CENTRE FOR COMMUNICATION GOVERNANCE AT NATIONAL LAW UNIVERSITY DELHI

COMMENTS TO THE DEPARTMENT OF TELECOM ON THE DISCUSSION PAPER ON THE FRAMEWORK OF AN INDIAN ARTIFICIAL INTELLIGENCE STACK[°]

nludelhi.ac.in | ccgdelhi.org | ccg@nludelhi.ac.in

[°]Authored by *Jhalak Kakkar* and *Nidhi Singh*. Reviewed and edited by *Smitha Krishna Prasad* and *Sarvjeet Singh*.

TABLE OF CONTENTS

EXECUTIVE SUMMARY 4

1. INTRODUCTION 11

2. ANCHORING AI REGULATORY PRINCIPLES WITHIN INDIA'S CONSTITUTIONAL FRAMEWORK 13

2.1. Discrimination 13

2.2. Privacy 15

3. GOVERNANCE AND REGULATORY CHALLENGES IN THE ADOPTION OF AI IN INDIA 17

3.1. Scope of AI Regulation 17

3.2. Heightened Threshold of Responsibility for Government or Public Sector Deployment of AI 17

3.3. Need for Overarching Principles based AI Regulatory Framework 19

3.4. Adaptation of Sectoral Regulation to Effectively Regulate AI 21

3.5. Contextualising AI systems for both their Safe Development and Use 23

4. PRINCIPLES FOR THE REGULATION OF AI 24

4.1. Principle of Safety and Reliability 26

4.1.1. Human Oversight 27

4.1.2. Assessment system for the safety and reliability of AI systems 28

4.2. Principle of Equality, Inclusivity and Non-Discrimination 28

4.2.1. Principle of Equality 29

4.2.2. Principle of Non-Discrimination 30

4.2.3. Principle of Inclusivity 34

4.2.4. Checklist Model to Ensure Equality, Inclusivity and Non-Discrimination in AI Systems 39

4.3. Principle of Privacy 40

4.3.1. Privacy and the Use of Data 43

4.3.2. Privacy in AI and the Draft Personal Data Protection Bill 2019 49
4.3.3. Use and Regulation of Non-Personal Data 51
4.4. Principle of Transparency 52
4.4.1. Transparency Challenges in AI 56
4.4.2. Adoption of Model Cards 57
4.5. Principle of Accountability 58
4.5.1. Pre-Deployment 60
4.5.2. During Deployment 62
4.5.3. Post Deployment Harms 63
APPENDIX: AI PRINCIPLES 65
COUNTRIES/ MULTILATERAL BODIES 65
Asia 65
Europe 66
North America 67
South America 68
Africa 68
Australia 68
Global 69
CIVIL SOCIETY, INTERNATIONAL AND ACADEMIC ORGANISATIONS 69
CORPORATIONS 70

EXECUTIVE SUMMARY

The AI Standardisation Committee of the Department of Telecommunications (DoT) has released a Discussion Paper on 'Indian Artificial Intelligence Stack' ('AI Stack Document') for comments. The AI Stack Document discusses the issues around developing a framework of an Indian AI stack and proposes an AI stack with an aim to remove the challenges to AI deployment and enable its productive adoption. The AI Stack Document proposes to divide the AI Stack in six different layers with horizontal and vertical integration. The vertical layer cutting across all the layers is the Security and Governance layer that ensures that AI systems are safe, secure and trusted. The framework of this Security and Governance layer of the stack directly relates to broader questions of the design of a regulatory and governance framework for AI in India.

We at the Centre for Communication Governance at National Law University Delhi (CCG) welcome the release of this AI Stack Document and commend the DoT for adopting an open and consultative approach and inviting comments from interested stakeholders. Recently, the NITI Aayog released the Working Document: Towards Responsible AI for All which attempts to delineate 'Principles for Responsible AI' and identify relevant policy and governance recommendations to regulate AI.[1] In our response to the NITI Aayog's Working Document we had highlighted the various challenges which must be considered in the framing of an AI governance and regulatory framework in India.[2] The same challenges and principles are relevant from the point of view of designing the Governance and Security Layer of the proposed AI Stack. Hence, in this document we draw from our previous response to the NITI Aayog and add in analysis that is particularly relevant from the perspective of the proposed AI Stack.

[1] NITI Aayog, 'Working Document: Towards Responsible AI for All' (2020) <https://niti.gov.in/sites/default/files/2020-07/Responsible-AI.pdf>.

[2] Centre for Communication Governance, 'Comments to the NITI Aayog on the Working Document: Towards Responsible AI for All' (September, 2020) <https://ccgdelhi.org/wp-content/uploads/2020/08/CCG-NLU-Comments-to-NITI-Aayog-on-the-Towards-Responsible-AI-for-All-Document.pdf>.

Our comments on the AI Stack Document revolve around three key challenges which must be considered in the framing of an AI governance and regulatory framework in India:

1. Anchoring AI Regulatory Principles within the Constitutional Framework of India

The adoption of AI technology in India would have to be adapted into the current constitutional framework. AI technology has vast implications on constitutionally protected rights such as the right against discrimination, the right to privacy and the right to freedom of speech and expression.[3] While the AI Stack Document discusses the challenges of bias in AI systems and the privacy implications arising from the use of AI systems, it does not comprehensively address the idea of framing AI governance principles in compliance with India's constitutional rights, such as the fundamental right to equality or privacy.

The deployment of various AI systems has raised concerns about their potential negative impact on constitutional values enshrined in the Indian Constitution. In particular, the adoption of AI governance frameworks to regulate the deployment and use of AI systems would have to strictly comply with the standards of anti-discrimination, privacy, the right to freedom of speech and expression, the right to assemble peaceably and the right to freedom of association as provided for in Part III of the Indian Constitution and interpreted by the Supreme Court. For instance, the large-scale deployment of AI systems such as

[3] Article 19 and Privacy International, 'Privacy and Freedom of Expression In the Age of Artificial Intelligence' (April 2018) <https://www.article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf>.

facial surveillance has raised several ethical and regulatory concerns across the world, in the context of both privacy[4] and equality.[5]

The use of vast datasets in the training of AI systems makes them particularly susceptible to bias and discrimination, and violation of constitutional rights. Removing bias and discrimination from AI datasets requires a mix of solutions including technical solutions for the system itself, as well as principle based regulation which guides the process of data collection, the design choices in the development of the AI system, and informs the values of the programmers and developers. The standard of equality and privacy sought to be achieved must be compatible with the constitutional thresholds for the protection of these rights. While, the AI Stack Document discusses the challenges of bias, privacy, security and inclusivity and refers to the need for a well-designed regulatory standard in the form of an open Indian stack in line with internationally agreed principles, it does not engage with the principles for responsible AI such as equality, inclusivity and non-discrimination, and privacy and security that have are being developed globally. There needs to be substantive discussion around the development of these AI principles in the Indian context and ensuring that they are in consonance with the constitutional guarantees for rights such as the right to equality and right to privacy.

2. REGULATORY CHALLENGES IN THE ADOPTION OF AI IN INDIA

As the DoT designs the 'Security and Governance' layer of the AI stack, they should consider the broader governance and regulatory regime that needs to be put in place to

[4] *See* Amnesty International, 'Amnesty International Calls for Ban on the Use of Facial Recognition Technology for Mass Surveillance' (11 June 2020) <https://www.amnesty.org/en/latest/research/2020/06/amnesty-international-calls-for-ban-on-the-use-of-facial-recognition-technology-for-mass-surveillance/#:~:text=In%20the%20context%20of%20racially,and%20the%20right%20to%20privacy>; Sangh Rakshita, 'The Proliferating Eyes of Argus: State Use of Facial Recognition Technology' (CCG Blog) <https://ccgnludelhi.wordpress.com/2020/09/23/the-proliferating-eyes-of-argus-state-use-of-facial-recognition-technology/>.

[5] *See* R (on the application of Bridges) v Chief Constable of South Wales Police [2020] EWCA Civ 1058 <https://www.judiciary.uk/wp-content/uploads/2020/08/R-Bridges-v-CC-South-Wales-ors-Judgment-1.pdf>.

govern the adoption and deployment of AI systems in India. The governance and regulatory challenges that need to be considered include:

i) Heightened Threshold of Responsibility for Government or Public Sector Deployment of AI Systems

Certain countries[6] are considering adopting a risk-based approach for regulation of AI, with heavier regulation for high-risk AI systems. The extent of risk concerning factors such as safety, consumer rights and fundamental rights is assessed by looking at the sector of deployment and the intended use of the AI system.

Drawing on this thinking, India must consider the adoption of a higher regulatory threshold for the use of AI by government institutions, especially where citizen's rights are directly impacted. Use of AI systems in processes of government decision making or functions can have a severe negative impact on the fundamental rights of a broad cross-section of Indian citizens. Government uses of AI systems that have the potential for such impact include the use of AI in the disbursal of government benefits, surveillance, law enforcement and judicial sentencing.

ii) Need for Overarching Principles Based AI Regulatory Framework

Different sectoral regulators are currently evolving regulations to address the specific challenges posed by AI in their sector.[7] While it is vital to encourage the development of sector-specific AI regulations, such piecemeal development of AI principles can lead to fragmentation in the overall approach to regulating AI in India. Therefore, it is crucial to put in place an overarching principles-based framework to regulate AI to ensure uniformity in the approach to regulating AI systems across sectors in India.

---

[6] European Commission, 'White Paper :On Artificial Intelligence - A European approach to excellence and trust' (2020) <https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf>.

[7] *See* Reserve Bank of India, 'Report of the Working Group on FinTech and Digital Banking' (November 2017) <https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/WGFR68AA1890D7334D8F8F72CC2399A27F4A.PDF>.

iii) Adaptation of Sectoral Regulation to Effectively Regulate AI

In addition to an overarching structure which forms the basis for the regulation of AI, it is equally important to envisage how these AI regulatory principles would work with sector-specific laws such as those in the financial sector or other horizontal laws such as consumer protection and product liability that may apply to various AI systems.

iv) Contextualising AI Systems for Both their Safe Development and Use

Finally, to ensure effective and safe use, AI systems have to be designed, adapted and trained on relevant datasets depending on the context in which they will be deployed. For instance, NITI Aayog's Working Document Towards Responsible #AIforAll[8] envisages India being the AI Garage for 40% the world - developing AI solutions in India which can then be deployed in other countries. Additionally, India will likely import AI systems developed in countries such as the US, EU and China to be deployed within the Indian context. Both scenarios involve the use of AI systems in a context distinct from one in which they have been developed. Regulatory standards and processes need to be developed in India to ascertain the safe use and deployment of AI systems that have been developed in contexts that are distinct from the ones in which they will be deployed.

3. PRINCIPLES FOR THE REGULATION OF AI

The AI Stack Document discusses the challenges of bias, privacy, security and inclusivity, and refers to the need for a well-designed regulatory standard in the form of an open Indian stack in line with internationally agreed principles. However, it does not specifically engage with the principles for responsible AI such as those of equality, inclusivity and non-discrimination, and privacy and security, that are being developed globally. There needs to be substantive discussion around the adoption of the principles for responsible AI such as the principle of safety and reliability, equality, inclusivity and non-discrimination, privacy and security, transparency, accountability, and the protection and reinforcement of positive human values and their integration into the Indian governance context.

[8] NITI Aayog, Towards responsible AI for All n(1).

As discussed above and in greater detail in our comments, any framework for the regulation of AI systems needs to be based on clear principles. In our comments, we elaborate on and discuss the constituent elements of each of these principles, which must be considered while incorporating these principles into the Indian context. Additionally, we provide flexible models which can be used to integrate these principles into India's AI governance framework effectively.

i) Principle of Safety and Reliability: We suggest employing mechanisms ensuring human oversight in the deployment of AI systems. The level of human involvement may vary depending upon a risk-based assessment and the circumstances relating to the deployment and impact of the AI system. Accordingly, a human in the loop, human on the loop, or any other oversight mechanism which is required may be adopted.

ii) Principle of Equality, Inclusiveness and Non-Discrimination: In order to comprehensively address issues surrounding bias, equality, inclusiveness and non-discrimination we have suggested a checklist model which has specific sections which test for direct bias, indirect bias, equity concerns, diversity, etc. The level of use of this checklist (whether internally, though external auditors or through a regulator) would be proportionate to the risk which the deployment of the AI system poses.

iii) Principle of Privacy and Security: We have examined the potential risks to the privacy of both individuals specifically, and the society at large which may be affected by the deployment of AI systems. We suggest the adoption of an incident investigation model which would enable organisations to share information on the potential failures in the deployment of AI systems.

iv) Principle of Transparency: For increasing transparency in AI systems we suggest the adoption of model cards, which are short documents which accompany a trained machine learning model, carrying the benchmarked evaluation of the system in a variety of conditions. These model cards are intended to clarify the scope of the AI systems deployment and minimise their usage in contexts for which they may not be well suited.

v) Principle of Accountability: We have identified that accountability must be maintained at three stages of an AI: pre-deployment, during deployment and post-

deployment of the system. We suggest the adoption of specific grievance redressal mechanisms, such as an AI ombudsperson, or a guided process for registering a complaint similar to the system for filing right to information requests.

We have based our comments and recommendations on an extensive review of AI principles developed and adopted by various countries around the world, international bodies, multinational companies and civil society organisations.[9] Additionally, we have drawn from academic[10] and policy[11] research undertaken around the world on the governance of AI. We draw on the literature that has developed globally around these AI principles and the governance of AI and attempt to discuss them from the perspective of embedding them into an Indian governance context. A compilation of the various sets of principles proposed by countries across the world, international bodies, multinational companies and civil society organisations that we have referred to in our comments is provided in the Appendix to this document.

[9] A comprehensive list of the principles can be found in the Appendix.

[10] Jessica Fjeld et al. 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI' Berkman Klein Center for Internet & Society, 2020. <http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420>.

