redress when AI systems produce outcomes that are perceived as unfair, biased, or harmful. Contestability is crucial for ensuring that AI systems are held accountable for their actions and that individuals have the opportunity to seek explanations and rectifications in cases of erroneous or undesirable outcomes.[70] It empowers users to challenge AI decisions and provides a means to address potential biases and discriminatory practices.

3.2 Sectoral Operationalisation

Amidst the rapid and dynamic progress of technological innovation, the incorporation of AI emerges as a beacon of transformative potential that spans across diverse sectors. The convergence of AI's computational prowess and human ingenuity holds the promise of elevating industries to new heights, amplifying efficiency, accuracy, and unleashing waves of innovation that were previously unimaginable.[71] This chapter discusses the proactive efforts required to harmonize the boundless potential of AI with its trustworthy implementation. Two pivotal sectors, Finance and Health, are spotlighted as domains where the integration of AI necessitates a profound consideration of trustworthy practices. These sectors are the lifeblood of economies and well-being, embodying the delicate balance between innovation and ethical considerations. The selection of these two specific sectors is guided by parameters like: the impact on livelihoods, the intricate interface with users, heightened risk levels, and deep government involvement. As the lifeblood of economies and well-being, a meticulous examination of AI applications in these sectors becomes imperative, given the potential consequences for both individuals and society. These sectors, characterized by heightened user interaction, increased risks in decision-making, and robust government oversight, provide invaluable insights into fostering responsible AI development, deployment, and governance. A steadfast commitment to trustworthy AI practices ensures a harmonious integration of innovation and ethics, paving the way for a future where AI's transformative potential aligns seamlessly with our shared values, elevating societies and unlocking human potential.

It is important to note here that various tools/strategies suggested to operationalize one principle can also be used to operationalize another. This means that some of our tools/strategies demonstrate the capacity to address multiple trustworthy AI principles concurrently, ensuring a comprehensive and integrated approach to the ethical adoption of AI technologies.

[69] Alfrink, K., Keller, I., Kortuem, G., & Doorn, N. (2022). Contestable AI by design: towards a framework. Minds and Machines. https://doi.org/10.1007/s11023-022-09611-z

[70] Rodrigues, R. (2020). Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. Journal of Responsible Technology, 4, 100005. https://doi.org/10.1016/j.jrt.2020.100005

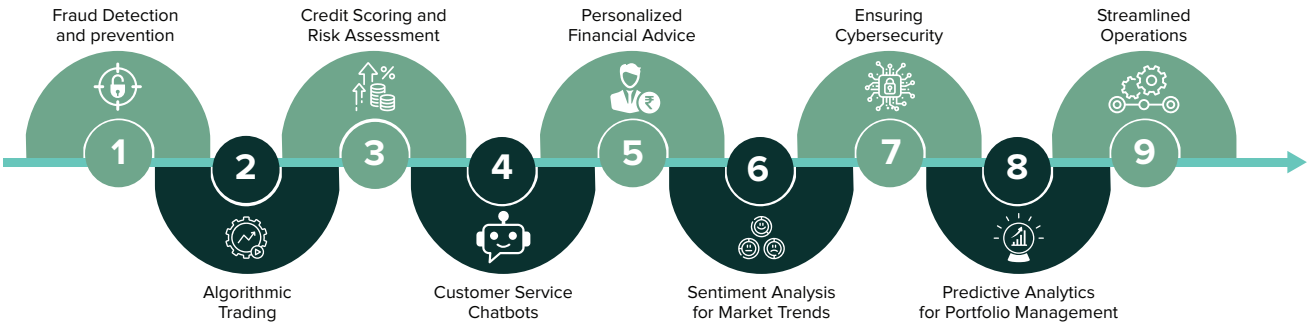
[71] MIT Technology Review. (2023, April 11). Technology and industry convergence: A historic opportunity. MIT Technology Review. Retrieved from https://www.technologyreview.com

3.2.1 Finance

As the financial industry continues to witness a transformative wave of technological advancements, the integration of AI becomes increasingly pervasive, revolutionizing the way financial institutions operate and serve their customers.[72] AI-driven systems promise a multitude of benefits, such as enhanced efficiency, personalized services, fraud detection, algorithmic trading, credit-lending, robo-advisory, and sophisticated risk assessment capabilities. However, the rapid adoption of AI technologies has also raised significant concerns regarding the trustworthiness of AI-driven decisions.

[72] How Artificial Intelligence is Transforming the Financial Services Industry. (n.d.). Retrieved from https://www2.deloitte.com/za/en/nigeria/pages/risk/articles/how-artificial-intelligence-is-transforming-the-financial-services-industry.html

Figure 3: Use cases of AI in Finance



This section focuses on devising strategies to effectively implement trustworthy AI principles at the level of AI developers, deployers and users in the financial sector. The integration of trustworthy AI principles in the financial sector is envisioned not only to bolster the industry's efficiency and profitability but also to foster a reliable and responsible financial ecosystem.

Figure 4: Finance Stakeholders

Developer

- Machine Learning Engineers
- Data scientists
- Financial Domain Experts

Deployer

- Financial Institutions
- Banks
- Investment firms
- Insurance companies
- Fintech Startups
- Regulatory Bodies

User

- Individual Consumers
- Businesses
- Investors
- Financial professionals etc

Our aim here is to outline indicative strategies for AI developers, deployers and users so as to help them harmonize their practices with trustworthy AI principles, fostering trust among stakeholders and promoting the responsible and ethical adoption of AI. Moreover, these strategies provide a cornerstone for establishing a secure, transparent, and user-centric financial ecosystem driven by AI.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Principle 1: Transparency and Explainability			
Technical	**Clear Documentation:** Develop comprehensive documentation for AI models used in financial services. This should include data lineage[73], model architecture, data sources, preprocessing steps. This documentation may be regularly updated to reflect any changes or updates to the model. Detailed documentation of AI model development provides crucial information about the model's architecture, training data, hyperparameters, and evaluation metrics[74]. **Interpretable Model Selection:** Prioritize the use of interpretable machine learning models, such as decision trees or linear regression, when possible, especially in areas where model explainability is critical.	**Regular Audits:** Embrace regular audits to ensure transparency and explainability in AI models, wherever possible. For instance, in the context of a bank's AI-driven credit scoring model, regular audits scrutinize the model's behavior over time. Auditors assess whether the model provides clear explanations for its credit decisions and whether these align with established principles of transparency and accountability. **Distinct Disclosure:** Deployers should prioritize providing comprehensive disclosure to consumers regarding the integration of AI systems in the delivery of financial products and services. This disclosure process should encompass detailed information on how AI is utilized, the	**Use of Explanatory Interfaces:** To enhance end-users understanding of transparency in AI systems, various tools can be adopted. These include intuitive interfaces providing user-friendly explanations of AI decisions, platforms displaying visual representations of the model's structure, user-friendly dashboards exhibiting key information on inputs and outputs, interactive pop-ups or notifications elucidating the rationale behind specific AI decisions during user interaction, etc.

[73] What is Data Lineage? (n.d.). Informatica. Retrieved August 25, 2023, from https://www.informatica.com/resources/articles/what-is-data-lineage.html

[74] Konigstorfer, F., & Thalmann, S. (2022). AI Documentation: A Path to Accountability. Journal of Responsible Technology, 11, Article 100043. https://www.sciencedirect.com/science/article/pii/S2666659622000208/

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	Further, more advanced models like XG Boost, etc can be considered in more critical areas of credit scoring, etc. **Feature Importance Analysis:** AI developers can leverage feature importance analysis techniques to automatically identify and highlight the variables exerting the most significant influence on model predictions during the AI development process. This information can be valuable for both model developers and users to understand the decision-making process. For eg, in the finance sector, AI developers can utilize feature importance analysis to automatically identify key variables influencing credit score predictions. This helps highlight factors such as income, credit history, and debt, providing transparency and aiding end-users in understanding the basis for their credit assessments.	decision-making processes involved, and the specific functionalities it serves. One way to do this is by developing educational modules within the platform to familiarize users with AI technology. These modules can explain how AI is integrated into financial processes, empowering users to make informed choices. **Post Deployment Functional and Performance Testing:** . Addressing the common issue of disparities between built and deployed models, post-deployment functional testing ensures the alignment of the deployed model with its intended functionalities, minimizing discrepancies. Simultaneously, performance testing evaluates the model's scalability under varying loads, ensuring it produces accurate outcomes. By implementing these testing procedures	

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	Model Explainability Tools: Integrate model explainability tools and libraries into the development process. Tools like LIME (Local Interpretable Model-Agnostic Explanations) or SHAP (SHapley Additive exPlanations) can provide insights into how specific predictions were made. LIME creates local interpretable models for specific predictions, clarifying factors influencing outcomes. For instance, in a credit scoring prediction, LIME reveals why an individual was classified as high risk by perturbing input features (e.g., credit history, income) to identify key influencers. On the other hand, SHAP employs a game-theoretic approach, attributing input feature contributions to predictions. In the same credit scoring example mentioned above, this tool will clarify the impact of each feature (e.g., credit history,	within the deployment phase, organizations can not only identify potential discrepancies but also provide insights into the model's operational capabilities. This strategy contributes to transparency by verifying that the deployed model accurately reflects the design, driving confidence in its functionality and performance.	

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	income) on the final credit score, aiding user understanding of model decisions.		
Non-technical	**Ethical AI Committees:** Establish internal committees or working groups dedicated to reviewing and ensuring the ethical and transparent use of AI in financial services. These committees can provide overall guidance and oversight. The constitution of the committee should include diverse experts, including women, individuals representing diverse demographics, external experts etc to provide a holistic perspective. They should actively contribute to the development process by providing inputs towards the creation of standard operating procedures (SOPs), conducting workshops to educate stakeholders, and offering insights on potential ethical concerns. **User Education:** Educate end-users,	**Regulatory Compliance:** Stay abreast of evolving regulations related to AI transparency and explainability in the financial sector. Work towards ensuring that AI systems maintain compliance with these regulations to meet the required standards.	**Seek Explanations:** End-users need to be educated to be able to seek clarification on financial service platforms that use AI for service delivery. This will help them seek explanations on how AI works and impacts their ability to avail services. Seeking explanations, where needed, goes a long way in ensuring that AI systems inbuilt transparency in its operations and don't function opaquely. Additionally, staying aware that certain details may not be available is essential, promoting a balance between seeking clarification and respecting the limitations of information accessibility.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	including both customers and employees, about how AI is used in financial services. Offer training and resources to help them understand AI-driven decisions and their implications.		
Principle 2: Accountability			
Technical	**Continuous Validation:** Implement rigorous validation protocols that continuously assess model performance. Regularly update models to adapt to changing data patterns and external factors, reducing the risk of unintended consequences. This practice is essential for timely identification of anomalies, biases, and unintended consequences, thereby facilitating prompt corrective actions. An alert system to notify stakeholders of impending periodic validations can be implemented, to ensure a uniform and timely assessment. Additionally, ad-hoc alerts for validations can	**Establish Feedback Mechanisms:** Establish channels for end-customers to provide feedback or raise concerns related to AI-based financial products or services. Actively seek and address customer input to enhance accountability. Certain innovative solutions such as integrating a responsive chatbot or employing Generative AI FAQs can be utilized to enhance engagement.	

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	be incorporated, which can be triggered by anomalies, enabling swift corrective actions and maintaining the model's reliability and ethical standards.		
Non-technical	**Robust Accountability Framework:** Developers should collaborate with senior management and compliance teams to define a clear framework for accountability. This framework should delineate responsibility at different levels of the organizational hierarchy and also provide protocols for addressing accountability challenges and adverse outcomes. **Stakeholder Consultation:** Developers should collaborate closely with business and compliance teams to fully comprehend the potential implications of AI-driven decisions.[75] Regular and open	**Internal risk management:** Internal risk management teams should assume accountability for the deployment of AI-based financial products and services. They must thoroughly understand the models they oversee and be prepared to explain their functioning to senior management and/or designated committee. However, it is also essential to acknowledge potential caveats, considering instances where mistakes may stem from AI end-users themselves. Consider a scenario where a financial institution deploys an AI-driven credit scoring system for loan approvals. The risk management team is responsible for	**Contribute to Feedback and Reporting Systems:** The active participation of end-users is integral to enhancing the effectiveness of feedback and reporting mechanisms established by developers and deployers. End-users occupy a unique vantage point, as they interact directly with the systems and applications in real-world scenarios. Soliciting their input creates a symbiotic relationship between developers and end-users, fostering a collaborative environment where feedback becomes a two-way street. Developers benefit from the firsthand experiences and

[75] Kirvan, P. (2020). How compliance provides stakeholders evidence of success. CIO. Retrieved from https://www.techtarget.com/searchcio/tip/How-compliance-provides-stakeholders-evidence-of-success

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	communication ensures that accountability considerations are woven into the fabric of the development lifecycle. Emphasizing stakeholder engagement in this communication process ensures that diverse perspectives are considered, enhancing accountability at every stage of AI development.	overseeing the model, ensuring its accuracy, and explaining its functioning to senior managers and the board. Despite the risk management team's efforts to design and monitor a robust AI model, errors in decision-making may occur if end-users lack a comprehensive understanding of how the AI system works or misinterpret its outputs; in this case, if loan officers or decision-makers, misinterpret or misapply the recommendations provided by the AI system. This could lead to incorrect loan approvals or rejections, potentially resulting in financial losses or missed opportunities for the institution. In such cases, there should be clear documentation on how to interpret AI outputs, and mechanisms for ongoing communication between the risk management team and end-users.	perspectives of end-users, gaining crucial insights into usability issues, bug identification, and feature enhancement suggestions. Moreover, involving end-users in the feedback loop not only ensures a more comprehensive understanding of system performance but also builds a sense of ownership and user engagement. This participatory approach establishes a feedback ecosystem that is more responsive, adaptable, and reflective of real-world usage scenarios.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		Comprehensive Governance Frameworks: Banks/Financial service providers should establish comprehensive governance frameworks that delineate unambiguous lines of accountability for the development and supervision of AI-based systems across their entire lifecycle, spanning from creation to implementation. This may necessitate enhancements to existing operational protocols related to AI. Internal model governance frameworks should be refined to more effectively encompass risks arising from AI utilization. **Imbibing accountability as a culture:** Financial Service Providers should institute a culture of accountability by emphasizing transparent communication and comprehensive documentation of AI systems' deployment.	