[11] Organisation for Economic Co-operation and Development, 'AI policies and initiatives' (OECD iLibrary) <https://www.oecd-ilibrary.org/sites/eedfee77-en/1/2/5/index.html?itemId=/content/publication/eedfee77-en&_csp_=5c39a73676a331d76fa56f36ff0d4aca&itemIGO=oecd&itemContentType=book>.

1. INTRODUCTION

> *"Digital technologies, especially Artificial Intelligence (AI), are transforming the world at an unprecedented speed. They have changed how we communicate, live and work."*[12]

Artificial Intelligence ('AI') has been the sphere of intense debate. While it has several benefits, there are several potential harms and unintended risks if the technology is not assessed adequately for its alignment with India's constitutional principles and its impact on the safety of individuals. Depending upon the nature and scope of the deployment of an AI system, its potential risks can include the discriminatory impact on vulnerable and marginalised communities, and material harms such as negative impact on the health and safety of individuals. In the case of deployments by the State, risks include violation of the fundamental rights to equality, privacy and freedom of speech and expression.

We welcome the Discussion Paper on 'Indian Artificial Intelligence Stack' ('AI Stack Document')[13] released by the AI Standardisation Committee of the Department of Telecommunications ('DoT'), as a significant step towards an informed discussion on the adoption of a governance framework for AI technology in India. Recently, the NITI Aayog released the Working Document: Towards Responsible AI for All which attempts to delineate 'Principles for Responsible AI' and identify relevant policy and governance recommendations to regulate AI.[14] In our response to the NITI Aayog's Working Document we had highlighted the various challenges which must be considered in the framing of an AI governance and regulatory framework in India.[15] The same challenges and principles are relevant from the point of view of designing the Governance and

---

[12] Ursula von der Leyen, 'A Union that strives for more: My agenda for Europe' <https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf> 13.

[13] AI Standardisation Committee and Department of Telecommunications, 'Indian Artificial Intelligence Stack' (September 2020) <https://ourgovdotin.files.wordpress.com/2020/09/paper-for-development-of -indian-artificial-intelligence-stack.pdf> (AI Stack Document).

[14] NITI Aayog, Towards responsible AI for All n(1).

[15] Centre for Communication Governance n(2).

Security Layer of the proposed AI Stack. Hence, in this document we draw from our previous response to the NITI Aayog and add in analysis that is particularly relevant from the perspective of the proposed AI Stack.

The AI Stack Document discusses the issues around developing a framework of an Indian AI stack and proposes an AI stack with an aim to remove the challenges to AI deployment and enable its productive adoption. The AI Stack Document proposes to divide the AI Stack in six different layers with horizontal and vertical integration. The five horizontal layers include the Infrastructure layer, the Storage layer, the Compute layer, the Application layer and the Data Exchange layer and the vertical layer is envisaged to be the Security and Governance layer. The vertical Security and Governance layer cutting across all the layers would ensure that AI systems are safe, secure and trusted. The framework of this Security and Governance layer of the stack directly relates to broader questions of the design of a regulatory and governance framework for AI in India.

The AI Standardisation Committee was formed by the DoT, to develop various interface standards and develop India's AI stack. The AI Stack Document discusses the need for "regulatory standards for data collection, interfaces, storage, analysis, application and customer use".[16] The AI Stack Document aims to "address some of these bottlenecks from standardisation point of view".[17] This stack is aimed to be structured across all sectors to ensure the protection of data, data federation, data minimisation, open algorithm framework, defined data structures, interfaces and protocols, proper monitoring, audit and logging, data privacy, ethical standards, digital rights, and trustworthiness.[18]

The AI Stack Document also discusses the need for adoption of internationally accepted principles for AI governance. As discussed in section 4 of our comments, the development of and adoption of these principles is an essential component of designing an effective governance framework for AI systems in India.

[16] AI Stack Document n(13) 18.
[17] ibid 15.
[18] Ibid.

Hence, the AI Stack Document not only envisages the creation of a Security and Governance layer in the Indian stack but also refers to the integration of principles for AI governance in India's governance framework. Consequently, in our response we seek to highlight the broader challenges and principles that need to be considered in the framing of an AI governance framework in India. We draw heavily on our previous response to the NITI Aayog[19] besides adding in analysis particularly relevant from the perspective of the proposed AI Stack.

Any governance structure and legislative framework for AI will need to balance the benefits and risks surrounding the deployment of AI and ensure its deployment aligns with Indian constitutional requirements, minimise the risks of potential harm and ensure the safe deployment of the technology. We explore these regulatory and governance challenges in the next section.

2. ANCHORING AI REGULATORY PRINCIPLES WITHIN INDIA'S CONSTITUTIONAL FRAMEWORK

The Constitution of India provides fundamental rights protecting an individual's right to equality[20], privacy[21] and freedom of speech and expression[22] (among others) and specifically protects individuals against various forms of discrimination arising from India's historical and cultural context. The use of AI systems has the potential to violate various fundamental rights enshrined in the Indian Constitution. To illustratively highlight the potential impact of the use of AI systems on fundamental rights, in this section, we specifically discuss the impact of AI systems on the right to equality and the right to privacy.

2.1. Discrimination

AI systems are trained on existing datasets. These datasets tend to be historically biased, unequal and discriminatory. Given that AI systems make decisions based on their training on existing datasets, we have to be cognizant of the propensity for historical bias' and

[19] Centre for Communication Governance n(2).
[20] Article 14, Constitution of India.
[21] Recognised under Article 21 and Part III of the Constitution of India.
[22] Article 19(1)(a), Constitution of India.

discrimination getting imported into AI systems. The AI Stack Document highlights how data can have embedded in it "unconscious and institutional biases" and other prejudices, which can get codified in the AI decision making matrix for years to come[23] and lead to flawed or biased AI decisions.[24]

Unless we attempt to tackle this challenge, due to the nature of AI technology and its potential for widespread impact, such discrimination will not only get further embedded in Indian society but also be severely exacerbated.[25] Given this, AI systems can have a disproportionate impact and consequences on marginalised and vulnerable communities. Additionally, marginalised and vulnerable communities have traditionally been at the margins of data collection and digital inclusion. We need to ensure that the deployment of AI systems in spaces such as FinTech and health do not end up further alienating and marginalising these communities.

Another policy document, the NITI Aayog's 'National Strategy for Artificial Intelligence', published in 2018, ('National Strategy for AI 2018')[26] elaborates upon the fairness concerns surrounding the use of AI systems. The strategy document acknowledges that bias is inherent in current datasets, and there is potential for such biases to get reinforced through the use of AI systems. The strategy suggests that fairer results can be achieved by identifying in-built biases, assessing their impact and finding strategies to reduce the bias in the datasets.[27] While such attempts are appreciable in their efforts to rectify the situation and yield fairer outcomes, such an approach disregards the fact that these datasets are biased because they arise from a biased, unequal and discriminatory world. As we seek to build effective regulation to govern the use and deployment of AI, we have to remember that AI systems are socio-technical systems that reflect the world around us and embed the biases, inequality and discrimination inherent in Indian society. We have

[23] AI Stack Document n(13) 16.
[24] ibid 17.
[25] Virginia Eubanks, *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor* (St Martin's Press 2018).
[26] NITI Aayog, 'National Strategy for Artificial Intelligence' (June, 2018), <https://niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf> (National Strategy for AI).
[27] ibid 85.

to keep this broader Indian social context in mind as we design AI systems and create regulatory frameworks to govern their deployment.

The Indian Constitution guarantees both citizens and non-citizens a fundamental right to equality before the law and the equal protection of the laws.[28] This prohibits discrimination and requires the State to give special treatment to persons in different situations to establish equality amongst all. No discrimination can be made by the State based on religion, race, sex, caste or place of birth.[29] Use of AI systems in criminal justice systems by the State for sentencing or decisions to prosecute or detain people could be problematic as there may be embedded biases in the algorithm which may re-enforce discrimination towards a particular religion, race, sex, caste or place of birth. While the AI Stack Document alludes to the need for relevant principles for responsible AI, there needs to be substantive discussion around how AI principles around equality, and inclusivity and non-discrimination will be developed to ensure compliance with India's constitutional right to equality.

2.2. Privacy

Increasingly we are seeing the global development and deployment of AI surveillance systems. For instance, the Indian government has put out a tender for the acquisition of facial surveillance technology.[30] In the context of smart cities, the NITI Aayog's National Strategy for AI 2018 discusses the use of AI-powered surveillance applications to predict crowd behaviour and for crowd management.[31] The use of AI powered surveillance systems has to be balanced with their impact on an individual's right to freedom of speech and expression and privacy. Decisions to deploy such technologies have to be made keeping in mind not only their adverse impact on fundamental rights guaranteed by the

[28] Article 14, Constitution of India.
[29] Article 15, Constitution of India.
[30] National Crime Record Bureau, "Request for proposal to procure National Automated Facial Recognition System (AFRS)" <https://ncrb.gov.in/sites/default/files/tender/AFRSRFPDate22062020UploadedVersion.pdf>.
[31] National Strategy for AI n(26).

Indian Constitution but also the current operational challenges around the extent of their accuracy and fairness.

Broader public consultation should be undertaken by the government, before deploying remote biometric identification AI systems such as facial recognition in public places. Given that the use of such systems poses significant risks to fundamental rights, their deployment by the government if at all, should only be done in specific contexts for a particular purpose and in compliance with the principles laid down by the Supreme Court in the case of *K.S. Puttaswamy vs. Union of India*.[32] In *Puttaswamy*, the Court laid down the standard of judicial review to be applied in cases of violation of an individual's right to privacy by the State. The Court held that the right to privacy may be restricted through actions of the government where such intrusion meets the three-fold requirement of (i) legality, which postulates the existence of an enabling law; (ii) need, defined in terms of a legitimate state aim; and (iii) proportionality which ensures a rational nexus between the objects and the means adopted to achieve them.[33] Additionally, a fourth prong to this test mandating "procedural guarantees against abuse of such interference" was provided by Justice Sanjay Kishan Kaul in his opinion.[34] Accordingly, any deployment of AI systems by the government would have to adhere to these tests safeguarding the right to privacy. While designing a governance framework for AI in India, there needs to be a detailed discussion on the contours of the AI governance principle of privacy and security to ensure it is in consonance with India's constitutionally recognised right to privacy.

In conclusion, the AI Stack Document discusses the challenges of bias, privacy, security and inclusivity and refers to the need for a well-designed regulatory standard in the form of an open Indian stack in line with internationally agreed principles. However, it does not engage with the principles for responsible AI such as equality, inclusivity and non-discrimination, and privacy and security that have are being developed globally. There needs to be substantive discussion around the development of these AI principles in the

[32] K.S. Puttaswamy vs. Union of India (2017) 10 SCC 1.
[33] Privacy Law Library, 'K.S. Puttaswamy vs. Union of India' <https://privacylibrary.ccgnlud.org/case/justice-ks-puttaswamy-ors-vs-union-of-india-ors?searchuniqueid=619122>.
[34] ibid [533] (Justice Kaul).

Indian context to ensure that they are in consonance with the constitutional guarantees for rights such as the right to equality and right to privacy.

3. Governance and Regulatory Challenges in the Adoption of AI in India

As the DoT designs the 'Security and Governance' layer of the AI stack, they should consider the broader governance and regulatory regime that needs to be put in place to govern the adoption and deployment of AI systems in India. The governance and regulatory challenges that need to be considered include:

3.1. Scope of AI Regulation

An essential aspect of developing a regulatory framework to govern AI systems is to chalk out the scope of its application. Conceptually the regulatory framework should apply to all products and services that use AI. Effective regulation would require a clear definition of what constitutes AI and will fall within the ambit of the regulatory framework for AI. A definition of what constitutes AI needs to be specific enough to be able to pinpoint with some certainty what constitutes an AI system but flexible enough to encompass the rapid technological progress happening in this space.

3.2. Heightened Threshold of Responsibility for Government or Public Sector Deployment of AI

To effectively regulate AI while not being excessively prescriptive, jurisdictions such as the EU have been contemplating adopting a risk-based approach. The European Commission ('Commission') believes this allows the regulatory intervention to be proportionate and not be unduly burdensome on SMEs.[35] The Commission is exploring the adoption of criteria to differentiate between various AI applications and determine whether an AI application is high risk or not. In particular, when evaluating whether an application is high risk they look at the sector of deployment and whether the intended use poses significant risks to safety, consumer rights and fundamental rights.[36] They have identified the use of AI in the public sector, or by the government as a high risk situation,

[35] European Commission, White Paper n(6).
[36] ibid 17.

and also identified sectors such as healthcare, transport and energy as high risk sectors. In conjunction with the sector, they assess the manner in which the AI system will be used in the sector and the risks likely to arise from such use and the extent to which they impact individuals. Along the same lines, the German Data Ethics Commission has proposed a risk-based regulatory system comprising five thresholds of risk ranging from no regulation for "innocuous" AI systems to a complete ban for dangerous AI systems.[37]

A potential model suggested in the AI Stack Document is the implementation of a certification service for all providers of critical AI services that are sector centric.[38] This would include certification requirements for the deployment of AI services in sectors such as financial, health, legal, etc., and would be implemented as a part of the Security and Governance layer.

Drawing on these proposed approaches, it is imperative that India at the very least, considers the adoption of a higher regulatory threshold for the use of AI by government institutions, especially where citizen's fundamental rights are directly impacted. Examples of such use by the government could include the use of AI by the government for law enforcement and surveillance or the use of AI by the judiciary for sentencing and predicting criminal recidivism or use of AI for disbursal of government benefits to citizens. Use of AI systems in any of these processes of government decision making or functions can have a severe negative impact on the fundamental rights of a broad cross-section of Indian citizens. For instance, tools such as the Correctional Offender Management Profiling for Alternative Sanctions ('COMPAS') used by US Courts to assess recidivism, a criminal defendant's likelihood of re-offending, have faced severe criticism and through detailed review have been demonstrated to be heavily biased against certain racial

[37] Data Ethics Commission, 'Opinion of the Data Ethics Commission' (2020) <https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3>; For example algorithms which are used in vending machines would qualify as level 1. Dynamic pricing on e-commerce portals may count as a level 2, whereas price algorithms for setting personalised prices would count as level 3. Algorithms which decide creditworthiness of applicants are classified as level 4 and finally algorithms used in lethal autonomous weapons systems may be classified as level 5.

[38] AI Stack Document n(13) 36.

groups.[39] The AI Stack document discusses an incident in 2016 around the use of COMPAS software by some US courts in predicting the likelihood of recidivism in criminal defendants. The use was demonstrated to be biased since the AI 'black box' was 'proprietary'. The AI Stack Document therefore encourages "openness in AI algorithms, an open algorithm framework and a need to enable clearly defined data structures".[40]

In the United States, automated decision making around eligibility and other aspects of the disbursal of government welfare benefits has been documented to wreck the social safety net, criminalise the poor and enhance discrimination against already marginalised groups.[41] Consequently, careful thought has to go into creating a regulatory system that has a higher standard of scrutiny and assessment of the design and deployment of an AI system to be used by the government.

3.3. Need for Overarching Principles based AI Regulatory Framework

Different sectoral regulators, including various financial regulators, have started to evolve regulations to address the challenges posed by AI. Given the expertise and technical depth of a sectoral regulator in navigating their specific domain of regulation, there is tremendous value in them examining how AI impacts their field. However, at the same time, divergent regulatory approaches and rules may create uncertainty and fragmentation in the overall approach to regulate AI in India. Additionally, sectoral regulators may not have the required technical expertise to fully comprehend the more technical aspects of regulating an AI application in their domain. Consequently, it is important to evolve an overarching and detailed policy framework to guide the direction of the development of the regulation of AI in India and develop mechanisms to support the technical needs of sectoral regulators as they navigate the impact of AI on their domain.

We draw on the global thinking that has emerged around the regulation of AI systems and explore the application of some of the key principles that have developed, in the

[39] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, 'Machine Bias' ProPublica (23 May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
[40] AI Stack Document n(13) 16.
[41] Virginia Eubanks n(25).

Indian context. Internationally, several countries and international bodies have developed principles around the use of AI. We discuss these principles and their potential application in the Indian context in section 4 of this policy response. We hope this discussion will be helpful for the DoT to draw on as they develop the foundational principles for the regulatory framework for AI in India. While it is useful to develop overarching principles for the regulation of AI, as underlined earlier, it is vital to think through how these principles will interact with constitutional principles and the fundamental rights enshrined in the Indian Constitution.

As discussed above, AI systems are susceptible to bias and discrimination. Inherent biases also impact human decision making. However, the challenge is that the same biases when present in an AI system, whether in the form of the dataset it is trained on or the factors the AI is designed to take into account, can have a more significant effect. They can impact human lives and discriminate against individuals without the legal oversight and social control mechanisms that seek to govern human behaviour.

This challenge is further exacerbated by relative opacity (black box effect) of AI systems and limited explainability of AI behaviour.[42] AI systems learn while in operation from the datasets they are fed and further refine the correlations and patterns that the system identifies in the dataset.[43] The regulatory challenge is to not only prevent design decisions that result in discriminatory impacts but also create an oversight mechanism over how the working of an AI system evolves. Given the challenges of opacity and explainability, serious thought will have to be given on how and to what extent mechanisms can be evolved to verify the basis on which a decision is made by the AI and gauge compliance with constitutional principles and other statutory rights. Additionally, well-developed mechanisms for accountability and liability for harm will have to be put in place.

[42] Yavar Bathaee 'The artificial intelligence black box and the failure of intent and causation' (2018) 31 Harvard Journal of Law & Technology 890.

[43] Katrina Wakefield, 'A guide to machine learning algorithms and their applications' (SAS) <https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html>; Tom Mohr, 'In The Loop — Chapter 26: AI, Machine Learning and Deep Learning' (Medium, 22 October 2019) <https://medium.com/ceoquest/in-the-loop-chapter-26-ai-machine-learning-and-deep-learning-140cc13a77b7>.

As previously discussed, the AI Stack Document underlines that the development of the AI stack would ensure uniformity across all sectors[44] and hence provide for an overarching structure for the deployment of AI systems in India. The AI Stack Document iterates that it is with a view to tackle bottlenecks from the perspective of standardisation that the DoT is seeking to develop various interface standards and India's AI stack.[45] While we welcome this initiative taken by the DoT, and the AI Stack Document, more public discussion is required around the building an overarching sectoral framework.

3.4. Adaptation of Sectoral Regulation to Effectively Regulate AI

It is important to envisage the manner in which these AI regulatory principles will interact with sector-specific laws such as the financial sector and other horizontal laws on subjects such as consumer protection and product liability that may apply to various AI systems. The relevant regulatory agencies may have to be empowered to intervene in the regulation of AI systems falling within their domain. Besides this, they will need the required technical expertise to inspect AI systems adequately. Mechanisms will have to be evolved to strengthen the capacity of sectoral regulators to navigate regulatory decision making in the context of the deployment of AI systems in their sectors. The opaqueness and lack of transparency of AI systems make it challenging to identify violations of laws and assess compliance with constitutional principles and other statutory rights and liabilities. Consequently, to ensure the effective application of laws and compliance with them, it may be necessary for existing legislation to be modified to address this change in technology. For instance, sectoral legislation regulating driving and drivers such as the Motor Vehicles Act, 1988 will have to be modified to enable and regulate the use of autonomous vehicles and other AI transport systems.

Horizontal laws such as those governing consumer protection and liability rules may have to be translated to stay relevant in the context of AI systems. To underline the importance of adequately adapting sectoral and horizontal legislation, we explore this example in some more detail.

[44] AI Stack Document n(13) 2.
[45] ibid 15.

The use of AI in various products and services can present safety concerns for users. These can arise from the quality and breadth of data relied on or design challenges in the AI system or problems related to machine learning. For instance, in autonomous vehicles, a defect in the object recognition technology can cause an accident causing damage and injuries. India has an existing body of legislation and jurisprudence that has developed around consumer protection, product safety and liability. This jurisprudence will be relevant and potentially applicable to many of the emerging AI applications. Clarity will have to be created concerning the application of these rules to the use of AI systems. Such clarity will not only address safety concerns for individuals but also develop standards for businesses using AI to comply with. Development of such standards will help dispel some of the regulatory uncertainty that surrounds the use of AI by businesses.

Challenges with regard to explainability and transparency of decision making by AI systems may make it difficult to apply current consumer protection and product liability statutory rules. Traditionally, consumer protection and product liability regulatory frameworks have been structured around fault-based claims. This involves the identification and proof of a defect and establishing a causal link between the defect and the damage caused.[46] However, given the nature of the functioning of AI systems, it may be challenging to establish the presence of a defect and for the individual who has suffered harm to provide the necessary evidence in court. A detailed study will have to be made on whether current consumer protection and product liability rules in India can be effectively applied to AI systems and the potential ways in which the current rules may have to be adapted to address concerns in the realm of AI systems. For instance, this could involve adapting the burden of proof requirements on consumers to establish a claim for damage caused by the operation of an AI system.[47] Additionally, while many of

[46] Chapter VI: Product Liability, Consumer Protection Act, 2019, <http://egazette.nic.in/WriteReadData/2019/210422.pdf; European Commission> ; 'Liability for Artificial Intelligence and other emerging digital technologies' (2019) <https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608 > 16.

[47] Under the product liability statutory framework, a consumer can make a claim for compensation under a product liability action for any harm caused by a defective product manufactured by a product manufacturer, serviced by a product service provider or sold by a product seller. For instance, two of the various grounds on which a product manufacturer is liable for product liability action is if either the product contains a

the provisions of the existing consumer protection legislative framework could potentially be extended to protecting consumers in relation to AI, additions may need to be made to address new risks posed by emerging digital technologies like AI.

3.5. Contextualising AI systems for both their Safe Development and Use

To ensure their effective and safe use, AI systems have to be designed, adapted and trained on relevant datasets, depending on the context in which they will be deployed. Without effectively contextualising AI systems to the environment they are to be deployed in, there are enhanced safety, accuracy and reliability concerns. Additionally, there is greater potential for negative impact on an individual's fundamental rights.

The AI Stack Document recognises the importance of deploying AI systems in sectors such as healthcare. The AI Stack Document discusses how the use of AI in healthcare could help in mitigating the problem of high barriers of access to healthcare facilities in rural areas which suffer from limited availability of healthcare professionals and facilities.[48] The implementation of AI could be used to drive diagnostics, provide personalised treatments, early identification of potential pandemics, and imaging diagnostics, among others.[49] However, a relevant consideration around the use of AI in healthcare is which context it has been trained to be deployed in.[50] Whether the AI has been trained on data relevant to such a low resource rural context or whether it has been trained on data based on high resource speciality hospitals in urban areas?

Another instance in which it is relevant to consider the importance of contextualising AI for safe use is the emphasis of the NITI Aayog on fashioning India as an AI Garage for the development and deployment of AI systems. The NITI Aayog envisages India being

---

manufacturing defect or the product is defective in design. As discussed, opaqueness and lack of transparency and explainability, may hinder a consumer's ability to establish a manufacturing defect or that a product is defective in design. Mechanisms will have to be evolved to address dissonance. Chapter VI: Product Liability, Consumer Protection Act, 2019, <http://egazette.nic.in/WriteReadData/2019/210422.pdf>.

[48] AI Stack Document n(13) 10.

[49] Ibid.

[50] W. Nicholson Price, 'Medical AI and Contextual Bias', Harv. J.L. & Tech. 33 (2019), 66.

the AI Garage for 40% the world,[51] especially to the global south - with the aim of developing AI solutions in India which can then be deployed in other countries. Additionally, it is likely that India will import various AI systems developed in countries such as the US, EU and China to be deployed within the Indian context.

Both scenarios involve the use of AI systems in a context distinct from one in which they have been developed. AI systems are socio-technical systems and cannot be divorced from the social context in which they are designed to function. When you transpose AI systems designed and trained on data outside India, into the Indian context, how do you translate them to function effectively in the Indian context? Similarly, when AI systems are trained on Indian datasets and designed keeping in mind the Indian context, how do you ensure their safe deployment in other countries? Regulatory standards and processes need to be developed in India to ascertain the safe use and deployment of AI systems that have been developed in contexts that are distinct from the ones in which they will be deployed.

4. PRINCIPLES FOR THE REGULATION OF AI

The AI Stack Document recognizes that the introduction of AI systems without proper safeguards may lead to bias, lack ethical governance, and limit transparency in its decision making process causing unfair outcomes and amplifying unequal access.[52] The AI Stack Document also refers to the need for a well-designed regulatory standard in the form of an open Indian stack in line with internationally agreed principles. However, it does not specifically engage with the principles for responsible AI and ethical AI that are being developed globally. AI ethics specifically deal with how human developers, manufacturers and operators should behave in order to minimise the ethical harms that can arise from AI in society, either arising from poor (unethical) design, inappropriate

[51] Niti Aayog, Towards responsible AI for All n(1) 3.
[52] AI Stack Document n(13) 18.

application or misuse.[53] It is important to underscore the discussion surrounding the adoption and deployment of AI systems in a framework of AI ethics.

There needs to be substantive discussion around the adoption of the principles for ethical and responsible AI into India's AI governance framework. These principles include those of safety and reliability, equality, inclusivity and non-discrimination, privacy and security, transparency and accountability.

These ethics must be developed in conjunction with relevant stakeholders and in line with the international standards for ethical AI and the Indian constitutional framework.

In this section, we discuss five key principles that should be embedded in any regulatory system that is adopted for governing AI. We draw from literature that has developed globally around these principles and attempt to discuss them from the perspective of embedding them into an Indian regulatory context. A compilation of the various sets of principles proposed by countries across the world, international bodies, multinational companies and civil society organisations that we have referred to in our comments is provided in the Appendix to this document. More recently, the NITI Aayog's Working Document Towards Responsible #AIforAll[54] has also discussed the principles for responsible AI that need to be embedded in an Indian governance framework and we draw from that discussion as well.

In these comments we will discuss five principles, namely principle of safety and reliability, equality, inclusivity and non-discrimination, privacy and security, transparency and accountability. These principles are the basis upon which India's regulatory framework for AI needs to be developed. In the following sections we discuss the components of these principles and mechanisms for their application.

[53] European Parliament, 'The ethics of artificial intelligence: Issues and initiatives' (2020) <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf > 2.

[54] NITI Aayog, Towards responsible AI for All n(1).

4.1. Principle of Safety and Reliability

The principle of reliability and safety aims to ensure that AI systems reliably operate in accordance with their intended purpose throughout their lifecycle.[55] This includes ensuring AI systems are reliable in relation to their roles and ensures the security, safety and robustness of an AI system.[56] AI systems should not pose unreasonable safety risks, should adopt safety measures which are proportionate to the potential risks, should be continuously monitored and tested to ensure compliance with their intended purpose, and should have a continuous risk management system to address any identified problems.[57]

Here, it is also important to note the distinction between the terms safety and reliability. The reliability of a system relates to the ability of an AI system to behave exactly as its designers have intended and anticipated. A reliable system would adhere to the specifications it was programmed to carry out. Reliability is, therefore, a measure of consistency, and it establishes confidence in the safety of a system.[58] Safety refers to an AI system's ability "do what it is supposed to do, without harming users (human physical integrity), resources or the environment."[59]

---

[55] Department of Industry, Innovation and Science, Australian Government, 'AI Ethic Principles' (2019) <https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles> , "Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose".

[56] G20 Ministerial Meeting on Trade and Digital Economy, 'G20 AI Principles' (2019) <https://www.g20-insights.org/wp-content/uploads/2019/07/G20-Japan-AI-Principles.pdf>, "Principle 1.4. Robustness, security and safety - (a) AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk. (b) To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system's outcomes and responses to inquiry, appropriate to the context and consistent with the state of art. (c) AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias".