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		This entails clearly defining roles and responsibilities for all stakeholders involved in AI decision-making and enhancing governance at multiple management layers and business units. Robust policies and procedures should be established for effective oversight, monitoring, and regulatory compliance. Training programs and awareness initiatives are essential to educate staff on ethical AI usage and its implications. Developing a culture that places accountability at its core will ensure alignment with trustworthy AI practices throughout the organization.	
Principle 3: Fairness and Non-Discrimination			
Technical	**Ethical Data Collection:** AI developers in finance should prioritize the ethical collection of financial data, ensuring it is representative and does not lead to bias or discrimination against any specific customer groups based on religion, caste, gender,	**Fairness Audits:** Auditing mechanisms of the model and the algorithm that checks the results of the model against baseline datasets can help ensure that there is no unfair treatment or discrimination by the technology. One of the	

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	low purchasing power, etc. Various tools that help in this process are differential privacy, data masking, data collection through consent management platforms, etc. **Data Quality:** Ensuring high data quality in AI applications, with a focus on representativeness and relevance, is paramount. Technical solutions such as data cleaning, data labelling, data augmentation, and annotation play a crucial role in achieving this[76]. Representativeness ensures a comprehensive portrayal of the studied population, preventing bias and under-representation, particularly in financial markets and credit scoring, impacting model training and financial inclusion. Relevance focuses on data contributing to understanding the	ways to do this could be by deploying a "first-order temporal logic" tool[77]. This tool can systematically analyze the model and algorithm over time, comparing current results with baseline datasets. By applying temporal logic, which deals with the sequencing of events and changes over time, this tool can provide a dynamic assessment, allowing for the identification and rectification of biases. Temporal logic tools contribute to a more comprehensive understanding of how the AI system evolves, ensuring ongoing fairness and minimizing the risk of discriminatory outcomes. **User Testing:** Conduct user testing sessions with diverse user groups to assess the AI system's impact on different demographics. This research can help	

[76] Ataman, A. (2024, January 3). Data Quality in AI: Challenges, Importance & Best Practices. AIMultiple. https://research.aimultiple.com/data-quality-ai/

[77] GeeksforGeeks. (2023, February 22). Artificial Intelligence Temporal logic. https://www.geeksforgeeks.org/aritificial-intelligence-temporal-logic/

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	subject without misleading information. For example, in credit scoring, assessing data related to individuals' behavior and reputation is vital. Though evaluating vast datasets on a case-by-case basis may be challenging, it's essential for data accuracy and appropriateness, even if it introduces complexities in AI deployment efficiency. **Anonymization and Masking:** Sensitive attributes can be anonymized or masked to remove any direct references to protected characteristics. For instance, through the use of attribute-based credentials, gender or race labels can be replaced with generic labels like "Group A" or "Group B." **Fairness-aware learning:** Developers should explicitly incorporate fairness constraints during the training process.[78] This	identify potential biases or discrimination issues and provide insights for technical adjustments.	

[78] Jin, D., Wang, L., He, Z., Zheng, Y., Ding, W., Xia, F., & Pan, S. (2023). A survey on fairness-aware recommender systems. Information Fusion, 100, 101906. https://doi.org/10.1016/j.inffus.2023.101906

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	approach involves considering fairness as an integral part of the AI model's objective function, ensuring that fairness is optimized alongside accuracy and other performance metrics. For eg., a bank wants to use an AI model to automatically approve or deny loan applications. Fairness-aware learning in this context would involve incorporating fairness constraints into the model's training process to avoid discrimination based on protected attributes like religion, race or gender. Banks in this case can set a constraint that the approval rate for qualified applicants should be similar across different religious, racial or gender groups. **Bias Mitigation Techniques:** Developers must implement techniques to reduce bias in financial AI models, particularly those used in credit underwriting. This involves re-sampling underrepresented		

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	groups and optimizing algorithms for fairness. **Transparent Data Usage:** AI developers should be transparent about the sources and usage of financial data in their models, enabling end-users to understand and trust the data-driven processes behind financial products and services. To facilitate the same, the implementation of technical solutions such as data mapping and digital watermarking can be pivotal. These technologies provide a clearer picture of how financial data is sourced and utilized in AI models, enhancing transparency and building confidence in the data-driven processes underpinning financial products and services.		
Non-technical	**Diversity and inclusion in development teams:** Revamping employment policies is pivotal to truly implementing diversity and inclusion within AI development teams,	**Customer-Centric Approach:** Financial service providers should adopt a customer-centric approach, prioritizing customer satisfaction and fairness when	**Co-Creation of fairness standards:** Co-creation also places a shared responsibility on end-users to actively contribute and participate in shaping the development and

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	transcending mere numerical benchmarks. It is crucial to move beyond tokenism and foster a work culture that facilitates meaningful involvement of employees from diverse backgrounds in critical discussions. By prioritizing genuine diversity, organizations can ensure a spectrum of perspectives and experiences that go beyond fulfilling quotas. This shift in culture not only addresses biases more effectively but also cultivates creativity and innovation within AI development, ultimately resulting in the creation of more inclusive and equitable AI systems.	delivering financial products and services powered by AI. This can be achieved by establishing a comprehensive ethical AI framework aligned with industry standards and regulatory guidelines. This framework should guide the deployment of AI models, emphasizing transparency, fairness, and non-discrimination. Further, regular assessments to evaluate the impact of AI algorithms on different customer segments can be conducted. This involves analyzing data to identify any disparate impacts on certain demographics and taking corrective measures to address potential biases. By placing customer atisfaction and fairness at the forefront, financial service providers can build trust, enhance user experience, and contribute to the creation of a more inclusive and ethical financial ecosystem. **Co-Creation of Fairness Standards:** Deployers	deployment of these technologies. Users play a crucial role in providing insights, feedback, and perspectives that contribute to the creation of systems aligned with their needs and values. Actively engaging in co-creation activities empowers users to voice concerns, share experiences, and influence the ethical considerations embedded in AI systems. Moreover, users bear the responsibility of actively monitoring and ensuring that co-creation processes remain transparent, inclusive, and responsive to their evolving expectations. **Financial Data Literacy:** Developing financial data literacy among end-users is crucial. To build this literacy is a collective shared responsibility of the industry, civil society and the government. This is crucial in empowering users to comprehend the implications of data

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		should collaborate with end users in establishing fairness standards for AI systems. They can conduct research to understand their views on fairness and non-discrimination, and incorporate their input into the overall evaluation process. Ethicists, legal experts, sociologists, and others can contribute their expertise to define fairness standards that align with societal values and norms.[79] This interdisciplinary approach ensures that AI systems are designed and deployed in a manner that considers ethical and societal considerations.	usage in AI systems, enabling them to make well-informed financial decisions and safeguard their interests. Education campaigns should prioritize financial data literacy, equipping individuals with the knowledge needed to navigate the intersection of finance and AI responsibly. **Fairness Advocacy:** End-users play a pivotal role in advocating for fairness and non-discrimination in AI-driven financial products. To empower users, a checklist outlining key fairness considerations can be instrumental. This checklist serves as a practical guide, allowing users to evaluate the fairness aspects of AI-driven financial products before usage. By actively engaging with financial institutions and regulators to raise concerns and seek

[79] Mantelero, A. (2022). The social and ethical component in AI systems design and management. In Information technology & law series (pp. 93–137). https://doi.org/10.1007/978-94-6265-531-7_3

Level/ Stakeholder	AI Developer	AI Deployer	AI User
			solutions, users contribute to fostering a fairness-first ecosystem, ensuring responsible and equitable deployment of AI technologies in the financial sector.
Principle 4: Reliability and Safety/Robustness			
Technical	**Blind Test Sets:** AI developers in the finance sector should create blind test datasets specific to financial applications. These datasets should not be part of the model selection and validation process, providing a more accurate estimate of the model's generalization performance for financial scenarios. **Synthetic Financial Data:** Developers can explore the use of synthetic financial datasets for validation. Synthetic financial data offers a valuable alternative for testing	**Financial Audits:** Financial AI deployers should conduct regular audits of AI models to ensure their reliability and safety in financial applications. These audits are essential for risk management and to maintain the reliability of financial models. Some ways to do the same would be through third-party attestation or detailed reviews of control environments.[80] **Financial Concept and Data Drift Monitoring:** Ongoing monitoring in the finance sector is critical to detect and address concept drifts and data drifts specific	**User Testing in Finance:** Users in the finance sector should actively participate in user testing of AI systems and provide feedback on the reliability and safety of financial applications. Their feedback is invaluable for identifying financial-specific issues and enhancing model performance.

[80] KPMG. (2023, November 17). AI's role in enhancing trust in financial reporting & Capital markets. https://info.kpmg.us/news-perspectives/advancing-the-profession/ai-in-audit-kpmg-2023.html

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	and improving the robustness of machine learning models, especially in scenarios where real financial data is scarce or expensive. **Financial Model Validation:** Continuous validation processes should be in place for financial AI models. This validation goes beyond back-testing and ensures that the model's outcomes are reproducible, aligning with the financial industry's standards and objectives. It includes identifying financial-specific limitations, assumptions, and assessing potential financial impacts. **Inbuilt mechanisms to flag limitations:** It is important to communicate the limitations of the technology. Developers play a pivotal role in achieving this by employing various means, such as	to financial data. Concept drifts may arise from changing financial market dynamics, while data drifts can impact the predictive power of financial models. Timely monitoring and adaptation are crucial in the financial sector. **Financial Control Mechanisms:** Deployers should implement financial control mechanisms, such as "kill switches," to quickly shut down AI systems in high-risk financial circumstances. Kill switches serve as a safeguard, enabling the swift shutdown of an AI-based system if it deviates from its intended functionality[81]. For instance, in Canada, financial institutions are mandated to incorporate "override" functionalities that can either automatically halt system operations or provide the firm with the capability to do so remotely, ensuring the immediate response to any high-risk scenarios.	