[57] Australian AI ethic principles n(55).

[58] D. Leslie, Alan Turing Institute, 'Understanding Artificial Intelligence Ethics and Safety' (2019) <https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf> 31.

[59] High Level Expert Group on Artificial Intelligence set up by the European Commission, 'Ethics Guidelines for Trustworthy AI' (2019) <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> 25.

Processes need to be put in place to assess the safety and reliability of the AI systems not only at the time of their deployment but during the entire life cycle of their deployment. AI systems need to be able to deal with direct attacks and attempts to manipulate the data or algorithms and flag errors that may arise. One of the potential models which may be incorporated to reduce concerns surrounding safety and reliability of AI systems would be to introduce human oversight in the deployment of AI systems.

4.1.1. Human Oversight

The AI Stack Document recognises the risks of passive adoption of AI systems which automate human decision. It identifies that "such delegation can lead to harmful, unintended consequences, especially when it involves sensitive decisions or tasks and excludes human supervision."[60]

Therefore, an important aspect of ensuring the safety and reliability of AI systems is the presence of human oversight over the system. Regulatory principles will have to specify the circumstances and degree to which human oversight is required. The purpose for which the system is deployed and impact it could have on individuals would be relevant factors in determining if human in the loop[61], human on the loop[62], or any other oversight mechanism is required. For instance, due to the sensitivity of the function and potential for significant impact on an individual's life, AI systems deployed in the context of the provision of government benefits, should have a high level of human oversight. Decisions made by the AI system should be reviewed by a human before being implemented. On the other hand, AI systems such as autonomous vehicles should have the ability for real time human intervention. There will be AI systems which are deployed in contexts that do not need constant human involvement but should have a mechanism in place for human review if a decision is subsequently raised for review by, say a user.

[60] AI Stack Document n(13) 15.
[61] Ge Wang, 'Humans in the Loop: The Design of Interactive AI Systems' Human Centred Artificial Intelligence (Stanford, 20 October 2019) <https://hai.stanford.edu/blog/humans-loop-design-interactive-ai-systems>.
[62] Ibid.

4.1.2. Assessment system for the safety and reliability of AI systems

To ensure compliance with the legal requirements and assess the potential impact of an AI system on society, the regulatory framework can have an assessment system in place for AI systems that are to be deployed in sensitive contexts. This will allow for assessment of the safety and reliability of the AI systems. For instance, in particularly sensitive sectors, methods analogous to the regulation of medical devices in India can be adopted. Medical Devices (Amendment) Rules 2020[63] require the registration of all medical devices with a central licensing authority in order to ensure that every medical device, either manufactured in India or imported, has quality assurance before they can be distributed / sold in the market.[64]

Certain AI systems, once deployed, continue to develop and learn from their experience. The regulatory framework will need to have an assessment system in place to conduct periodic evaluations of such AI systems. These assessments could be a combination of self-assessment, assessments by expert third parties and by regulatory bodies.

4.2. Principle of Equality, Inclusivity and Non-Discrimination

The principles of equality, inclusivity and non-discrimination are among the most common principles included in most AI principles. The principles are vast in scope and include aspects relating to fairness and human centred values.[65] The principle also refers to the

[63] Medical Devices (Amendment) Rules 2020 <https://cdsco.gov.in/opencms/opencms/system/modules/CDSCO.WEB/elements/download_file_division.jsp?num_id=NTU0OQ==>.

[64] 'Medical Devices (Amendment) Rules 2020: Impact on new Government regulations' (Express Healthcare, 10 April 2020) <https://www.expresshealthcare.in/blogs/medical-devices-amendment-rules-2020-impact-on-new-government-regulations/418451/>.

[65] G20 AI Principles n(58), "1.2. Human-centred values and fairness - (a) AI actors should respect the rule of law, human rights and democratic values, throughout the AI system lifecycle. These include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised labour rights. (b) To this end, AI actors should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art".

inclusion of aspects of equity,[66] diversity,[67] and the promotion of human rights.[68] The principle seeks to address pertinent concerns surrounding the implementation of the human rights of equality, non-discrimination and inclusivity. In the following sections we examine the constituent principles in detail.

4.2.1. Principle of Equality

The principle of equality holds that everyone, irrespective of their status in the society, should get the same opportunities and protections with the development of AI systems.[69]

---

[66] University of Montreal, Canada, 'Montreal Declaration for Responsible AI' (2018) <https://www.montrealdeclaration-responsibleai.com/the-declaration>, "Equity Principle - (i) AIS must be designed and trained so as not to create, reinforce, or reproduce discrimination based on — among other things — social, sexual, ethnic, cultural, or religious differences. (ii) AIS development must help eliminate relationships of domination between groups and people based on differences of power, wealth, or knowledge. (iii) AIS development must produce social and economic benefits for all by reducing social inequalities and vulnerabilities. (iv) Industrial AIS development must be compatible with acceptable working conditions at every step of their life cycle, from natural resources extraction to recycling, and including data processing. (v) The digital activity of users of AIS and digital services should be recognised as labor that contributes to the functioning of algorithms and creates value. (vi) Access to fundamental resources, knowledge and digital tools must be guaranteed for all. (vii) We should support the development of commons algorithms — and of open data needed to train them — and expand their use, as a socially equitable objective.

[67] ibid "Diversity and Inclusion Principles - (i) AIS development and use must not lead to the homogenisation of society through the standardisation of behavior and opinions. (ii) From the moment algorithms are conceived, AIS development and deployment must take into consideration the multitude of expressions of social and cultural diversity present in the society. (iii) AI development environments, whether in research or industry, must be inclusive and reflect the diversity of the individuals and groups of the society. (iv) AIS must avoid using acquired data to lock individuals into a user profile, fix their personal identity, or confine them to a filtering bubble, which would restrict and confine their possibilities for personal development — especially in fields such as education, justice, or business. (v) AIS must not be developed or used with the aim of limiting the free expression of ideas or the opportunity to hear diverse opinions, both being essential conditions of a democratic society. (vi) For each service category, the AIS offering must be diversified to prevent de facto monopolies from forming and undermining individual freedoms."

[68] Institute of Electrical and Electronics Engineers, 'Ethically Aligned Design: Version 2' <https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf>, "Principle 1 - Human Rights–A/IS shall be created and operated to respect, promote, and protect internationally recognised human rights."

[69] European Commission's High-Level Expert Group on Artificial Intelligence n(59); Amnesty International and Access Now, 'The Toronto Declaration' (2018) <https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf> "Equality of human beings goes beyond nondiscrimination, which tolerates the drawing of distinctions between dissimilar situations based on objective justifications. In an AI context, equality entails that the same rules should apply for everyone to access information, data, knowledge, markets and a fair distribution of the value added being generated by technologies".

Equality in respect of AI should exist "in terms of human rights, access to technology, and guarantees of equal opportunity through technology".[70] The principle of equality has several corollary principles such as the principle of fairness, non-discrimination, inclusiveness, etc. Different AI instruments have chosen to deal with these principles in different ways.

Implementing equality in AI systems essentially requires three components:

(i) Protection of Human Rights: AI instruments developed across the globe have highlighted that the implementation of AI would pose risks to the right to equality, and countries would have to proactively take steps to mitigate such risks.[71]

(ii) Access to Technology: The AI systems should be designed in a way to ensure widespread access to technology, so that people may derive benefits from AI technology.

(iii) Guarantees of Equal Opportunities through Technology: The guarantee of equal opportunity relies upon the transformative power of AI systems to "help eliminate relationships of domination between groups and people based on differences of power, wealth, or knowledge" and "produce social and economic benefits for all by reducing social inequalities and vulnerabilities."[72]

4.2.2. Principle of Non-Discrimination

The idea of non-discrimination on the other hand mostly arises out of technical considerations in the context of AI. It holds that non-discrimination and the prevention of bias in AI should be mitigated in the training data, technical design choices, or the technology's deployment to prevent discriminatory impacts.[73] The AI Stack Document

[70] Jessica Fjeld n(10).
[71] Ibid.
[72] Montreal Declaration for Responsible AI n(66).
[73] Smart Dubai, 'Dubai's AI Principles' (2019) <https://www.smartdubai.ae/initiatives/ai-principles> "Data ingested should, where possible, be representative of the affected population, and Algorithms should avoid non-operational bias".

also considers the potential discriminatory impacts which may be caused by the use of computing analytics, in the compute layer of the proposed AI stack.[74]

A brief literature review of non-discrimination in AI decision-making demonstrates that AI can lead to discrimination in at least six different ways:[75]

(i) Target Variables and Class Labels:[76] Target variable and class labels are used in the process of data mining to filter information. A target variable is the result sought to be achieved whereas the class labels help to sort the values in relation to the target variables. An example of this can be an AI system which hopes to classify 'good' employees (target variable). Here the potential class labels could be objectives achieved, punctuality, etc. However, for an office situated in the middle of a major city, it is possible that people from the suburbs (potentially from a lower income class) would be in the category which is late more often. As a result of this, the AI system may learn to exclude people from specific neighbourhoods, from the filter of 'good' employees due to their address. Therefore, the selection of target variables and class labels can inadvertently lead to discrimination by the AI system.

(ii) Training Data:[77] AI decision-making can also have discriminatory results if the system "learns" from discriminatory training data. This may occur in two ways, the AI system might be trained on biased data or problems may arise when the AI system learns from a biased sample. In both these cases, the AI system would reproduce this bias.

(iii) Collecting Training Data:[78] The procedure of sampling, or the collection of training data upon which the AI system is to be trained, can also be biased. The success of an AI system is directly dependent on the training data and therefore, the training data collection

[74] AI Stack Document n(13) 30. The AI Stack Document states that computing analytics "involves analysis to mine troves of personal data and find correlations, which will then be used for various computations".
[75] Council of Europe, 'Discrimination, artificial intelligence, and algorithmic decision-making' (2018) <https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73>.
[76] ibid 10.
[77] Solon Barocas & Andrew D. Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671, 681.
[78] Can Yavuz, 'Machine Bias: Artificial Intelligence and Discrimination' (2019) <https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8987035&fileOId=8987040>.

process has a large impact on the output of the AI system. As AI becomes increasingly ubiquitous in all aspects of life, it would be important to ensure that the development and training of these systems is undertaken with data that is fair, interpretable and unbiased.[79]

Examples of this can be seen in data collection in policing. Given that the police force has historically paid extra attention to neighbourhoods with minorities, this typically results in police data showing higher incidences of crime in minority neighbourhoods.[80] If an AI system is trained on such biased data it would result in skewed output. Use of AI systems becomes safer when they are trained on datasets that are sufficiently broad and the datasets encompass the various scenarios in which the system is envisaged to be deployed. Additionally, datasets should be developed to be representative of the whole population and must not over-represent or under-represent any section of the populace to avoid discriminatory outcomes from the use of the AI system. Rules will have to be developed within India's regulatory framework on AI to ensure this.

(iv) Election of Features/ Technical Design Choices:[81] AI systems often solve complex problems by simplifying the problem into simpler issues, using a process called feature selection. Feature selection refers to the attributes or the technical design choices which an AI system observes and subsequently considers in its analyses. Feature selection may generate discriminatory treatment on protected grounds, as it may not consider that the details necessary to achieve non-discriminatory determinations would reside at a level of granularity which is not considered by the selected features.

The importance of feature selection has also been echoed in the AI Stack Document. The document recommends changing the overall culture so that coders and developers may themselves recognise the "harmful and consequential" implication of biases. We welcome the DoT's broad perspective which insists on looking beyond the standardisation of the algorithmic code and focuses on the programmers of the code. The AI Stack Document iterates that since much of coding is outsourced, this would place the onus on the

[79] 'Bias in AI: How we Build Fair AI Systems and Less-Biased Humans' IBM (Think Policy Blog, 1 February 2018) <https://www.ibm.com/blogs/policy/bias-in-ai/>.
[80] Discrimination, artificial intelligence, and algorithmic decision-making n(75) 11.
[81] Solon Barocas & Andrew D. Selbst n(77) 688.

company developing the software product to enforce such standards.[82] They are of the view that such a comprehensive approach would tackle the problem across the industry as a whole, and enable AI software to make fair decisions made on unbiased data, in a transparent manner.

(v) Proxies:[83] While making decisions AI systems may require access to sensitive data on protected characteristics such as race, ethnicity, political opinions, etc. Access to these protected characteristics can bias the decision making of the AI system. While developers may remove such sensitive data from the datasets to reduce bias, sometimes AI systems may develop proxies for such data points. A proxy may allow an AI system to include certain categories of data that correspond to other variables which have been removed from the system for having protected characteristics or are sensitive in nature.

(vi) Organisations/ Nations using AI Systems to discriminate on Purpose:[84] Finally, discrimination may also occur on purpose if the system was designed to discriminate on the basis of certain characteristics or choices surrounding the deployment of AI technology.

Examples of Discrimination by AI Systems

To ensure effective non-discrimination, AI policies must mitigate against these factors. An example of this can be semi-autonomous vehicles which experience higher accident rates among dark-skinned pedestrians due to the software's poorer performance in recognising darker-skinned individuals.[85] This can be traced back to training datasets, which contained mostly light skinned people. The lack of diversity in the data set can lead to discrimination against specific groups in society.

In the National Strategy for AI 2018, NITI Aayog highlights the fallibility of AI systems.[86] The AI systems developed using biased training data is likely to be biased. It is proposed

[82] AI Stack Document n(13) 18.
[83] Solon Barocas & Andrew D. Selbst n(77) 688.
[84] Discrimination, artificial intelligence, and algorithmic decision-making n(75) 13.
[85] Brady McCombs, 'Utah driver sues Tesla after crashing in autopilot mode', (Associated Press, 6 September 2018) <https://apnews.com/3f1ac72f186d45cdbfac7dbb04907b11>.
[86] National Strategy for AI n(26).

in the said document that in-built biases must be identified and assessed in order to reduce them. While this is a useful starting point, broader thinking has to go into the fact that datasets are biased because they arise from a biased, unequal and discriminatory world. As we seek to build effective regulation to govern the use and deployment of AI, we have to remember that AI systems are socio-technical systems that reflect the world around us and embed the biases, inequality and discrimination inherent in Indian society. We have to keep this broader Indian social context in mind as we design AI systems and create regulatory frameworks to govern their deployment.

Given the concerns around opacity of AI systems and the related challenges with verifying adherence to rules and regulations, suggestions have been made to maintain records on the programming of the algorithm and the data used to train the AI systems (especially high risk systems).[87] Requiring the documentation of the main characteristics of the data and the process of selection of the dataset would be useful in the context of examining problematic decisions by an AI system.[88] Additionally, it is useful to mandate the documentation of the programming and training methods as well as processes used to build, test and validate the AI system to ensure the effective functioning of the AI including safety and non-discriminatory decision making.[89] The regulatory framework can put in place reasonable time periods for which these records should be maintained, the authorities that are empowered to access and audit these records, the process by which they can access these records and safeguards to protect the intellectual property and confidential information of the developers.

4.2.3. Principle of Inclusivity

The idea of inclusivity is based on the just distribution of the benefits of AI and diverse participation in the process of development of AI.[90] Inclusivity can be implemented in two parts:

---

[87] European Commission, White Paper n(6).

[88] ibid.

[89] Ibid.

[90] Jessica Fjeld n(10) ; Microsoft, 'Microsoft AI principles' <https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6>, "AI systems should empower everyone and engage people. If we are

(i) Inclusivity in the Impact: Inclusiveness in impact refers to the distribution of AI benefits to all the intended users, particularly segments of the population which have been historically discriminated against. Inclusivity in AI seeks to achieve welfare, increase citizen's mental autonomy, and provide equal distribution of economic, social and political opportunity.[91]

(ii) Inclusivity in the Design Process: Inclusiveness in design, on the other hand, relates to diversity in the process of designing AI systems. This can be implemented in two ways, by including diversity in the teams which design AI systems, and by including diversity in the process of deciding the aims of AI deployment in society.[92]

In order to implement inclusivity in AI, the diversity of the team involved in design as well as the diversity of the training data set would have to be assessed.[93] This would involve the creation of guidelines to help researchers and programmers in designing data sets, measuring product performance, asking the right questions and testing new systems through the lens of inclusivity.[94]

Exclusion can often be traced back to five major types of bias arising in data sets, these are:[95]

- **Dataset Bias**: This occurs when the data used to train the machine learning models is not representative of the diversity of the customer base.

---

to ensure that AI technologies benefit and empower everyone, they must incorporate and address a broad range of human needs and experiences. Inclusive design practices will help system developers understand and address potential barriers in a product or environment that could unintentionally exclude people. This means that AI systems should be designed to understand the context, needs and expectations of the people who use them".

[91] Ethics Guidelines for Trustworthy AI n(59).

[92] Jessica Fjeld n(10) 51.

[93] Steven Aldrich 'The Need for Inclusion in AI and Machine Learning' Information Week (2017) <https://www.informationweek.com/big-data/ai-machine-learning/the-need-for-inclusion-in-ai-and-machine-learning/a/d-id/1330464>.

[94] Facebook, 'Building inclusive AI at Facebook' (2019) <https://tech.fb.com/building-inclusive-ai-at-facebook/>.

[95] Joyce Chou, Oscar Murillo, and Roger Ibars 'How to Recognise Exclusion in AI' (Medium, 26 Sep 2017) <https://medium.com/microsoft-design/how-to-recognize-exclusion-in-ai-ec2d6d89f850>.

- **Association Bias**: These biases may occur when the data which is being used to train a model reinforces and multiplies a cultural bias. In such cases, human biases can make their way to machine learning. An example of this may be common associations, such as language translation tools which associate terms like pilots with men, and flight attendants with women.
- **Automation Bias**: This includes bias which occurs when automated decisions override social and cultural considerations. AI systems may give results which go against human diversity. An example of this may be beauty filters, which automatically default to European features.
- **Interaction Bias**: This is one of the most commonly seen biases which occur in chatbots. AI interactions with humans without safeguards may result in introducing bias which may result in infecting the system with human bias. An example of this is when humans deliberately input racist or sexist language into a chatbot to train it to say offensive things.[96]
- **Confirmation Bias**:[97] Confirmation bias interprets information in a way that confirms preconceptions. In this case, the AI systems only serve content which matches a profile created by the system. As the individual only receives information provided by the system, they do not see contrasting points of view and are blocked from seeing alternatives and diverse ideas.

In order to ensure true inclusivity, these factors must also be considered while designing datasets. One way to implement this would be to make the process of research and development in AI inclusive by including social scientists, checking for potential biases in algorithms, exploring the complexities of human-machine interactions, and providing for gender equality in technical sectors.[98] Another potential solution would be to monitor the use of AI after its release and assess its impact amongst different cultures and

[96] James Vincent, 'Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day' The Verge (24 March 2016) <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>.
[97] Joyce Chou, Oscar Murillo, and Roger Ibars 'How to Recognise Exclusion in AI' (Medium, 26 Sep 2017) <https://medium.com/microsoft-design/how-to-recognize-exclusion-in-ai-ec2d6d89f850>.
[98] European Commission, 'Building Trust in Human-Centric Artificial Intelligence', (2019) <https://ec.europa.eu/transparency/regdoc/rep/1/2019/EN/COM-2019-168-F1-EN-MAIN-PART-1.PDF>

communities.[99] The launch of a truly inclusive AI system would require continuous testing on datasets, model outcomes for mitigation of bias and adjustments to AI systems in real time.

Examples of Non-Inclusiveness by AI Systems

Another potential error may be inconsistencies in the decisions being made by the AI across stakeholders, who should have ordinarily been treated alike. This can usually be traced back to errors relating to and bias in datasets which results in discriminatory behaviour. One of the most common examples which have been cited relates to the AI system used as a part of Amazon's hiring process, which was found to be fraught with bias, due to discriminatory data sets. The AI preferred male candidates over female candidates, as the data sets used to train the model was based upon 10 years of recruitment data, which reflected the dominance of male candidates in the tech field.[100] While this project was shelved eventually, it is indicative of the major faults which lie in introducing AI which may rely on incorrect or flawed data sets.

Another case of algorithmic bias identified in the AI Stack Document was when Microsoft researchers found that the word-embedding algorithm had problematic biases, like associating "computer programmer" with male pronouns and "homemaker" with female ones. The AI Stack Document highlights that this debunks the myth of AI neutrality and sheds light on algorithmic bias, a phenomenon that can reach critical dimensions as algorithms become increasingly involved in each decision in AI. The AI Stack Document also underlines that this also increases the need for trustworthiness in use of AI systems.[101] Consequently, the AI Stack Document advises against the adoption of black box type solutions, which lack transparency and ethical values.

---

[99] Alex Campolo et al. 'AI Now 2017 Report' (2017). <https://assets.ctfassets.net/8wprhhvnpfc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf>.

[100] Jeffrey Dastin, 'Insight - Amazon scraps secret AI recruiting tool that showed bias against women' (Insight , 10 October 2018) <https://in.reuters.com/article/amazon-com-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idINKCN1MK0AH>.

[101] AI Stack Document n(13) 16.

Exclusion by AI Systems

Another systems consideration which relates to these principles is incorrect decision making by AI leading to exclusions in the provision of benefits. A common example provided for this is the deployment of AI in health insurance to check for fraud. An incorrect judgment in this scenario would lead to the exclusion from the benefits of health insurance.

In general, the efficacy of an AI system would depend upon the data that has been used for training the machines, and collecting and compiling data that is unbiased and relevant. An example of a possible violation of this principle could be in the deployment of AI in the PDS system (Public Distribution System) in India. The state of Telangana has an Aadhar-based Public Distribution System, and the state government has also started using AI and mainstream data analytics linked with ePOS and IRIS devices to make decisions on public distribution.[102] While the system is still in its preliminary stages, it has shown some potential. However, the distribution of an essential service such as the PDS through an AI has also raised concerns surrounding exclusion. These could include considerations such as caste and class bias,[103] incorrect classification of the poor as non-poor or even digital exclusion due to internet connectivity issues.[104]

The AI Stack Document recognizes the potential biases, which may be transferred by developers and programmers into an AI system.[105] They acknowledge that the data from which the AI learns can itself be flawed or biased and this may lead to flawed AI decisions. The AI Stack Document holds that this would be against the intention of algorithmic decision-making, which is "perhaps a good-faith attempt to remove unbridled discretion — and its inherent biases."[106] To overcome this challenge, it suggests that there is a need

---

[102] Akun Sabharwal 'Better PDS with data analytics' (Telangana Today, 1 March 2019) <https://telanganatoday.com/better-pds-with-data-analytics>.

[103] Ashit Kumar Srivastava 'AI, bias the Law' The Statesman (11 July 2019) <https://www.thestatesman.com/supplements/law/ai-bias-law-1502776178.html>.

[104] Shiv Sahay Singh, 'Death by digital exclusion? : on faulty public distribution system in Jharkhand' The Hindu (13 July 2019) <https://www.thehindu.com/news/national/other-states/death-by-digital-exclusion/article28414768.ece>.

[105] AI Stack Document n(13) 15.

[106] ibid 17.

to implement a system to ensure that the data is centrally controlled, using a single or multiple cloud controllers.[107]

This centrally controlled data model forms a part of the proposed Infrastructure layer of the AI Stack, which mandates the setting up of a common Data Controller for multiple cloud scenarios including both private and public information.[108] The Infrastructure layer would also contain the model for data collection by the controllers. It is unclear how the central control of data would reduce bias and how centralised controllers for both public and private data would increase the accuracy of the AI systems and reduce inconsistencies or bias in the data.

4.2.4. Checklist Model to Ensure Equality, Inclusivity and Non-Discrimination in AI Systems

A potential model which India could adopt in this regard would be the 'checklist' model. The European Network of Equality Bodies (EQUINET), in its recent report on 'Meeting the new challenges to equality and non-discrimination from increased digitisation and the use of Artificial Intelligence' provided a checklist to assess whether an AI system is complying with the principles of equality and non-discrimination.[109] Looking at the approach in this checklist would give us an idea as to how a similar checklist can be evolved in the Indian context.. The checklist consists of several broad categories, with a focus on the deployment of AI technology in Europe. This includes heads such as direct discrimination, indirect discrimination, transparency, other types of equity claims, data protection, liability issues, cross over jurisdictions and identification of the liable party.

The list contains a series of questions which judges whether an AI system meets standards of equality, and identifies any potential biases it may have. For example, the question "Does the artificial intelligence system treat people differently because of a protected characteristic?" includes both direct data and proxies. If the answer to the question was yes, the system would be executing an indirect bias. On the other hand,

---

[107] Ibid.

[108] ibid 22.

[109] EQUINET, 'Regulating for an Equal AI: A New Role for Equality Bodies' (2020) <https://equineteurope.org/wp-content/uploads/2020/06/ai_report_digital.pdf>.

other considerations have a graded scale upon which they are tested. For instance, on the issue of indirect discrimination, the checklist provides multiple questions in an 'if yes, then' format as follows:

- Does the artificial intelligence system consist of an algorithm and / or is it trained on a data set that places certain protected groups at a disadvantage?

  If the answer is yes, there is prima facie indirect discrimination.

- If so, can the body using the artificial intelligence system point to a legitimate aim to justify the use of the algorithm and / or data set?

  In this case, there would have to be an assessment on the extent to which there is a defence to prima facie indirect discrimination.

Such a checklist is one potential tool that could enable the execution of the principles of equality, inclusivity and non-discrimination.

4.3. Principle of Privacy

Privacy is an essential human right and can be found in several international law instruments such as the Universal Declaration of Human Rights,[110] and the International Covenant on Civil and Political Rights.[111] The right to privacy is supported by both regional and national laws and these laws highlight the importance of privacy for the development of an individual. Privacy is understood to be the right to be left alone, from unwanted intrusion or interference.[112] In the context of AI, the principle of privacy has been considered to include aspects of intimacy[113] and agency.[114] AI principles typically recognise privacy in AI systems in the context of national and international human rights

[110] Article 12, Universal Declaration of Human Rights, 1948.
[111] Article 17, International Covenant on Civil and Political Rights, 1976.
[112] Samuel D. Warren, Louis D. Brandeis, 'The Right to Privacy' (1890) 4(5) Harvard Law Review 193, <https://www.cs.cornell.edu/~shmat/courses/cs5436/warren-brandeis.pdf>.
[113] Montreal Declaration for Responsible AI n(66), "Principle 3 - Protection of Privacy and Intimacy - Privacy and intimacy must be protected from AIS intrusion and data acquisition and archiving systems".
[114] Institute of Electrical and Electronics Engineers n(68), " Principle 3 - Data Agency – A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity".

regimes.[115] In the context of AI systems, discussions around the principle of privacy usually relate to how these aspects inter-connect with informational privacy.

In the digital age, individuals share vast amounts of information, whether directly or indirectly, with corporations and governments. Robust data protection regimes which seek to protect the personal information of users, and supplement their right to privacy are more relevant today than ever before. AI models specifically, are heavily dependent on data and the introduction of AI relevant regulation into the existing data protection frameworks raises new challenges. In the past, there has been a higher degree of human oversight and control over digital technology, and the legal framework on information privacy has also been designed around this assumption, however, the increased use of AI means this may no longer be the case.[116] The application of AI to existing technologies stands to profoundly alter their current use and privacy considerations.