[81] Fyler, T. (2023, August 22). The big red button: why do we need an AI kill switch? - TechHQ. TechHQ. https://techhq.com/2023/08/will-a-big-red-button-add-security-to-generative-ai/

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	incorporating alerts and pop-ups at the technical level. These mechanisms serve as proactive indicators, ensuring that users are informed about the constraints of the technology they are interacting with.		
Non-technical	**Certification and Accreditation Mechanism:** Developers should actively seek certification and accreditation mechanisms to demonstrate the reliability and robustness of their AI systems. Certifications such as ISO standards for AI[82] can establish adherence to globally recognized best practices, serving as a benchmark for excellence. Accreditation from reputable institutions or industry-specific bodies adds credibility, providing tangible assurances of the system's robustness. By proactively pursuing	**Human Preparedness:** Excessive dependence on fully automated AI-based systems poses a heightened risk of service disruption, potentially leading to systemic impacts within financial markets. In scenarios where these markets encounter technical or other disturbances, financial service providers must maintain preparedness in terms of human resources. This requires well-trained human counterparts who can promptly step in to replace automated AI systems, serving as a human safety net to avert any market disruptions. It's essential to acknowledge,	**Prioritize Financial Explainability:** Users should prioritize the use of financial AI systems that provide human-meaningful explanations. Having said that, it is crucial for explanations to be understandable; users should also be mindful of the complexity inherent in financial processes. Striking a balance between simplicity and necessary detail is key. Research indicates that understandable explanations positively influence the perception of system accuracy, even in the finance sector. This focus on financial explainability enhances

[82] ISO. (2023, September 21). Artificial intelligence (AI) standards. https://www.iso.org/sectors/it-technologies/ai

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	these mechanisms, developers not only showcase their commitment to quality but also contribute to building trust and confidence among users and stakeholders. **Outline Limitations:** Explicitly outlining limitations in procurement contracts with deployers, both in technical and non-technical terms, fosters a shared understanding of the system's capabilities and constraints. This proactive communication helps manage expectations, avoids misuse, and establishes a foundation for responsible deployment.	though, that not all fundamental financial services delivery systems can be replaced by human counterparts. **Clear Communication:** Promoting reliability and safety in AI adoption can be achieved through clear communication regarding the integration of AI and the protective measures in place for the system and its users. This communication can be established through channels like workshops, webinars, blogs, etc. In the context of easily accessible domestic and cross-border financial services, it's vital to establish and uphold a multidisciplinary dialogue between policymakers and industry stakeholders at both the national and international levels. This collaborative effort enhances understanding and cooperation while facilitating the adoption	reliability and user confidence.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		of innovative AI techniques.	
Principle 5: Human Autonomy and Determination			
Technical	**Establish feedback mechanisms:** AI developers in the finance sector should create user-friendly interfaces that allow customers to provide feedback and challenge AI model outcomes. For example, in the case of a robo-advisory service, if a customer disagrees with investment advice provided by an AI-driven system, there should be a straightforward process for the user to voice their concerns. The feedback should be carefully documented and used to enhance model performance and customer satisfaction. AI-enabled Interactive Voice Response (IVR) systems can be deployed to enable this feedback loop. **Enable user-driven adjustments:** Developers should offer financial service platforms that allow users to customize their	**Establish clear dispute resolution procedures:** Financial institutions and deployers of AI systems must create well-defined technical protocols for addressing customer challenges and seeking redress when AI model outcomes lead to disputes or dissatisfaction. These procedures should be prominently communicated to customers to ensure they are aware of the mechanisms available for intervention. In the event of discrepancies, deployers should have mechanisms in place to investigate, mediate, and resolve issues. **Maintain human oversight:** While AI models may be entrusted with certain financial tasks, there should be a structured framework for human intervention when necessary. This oversight ensures that	**Utilization of Feedback Mechanisms:** End-users can take advantage of feedback mechanisms provided by banks/financial institutions. When using AI-based financial tools, users should engage with the feedback and dispute resolution channels to express concerns, provide input, or challenge AI model outcomes. Constructive feedback can influence system improvements and help users feel more in control of their financial activities.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	preferences and adjust risk tolerance, investment strategies, or other relevant parameters. For example, financial service platforms can provide users with real-time portfolio customization, ethical investment preferences, goal based investing, etc. as means to exercise greater control and customization. By providing such customization options, customers retain control over AI-driven financial decisions, and this feature not only empowers users but also ensures a more personalized experience.	AI systems do not make significant financial decisions in isolation, reducing the potential for errors and losses.	
Non-technical		**Promote user empowerment:** Financial institutions should encourage their customers to actively engage with AI-driven systems and take an active role in decision-making. By fostering a sense of empowerment and responsibility, end-users can feel more in control of their financial	**Financial Education:** End-users should invest time in educating themselves about AI-based financial tools and their implications. Understanding the basics of AI, machine learning, and the specific AI-driven services they use empowers individuals to make informed decisions. Knowledge

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		activities. This can be achieved through educational campaigns, user guides, and providing accessible resources for understanding AI-powered tools. **Utilize accessible feedback mechanisms:** Deployers should inform end-users about the various channels available to them for challenging AI model outcomes or expressing concerns. For instance, if a customer believes that their loan application was unfairly rejected by an AI-driven credit scoring model, they should be encouraged to utilize the provided channels to report the issue. In the finance sector, accessible feedback mechanisms are instrumental in ensuring that AI systems are continuously improved to meet customer expectations and regulatory standards. **Empower human decision-making:** Deployers should establish internal	equips users to better interpret AI model outcomes and assess their alignment with personal financial goals. **Active Engagement:** Users can take an active role in their financial decision-making processes. While AI systems can provide valuable insights and assistance, individuals should not relinquish all responsibility. By staying engaged and continuously monitoring their financial activities, users can exercise their autonomy and make adjustments when necessary. This includes regularly reviewing their investments and financial plans. **Leverage Regulatory Protections:** Users can stay informed about financial regulations and consumer protections related to AI applications. By understanding their rights and the regulatory landscape, individuals can leverage protections provided by authorities in case of

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		organizational policies that incorporate a human-in-the-loop approach, granting autonomy to individuals involved in decision-making. In instances where AI-generated suggestions may malfunction, it is imperative to empower humans on the ground with the autonomy to deviate from automated decisions, albeit within a framework of checks and balances. This policy ensures that human judgment prevails, allowing for course corrections and mitigating the potential risks associated with over-reliance on AI systems. Striking a balance between AI assistance and human intervention enhances the adaptability and resilience of decision-making processes within an organization.	disputes or concerns related to AI-driven financial services.
Principle 6: Privacy and Data Protection			
Technical	**Privacy-Focused Data Handling:** Developers should implement robust data encryption	**Secure Data Sharing Protocols:** Financial service providers may need to share data for	**Informed Consent:** Financial AI users should carefully read consent agreements

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	techniques to protect sensitive financial information. For example, in a banking AI application, all customer data, including account numbers and transaction details, should be encrypted both in transit and at rest to prevent unauthorized access, even if the data is intercepted or stored improperly.[83] **Data Minimization:** Developers can reduce the risk of data breaches by only collecting and using the minimum amount of data required for AI models. For instance, in a credit scoring model, developers should only consider relevant financial information like credit history, income, and debt, rather than collecting extensive personal data that isn't necessary for the model's purpose.	services like credit checks or fraud detection. Implement secure APIs and data sharing protocols to ensure that data is encrypted during transmission. For example, when a bank shares financial data with a credit bureau, they should use encrypted connections to protect customer data during the transfer. **Transparent Data Usage Policies:** Deployers should provide clear data usage policies to their customers. For example, an online investment platform should offer a transparent data usage policy that explains how customer financial data is handled, stored, and shared, processed to instill trust in the AI-powered service. **Risk Management Strategies:** Develop documented risk management strategies focused on mitigating	before using AI-driven financial services. For instance, when signing up for a mobile payment app, users should understand and agree to the app's data handling practices, especially how their financial data will be used and shared. **Data Monitoring:** Users can actively monitor how financial institutions handle their data. For example, customers can inquire about a bank's data-sharing practices, such as whether the bank shares their financial data with third parties for marketing purposes.

[83] For instance, 'Guardrails in Amazon Bedrock' automatically detects and prevents queries and responses that fall into restricted categories. With its help, developers can tailor AI systems to detect Personally Identifiable Information (PII) in user inputs and FM responses and selectively reject inputs containing PII or redact PII in FM responses. https://aws.amazon.com/bedrock/guardrails/

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	Privacy Enhancing Technologies (PETs): PETs are gaining traction as a means to protect data privacy. PETs aim to maintain the fundamental properties and characteristics of original data while concealing individual data samples. This encompasses techniques like differential privacy, federated analysis, homomorphic encryption, and secure multi-party computation. Notably, differential privacy provides rigorous mathematical guarantees for achieving the desired level of privacy without compromising accuracy, outperforming synthetic datasets. The key advantage of these methods is that models trained on synthetic data, as opposed to actual data, exhibit minimal performance loss, ensuring data privacy while maintaining AI model efficacy.	risks related to data quality and trading algorithm vulnerabilities in response to regulatory changes. Documented risk management strategies refer to well-defined plans and protocols that outline how an organization intends to identify, assess, and address risks associated with specific aspects of its operations. These strategies involve creating clear documentation that articulates the steps and measures to be taken to ensure data quality, address algorithm vulnerabilities, and adapt to regulatory shifts. The documentation may include detailed risk assessment procedures, preventive measures, and response protocols to minimize the impact of potential threats, providing a systematic and organized approach to risk management in the specified domains.	

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Non-technical	**Privacy as a Value Proposition:** Businesses must prioritize privacy as a fundamental value proposition, aligning with a privacy-first culture. Developers should stay informed about data privacy regulations, such as the Digital Personal Data Protection Act in India, and ensure AI models comply with these standards. This approach not only helps avoid legal issues related to financial data privacy but also contributes to building trust with users over the long run.	**Prioritizing privacy-safe technologies:** To reinforce the importance of privacy, deployers should prioritize and demand privacy-safe AI technologies during procurement. By making privacy a key criterion, deployers create market incentives for developers to prioritize and integrate robust privacy measures in their AI solutions. This approach not only safeguards user privacy but also encourages developers to proactively address privacy concerns in their technology, aligning with ethical standards and legal regulations. **Data Ethics Training:** Offering data ethics training for developers is essential[84]. For example, financial institutions can conduct workshops and seminars on data ethics and privacy policies to educate their AI	**Championing Privacy Protections:** End-users have the opportunity to actively engage in conversations surrounding data privacy in financial services. By connecting with advocacy groups and participating in dialogues with policymakers, they can help shape regulations that emphasize robust data privacy and security within the finance sector. **Awareness:** Individuals must possess a comprehensive understanding of their digital rights and exercise them responsibly. This knowledge empowers them to navigate the digital landscape adeptly, make informed choices regarding AI interactions, and advocate for their rights in the evolving technological landscape. Going beyond mere acknowledgment, this awareness fosters

[84] Data Security Council of India (DSCI). (2021, July). Handbook on Data Protection and Privacy for Developers of Artificial Intelligence (AI) in India: Practical Guidelines for Responsible Development of AI. https://www.dsci.in/files/content/knowledge-centre/2023/AI-Handbook.pdf

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		development teams about best practices in handling financial data.	active participation in shaping the ethical development and deployment of AI.
Principle 7: Social and Environmental Sustainability			
Technical	**Establish energy-efficient algorithms:** Developers should consider the energy consumption and carbon footprint of AI systems. Energy-efficient algorithms and hardware choices can contribute to reducing the environmental impact of AI operations. By optimizing resource usage, developers can mitigate the contribution of AI to energy consumption and greenhouse gas emissions. **ESG Ratings Algorithms:** Developers should establish AI algorithms specialized for ESG ratings. These algorithms should be optimized to analyze financial data while considering sustainability metrics,	**AI-Enhanced Due Diligence:** Implement AI-based due diligence tools for the finance sector, focusing on ESG-centric investments. These tools should utilize AI capabilities to analyze financial data and ESG metrics simultaneously, helping in the selection of sustainable investment opportunities. This entails evaluating potential AI vendors based on their commitment to ethical practices, transparency, and environmental responsibility. By choosing AI systems that align with sustainability standards, financial institutions can demonstrate their dedication to responsible technology adoption.[85]	

[85] Aldboush, H. H. H., & Ferdous, M. (2023). Building Trust in Fintech: An analysis of ethical and privacy considerations in the intersection of big data, AI, and customer trust. International Journal of Financial Studies, 11(3), 90. https://doi.org/10.3390/ijfs11030090

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	ensuring a balance between profitability and social responsibility. **Energy Efficient Models:** AI developers can focus on developing AI models that prioritize energy efficiency and minimize their carbon footprint.[86] This can be achieved by adopting energy-efficient computing resources and optimizing algorithms to reduce energy consumption during inference and training. Such measures help mitigate the environmental impact of AI systems and contribute to overall sustainability.		
Non-technical	**Collaboration with relevant stakeholders:** Collaboration with environmental organizations and stakeholders is another key strategy. By working together, AI developers and users can establish sustainability standards	**Social Impact Assessments:** Considering social impact assessments before deploying AI systems is crucial for promoting social sustainability. Social impact assessments involve analyzing the potential effects of AI	**Public Awareness Campaigns:** End-users can actively contribute to promoting the significance of social and environmental sustainability in AI through their actions and choices. By staying informed about the impact of AI

[86] Van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. AI And Ethics, 1(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	for AI development and deployment.[87] These standards can include guidelines for minimizing carbon emissions, reducing energy consumption, etc[88]. Having clear sustainability standards in place ensures that AI technologies are developed and used in a manner that aligns with environmental and societal goals. Incorporating public input into the development process allows developers to identify potential ethical or sustainability challenges early on and make necessary adjustments to ensure that AI solutions are socially responsible and environmentally conscious.	technologies on individuals and society as a whole, with a focus on fairness, inclusivity, and equal treatment. By conducting these assessments, AI developers can identify and address biases, discriminatory outcomes, and potential harm to vulnerable populations. This ensures that AI systems are designed to enhance human welfare, protect human rights, and align with societal values. **Corporate Social Responsibility (CSR):** Integrate AI-driven ESG strategies with corporate social responsibility initiatives[89] within the financial sector. Align the finance sector's values and sustainability goals with AI-enhanced investment practices that cater to ESG considerations.	technologies on society and the environment, individuals can raise awareness and encourage the adoption of responsible AI practices by organizations. Sharing knowledge about sustainability and advocating for responsible AI within their communities and social networks can create social pressure on financial institutions, leading to increased demand for ethical and sustainable AI solutions.