The diversity of AI systems means that the risk which they pose to the individuals, and society as a whole are also varied. Some privacy concerns which are applicable in this context include:

(i) Re-identification and De-Anonymisation: AI applications can be used to re-identify anonymised data. Datasets are often anonymised through a de-identification and sampling process before they are shared to address privacy concerns. However, current technology makes it possible for AI systems to reverse this process to re-identify people.[117] The AI Stack Document refers to the lack of regulations surrounding the anonymisation of data as one of the barriers of the adoption of AI systems.[118]

[115] The Toronto Declaration n(69), "Principle 23 - States must adhere to relevant national and international laws and regulations that codify and implement human rights obligations protecting against discrimination and other related rights harms, for example data protection and privacy laws".
[116] Office of the Victorian Information Commissioner, 'Artificial intelligence and privacy' (June 2018) <https://ovic.vic.gov.au/wp-content/uploads/2018/08/AI-Issues-Paper-V1.1.pdf>.
[117] Luc Rocher, J.M. Hendrickx & Y. de Montjoye, 'Estimating the success of re-identifications in incomplete datasets using generative models' Nature Com (2019) <https://www.nature.com/articles/s41467-019-10933-3#citeas>.
[118] AI Stack Document n(13) 15.

(ii) Data Exploitation: AI systems are complex and with the added issues surrounding transparency and explainability, people are often unable to fully understand the quantum and type of data which their devices, networks, and platforms generate, process, or share. With the introduction of the Internet of Things (IoT), more data is being generated about people than ever before. It is therefore increasingly necessary to enforce limits on data exploitation.[119]

Examples of Sector /Community Specific Privacy Concerns

AI systems are rapidly entering the field of healthcare where they serve major roles from automating routine tasks in medical practice to managing patients and medical resources. The AI Stack Document highlights some of the benefits of the adoption of AI systems in the healthcare sector.[120] As AI systems are created to handle these roles, new challenges emerge in relation to privacy concerns and the risk to patient privacy of individuals. In order to train AI systems for healthcare, developers would require large and diverse datasets from patients. This may lead to concerns about the violation of the privacy of patients, especially since with technology it is possible to re-identify anonymised data. Another privacy implication in AI systems may relate to revealing information which the patient themselves were unaware of. With predictive AI systems, it may be possible to predict and disclose medical conditions of patients, such as Parkinson's disease, before patients are even aware of it. Patients might consider this a violation of their privacy, especially if the AI system's inference were available to third parties, such as banks or life insurance companies.[121]

Use of AI in facial recognition systems allows individuals to be tracked and identified in public spaces by the government, and can facilitate heightened surveillance of individuals. The impact on an individual's privacy will have a broader chilling effect in society as a whole. Another field where privacy could be impacted by AI systems is online

[119] Privacy International, 'Artificial Intelligence' <https://privacyinternational.org/learn/artificial-intelligence>
[120] AI Stack Document n(13) 10.
[121] W. Nicholson Price, 'Risks and remedies for artificial intelligence in healthcare' (Brookings, 14 Nov 2019) <https://www.brookings.edu/research/risks-and-remedies-for-artificial-intelligence-in-health-care/>.

advertising.[122] With numerous scandals breaking out in the recent past, including Cambridge Analytica[123], there is greater awareness and concern about the role of profiling and targeted advertising using smart systems for political purposes and its detrimental impact on democracy. While legislative safeguards are being considered, such as the obligations of social media intermediaries with regards to actions that may impact democracy, public order or sovereignty and security of the State, it is also important to consider these matters while defining AI ethics.

4.3.1. Privacy and the Use of Data

The use of personal data by companies and governments alike can lead to specific information privacy harms. For several decades now data protection principles have evolved to address these harms as both technology and business models develop and new uses of personal data are found. Data protection principles and laws have traditionally focused on ensuring that the individual whose data is being collected and processed i.e. the data subject, is in control of such collection and processing of their data. With the evolution of technological models, and specifically AI systems which use such data at scale, there has been concern as well as considerable effort to find suitable data protection frameworks. In this section we address some of the principles and concepts that provide the foundation for data protection frameworks, and are most relevant in the context of AI systems. The AI Stack Document also recognizes the absence of a clear data protection framework in India, as one of the challenges in the adoption of AI systems. The document recommends the adoption of the standards in other data protection laws such as the EU's GDPR in the interim.[124]

In the following sections, we also briefly look at the various principles that have been incorporated in the proposed Personal Data Protection Bill, 2019 and their implications for AI systems.

[122] Privacy International, 'Adtech' <https://privacyinternational.org/learn/adtech>.
[123] Issie Lapowsky, "How Cambridge Analytica Sparked the Great Privacy Awakening" Wired (17 March 2019) <https://www.wired.com/story/cambridge-analytica-facebook-privacy-awakening/>.
[124] AI Stack Document n(13) 23.

(i) Notice and Consent:[125] Notice relates to providing clear, concise and comprehensible information regarding what data is collected, its purpose, details of third parties it is shared with, etc. Consent is premised on the provision of such notice, and is measured in the context of the individual's ability to provide meaningful and explicit consent.

In the context of AI specifically, the notice and consent principle becomes difficult to comply with when an organisation itself is not fully aware of details relating to data collection, its use, subsequent purposes of processing, or even the volume of data collected and derived[126]. Often where organisations are aware of such details, they may be unable to or unwilling to provide such information to the individual in simple, understandable formats. This problem is exemplified through AI applications such as smart traffic signals and other sensors needed to support self-driving cars —it would become impossible to provide accurate and unambiguous notice in these cases as it is difficult to explain the use of data by the other AI systems the data is being shared with.[127]

The AI Stack Document notes that until an Indian data protection law is put in place, the EU's General Data Protection Regulation (GDPR) standards can be applied.[128] It notes that "consent for use of data from customers will be taken through a properly framed consent framework".[129] The AI Stack Document focuses on the need to ensure proper ethical standards to maintain digital rights.[130]

The AI Stack Document also focuses on the need to ensure proper monitoring and data privacy of the data stored. In fact, it envisages that the Data/Information Exchange Layer will have to support an adequate consent framework for access of data by/for the customer.[131] Additionally, it is envisaged that the provision of consent can be for individual or collective data fields. The AI Stack Document proposes that there could be different

[125] Jessica Fjeld n(10) 22.
[126] Fred H. Cate & Rachel Dockery, 'Artificial Intelligence and Data Protection: Observations on a Growing Conflict' <https://ostromworkshop.indiana.edu/pdf/seriespapers/2019spr-colloq/cate-paper.pdf> 11.
[127] Cameron F. Kerry, 'Protecting privacy in an AI-driven world' (Brookings Institute, 10 Feb 2020) <https://www.brookings.edu/research/protecting-privacy-in-an-ai-driven-world/>.
[128] AI Stack Document n(13) 23.
[129] ibid 20.
[130] ibid 23.
[131] ibid 22-23.

tiers of consent made available to accommodate different tiers of permission.[132] More clarity is required from the DoT on the various tiers of consent and permission that are being referred to here.

The definition and standards applicable for the provision of notice and consent are important. Google's AI principles[133] merely define consent as permission in a standard manner, whereas the NITI Aayog's National Strategy on AI 2018[134] warns against unknown and uninformed consent. The Chinese White Paper on AI Standardisation[135] also highlights the need to redefine standard consent principles and consciously regulate AI given that AI systems could derive and process more data than was initially consented to by the data principal.

(ii) Control Over the Use of Data:[136] Emanating from the principle of consent and choice with respect to how one's data is processed, control over the use of data requires the data subject to have a degree of influence on how and why the information about them is used. As with notice and consent, it is important to identify the right standard for defining and ensuring that the individual has the control that is required and provided for under privacy and data protection laws. The understanding of meaningful control over the use of data also differs based on the set of principles put forward by an organisation– for example, Microsoft's AI principles simply state that data subjects should be provided with appropriate control over their data[137] whereas others such as the IBM AI principles state more explicitly that data subjects must always have control over what data is used and in what context.[138]

---

[132] ibid 23.
[133] Google, 'Artificial Intelligence at Google: Our Principles' <https://ai.google/principles>.
[134] National Strategy for AI n(26).
[135] China Electronics Standardisation Institute, 'Artificial Intelligence Standardisation White Paper' (2018) (English Translation) <https://cset.georgetown.edu/research/artificial-intelligence-standardization-white-paper/>.
[136] Jessica Fjeld n(10) 22.
[137] Microsoft AI principles n(90).
[138] IBM, 'Everyday Ethics for Artificial Intelligence: Five Areas of Ethical Focus' <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>.

Anonymisation of data, however, is seen as a way to exploit the economic opportunities of big data without violating control principles. The German AI strategy suggests the use of "pseudonymised and anonymised data" as potential tools to "help strike the right balance between protecting people's right to control their personal data and harnessing the economic potential of big-data applications."[139]

Another model for consideration is the creation of data trusts. Data trusts are flexible and global, and they can be designed in ways that create legally accountable governance structures. Data trusts steward, maintain and manage how data is used and shared, this includes who is allowed access, under what terms and how.[140] There have also been recommendations made for the adoption of data trusts to facilitate the sharing of data between organisations holding data and organisations looking to use data to develop AI systems.[141] The UK has advocated for the creation of data trusts which would allow individuals to make their views heard, however no concrete mechanism has been prescribed for the same.[142] The creation of data trusts may be done through some combination of consultative procedures, "personal data representatives," or other mechanisms.[143] However, the "black box" associated with the use of AI and the lack of complete knowledge about how the algorithm often works means that exercising control over the data and how the AI systems use it becomes increasingly difficult. The AI Stack Document discourages the use of black box algorithms, as they lack transparency and ethical values.[144]

---

[139] German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs, 'Artificial Intelligence Strategy' <https://www.ki-strategie-deutschland.de/home.html>.

[140] Bianca Wylie and Sean McDonald, 'What Is a Data Trust?' Centre for international governance innovation (9 Oct 2018) <https://www.cigionline.org/articles/what-data-trust>.

[141] Dame Wendy Hall and Jérôme Pesenti, 'Growing the Artificial Intelligence Industry in the UK' (2017) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/652097/Growing_the_artificial_intelligence_industry_in_the_UK.pdf>; BPE Solicitors, Queen Mary University of London and Pinset Masons, 'Data trusts: legal and governance considerations' (April, 2019) <https://theodi.org/wp-content/uploads/2019/04/General-legal-report-on-data-trust.pdf>.

[142] House of Lords, Select Committee on Artificial Intelligence, 'AI in the UK: ready, willing and able?' (2018) <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf> 35.

[143] ibid.

[144] AI Stack Document n(13) 17.

(iii) Ability to Restrict Processing: The ability to restrict the processing of personal data allows individuals to limit how an organisation is making use of their data. This is often provided for under the data protection regime of a country.[145] The extent to which individuals should be allowed to restrict processing differs in different AI documents, while the Montreal Declaration holds that people have a "right to digital disconnection" and encourages AI systems to "explicitly offer the option to disconnect at regular intervals, without encouraging people to stay connected",[146] the European High Level report on Ethics Guidelines for Trustworthy AI, by the European Commission, recommends allowing people to opt-out of citizen scoring AI systems.[147]

One of the most commonly recommended solutions for the use of large data sets is anonymising data to avoid regulatory hurdles.[148] However, studies have shown that it is possible to re-identify individuals despite anonymisation techniques using machine learning.[149] It is therefore important that privacy considerations also evolve with technological advances. Thus, even though requests to restrict processing of data may be harder to fulfil in cases where the data is part of a larger data set, clear mechanisms should be developed so that an organisation can do so. At the same time, we may need to recognise and provide for times where an organisation engaged in the usage of an AI system will not be able to fulfil a request to restrict processing of data because it cannot conclusively identify and isolate an individual's data from amongst the training data.

(iv) Right to Rectification and Erasure: Data protection frameworks typically provide for a right to rectification, whereby an individual can ask to have any data about them be rectified or updated with the aim of preventing any adverse consequence of inaccurate data being held about them.[150] The right to rectification may exist in the context of training

[145] Article 18, General Data Protection Regulation (EU).
[146] Montreal Declaration n(66).
[147] Ethics Guidelines for Trustworthy AI n(59) 12.
[148] Macy Bayern, 'DeepMind, NHS use anonymised patient data in AI to avoid regulatory hurdles' (TechRepublic, 2 July 2018) <https://www.techrepublic.com/article/deepmind-nhs-use-anonymized-patient-data-in-ai-to-avoid-regulatory-hurdles/>.
[149] Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye, 'Estimating the success of re-identifications in incomplete datasets using generative models' Nature Communications (2019) <https://www.nature.com/articles/s41467-019-10933-3>.
[150] Jessica Fjeld n(10) 24.

data in AI systems. However, training data is often used to train AI models by using large datasets which means that any inaccuracy in an individual's data will likely not have any direct impact on an individual data subject. It has been recommended that organisations should prioritise rectifying that data which holds the potential of having a direct effect on the individual data subject rather than that data whose accuracy is less likely to have an effect on the individual data subject.[151]

The right to erasure of data may also be exercised by individuals and an organisation could receive requests from individuals for erasure of data when it is no longer necessary, the information has been misused, or the relationship between the user and the entity is terminated.[152] However, sometimes datasets are retained to re-train, refine, or evaluate the AI system on an ongoing basis. Data protection frameworks that address AI systems may need to provide for a case by case determination of a companies obligations in such a situation.[153]

(v) Privacy by Design: Privacy by design ('PbD') is an approach that embeds privacy and security into the design and operation of a product and its network[154]. It seeks to make privacy the 'default setting' rather than a post-facto consideration. PbD is relevant to the development of AI systems.[155] In addition to this, certain forms of encryption can also be included in the process. Encryption serves a twofold function – it reduces the ability to identify individuals to a large extent and also acts as a precautionary measure against reducing harms of data leakage or breach.[156]

---

[151] Cynthia O'Donoghue and Daniel Millard, 'UK: ICO Blogs On AI And Data Subject Rights' (Mondaq, 13 Nov 2019) <https://www.mondaq.com/uk/data-protection/863632/ico-blogs-on-ai-and-data-subject-rights>.

[152] Access Now, 'Human Rights in the Age of AI' (2018) <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf> 31.