[87] González-Gonzalo, C., Thee, E. F., Klaver, C. C. W., Lee, A., Schlingemann, R. O., Tufail, A., . . . Sánchez, C. I. (2022). Trustworthy AI: Closing the gap between development and integration of AI systems in ophthalmic practice. Progress in Retinal and Eye Research, 90, 101034. https://doi.org/10.1016/j.preteyeres.2021.101034

[88] Accenture. (2023, October 18). How do we make generative AI green? https://www.accenture.com/us-en/blogs/consulting/making-generative-ai-green

[89] PricewaterhouseCoopers. (2022, June 2). Responsible AI and ESG: The power of trusted collaborations. PwC. https://www.pwc.com/us/en/tech-effect/ai-analytics/the-power-of-pairing-responsible-ai-and-esg.html; See Walsh, B. (2023, November 28). Revolutionizing ESG Reporting with AI: A Critical Move for Today's Businesses. https://www.wwt.com/article/revolutionizing-esg-reporting-with-ai-a-critical-move-for-todays-businesses

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Principle 8: Governance and Oversight			
Technical	**Model Governance Frameworks[90]:** It's essential to create dedicated model governance frameworks to ensure stringent compliance with local financial regulations. Such governance frameworks should be tailored to align seamlessly with the Reserve Bank of India (RBI) and Securities Exchange Board of India (SEBI) guidelines, ensuring that AI applications in finance fully comply with Indian financial laws and regulations.	**Risk Assessment:** Conduct in-depth risk assessments, considering scenarios like market crashes, to evaluate the impact of AI-driven trading systems on market stability. For example, SEC in the US proposed new requirements to address risks to investors from conflicts of interest associated with the use of predictive data analytics[91]. By systematically evaluating the potential impacts, deployers can establish contingency plans, risk mitigation strategies, and regulatory measures tailored to ensure the responsible and stable operation of AI-driven trading systems.	**Vigilance and Reporting:** Financial end-users should stay vigilant when using AI-driven services and promptly report any irregularities or unfair practices. For instance, reporting algorithmic trading anomalies that may affect their investments. **Cybersecurity Awareness:** Stay vigilant about AI-driven cybersecurity practices, especially when handling online banking and investment portfolios, to guard against potential cyberattacks. Recognize the importance of self-governance as an additional layer of protection, acknowledging that despite all implemented measures, there may be vulnerabilities that require individual attention.

[90] Micagni, A. (2023, October 31). AI in Financial Services: Regulatory Frameworks. Grand. https://blog.grand.io/ai-in-financial-services-regulatory-frameworks/

[91] U.S. Securities and Exchange Commission (SEC). (2023, July 26). SEC Proposes New Requirements to Address Risks to Investors From Conflicts of Interest Associated With the Use of Predictive Data Analytics by Broker-Dealers and Investment Advisers. https://www.sec.gov/news/press-release/2023-140

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Non-technical	**Voluntary Governance:** Incentivize employees, especially technology professionals, to develop trustworthy AI systems. This approach involves encouraging individuals within an organization to voluntarily adhere to ethical principles and standards in AI development. By offering incentives, such as recognition, career advancement opportunities, or financial rewards, organizations motivate their tech teams to prioritize the creation of AI systems that align with ethical considerations, user privacy, and societal well-being. This voluntary governance model relies on the intrinsic motivation of employees to contribute to responsible AI practices, fostering a culture of ethical innovation and responsible technology development within the organization. It complements	**Oversight Committee:** Establish a dedicated oversight committee with a deep understanding of financial regulations and compliance standards to ensure rigorous adherence. **Regulatory Liaison:** Maintain a robust relationship with regulatory authorities to remain up-to-date on evolving financial regulations and ensure alignment with compliance standards, particularly in AI-driven trading.	**Consumer Choice:** Opt for financial service providers with a track record of ethical AI practices, regulatory compliance, and transparent operations.

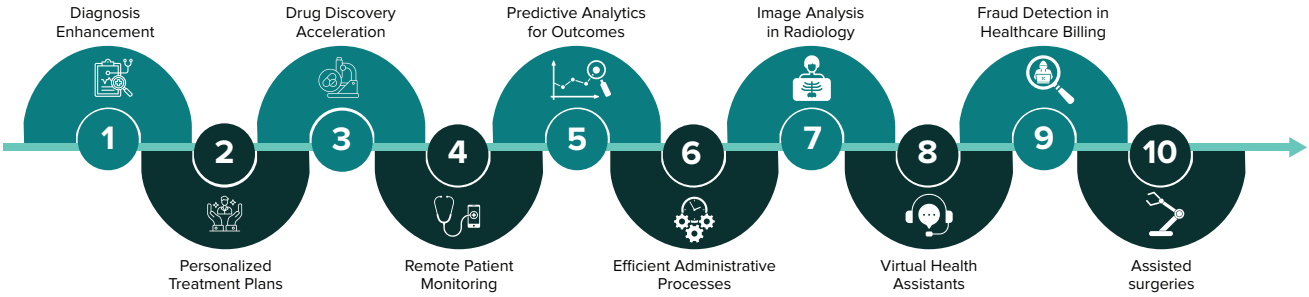
Level/ Stakeholder	AI Developer	AI Deployer	AI User
	regulatory frameworks and formal governance structures by tapping into the commitment and values of individual employees to ensure the responsible deployment of AI technologies.		
Principle 9: Contestability			
Technical		**Standardized Redress Mechanisms:** Financial institutions, such as banks and credit-lending institutes, should create accessible channels through which customers can initiate inquiries, appeals, or reviews of AI-driven decisions that directly impact them. This mechanism empowers customers to contest points in decision-making where they have concerns or disagreements.	
Non-technical			**Awareness and Education:** End-users should proactively educate themselves about AI's role in their financial interactions. It's crucial to understand

Level/ Stakeholder	AI Developer	AI Deployer	AI User
			their legal rights under privacy regulations and consumer protection laws, as these regulations often provide mechanisms for contesting AI-based decisions. Being familiar with privacy policies and the terms of service related to AI-driven financial products ensures that users are aware of how their data is handled. Additionally, end-users should acquaint themselves with the specific redress procedures offered by financial institutions or service providers and provide feedback when issues arise. Seeking assistance from consumer protection agencies or legal advisors with expertise in AI-related matters can also be valuable if challenges occur during the contestation process.

3.2.2 Healthcare

In the last few years, AI has achieved great strides in the field of healthcare and is further anticipated to revolutionize the field. This has largely been enabled by the exponential rise in the volume of electronic health data. With the digitization of medical records, the healthcare industry now has access to vast amounts of patient information, including medical histories, diagnostic images, and treatment outcomes. Access to this vast repository of data allows developers and researchers to efficiently process and analyze the available data to derive valuable insights and patterns.

Figure 5: Use cases of AI in Healthcare



AI in healthcare has a wide range of applications. One significant application is in clinical decision-making, where AI-powered algorithms help healthcare professionals by providing data-driven recommendations, assisting with diagnoses, and offering personalized treatment plans based on individual patient characteristics.[92] This leads to more accurate diagnoses and optimized treatment strategies, ultimately improving patient outcomes. Furthermore, AI is invaluable in advancing biomedical research and drug development. AI algorithms have the ability to analyze large datasets and identify potential drug candidates, which greatly speeds up the drug discovery process. Here, the existing medical records play an integral role in developing healthcare AI systems. The secondary use of medical records involves repurposing patient data collected for clinical purposes to train artificial intelligence models. This practice leverages diverse and extensive datasets to improve AI accuracy, identify patterns, and develop predictive models for personalized medicine. It contributes to quality improvement, research, and better patient outcomes.

Box 8: Case Study - Using AI to Identify Tissue Growth from CT scans[93]

In collaboration with the NHS AI Lab Skunkworks team, George Eliot Hospital embarked on a groundbreaking project to harness the power of AI in the analysis of CT scans. The central objective of this initiative was to streamline the assessment of patient scans, making it quicker and more precise, while also automating the identification of organs, growths, and the detection of anomalies. The project involved a multifaceted approach, including tissue sectioning, anomaly detection, and scan alignment.

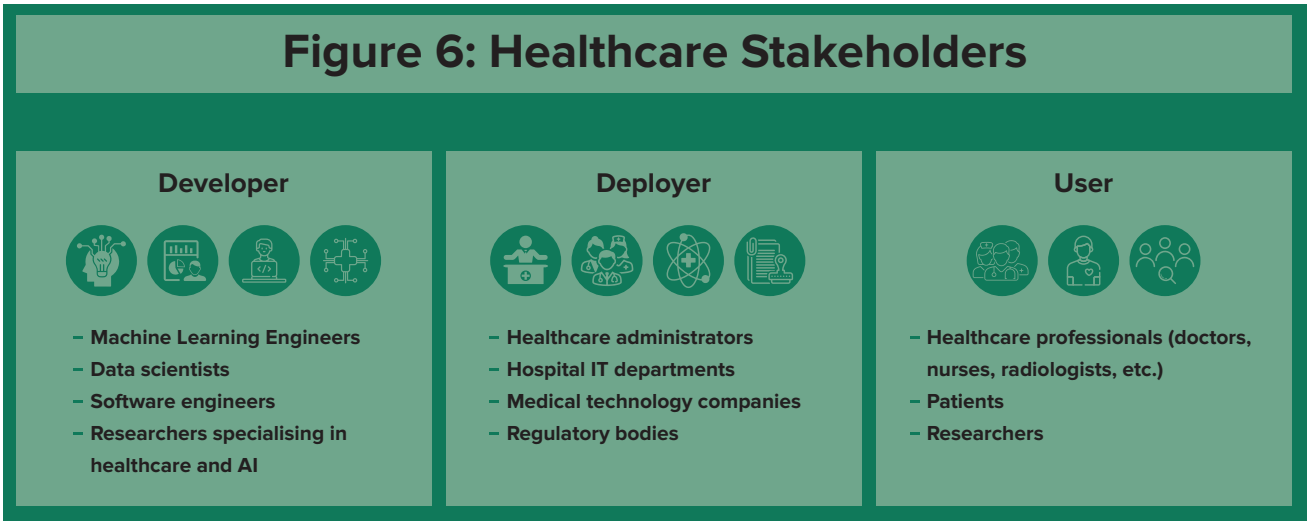
The results of this venture showed considerable promise. The team achieved significant success in automating scan alignment, producing both rigid and non-rigid 3D alignments. Although the alignment methods were an improvement on manual alignment, they weren't flawless, especially when patients experienced variations in their body positions during scans, such as inhaling or exhaling. Furthermore, the project delivered precise overlay capabilities for both 3D and 2D images, even when manipulating the images

92. Artificial Intelligence (AI) In Healthcare & Hospitals. (n.d.). Retrieved from https://www.foreseemed.com/artificial-intelligence-in-healthcare
93. NHS Transformation Directorate. (2021, November 17). Using AI to identify tissue growth from CT scans. https://transform.england.nhs.uk/ai-lab/explore-all-resources/develop-ai/using-ai-to-identify-tissue-growth-from-ct-scans/

through zooming, rotation, or panning. This feature enhances the radiologist's ability to make comparisons between scans efficiently and effectively. In terms of new tissue growth detection, the project managed to measure anomalies in 3D. The tool offers a less manual and potentially time-saving process for medical professionals.

However, despite its benefits, the widespread use of AI in healthcare also comes with challenges and risks. One major concern is the potential for unintended consequences and biases when AI systems are applied on a large scale.[94] While AI models may perform exceptionally well in controlled settings with curated and standardized data, they can face difficulties when exposed to real-world health data, which is often diverse and not standardized.[95] Additionally, AI models may inadvertently perpetuate biases present in the data they were trained on, resulting in unfair or unequal treatment for certain patient groups. Therefore, to harness the full potential of AI in healthcare while mitigating associated risks, it is crucial to adopt a thoughtful and collaborative approach.

Figure 6: Healthcare Stakeholders



Development of trustworthy AI in the healthcare sector would involve three key stakeholders: *first*, AI developers and *second*, AI Deployers and *third*, AI users, i.e. healthcare experts, medical practitioners, hospitals, government etc.[96] Unlike the first group, these users do not possess extensive expertise in AI development. Instead, they interact with AI technologies in their respective fields, leveraging the capabilities of AI without being directly involved in creating or refining the underlying algorithms. In the following section, we explore how these groups can help implement trustworthy AI in the healthcare sector.

[94] Obermeyer, Z., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366, 447-453. https://doi.org/10.1126/science.aax2342; Mittermaier, M., Raza, M. M., & Kvedar, J. C. (2023). Bias in AI-based models for medical applications: Challenges and mitigation strategies. npj Digital Medicine, 6, 113. https://doi.org/10.1038/s41746-023-00858-z

[95] Khalid, N., Qayyum, A., Qayyum, A., Al-Fuqaha, A., & Qadir, J. (2023). Privacy-preserving artificial intelligence in healthcare: Techniques and applications. Computers in Biology and Medicine, 158, 106848. https://doi.org/10.1016/j.compbiomed.2023.106848

[96] Survey of Explainable AI Techniques in Healthcare. (2023, January 5). NCBI. Retrieved August 29, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9862413/

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Principle 1: Transparency and explainability			
Technical	**Interpretable systems:** Healthcare AI systems use patients' clinical history to predict future diagnoses, making them valuable for everyday clinical practice. However, the complex nature of AI models makes them unexplainable and uninterpretable, thus making it hard to understand how and why a particular choice was made, limiting their use in real-world healthcare.[97] To avoid the lack of interpretability, the AI models should be built to provide explanations for the decisions in the form of decision trees or rule based thus making it easier to comprehend by daily AI user groups i.e. healthcare professionals. Further, attention mechanisms should be used to highlight specific regions or elements in medical images or patient records that	**Third-Party Audits:** Seek third-party audits and certifications to validate the transparency and fairness of AI systems.	**User-friendly interface:** For healthcare professionals engaging with AI systems, the adoption of explanatory interfaces is pivotal for enhancing transparency and understanding. These interfaces, characterized by user-friendly designs and intuitive displays, serve as conduits for clear explanations of AI decisions. User-friendly dashboards offer succinct displays of crucial information, encompassing inputs, outputs, and performance metrics, accompanied by indicators of model confidence levels. Educational resources, seamlessly integrated within the interface, can guide healthcare professionals through the intricacies of AI workings. Customizable settings empower users to tailor the depth of explanations based on their expertise, while feedback mechanisms and collaboration with

[97] Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	contribute most to the AI's decision-making process. This enables clinicians to focus on crucial areas.[98] **Comprehensive documentation:** AI developers should provide comprehensive documentation that details the AI model's development, the data used, the algorithms employed, and the rationale behind design choices. This documentation helps users understand the AI system's functioning.		UX experts ensure continuous refinement of these interfaces. In this symbiotic relationship between healthcare professionals and AI, explanatory interfaces not only promote transparency but also contribute to the seamless integration of AI technologies, ultimately improving patient outcomes and decision-making within the healthcare sector.
Non-technical	**Collaborate with Experts:** Work closely with healthcare professionals to ensure that the AI models align with clinical insights and requirements.	**Informed consent:** User groups that are providing data sets to developers for training and building of models should ensure that the patient is informed about the same and has consented to the use of his sensitive personal data for this purpose specifically. However, in some cases, data protection laws of a	**Training healthcare professionals and students:** To ensure transparency in AI systems used by medical practitioners, they must understand how the system functions, its potential impact, and its advantages and disadvantages. This requires education and training of healthcare

[98] Chen, P., Dong, W., Wang, J., Lu, X., Kaymak, U., & Huang, Z. (2020). Interpretable clinical prediction via attention-based neural network. BMC Medical Informatics and Decision Making, 20(S3). https://doi.org/10.1186/s12911-020-1110-7

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		country may absolve users from this obligation. For instance, where due to urgency like epidemic, taking individual consent for research and development purposes is not feasible. In such cases, steps should be taken to apprise the population as soon as possible to ensure transparency right from the beginning. Further, once the AI system is being used in the medical decision-making process, patients must be informed about the intention, outcome and limitations of using AI technologies. **Foster a culture of transparency:** It is imperative to establish a culture and policy framework that explicitly communicates to end-users and affected populations when AI technology has been employed for medical diagnosis or other health-related outcomes. Transparency is a fundamental principle in	personnel through internal workshops and orientation programs. Additionally, students in medical colleges should be imparted adequate education to equip them with understanding of ethical considerations like data privacy, bias mitigation, and transparency in AI algorithms. **Advocate for Transparency:** Encourage transparency initiatives within the healthcare organization, promoting open discussions about AI systems and their implications. **Collaborate with Developers and Deployers:** Foster collaboration with developers and deployers of AI systems in a healthcare organization. Provide constructive feedback on the usability, accuracy, and relevance of AI applications. Engage in open communication to ensure that developers understand the unique

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		the ethical deployment of AI in healthcare, as individuals often remain unaware that their health-related decisions are influenced by AI systems. This communication can take various forms, such as clear and easily understandable notifications in patient portals, informed consent procedures, or informational materials distributed by healthcare providers.	challenges and requirements of clinical practice.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Principle 2: Accountability and Responsibility			
Technical	**Review and Redressal Mechanisms:** To address concerns and issues related to the AI system's performance, developers should establish robust review and redress mechanisms tailored for the healthcare sector. These mechanisms should facilitate constructive feedback from healthcare practitioners, patients, and stakeholders, creating a dynamic feedback loop for continuous improvement in AI technology. Within the healthcare setting, these mechanisms can be implemented through user-friendly interfaces, such as dedicated portals or applications, where practitioners and patients can provide feedback on AI-driven diagnostic or treatment suggestions. Additionally, incorporating advanced analytics tools can aid in analyzing patterns in feedback to identify recurring issues and areas for enhancement.	**Feedback Loops:** Establish channels for patients and healthcare professionals to provide feedback on AI-assisted diagnosis or treatment recommendations. Utilize innovative solutions to enhance user engagement and gather valuable insights for continuous improvement.	

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	Regular forums, workshops, or training sessions can also be organized to ensure effective communication between developers and healthcare professionals, fostering a collaborative environment for ongoing refinement of AI applications in healthcare. **Adverse Outcome Protocols:** Work closely with senior healthcare management to develop protocols for addressing accountability challenges and adverse outcomes. In healthcare, this might involve defining clear procedures for cases where AI recommendations conflict with established medical protocols. Ensure that healthcare professionals can easily report discrepancies and that there are processes in place to investigate and rectify any issues.		

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Non-technical	**Code of conduct:** To ensure accountability and responsibility in AI systems deployed in the healthcare sector, AI developers must take several crucial steps. AI developers should draft a thorough and well-defined Code of Conduct that outlines the guiding principles and intentions behind the design, development, and deployment of AI systems in healthcare. This code should explicitly state the commitment to uphold fundamental rights, accuracy, safety, transparency, fairness, and the minimization of harm. The code acts as a moral compass for developers, setting the ethical foundation for their AI applications. These standards should be tailored to the healthcare domain and take into account the unique challenges and sensitivities associated with patient data and well-being. The AI system must adhere to these principles to ensure that it	**Code of conduct:** Hospitals and healthcare organizations should create a code of ethics for AI adoption, outlining the principles that guide the responsible use of AI technologies in patient care. This code should prioritize patient welfare, ensure transparency in decision-making, and address any potential conflicts of interest. **Liability Clauses in Contracts:** When healthcare institutions procure AI systems from vendors, the contract should include liability clauses and service-level agreements and delineate the accountability and liability of each stakeholder precisely. **Ethics Boards:** Establish ethics review boards within healthcare organizations to evaluate the ethical implications of AI applications. These boards can include	**Education on AI in Healthcare:** Healthcare Professionals and Patients should strive to stay informed about the application and implications of AI in healthcare. Understand how AI systems are used in diagnostic processes, treatment recommendations, and other clinical workflows. This knowledge empowers healthcare professionals and patients to make informed decisions regarding AI-assisted care. **Ethical Guidelines:** Healthcare professionals should familiarize themselves with the ethical guidelines governing the use of AI in healthcare. This understanding helps users assess the ethical implications of AI applications and advocate for adoption of transparent and accountable AI systems that prioritize patient well-being.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	contributes positively to patient care and the overall healthcare ecosystem. **Transparent Decision Framework:** Collaborate with senior healthcare management and compliance teams to establish a robust accountability framework. Define a transparent decision-making framework for AI-driven clinical applications, clearly outlining responsibilities at different organizational levels. For example, in the context of a diagnostic AI tool, articulate the roles of healthcare professionals, data scientists, and administrators in ensuring accurate and ethical diagnoses. **Stakeholder Consultation:** Collaborate closely with healthcare business and compliance teams to understand the potential implications of AI-driven decisions. In healthcare, involve	representatives from various departments, ensuring that ethical considerations are thoroughly examined. For example, in the context of AI-assisted surgeries, an ethics review board may assess the impact on patient safety and consent. **Clinical Risk Management Teams:** Establish internal clinical risk management teams responsible for overseeing the deployment of AI-based diagnostic and treatment decision support systems. These teams should thoroughly understand the AI models they oversee. **Acknowledging User Misinterpretation:** Recognize potential caveats where mistakes may arise from end-users, such as healthcare professionals misinterpreting AI outputs. For example, in a scenario where an	**Participate in User Feedback Mechanisms:** Healthcare professionals should actively participate in user feedback mechanisms established by healthcare organizations. They should share their experiences with AI systems, including any challenges faced or successes observed. This engagement contributes to the continuous improvement of AI applications and promotes accountability.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	clinicians, nurses, and other frontline healthcare professionals in the development process to gain insights into the real-world impact of AI applications on patient care. Emphasize stakeholder engagement in the decision-making process. This involves actively seeking input from clinicians, patients, and other relevant stakeholders to incorporate diverse perspectives. For instance, when developing AI tools for treatment planning, involve oncologists, nurses, and patients to ensure the technology aligns with the holistic needs of cancer care.	AI-driven diagnostic system is deployed, errors in clinical decisions may occur if healthcare practitioners misunderstand the system's recommendations. Develop mechanisms to address such situations, including clear documentation on how to interpret AI outputs and ongoing communication between the risk management team and healthcare professionals.[99] **Imbibing Accountability as a Culture:** Institute a culture of accountability by emphasizing transparent communication and comprehensive documentation of AI systems' deployment in healthcare settings. Clearly define roles and responsibilities for healthcare professionals, data scientists, and	

[99] For instance, AWS HealthScribe automatically creates clinical notes from patient-clinician conversations using generative AI. Every AI-generated summary statement comes with traceable transcript references that make it easier for clinicians or scribes to quickly verify accuracy and locate the source of the insight. https://aws.amazon.com/healthscribe/

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		administrators involved in AI decision-making.	
Principle 3: Fairness and Non-discrimination			
Technical	**Clean and inclusive data sets:** It is crucial to address potential biases in the design and deployment of adaptive AI systems in healthcare. Most algorithms are trained on electronic health records (EHRs). EHRs mostly contain data of people who have access to healthcare thus leaving a large section of society with limited healthcare access out of the datasets. EHRs therefore may not capture information from certain individuals or have consistent data structures, leading to possible replication of human cognitive errors by the AI model.[100] Consequently, inaccurate medical diagnoses may happen due to the absence or inadequacy of data	**Feedback mechanisms:** Patients and healthcare professionals should have accessible and user-friendly channels to report any issues or biases they observe in AI-driven healthcare systems. These feedback mechanisms should be integrated into user interfaces, such as dedicated portals, mobile apps, or interactive platforms, ensuring a seamless and straightforward process for users to share their observations. **Fairness Metrics and Monitoring:** Implement fairness metrics during the development phase and establish continuous monitoring mechanisms in production. Regularly assess the model's	

[100] International Medical Device Regulators Forum. (2013, December 18). Software as a Medical Device (SAMD). https://www.imdrf.org/working-groups/software-medical-device-samd

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	pertaining to certain vulnerable groups. The unfair data sets can further lead to AI models that propagate discriminatory treatment,[101] especially AI deployed in the medical insurance sector. Addressing these biases is therefore vital to ensure fair and personalized healthcare outcomes. The developers should adequately educate themselves of the nature of the datasets being used to train AI models and should further ensure that data sets correspond to diverse sections. The developers should ensure that the characteristics of the training dataset account for variations in age, gender, ethnicity, and geographic locations to avoid algorithmic biases that may disproportionately affect certain patient groups.[102]	outputs across demographic groups to identify and address any disparities in predictions, ensuring fair treatment for all patients.	