[153] 'ICO issues blog post on enabling access, erasure, and rectification rights in AI systems' (Tech law for everyone) <https://www.scl.org/news/10703-ico-issues-blog-post-on-enabling-access-erasure-and-rectification-rights-in-ai-systems>.

[154] Ann Cavoukian, 'Privacy by Design - The 7 Foundational Principles' (IAPP) <https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf>.

[155] Information Commissioner's Office, 'Enabling access, erasure, and rectification rights in AI systems' (15 Oct 2019) <https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/>.

[156] Maria Klyus, "Privacy By Design: GDPR in AI Technologies" (Legal IT group) <https://legalitgroup.com/en/privacy-by-design-en/>.

4.3.2. Privacy in AI and the Draft Personal Data Protection Bill 2019

From a regulatory perspective, the safe deployment of AI systems depends on ensuring that the privacy and personal data of individuals is protected. The Personal Data Protection Bill, 2019 ('PDP Bill 2019') is currently under consideration in Parliament. Some of the implications of the PDP Bill 2019 on the regulation of AI systems are discussed below:

i) Privacy by Design has been incorporated in the PDP Bill 2019 in clause 22[157] which requires data fiduciaries to incorporate the privacy by design principles described in the provision. While the clause deals with the application of Privacy by Design principles, this is limited to the seven principles described in the section. With the rapid advent in technology, it is recommended that the provision mandates incorporation of 'Privacy by Design' as a concept, in order to ensure that the law takes into account any evolution of these principles.

ii) Right to correction and erasure of data has been included in clause 18[158] of the PDP Bill 2019. The right is currently restricted to the four grounds listed under the provision.

[157] Personal Data Protection Bill, 2019 - Clause 22: Privacy by Design - (1) Every data fiduciary shall prepare a privacy by design policy, containing-
(a) the managerial, organisational, business practices and technical systems designed to anticipate, identify and avoid harm to the data principal;
(b) the obligations of data fiduciaries;
(c) the technology used in the processing of personal data is in accordance with commercially accepted or certified standards;
(d) the legitimate interests of businesses including any innovation is achieved without compromising privacy interests;
(e) the protection of privacy throughout processing from the point of collection to deletion of personal data;
(f) the processing of personal data in a transparent manner; and
(g) the interest of the data principle is accounted for at every stage of processing of personal data.

[158] Personal Data Protection Bill, 2019 - Clause 18, PDP BIll 2019: Right to correction and erasure - 1) The data principal shall where necessary, having regard to the purposes for which personal data is being processed, subject to such conditions and in such manner as
may be specified by regulations, have the right to-
(a) the correction of inaccurate or misleading personal data;
(b) the completion of incomplete personal data;
(c) the updating of personal data that is out-of-date; and
(d) the erasure of personal data which is no longer necessary for the purpose for which it was processed.

This right allows the data principal to have inaccurate or incomplete personal data corrected.

iii) The right to be forgotten is provided for under Clause 20 of the PDP Bill 2019. This right enables an individual (i.e. the data principal) to restrict or prevent the continuing disclosure of their personal data, if the purpose for such disclosure has been served, or if they withdraw consent. The jurisprudence surrounding the right to be forgotten is currently at a nascent stage, however, the inclusion and implementation of this provision is important to preserve the autonomy of an individual with regard to the processing or disclosure of their data by an AI system.

iv) PDP Bill's regulatory sandbox for AI: The PDP Bill 2019 empowers the Data Protection Authority (DPA) to create a regulatory sandbox to encourage innovation in artificial intelligence, machine learning or any other emerging technology in public interest.[159] The DPA can modify the application of user protection obligations on the data fiduciary such as purpose limitation and retention of personal data to a data fiduciary that qualifies for inclusion in the regulatory sandbox. This raises questions around the risk this exposes users to and highlights the need for us to think through the regulatory frameworks that commonly accompany sandboxes and the manner in which they should be used in the context of AI.

The PDP Bill 2019 has omitted several other rights that we see incorporated in other legislations such as the European Union's General Data Protection Regulation (GDPR). Many of these rights are integral to the rights of data principals in the digital age, e.g., the right to not be subject to a decision based solely on automated processing.[160] The right to object to a decision solely based on automated decision making is important not only to ensure that no discriminatory decisions are made about the data principal but also to allow data principals the right to know how decisions about them are being made. In situations where even the makers of these automated systems often do not understand

[159] Personal Data Protection Bill, 2019 - Clause 40(1).
[160] Article 22, General Data Protection Regulation (EU).

or are unable to explain the decision making processes of these AI systems, the rights of individuals are significantly impaired by the absence of this right.

Any regulation introduced to govern the processing of personal data must ensure that individuals are adequately protected in the context of automated decisions being made by AI systems. This will ensure that the law would continue to be relevant in a future which includes rapid adoption of automated processes. We recommend that the right to object to a decision solely based on automated decision making is provided for under an appropriate framework such as the data protection legislation in order to ensure that the data principal is able to exercise their autonomy in full.[161]

4.3.3. Use and Regulation of Non-Personal Data

Additionally, the government has recently released the Report by the Committee of Experts on Non-Personal Data Governance Framework.[162] One of the key objectives of the report is to enable the creation of datasets of non-personal data to facilitate the development of the AI industry in India. The report envisages the creation of such datasets from anonymised personal data to be made available to the government, private sector and communities. There are significant concerns around ensuring the continued anonymisation of anonymised personal data and the technical capacity currently available to re-identify anonymised personal data. Consequently, significant concerns may be raised around the violation of an individual's privacy. Leaving aside concerns around the current objectives and structuring of the proposed regulatory framework around NPD, it is imperative that significant consideration is given to the kind of data that is being used to build datasets for use by AI developers, access to these datasets and the regulations that need to put in place to ensure the privacy and security of this data.

Model for Consideration: Incident Investigation Report

[161] Centre for Communication Governance, 'Comments on the Draft Personal Data Protection Bill, 2018' <https://ccgdelhi.org/wp-content/uploads/2018/10/CCG-NLU-Comments-on-the-PDP-Bill-2018-along-with-Comments-to-the-Srikrishna-Whitepaper.pdf> 25.

[162] Ministry of Electronics and Information Technology, 'Report by the Committee of Experts on Non-Personal Data Governance Framework' <https://static.mygov.in/rest/s3fs-public/mygov_159453381955063671.pdf>.

Annexure 13 of the Convention on International Civil Aviation (Chicago Convention)[163] states that, in the event of an accident to an aircraft of a contracting nation occurring in another contracting nation, and involving either death, serious injury, or serious technical defect in the aircraft or air navigation facilities, the nation in which the accident occurs will institute an inquiry into the circumstances of the accident.

The aim of this accident investigation report is not to apportion blame or liability from an accident investigation, but rather to extensively study the cause of the accident to prevent future incidents. A similar incident investigation mechanism may be employed for AI incidents involving significant breaches of privacy or security, which may allow for reports on any specific failures in the large-scale deployment of AI systems to be shared between countries. With many countries now entering the AI race, the proposal of such a system would ensure that the countries can learn from other's experiences and better the deployment of AI systems. The adoption of such an investigative mechanism would allow for better security of AI systems as a whole. Incident report may help nations study the problems surrounding the deployment of AI systems, and prevent future incidents.

4.4. Principle of Transparency

The concept of transparency is a recognised prerequisite for the realisation of 'trustworthy AI'.[164] Transparency requires AI systems and technology to be designed and executed in a manner so as to allow for oversight with respect to converting their operations into intelligible outputs and expressing where, when and how they are being used.[165] This is a multifaceted concept and covers algorithmic transparency in terms of the nature of data inputs, influences, compliances, and outcomes and decision-making.[166] Transparency relates to several different aspects including the provision of meaningful information,[167]

---

[163] Annex 13 to the Convention on International Civil Aviation, 1944 <https://www.emsa.europa.eu/retro/Docs/marine_casualties/annex_13.pdf>.

[164] Ethics Guidelines for Trustworthy AI n(59) 18.

[165] Jessica Fjeld n(10) 5.

[166] Stefan Larsson and Fredrik Heintz, 'Transparency in artificial intelligence' (2020) Internet Policy Review 9(2).

[167] The Organisation for Economic Co-operation and Development (OECD), 'OECD Principles on AI' (2019) <http://www.oecd.org/going-digital/ai/principles/>, "1.3. Transparency and explainability - AI Actors should commit to transparency and responsible disclosure regarding AI systems. To this end, they should provide

discoverable systems,[168] information asymmetry,[169] explainability of AI[170] and democratic participation.[171]

Implementation of transparency principles would involve open source data and algorithms, notifications when interacting with an AI system and when such AI systems make decisions, regular reporting of outputs, right to information mechanisms, and open procurement for the government.[172] It also necessitates transparency in terms of understanding the internal growth of AI and machine learning tools, and regulation of any information asymmetry in the growth of technology.[173]

The goal of transparency in ethical AI is to make sure that the functioning of the AI system and resultant outcomes are non-discriminatory, fair and bias mitigating, and that the AI system inspires public confidence in the delivery of safe and reliable AI innovation and development.[174] Additionally, transparency is also important in ensuring better adoption of AI technology, the more that users feel they understand the overall AI system, the more

---

meaningful information, appropriate to the context, and consistent with the state of art: (i) to foster a general understanding of AI systems; (ii) to make stakeholders aware of their interactions with AI systems, including in the workplace; (iii) to enable those affected by an AI system to understand the outcome; and, (iv) to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision".

[168] Ethically Aligned Design: Version 2 n(68), "Principle 5 - The basis of a particular A/IS decision should always be discoverable".

[169] Ethics Guidelines for Trustworthy AI n(59) "10. Transparency - Transparency concerns the reduction of information asymmetry. Explainability – as a form of transparency – entails the capability to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environments, as well as the provenance and dynamics of the data that is used and created by the system. Being explicit and open about choices and decisions concerning data sources, development processes, and stakeholders should be required from all models that use human data or affect human beings or can have other morally significant impact".

[170] Artificial Intelligence Strategy n(139), "We will promote research regarding explainability and accountability of algorithm-based forecasting and decision-making systems" <https://ec.europa.eu/knowledge4policy/publication/germany-artificial-intelligence-strategy_en>

[171] Montreal Declaration for Responsible AI n(66), "Principle 5 – Democratic Participation Principle - In accordance with the transparency requirement for public decisions, the code for decision-making algorithms used by public authorities must be accessible to all, with the exception of algorithms that present a high risk of serious danger if misused".

[172] Jessica Fjeld n(10) 44.

[173] Keng Siau and Weiyu Wang, 'Artificial Intelligence Ethics: Ethics of AI and Ethical AI' (2020) Journal of Database Management 31(2) 73, 80.

[174] D. Leslie n(58).

inclined and better equipped they are to use it. The AI Stack Document holds that AI algorithms should consequently be open, and should have an open algorithm framework and clearly defined data structures.[175]

Presently, AI algorithms are black boxes where automated decisions are taken based on machine learning over training data. There is little understanding of how these decisions are taken by AI systems. This results in a lack of transparency. The Document deals with the black box paradox which plagues AI systems. The AI Stack Document explains that as AI systems become more intelligent, they become more effective at its tasks of prediction and decision making, but conversely its processes also become less transparent to humans.[176] According to the Document this "opacity" problem leads to a lack of control and supervision by controllers and users of AI, and ultimately risks the progress of the AI system.[177] The AI Stack Document therefore suggests the adoption of unbiased open architecture at the Application layer to ensure transparency and openness.[178]

The Institute of Electrical and Electronics Engineers has suggested that different stakeholder groups require varying levels of transparency in respect of decisions made by an AI system,[179] and the developers and creators of an AI system must provide the requisite level of details in accordance with the target group. This has also been endorsed by the Australian AI Ethics Principles.[180] There are five categories of stakeholders who have been identified:

(i) Users: For the users of an AI system, it is particularly important to focus broadly upon what the AI system is doing and why. This would require a non-technical explanation which explains the purpose of the system and the steps taken by it to achieve this purpose.

[175] AI Stack Document n(13) 16.
[176] ibid 17.
[177] ibid.
[178] ibid 18.
[179] Ethically Aligned Design: Version 2 n(68) 29.
[180] Australian AI ethic principles n(55).

(ii) Validators: This category refers to the experts who will be undertaking the validation and certification of AI systems. In terms of transparency, the validators would have to be given an explanation of the systems' processes, the input data being used and the technical design choices.

(iii) Incident Investigator: It is also important to set standards of transparency which would apply to incident investigators, who would be required to check AI systems in case of a harmful outcome from the use of an AI system. This would require a two-fold mechanism for transparency, (i) setting up transparency standards in respect of the information to be disclosed to the incident investigator by the creator to allow for an effective investigation, and (ii) standards of transparency to be followed by the investigator for the duration of the investigation and reporting of their findings. A potential solution would be to set up an entire framework for incident investigations, drawing from the Chicago Convention,[181] which lays down the procedure for accident investigations in an aviation incident. The accident investigation model has been suggested in section 4.3.4 of this document.

(iv) For those in the Legal Process: A certain level of transparency would also be required to inform evidence and decision-making, in the legal process. It is important to note here that the legal process may include multiple stakeholders with differing levels of familiarity with digital technology and particularly AI systems. This not only includes lawyers in adjudication, negotiation and court process but would also include judges.

(v) For the Public: Finally, one of the most important stakeholder groups to be considered in the general public. It is important to maintain a standard of transparency in order to build confidence in the general public in AI technology.

There have been some concerns about the mechanisms for enforcement of transparency in AI systems. The NITI Aayog's Working Document Towards Responsible #AIforAll holds that disclosing the algorithm is not a solution and instead, we should aim towards explaining how the decisions are taken by AI systems.[182] On the contrary the AI Stack

[181] International Civil Aviation Organisation, 1944.
[182] National Strategy for AI n(26) 86.