[101] Grant, C. (2023, February 24). Algorithms are making decisions about health care, which may only worsen medical racism. American Civil Liberties Union. https://www.aclu.org/news/privacy-technology/algorithms-in-health-care-may-worsen-medical-racism

[102] For example, 'Amazon SageMaker Clarify' helps in mitigating bias by detecting potential bias during data preparation, after model training, and in the deployed model by examining specific attributes. https://aws.amazon.com/sagemaker/clarify/

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Non-technical	**Diverse and Inclusive Development Teams:** Foster diversity and inclusivity within AI development teams. Include professionals with diverse backgrounds, experiences, and perspectives, as this diversity can contribute to the identification and mitigation of biases in AI models.	**Due Diligence:** Biases present in training data, whether related to ethnicity, gender, age, or socio-economic factors, can impact the model's performance differently for distinct demographics. The practitioner should ensure that the model has been tested in different settings and on different user groups for bias before deploying the system at large. By subjecting the model to diverse user groups, practitioners can uncover and address such biases, promoting accuracy and equity across all patient populations. **Ethical Guidelines and Governance:** Establish and adhere to ethical guidelines specific to healthcare AI development. Implement governance structures that prioritize fairness, and regularly review and update these guidelines to stay aligned with evolving	**Advocate for Inclusive Data:** End users can advocate for healthcare systems to use diverse and representative datasets during the development and training of AI models. Inclusive data helps mitigate biases and ensures that AI systems are applicable to a wide range of patient populations. **Seek information:** Users can inquire about the steps taken to mitigate biases in AI algorithms and understand whether the developers have implemented strategies to address potential biases related to demographics, socio-economic factors, or other variables. **Give feedback:** End-users should use various avenues for reporting feedback, including chatbots, online forms, or direct communication with support teams. This

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		ethical standards in the healthcare sector.	approach encourages active user participation, contributing to the continuous improvement and refinement of AI systems in healthcare. **User Education and Engagement:** Engage in educational programs that help healthcare professionals and patients understand how AI systems work, including their limitations and potential biases.
Principle 4: Reliability and Safety/Robustness			
Technical	**Unique safety assessment standards:** AI can be deployed in the healthcare sector for a multitude of purposes with some being more severe in nature than others. In case of an inaccurate result or a non-robust AI, the degree of resulting potential harm	**Monitoring and Updation:** To ensure the reliable use of AI in healthcare, deployers should prioritize continuous monitoring and improvement of AI models. Towards this, regular surveys and assessments of developments in	**Incident Reporting and Analysis:** Encourage incident reporting from healthcare professionals and patients. Analyze reported incidents to identify potential safety risks, learn from experiences, and implement

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	would vary with each purpose. Therefore, different AI systems would require different levels of safety standards that would be proportional to the risk of potential harm that might occur in case of an unsafe system. For example, an AI system designed to detect arm fractures might require less stricter security standards than a system designed to detect cancer in superlative degrees. Therefore, developers should undertake requisite action to implement a set of safety assessment standards that would be unique to each system. Technical tools such as risk matrices, failure mode and effects analysis (FMEA), or probabilistic risk assessment (PRA) can be utilized to systematically analyze and quantify potential risks[103].	medical research can play a crucial role in staying up-to-date with the latest advancements and breakthroughs in the field. Deployers should identify emerging trends, novel treatment methods, and changes in best practices and ensure timely updation of AI models in line with these developments. Timely updates to AI models based on new research findings would help enhance the accuracy and effectiveness of the AI system in diagnosing, treating, and managing various medical conditions. **Interoperability Standards:** Adhere to interoperability standards to ensure seamless integration with existing healthcare systems. This includes using standardized data formats and communication protocols, contributing	improvements to enhance the overall reliability and safety of AI systems in healthcare. **Real-world Simulation Scenarios:** Participate in user testing scenarios that simulate real-world healthcare situations. This provides a practical understanding of how AI systems perform in dynamic clinical settings, allowing for the refinement of models to address challenges encountered in actual healthcare practice.

[103] Qin, J., Yan, X., & Pedrycz, W. (2020). Failure mode and effects analysis (FMEA) for risk assessment based on interval type-2 fuzzy evidential reasoning method. Applied Soft Computing, 89, 106134. https://doi.org/10.1016/j.asoc.2020.106134

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	Regulatory sandbox: The implementation of regulatory sandboxes, initially prominent in fintech, is gaining traction in the healthcare sector, exemplified by the United Kingdom's Care Quality Commission and the success of Ayushman Bharat Digital Mission (ABDM) Sandbox[104]. ABDM's Sandbox provides a valuable case study for understanding how this approach fosters collaboration, innovation, and risk mitigation in healthcare. Regulatory sandboxes offer developers and users the opportunity to test AI systems in a live setting, assessing their robustness and identifying potential concerns. By extending the sandbox concept to healthcare, organizations like the UK's Care Quality Commission aim to drive improvements in health and social care services[105]. The ABDM	to the overall safety and reliability of the AI system within the healthcare infrastructure. **Emergency Shutdown Protocols:** Implement emergency shutdown protocols, akin to "kill switches," for AI systems used in critical healthcare scenarios. These protocols serve as a safety net, allowing the immediate shutdown of an AI-based system in high-risk medical circumstances. Ensure that healthcare professionals and administrators are trained to recognize high-risk situations and can initiate immediate shutdown procedures when necessary.	

[104] Ayushman Bharat Digital Mission Sandbox. National Health Authority. https://sandbox.abdm.gov.in/abdm-docs/AboutABDMSandbox
[105] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., . . . Herrera, F. (2023b). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	Sandbox's success underscores the potential for learning and adapting sandbox methodologies to ensure the trustworthy deployment of AI in healthcare, ultimately enhancing healthcare experiences and outcomes. **Result testing methods:** Developers can use several methods to check if results are reliable. Some methods like GRADE (Grading of Recommendations, Assessment, Development and Evaluation) allow to rate the quality of evidence and recommendations.[106] Another framework, Software as a Medical Device (SaMD), helps decide how much proof is needed based on the impact and situation.[107]		

[106] Trustworthy Augmented Intelligence in Health Care - PMC. (2022, January 12). NCBI. Retrieved August 29, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8755670/#CR40, See About GRADE. (n.d.). Retrieved from https://cebgrade.mcmaster.ca/aboutgrade.html

[107] Trustworthy Augmented Intelligence in Health Care - PMC. (2022, January 12). NCBI. Retrieved August 29, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8755670/#CR40, See Software as a Medical Device (SAMD). (2013, December 18). Retrieved from https://www.imdrf.org/working-groups/software-medical-device-samd

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Non-technical	**Ethical Guidelines and Governance:** Adhere to and promote ethical guidelines for the development of AI systems in healthcare. Establish robust governance structures that prioritize patient safety, privacy, and ethical considerations throughout the development lifecycle.	**User Training Programs:** Implement comprehensive training programs for healthcare professionals using AI systems. Ensure that users are well-equipped to understand and interact with AI tools, emphasizing the importance of following safety protocols and recognizing potential issues. **Incident Reporting Mechanisms:** Institute incident reporting mechanisms that allow healthcare professionals to report any safety concerns or unexpected behaviors promptly. Use reported incidents as opportunities for learning and improvement in the deployment process.	**Active Participation in Training:** Actively participate in training programs offered by deployers. Ensure a deep understanding of AI systems' functionalities, limitations, and safety measures, empowering users to make informed decisions during system interactions.
Principle 5: Human Autonomy and Oversight			
Technical	**Human in the loop:** Healthcare AI should implement a 'human in the loop' approach wherein an AI system by design leaves room for healthcare professionals to	**Establish Clear Dispute Resolution Procedures:** Healthcare institutions deploying AI systems should establish transparent procedures for addressing patient challenges and seeking	**Oversight:** Rather than blindly relying on the results generated by the AI system, medical practitioners should actively evaluate the results from a non-biased perspective.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	supervise not only the final result but also assess different stages of the AI process and suggest and tailor results at each stage to cure any inaccuracies.[108] This would ensure that the AI does not override the decision of the medical experts.	redress when AI model outcomes lead to disputes or dissatisfaction. Clearly communicate these procedures to patients, ensuring they are aware of the mechanisms available for intervention in case of discrepancies in healthcare AI applications.	However, to ensure informed human oversight and autonomy over AI decisions, medical practitioners should work in tandem with the patients.
Non-technical	**Educational training:** Educational training for AI developers in medical knowledge is integral to realizing human autonomy and oversight in healthcare AI development. A basic understanding of medical concepts empowers developers to detect and rectify clinical inaccuracies early in the development phase, ensuring meaningful oversight of AI systems. Additionally, developers equipped with medical knowledge can strike a nuanced balance between automation	**Empower Human Decision-Making:** Healthcare AI deployers should establish internal policies that incorporate a human-in-the-loop approach. This approach grants autonomy to healthcare professionals involved in decision-making processes. In instances where AI-generated suggestions may be questionable, empowering healthcare professionals with the autonomy to deviate from automated decisions, within a framework of checks and balances, ensures	

[108] Bajwa, J., Munir, U., Nori, A. V., & Williams, B. (2021). Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthcare Journal, 8(2), e188–e194. https://doi.org/10.7861/fhj.2021-0095

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	and human decision-making, designing AI systems that complement rather than replace healthcare professionals. Towards this several initiatives are being undertaken. For instance, a partnership between Gustave Roussy, a prominent cancer hospital in Europe, and two engineering schools in Paris, École des Ponts ParisTech and CentraleSupelec, is educating young computer scientists in medicine.[109] To involve physicians in health innovation, the American Medical Association established the Physician Innovation Network, connecting healthcare solution developers and physicians to integrate their input into AI system design and uphold ethical values in medicine for improved outcomes.[110]	that patient care remains at the forefront.	

[109] OECD. (2020, July 24). TRUSTWORTHY AI IN HEALTH. Retrieved August 29, 2023, from https://www.oecd.org/health/trustworthy-artificial-intelligence-in-health.pdf
[110] American Medical Association. (n.d.). Physician Innovation Network. Retrieved August 21, 2020, from https://innovationmatch.ama-assn.org

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Principle 6: Data Privacy and Security			
Technical	**Secure systems:** Developing Healthcare AI entails handling vast datasets encompassing critical information such as medical history, genetic profiles, and biological characteristics. Given the sensitive nature of this data and the severe consequences of a potential breach, it is imperative to prioritize the construction of secure systems. When transmitting healthcare data between systems, employing secure communication protocols like HTTPS is essential.[111] This guarantees that data is encrypted during transit, mitigating the risk of unauthorized access or interception. Furthermore, healthcare data should be stored in secure databases or cloud storage solutions that align with industry standards for security. This involves implementing robust	**Anonymised data sets:** Prior to the stage of AI development, user groups, including medical institutions or government bodies that provide developers with access to datasets in the form of EHRs should ensure that the sensitive personal data is anonymised before providing access to developers. This can be done through techniques like masking where specific characters in data are replaced with non-sensitive characters, and tokenization, where sensitive data is replaced with unique tokens. This ensures that the data remains useful for analysis without compromising privacy. **Access and verification:** Post AI deployment, healthcare institutions must ensure that only authorised	

[111] Hypertext transfer protocol secure (HTTPS). (n.d.). Default. https://www.csa.gov.sg/Tips-Resource/internet-hygiene-portal/information-resources/https

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	access controls, encryption measures, and conducting regular security assessments to ensure the ongoing integrity and confidentiality of the data. **Differential Privacy:** Differential privacy adds mathematical noise to data, making it challenging to determine whether any particular individual's information is included in a dataset. By incorporating differential privacy mechanisms into healthcare AI systems, developers can strike a balance between extracting valuable insights from patient data and preserving individual privacy. Techniques like adding noise to query responses or employing privacy-preserving data aggregation methods can be applied. This would be especially useful in preventing insurance companies	medical personnel have access to the technology. Requisite security verification measures should be put in place. For instance, multi-factor authentication (MFA) should be implemented to enhance user authentication.[112] This adds an extra layer of security by requiring users to provide multiple forms of identification before accessing sensitive healthcare data. Further, organizations can implement role-based access control (RBAC) to restrict access to EHRs based on job responsibilities.[113] This ensures that only authorized personnel with a legitimate need can access specific sets of data. Audit logs can also be used to track who accesses the data and for what purpose.	

[112] Suleski, T., Ahmed, M., Yang, W., & Wang, E. (2023). A review of multi-factor authentication in the Internet of Healthcare Things. Digital health, 9, 20552076231177144. https://doi.org/10.1177/20552076231177144

[113] Tiwari, B., & Kumar, A. (2015). Role-based access control through on-demand classification of electronic health record. International journal of electronic healthcare, 8(1), 9–24. https://doi.org/10.1504/ijeh.2015.071637

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	from de-identifying patient data and profiling patients to discriminately charge premiums. **Adequate safeguards:** The developer should put safeguards to prevent re-identification from datasets and data leakages. For instance, developers can use techniques to anonymise data. Anonymizing and de-identifying data involve removing or encrypting personally identifiable information (PII) from datasets. PII includes information such as names, addresses, and social security numbers. By using anonymized data, developers can reduce the risk of exposing sensitive information and still derive valuable insights from the data. **Certification and accreditation:** Obtaining technical certifications and accreditation is crucial	**Enabling Personal Health Record System (PHR):** A PHR empowers individuals by giving them control over their health data, distinguishing it from broader institutional EHRs. In this context, the integration of the Ayushman Bharat Health Account (ABHA) system within the Ayushman Bharat Digital Mission (ABDM) becomes pivotal. The ABHA system under ABDM serves as the backbone for a comprehensive, patient-centric health data management approach. Health facilities acting as AI deployers should prioritize the establishment of systems that seamlessly support ABHA. It facilitates the seamless aggregation and retrieval of personal health data, streamlining the decision-making process for both healthcare providers	

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	for ensuring the reliability and compliance of the AI system. Certification bodies often evaluate the performance of AI models against predefined benchmarks, ensuring that they meet or exceed the necessary criteria for medical applications. In the USA, the Food and Drug Administration (FDA) is currently overseeing and authorizing AI/ML enabled medical devices.[114]	and patients. To successfully deploy ABHA, AI deployers must invest in robust and secure information systems. Additionally, user-friendly interfaces should be developed to encourage active participation from individuals in managing their health records.	
Non-technical	**Legal compliance:** Incorporating privacy should be seen as a value proposition and not merely a legal obligation but a fundamental commitment to user trust and ethical practices. The developer should ensure that the technology complies with data protection laws of a country. Since	**Employee Training:** Deployers like hospitals should provide comprehensive training to staff on security best practices, including the importance of safeguarding patient information and recognizing potential security threats. **Incident Response Plan:** Deployers should develop and regularly	**Informed Use of PHRs:** Effectively using Personal Health Records (PHRs) can empower individuals to take control of their health information, improve communication with healthcare providers, and make informed decisions about their well-being. Users should understand the privacy and security settings of

[114] Center for Devices and Radiological Health. (2023, December 6). Artificial Intelligence and Machine Learning (AI/ML)-Enabled medical devices. U.S. Food And Drug Administration. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	healthcare AI deals with sensitive personal data, the developers would normally be subjected to different and more onerous obligations under a law. Additionally, AI developers should educate themselves on various rules, regulations and guidelines released by regulatory and non-regulatory medical bodies.[115] In India, among other bodies, these may include the Central Drugs Standard Control Organisation and Indian Council of Medical Research.	update an incident response plan to effectively respond to and mitigate security incidents. This should include procedures for identifying, reporting, and responding to security breaches.	the PHR platform and set appropriate levels of access to ensure that health information is shared only with authorized individuals. Users should further take the time to review the privacy policies of the PHR platform and ensure that they are comfortable with how data is handled and stored.
Principle 7: Social and Environmental Sustainability			
Technical	**Impact Assessment:** To ensure healthcare AI is socially and environmentally sustainable, AI developers must conduct impact assessments to identify potential social and environmental implications of AI implementation in	**Cloud Computing Efficiency:** If the AI system is deployed on cloud infrastructure, choose cloud providers that prioritize energy efficiency in their data centers.	

[115] Indian Council of Medical Research. (2023). Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare. Retrieved August 29, 2023, from https://main.icmr.nic.in/sites/default/files/upload_documents/Ethical_Guidelines_AI_Healthcare_2023.pdf

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	healthcare. **Consultation and certifications:** Engaging stakeholders and exploring sustainability certifications can further contribute to creating AI solutions that align with social and environmental values. Relevant sustainability certifications such as ISO 14001 (Environmental Management) and ISO 26000 (Social Responsibility) should be explored and pursued. **Monitoring and auditing:** Medical AI systems should be regularly audited to identify and address any social or environmental issues.[116] **Sustainable Procurement:** When procuring hardware, software, or services for AI development, consider sustainability criteria, such as the environmental footprint of suppliers.		

[116] Indian Council of Medical Research. (2023). Ethical Guidelines for Application of Artificial Intelligence in Biomedical Research and Healthcare. Retrieved August 29, 2023, from https://main.icmr.nic.in/sites/default/files/upload_documents/Ethical_Guidelines_AI_Healthcare_2023.pdf

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Non-technical	**Collaborative Partnerships:** Collaborate with organizations and initiatives that prioritize environmental and social sustainability in healthcare, sharing best practices and resources.	**Educational training:** Continuous education and training should be provided to healthcare professionals and AI developers to inform them of the potential environmental risks and their impact. **Regulatory Compliance:** Stay informed about environmental regulations and standards related to AI development. Ensure compliance with any guidelines or requirements set by regulatory bodies. **Corporate Social Responsibility (CSR) in Healthcare:** Integrate AI-driven Environmental, Social, and Governance (ESG) strategies with corporate social responsibility initiatives within the healthcare sector. Ensure that AI applications align with the core values and sustainability goals of healthcare organizations.	**Advocacy for Sustainable Healthcare:** Share knowledge about the sustainability aspects of healthcare AI and advocate for responsible AI practices within communities and social networks.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Principle 8: Governance and Oversight			
Technical	**Internal Governance Framework:** Developers should establish frameworks that adhere to healthcare privacy regulations (e.g., DPDPA 2023) and implement measures for secure data storage, access control, and data sharing protocols.	**Long-Term Monitoring:** Implement long-term monitoring of patient outcomes and AI system performance to identify any adverse effects or unintended consequences. This can help in making continuous improvements and adjustments. **Reporting Cybersecurity Concerns:** Establish clear reporting mechanisms for cybersecurity concerns related to AI-driven healthcare systems. Healthcare IT teams and users should collaborate to report and address potential vulnerabilities, ensuring that cybersecurity protocols are continually strengthened to protect against evolving threats.	**Reporting Healthcare Anomalies:** Healthcare professionals should report any unexpected patterns or irregularities in the AI-driven decision-making process, contributing to the continuous improvement of healthcare AI models. This could include anomalies in diagnostic results or unexpected outcomes from AI-assisted medical procedures. **Health Data Protection Vigilance:** Stay vigilant about AI-driven cybersecurity practices, especially when handling EHR and sensitive patient data. Recognize the critical importance of safeguarding health information against potential cyber threats to maintain the integrity and confidentiality of patient records.
Non-technical	**Awareness:** AI developers engaged in the creation of medical applications must stay	**Regulatory Liaison:** Maintain a robust relationship with regulatory authorities to	**Advocacy for Regulatory intervention:** Users should advocate for

Level/ Stakeholder	AI Developer	AI Deployer	AI User
	informed about technical standards and certification requirements. Adherence to these standards is paramount for several reasons. Firstly, it ensures patient safety and regulatory compliance, particularly in regions like the United States where the FDA sets guidelines for medical device development, including AI systems. Secondly, privacy and security considerations, encapsulated in standards like HIPAA, are equally crucial to safeguarding sensitive patient information. Staying current on international standards for medical device software development such as ISO 13485 and IEC 62304 is further imperative for global compliance.	remain up-to-date on evolving medical regulations and ensure alignment with compliance standards. **Protocols for Adverse Events:** Develop protocols for handling adverse events or errors caused by AI systems, including reporting, investigation, and resolution procedures. Ensure that healthcare professionals know how to respond in such cases. **Independent Oversight:** Consider establishing an independent body or review board that oversees the development and deployment of AI systems in healthcare. This body can provide an additional layer of accountability and ensure ethical and responsible practices. This imperative action should follow a transparent formation process, ensuring a diverse composition of experts spanning	clear regulations and standards for AI deployment in healthcare, with robust enforcement mechanisms to hold AI developers and healthcare institutions accountable for adhering to ethical guidelines and policies. In case of an AI induced harm, a mechanism should identify the roles of stakeholders, including manufacturers and users, and establish legal liability. This push from end-users will help the government form policies that minimize harm.

Level/ Stakeholder	AI Developer	AI Deployer	AI User
		medicine, ethics, law, technology, and patient advocacy. The board's members, selected without conflicts of interest, should serve with term limits and undergo periodic rotations to maintain a dynamic and unbiased perspective. The primary objective of this oversight board is to conduct rigorous ethical reviews, assess compliance with regulations, and uphold transparency by publicly reporting assessments. Granting the board the authority to enforce compliance and recommend corrective actions can be crucial for ensuring accountability to ethical standards.	

Level/ Stakeholder	AI Developer	AI Deployer	AI User
Principle 9: Contestability			
Technical		**Standardized Redress Mechanisms:** Healthcare institutions, such as hospitals, should create accessible channels through which patients can initiate inquiries, appeals, or reviews of AI-driven decisions that directly impact them. This mechanism empowers patients to contest points in decision-making where they have concerns or disagreements.	
Non-technical		**Whistleblower Protection:** Institutions can implement whistleblower protection mechanisms for individuals who raise concerns about the ethical or responsible use of AI in healthcare, ensuring they can do so without fear of retaliation.	**Patient Advocacy:** Patient advocacy groups should participate in discussions surrounding AI in healthcare and represent patient interests in AI system development and accountability discussions.

4 Implementation of Principle-based Governance of AI

In the preceding chapter, the focus was on the practical implementation of ethical principles in AI, outlining the responsibilities of AI developers, deployers, and end-users. However, for these operationalizations to genuinely take root, the active involvement of the government becomes imperative. This chapter delves into the intricacies of implementing a principle-based framework for trustworthy AI, identifying key drivers for responsible adoption. At the domestic level, we navigate the complexities of aligning with existing laws and advocate for regulations adaptable to the evolving AI landscape. As we shift to the international stage, the emphasis on cross-border cooperation becomes paramount. Harmonizing global AI regulations is vital, providing a unified and ethical foundation transcending national borders. We explore the dynamics of public-private partnerships as powerful catalysts for change, leveraging market mechanisms to incentivize developers towards consumer protection and safety. The chapter underscores the interplay of these multifaceted levers, emphasizing their collective importance in forging a path toward the trustworthy and ethical adoption of AI, across nations, sectors, and the public-private landscape.

4.1 Domestic Coordination

The integration of AI into various sectors in India represents a transformative shift that promises innovation and efficiency. However, this technological evolution necessitates a delicate balance between promoting AI-driven advancements and ensuring compliance with the regulatory norms that already govern these sectors. In this critical context, the harmonization of AI-specific regulations with existing sectoral regulations emerges as a pivotal imperative to ensure trustworthy AI integration.

4.1.1 Indian Regulatory Landscape for the Finance Sector

India's finance sector, comprising banks, insurance companies, and stock exchanges, is governed by a robust regulatory framework. Key authorities like the Reserve Bank of India (RBI), the Securities and Exchange Board of India (SEBI), and the Insurance Regulatory and Development Authority of India (IRDAI) play crucial roles. The RBI oversees banking and monetary policies, ensuring system stability. SEBI regulates the securities market, focusing on investor protection and transparent operations. IRDAI governs the insurance sector, safeguarding policyholders' interests and ensuring financial soundness. These bodies collectively protect consumers, enforce regulations, and provide dispute resolution mechanisms to address grievances. This comprehensive and intertwined regulatory ecosystem in India's finance sector serves the dual purpose of fostering financial innovation while safeguarding the interests of consumers and the stability of the financial system.

As the finance sector undergoes a transformative wave with the integration of AI,

its potential to revolutionize financial services brings forth the need for cohesive regulation. Ensuring regulatory consistency is paramount to safeguard consumer interests, maintain financial market stability, and protect the integrity of sensitive data. Achieving this balance requires the development and implementation of AI-specific guidelines in India that are intricately intertwined with existing financial regulations.

The cornerstone of this harmonization effort lies in the emphasis on adherence to established financial regulations. The AI-specific guidelines should be meticulously crafted to ensure that AI applications in the finance sector comply with sector-specific rules. This entails setting thresholds for AI-driven decisions, meaning that AI systems should operate within predefined boundaries to avoid extreme or unsanctioned outcomes. For instance, algorithms used in trading should adhere to strict limits to prevent market manipulation. This transparency is essential to ensure that AI-based financial decisions are not driven by opaque, unexplainable processes.

Further, mandating compliance auditing is another pivotal aspect of ensuring regulatory consistency. This process involves rigorous assessments to verify that AI applications adhere to the established guidelines and financial regulations. Compliance auditing not only acts as a safeguard but also incentivizes financial institutions to proactively align their AI systems with regulatory norms. It plays a significant role in maintaining transparency, accountability, and the overall ethical conduct of AI applications in finance.

Achieving this harmonization isn't solely the responsibility of individual organizations or regulatory bodies. Domestic coordination is of paramount importance. This involves collaborative efforts between regulatory authorities, financial institutions, and data protection authorities to align AI applications with existing financial regulations. These stakeholders must work cohesively, sharing expertise, insights, and resources to prevent AI applications from undermining the integrity of financial markets and the security of financial data. This domestic coordination is not only about adhering to rules but also about fostering innovation responsibly within the financial sector.

4.1.2 Indian Regulatory Landscape for the Health Sector

India's healthcare sector operates within a multifaceted regulatory framework designed to uphold the highest standards of patient care, data privacy, and healthcare quality. The regulatory apparatus is overseen by three paramount authorities: the Indian Council of Medical Research (ICMR), National Health Authority (NHA), and the Central Drugs Standard Control Organisation (CDSCO) headed by the Drugs Controller General of India (DCGI), play instrumental roles in shaping and preserving these standards. ICMR, a preeminent research institution, assumes a pivotal role in setting and monitoring healthcare norms and ethical standards in clinical research, patient care, and medical education.

With the current wave of AI integration, this sector stands on the precipice of a revolutionary transformation. AI is permeating the sector, presenting innovative solutions in

areas like diagnostic tools, telemedicine, and drug discovery. This paradigm shift, while promising tremendous benefits, necessitates regulatory alignment to ensure that AI applications in healthcare comply with sector-specific regulations. Specifically, harmonizing or ensuring regulatory interoperability becomes crucial with healthcare regulations.

The critical importance of regulatory interoperability, be it in legal frameworks or administrative processes, cannot be overstated in effectively implementing and upholding healthcare laws and policies. In the era of digital healthcare, where a cohesive global network of data is envisioned, the seamless integration of regulatory systems is essential to navigate the complexities of the healthcare landscape. This approach helps avoid a fragmented regulatory environment, fostering a more unified and comprehensive oversight.

Legal interoperability serves as the cornerstone of this regulatory cohesion, requiring the development of harmonious and complementary legal and policy instruments. The goal is to ensure these instruments not only align with each other but also avoid contradictions that could impede the effective implementation of healthcare regulations. Achieving such legal interoperability may involve a spectrum of actions, from enacting new laws to amending or reinterpreting existing ones, all geared towards establishing a framework that promotes synergy.

For instance, India's data protection landscape has undergone significant changes, marked by the recent enactment of the DPDP Act 2023. This legislative development is a pivotal move towards establishing privacy and data protection in the country. Concurrently, the draft Health Data Management Policy 2.0 (HDMP) empowers the National Health Authority (NHA) to formulate rules within the ABDM, granting authority over entities in the National Digital Health Ecosystem (NDHE). However, challenges can arise due to divergences between the draft HDMP and the DPDP Act, 2023 especially concerning the rights of data principals and the overlapping powers of the NHA and the DPB. This creates an unclear legal and regulatory landscape, causing uncertainty.

It's further crucial to recognize that legal interoperability extends beyond the immediate realm of health-related regulations. It encompasses a broader spectrum, encompassing other sectoral regulators like the Data Protection Board, etc. have the potential to impact the healthcare sector. A reference to the Interoperable Europe Board[117] highlights the importance of harmonizing policy frameworks and solutions across sectors. This holistic approach acknowledges the interconnectedness of various regulatory domains and emphasizes the need for a comprehensive and cohesive legal foundation to navigate the intricate landscape of healthcare governance effectively.

Therefore, domestic coordination between health authorities, medical institutions, and data protection agencies is essential for crafting AI regulations that enhance patient outcomes while also respecting the critical tenets of privacy and data security. The collaboration of these entities is instrumental

[117] Interoperable Europe Act Proposal. (2022, November 30). European Commission. https://commission.europa.eu/publications/interoperable-europe-act-proposal_en

in striking the right balance between harnessing AI's potential in healthcare and upholding the ethical and legal norms that are integral to the healthcare sector.

4.2 International Coordination

In today's rapidly evolving technological landscape, the relentless progress of AI presents a host of complex and interconnected challenges that transcend borders. The transformative potential of these technologies, while promising immense benefits, also raises critical questions concerning ethics, safety, and governance. As AI systems become more integrated into our daily lives and across industries, they impact everything from healthcare and finance to national security and social dynamics. This profound impact necessitates a global response. This urgency for international coordination stems from the recognition that many of these challenges posed by AI are not confined within the borders of a single nation. They have the potential to ripple across the globe, affecting countries, economies, and societies in profound ways. This pressing need for a collective approach to AI governance extends to establishing global standards, principles, and frameworks that ensure responsible development, deployment, and use of AI. It requires countries, organizations, and experts to work collaboratively, leveraging their combined knowledge and resources to create a cohesive response to these shared challenges.

The call for international coordination reflects a collective awareness that the solutions to AI's global challenges are not the domain of any single nation or entity but require a united, global effort to address effectively. This approach seeks to strike a balance between fostering innovation and harnessing AI's full potential while safeguarding against the potential risks and ethical concerns that AI technologies may introduce into the world. By collaborating on a global scale, nations can collectively shape a future in which AI is a force for positive change, driving advancements that benefit all of humanity. Recognizing this imperative, the G7 leaders have proactively taken steps to champion the establishment of regulatory frameworks for AI. Their efforts took center stage during their meeting in Japan in May 2023, where they collectively called for the formulation of rules governing AI, GenAI included. To drive this vision forward, they set in motion an intergovernmental forum, named the Hiroshima AI Process[116].

This forum assumes a central role in shaping the landscape of AI governance on a global scale. Its primary mission is to engage in extensive deliberations concerning the development and adoption of technical standards. These standards are designed to address the multifaceted challenges presented by GenAI tools and to establish a robust framework that ensures the trustworthiness of AI technology. The forum's purview extends across a wide spectrum of crucial domains, including countering disinformation, safeguarding intellectual property, and instituting regulations governing AI applications.

Further, countries around the world have also recognized the need to establish regulations to govern various facets of AI. These regulations are at different stages of development, with some still in the preliminary draft stage, while others have already been enacted into law. The recently concluded AI Safety summit 2023 also reiterated the same, where several countries came up with a Declaration[117] that highlighted the need to support an internationally inclusive network of scientific research on frontier AI safety that encompasses and complements existing and new multilateral, plurilateral and bilateral collaboration, including through existing international fora and other relevant initiatives, to facilitate the provision of the best science available for policy making and the public good.

Given the varied and fragmented nature of these regulatory frameworks, it has become increasingly important to establish guiding principles for international coordination. This is essential to ensure a harmonized and consistent approach to AI regulation across borders. AI technologies transcend national borders and are often developed, deployed, and used across multiple countries simultaneously. Without universal consensus, conflicting regulations can hinder the smooth operation of AI systems and international collaboration. Further, AI has the potential to raise ethical dilemmas, from bias in algorithms to autonomous weapons. A global consensus on ethical principles can help prevent misuse and ensure that AI is developed and used in ways that align with shared values and norms. In the absence of a coordinated approach, conflicting regulations can disrupt the seamless operation of AI systems that operate across nations, impeding international collaboration. This, in turn, may hinder the responsible and effective adoption of AI technologies that have the potential to drive positive changes on a global scale.

Establishing guiding principles for international coordination is not merely a matter of convenience; it is an essential step towards a future where AI technologies can be developed and harnessed in a manner that aligns with shared values and ethical norms. A global consensus on ethical principles safeguards against the misuse of AI, ensuring that it is applied in ways that resonate with the broader global community. By aligning on these principles, nations can collectively navigate the intricate and rapidly evolving landscape of AI, fostering responsible and ethically sound innovation that benefits societies worldwide.

Against this backdrop, our paper seeks to contribute to the global discourse on AI regulation by taking a pivotal step towards harmonizing the diverse and fragmented regulatory landscapes. In doing so, it endeavours to map out a set of principles that hold the potential for universal applicability across various sectors. The fundamental premise underlying this effort is the recognition that AI technologies, with their transformative potential and wide-reaching impact, transcend sectoral boundaries and geographical borders.

The mapping of these principles represents a significant endeavor to identify a common ground upon which nations, organizations, and industries can converge. By extracting the core principles that underpin trustworthy

AI regulation, our paper aims to facilitate international dialogue and collaboration. This approach recognizes that while the nuances of AI applications may differ from one sector to another, there exists a shared foundation of ethical and responsible conduct that can, and should, be universally upheld.

Furthermore, this mapping of principles acknowledges that as AI continues to evolve, its applications will traverse various sectors, from finance and healthcare to education and transportation. Therefore, the establishment of these overarching principles is vital to create a flexible framework that can be adapted and extended to suit the specific needs of different industries.

4.3 Public-Private Coordination

This section underscores the importance of exploring alternative approaches for regulating AI that leverage market mechanisms and encourage public-private collaboration. The core objective here is to craft mechanisms and incentives that stimulate AI developers to place consumer protection and safety at the core of their value proposition. By achieving this, we aim to cultivate an ecosystem where trustworthiness becomes an intrinsic and non-negotiable element in the development and deployment of AI systems. This shift is imperative for ensuring that AI technologies align with the highest ethical, security, and reliability standards, ultimately serving the best interests of consumers and society as a whole. This approach not only propels the transformation and enhancement of various sectors but also safeguards fundamental ethical and security standards.

Realizing this objective necessitates the effective coordination of public and private entities, driven by a well-defined set of strategies. Stakeholder engagement stands as a pivotal strategy, fostering collaborative partnerships between public and private sectors through dialogues, industry forums, and dedicated working groups. Furthermore, incentivization assumes a critical role, wherein structured frameworks are designed to reward AI developers for upholding ethical and safety standards. These incentives could include tax benefits, access to funding opportunities, or preferential treatment in government contracts. Additionally, investments in research and development are deemed vital to bolster the safety and trustworthiness of AI systems. Encouraging public-private research partnerships can significantly accelerate progress in this domain, ensuring that AI remains a force for positive change while upholding fundamental standards of ethics and security.

5 Conclusion

In the dynamic landscape of global AI advancements, the imperative to ensure ethical and responsible adoption of AI technologies cannot be overstated. Trustworthy AI guidelines stand as a linchpin in this endeavor, offering a structured framework that encompasses ethical considerations, transparency, and accountability. Acting as a guiding beacon, these guidelines are indispensable tools for navigating the intricate terrain of artificial intelligence. When applied to pivotal sectors like finance and health, where the impact on societal well-being is profound, responsible adoption becomes not only a strategic imperative but a moral obligation.

By embracing and prioritizing these guidelines, industries can fortify themselves against potential risks inherent in the development and deployment of advanced AI technologies. The application of ethical practices ensures that AI becomes a force for good, contributing positively to the enhancement of human life rather than posing unforeseen challenges. Trustworthy AI is not a luxury but a necessity, a compelling call to action in a world where technological progress requires a deliberate and ethical approach. The urgency lies in our collective responsibility to foster harmonious integration of AI into these critical sectors, laying the groundwork for a future where innovation aligns seamlessly with ethical considerations.

Authors

RAMA VEDASHREE

Former CEO, Data Security Council of India (DSCI)

Rama Vedashree is a former CEO of Data Security Council of India (DSCI). With over 35 + years in the Tech Industry, she brings varied expertise with long stints at NIIT Technologies, Microsoft and NASSCOM. She has published an Edited Volume titled “Digital++, Reimagining Security & Privacy”.

JAMEELA SAHIBA

Senior Programme Manager, Emerging Technologies (AI)

Jameela is a law and policy professional with 5+ years of experience with a strong understanding of the Parliamentary/policy framing ecosystem. Prior to this, she worked as the Chief of Staff for the Office of Dr Amar Patnaik, Member of Parliament, where she led parliamentary and legislative research work. She has experience of working with a diverse cohort of MPs while displaying excellent research, analytical and organizational skills. At the Dialogue, she is leading the vertical on emerging technology and also managing government/parliamentary outreach for the organization.

BHOOMIKA AGARWAL

Senior Research Associate

Bhoomika completed her B.A.LLB (H) from Guru Gobind Singh Indraprastha University and has several publications and paper presentations to her name. She has done internships with several reputed organisations where she worked on a project to make internet space safer for women and other communities. Her focus areas include technology policy and competition law.

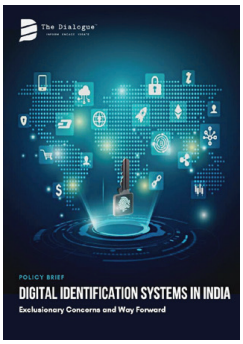
KAMESH SHEKAR

Senior Programme Manager, Data Governance and Emerging Technologies (AI)

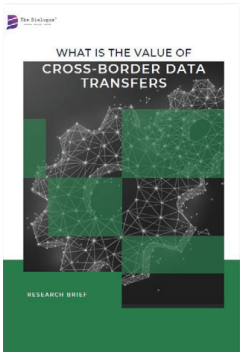
His area of research covers informational privacy, surveillance technology, intermediary liability, safe harbour, issue of mis/disinformation on social media, AI governance etc. Prior to this, Kamesh has worked as a communication associate at Dvara Research. Kamesh holds a PGP in Public Policy from Takshashila Institution and holds an MA in media and cultural studies and a BA in social sciences from the Tata Institute of Social Sciences.

More From Our Research

Research Paper
Privacy Technologies in India – Strategies to Enhance the Ecosystem

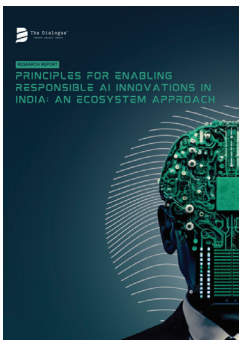


Policy Brief
Digital Identification Systems in India – Exclusionary Concerns and Way Forward



Policy Brief
What is the Value of Cross-border Data Transfers

Research Paper
Digital Identification Systems in India – Exclusionary Concerns and Way Forward



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

Research Report
The Institutionalisation of India's Data Protection Authority

LinkedIn | The Dialogue

Twitter | The Dialogue

Facebook | The Dialogue

Instagram | The Dialogue