Document holds that the main effect of opening existing AI through open sourcing code and placing related intellectual property into the public domain, would be to hasten the diffusion and application of current state of the art techniques.[183] The document considers software and knowledge about algorithms to be non-rival goods.[184] The DoT holds that making the algorithms freely available would enable more people to use them, at low marginal cost.[185]

4.4.1. Transparency Challenges in AI

As highlighted above, one of the major challenges with transparency in AI is that the technology operates in what is commonly referred to as a black box. AI systems often use extremely complicated algorithms which can only be understood by other computers. When conventional AI systems produce a decision, human end users don't know how it arrived at its conclusions.[186] This brings us to two major transparency problems, the first deals with public perception and understanding of how AI works, which may be addressed through increased transparency in different phases. Dr. David Leslie of the Alan Turing Institute suggests the following steps towards ensuring transparency in an AI process:[187]

(i) Justify Process: This process relates to explaining the mechanism of the AI. At this stage, the affected stakeholders would have to be informed of the safety, trustworthiness, non-discrimination in an AI system. This part would also have to be supplemented with providing explanations on the existence of a reasonable audit process to ensure that the process employed by the AI process is mindful of the ethical principles of AI.

(ii) Clarify Content and Explain Outcome: This involves providing an explanation of why an AI system performed the way it did in a specific decision-making or behavioural

[183] AI Stack Document n(13) 24.
[184] ibid 24.
[185] ibid.
[186] AJ Abdallat, 'Explainable AI: Why We Need To Open The Black Box' Forbes (22 Feb 2019) <https://www.forbes.com/sites/forbestechcouncil/2019/02/22/explainable-ai-why-we-need-to-open-the-black-box/#322299d11717>.
[187] D. Leslie n(58).

context. This would involve using non-technical knowledge to explain the rationale behind the decision or behaviour of an AI system.

(iii) Justify Outcome: This involves justifying the outcome produced by an AI system in the given context of AI values and ethics. The outcome of AI systems must also be fair, non-discriminatory and uphold AI principles.

The second major transparency problem deals with how much developers actually understand about their own AI. In many cases, developers may not know, or be able to explain how an AI system makes logical conclusions or how it has arrived at certain solutions.[188] It's a black box system, where trust is based on the accuracy and predictability of the system. Thus, while it is important to aim towards transparency in AI processes, it is also important to recognise that a 100% transparent AI system may not be possible.

Thus, in order to comprehensively address issues surrounding transparency it is important to develop continuous processes, which may analyse and provide real time solutions to issues as and when they arise. . Powers may be left for regulators to introduce transparency measures as necessary, while broadly including an overarching principle of transparency in AI systems.

4.4.2. Adoption of Model Cards

While many countries and organisations are researching different techniques which may be useful in increasing the transparency of an AI system in each of these steps, one of the common suggestions which has gained traction in the last few years is the introduction of labelling mechanisms in AI systems. An example of this is Google's proposal to use 'Model Cards'[189].

Model cards are short documents which accompany a trained machine learning model, carrying the benchmarked evaluation in a variety of conditions, including across different

[188] Will Knight, 'The Dark Secret at the Heart of AI' MIT Technology Review (11 April 2017) <https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/>.
[189] Google, 'Model Cards' <https://modelcards.withgoogle.com/about>.

cultural, demographic, and intersectional groups which may be relevant to the intended application of the AI system.[190] These model cards are intended to clarify the scope of the AI systems deployment and minimise their usage in contexts for which they may not be well suited. The model cards would also be accompanied with comprehensive documentation which details the full performance characteristics of an AI system. They also seek to inform the users of the contexts in which the AI systems may be used.[191] Adopting model cards and other similar labelling requirements in the Indian context may go a long way in introducing transparency into AI systems.

The regulations governing the use of AI should specify transparency requirements in relation to the deployment and use of the system. For instance, clear information should be available on the AI system's capabilities, the intended purpose for which it is being deployed, conditions under which it has been designed to function, expected accuracy and limitations. Additionally, there should be disclosure requirements and individuals should be notified when they are interacting with an AI system.

4.5. Principle of Accountability

The Principle of Accountability implies that the different stages and actors of an AI system should be identifiable and accountable for its outcomes.[192] It aims to recognise the responsibility of different organisations and individuals that develop and deploy AI systems.[193] Accountability is about responsibility, answerability as well as trust. There is

[190] Margaret Mitchell et al. 'Model Cards for Model Reporting' (2019) <https://arxiv.org/pdf/1810.03993.pdf>.
[191] ibid.
[192] G20 AI Principles n(58), "AI actors should be accountable for the proper functioning of AI systems and for the respect of the above principles, based on their roles, the context, and consistent with the state of art".
[193] Montreal Declaration for Responsible AI n(66), "Principle 9 - Responsibility Principle - (1) Only human beings can be held responsible for decisions stemming from recommendations made by AIS, and the actions that proceed therefrom. (2) In all areas where a decision that affects a person's life, quality of life, or reputation must be made, where time and circumstance permit, the final decision must be taken by a human being and that decision should be free and informed. (3) The decision to kill must always be made by human beings, and responsibility for this decision must not be transferred to an AIS. (4) People who authorise AIS to commit a crime or an offense, or demonstrate negligence by allowing AIS to commit them, are responsible for this crime or offense. (5) When damage or harm has been inflicted by an AIS, and the AIS is proven to be reliable and to have been used as intended, it is not reasonable to place blame on the people involved in its development or use".

no one standard form of accountability, but rather this is dependent upon the context of the AI and the circumstances of its deployment.[194] Accountability of AI designs comes into play at three stages in the design of AI ethics- the features of AI software, the use of the AI software and finally, the broader socio-technological system that has deployed AI software.[195] This means that accountability has to be ensured (i) pre-deployment, (ii) during deployment and (iii) harm at the post-deployment stage.[196]

Holding someone accountable has great ramifications as the AI systems generally involve multiple parties and have wide-spread impacts. The adverse impacts caused by AI software also go beyond the existing regimes of tort law, privacy law or consumer protection law.[197] Some amount of accountability can be achieved by enabling greater human oversight. It would also require the use of AI software to be made more transparent.[198] In order to foster trust in AI as well as to correctly determine the party who is accountable, it is necessary to build a set of shared principles that clarify responsibilities of each stakeholder including the developers, service providers and end users in the research, development and implementation of AI.[199]

During the lifecycle of AI systems, many decisions have an impact on users and others interacting with the AI systems. The deployment environment also affects a self-learning AI. Assigning accountability for specific decisions becomes difficult in a scenario with

[194] Ethics Guidelines for Trustworthy AI n(59), "Accountability - Good AI governance should include accountability mechanisms, which could be very diverse in choice depending on the goals. Mechanisms can range from monetary compensation (no-fault insurance) to fault finding, to reconciliation without monetary compensations. The choice of accountability mechanisms may also depend on the nature and weight of the activity, as well as the level of autonomy at play. An instance in which a system misreads a medicine claim and wrongly decides not to reimburse may be compensated for with money. In a case of discrimination, however, an explanation and apology might be at least as important".

[195] G7 Multistakeholder Conference on Artificial Intelligence "Accountability in AI Promoting Greater Societal Trust" (6 Dec 2018, Montreal) <https://www.ic.gc.ca/eic/site/133.nsf/vwapj/3_Discussion_Paper_-_Accountability_in_AI_EN.pdf/$FILE/3_Discussion_Paper_-_Accountability_in_AI_EN.pdf>.

[196] Jessica Fjeld n(10) 28.

[197] Mission assigned by the French Prime Minister, 'For a Meaningful Artificial Intelligence: Toward a French and European Strategy' (2018) <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf>.

[198] Samuele Lo Piano, 'Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward' (2020) Humanities and Social Sciences Communications <https://www.nature.com/articles/s41599-020-0501-9>.

[199] ibid.

multiple players. In the absence of any consequences for decisions harming others, no one party would feel obligated to take responsibility or action.[200] Additionally, the lack of accountability also affects the process of grievance redressal which can be used to address such issues.[201]

One of the ways to improve accountability in AI systems is through the adoption of an overarching framework. One such framework could be the security and governance layer envisioned in the AI Stack Document. As suggested in Section 3.3 and 3.4 of this document, it may be beneficial to introduce an overarching legislation or framework which can translate AI systems into the current legal framework including how product liability and consumer protection apply to AI. For example, this can be implemented by including distinct definitions for actors in the AI deployment system such as creator, developers or consumers of an AI system. Additionally, addressing the classification of AI systems as products or services will potentially facilitate the application of the current consumer protection laws. Such clarifications will allow AI systems to be regulated under the broader Indian legal framework.

Accountability and the associated regulatory obligations will have to be placed on the various stakeholders involved such as the developer and the deployer of the AI system (this is distinct from liability obligations). Regulations will have to be designed to place the obligation on the stakeholder best positioned to address the risk and mitigate its impact.

4.5.1. Pre-Deployment

The AI governance framework should seek to implement an audit process as has been envisioned in the AI Stack Document. The Document envisions “audit logging for promotion of accountability, reconstruction of events, security and forensics applications”[202] and suggests that algorithmic auditing will be a mechanism to ensure that AI developers and coders are adhering to regulatory standards and principles for AI[203]. A potential mechanism for implementing this could be a multi stage audit process which is

[200] NITI Aayog, Towards responsible AI for All n(1) 12.
[201] ibid.
[202] AI Stack Document n(13) 35.
[203] ibid.

undertaken post design, but before the deployment of the AI system.[204] The SMACTR model was first proposed in the Conference on Fairness, Accountability, and Transparency in January 2020 at Barcelona. The model has also been adopted by Google in its audit framework to close the AI accountability gap.[205]

Depending on the nature of the AI system and the potential for risk, regulatory guidelines can be developed prescribing the various categories of audits. While there can be no one size fits all model which will be applicable here, it is advisable to follow a scaled approach, where AI deployment in riskier or crucial systems is held up to more stringent requirements to ensure accountability. There can be three basic types of audits, (i) internal audit, (ii) external audit, and (iii) audit by regulatory bodies. The suggested multi-stage audit process envisages:

(i) The Scoping Stage:[206] In this stage, the risk of the AI system is assessed and auditors would produce assessments of social impact and an ethical review of the system. The goal of the scoping stage is to clarify the objective of the audit by reviewing the motivations and intended impact of the system. At this stage, the audit is implemented by mapping out the intended use cases and identifying analogous deployments of similar AI systems.

(ii) The Mapping Stage:[207] This stage refers to the creation of a map of stakeholders, and identifying key collaborators who would be necessary for the execution of the audit. For example, if the system was mostly meant to be effected within a company, the stakeholders and collaborators would mostly be employees within the company, however, if this was an AI system being deployed by the government, it would involve various government departments in the process. Creating a map would also help in creating a

[204] Inioluwa Deborah Raji et al., 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing' (3 January 2020) <https://arxiv.org/pdf/2001.00973.pdf>.
[205] Khari Johnson, 'Google researchers release audit framework to close AI accountability gap' (Venture Beat, 30 Jan 2020) <https://venturebeat.com/2020/01/30/google-researchers-release-audit-framework-to-close-ai-accountability-gap/>; Jessica Newman, 'CLTC white paper series: Decision Points in AI Governance' UC Berkeley, Centre for Long Term Cybersecurity <https://cltc.berkeley.edu/wp-content/uploads/2020/05/Decision_Points_AI_Governance.pdf>.
[206] Inioluwa Deborah Raji n(204) 7.
[207] Khari Johnson n(205).

record of individual accountability with respect to participation towards the final outcome of making the AI system.

(iii) The Artifact Collection Stage:[208] At this stage, the data on the AI system would be collated, this includes the creation of audit checklists, and datasheets on the creation and development of the model, assumptions and conditions made during the development process and its intended use. This often includes the collection of documents from across teams and developers. At this stage of the audit process, it would become easier to recognise missing documentation or permits and collate the same.

(iv) The Testing Stage:[209] This stage uses technology and other mechanisms to assess the performance of the AI system using methods like adversarial training. Adversarial testing pushes at the boundaries of the AI system by simulating the behaviour of a hostile or bad actor who may have gained access to the system. The process also includes stress testing wherein the system may be presented with extreme data sets which may be very improbable but may carry a higher rate of failure. This stage aims at risk analysis of the system highlighting the likelihood and severity of failure or other risks. This risk analysis would help in identifying high, mid and low risk tasks, depending upon the degree of failure, and the risk associated with the task.

(v) The Reflection stage:[210] This is the final stage of the process where the internal development team of the organisation or the above mentioned external expert or regulatory auditors can evaluate their internal design recommendations and create mitigation plans for the risks identified in the process. This could include more diversity in training data sets, or excluding certain proxies from the evaluation process.

4.5.2. During Deployment

Once the AI system has been deployed, it is important to keep up the process of audit by continuously noting the changes being made in the AI system as it is deployed In all real-world deployments of AI solutions, there will be a subset of the input data which cannot

[208] Inioluwa Deborah Raji n(204) 8.
[209] ibid 9.
[210] ibid.

be predicted or prepared for and may not have a mitigation plan in place. In such a case, it is important that the development team is continuously monitoring the system to capture these errors and address them promptly.

4.5.3. Post Deployment Harms

Finally, post the deployment process, there is a need to provide for grievance redressal mechanisms,[211] to mitigate and correct any irregularities in the deployment process. This would require timely, accurate, and complete assessment by the independent oversight bodies for the purpose of redressing the accountability concerns around AI systems. The difficulty in redressing grievances particularly arises in automated decisions. There is an accountability gap in prescribing human answerability to decisions assisted or produced by an AI system.[212] The complex and multi-agent character of AI poses a complex challenge in answering the question of who among the parties involved in the system should bear responsibility if these systems produce negative consequences.

The Council of Europe, in its guidelines on the human rights impacts of algorithmic systems, has highlighted the need for effective remedies in order to address human rights concerns caused by the deployment of AI systems.[213] A potential model for grievance redressal would be the redressal mechanism suggested in the Atomium - European Institute for Science, Media and Democracy report.[214] A grievance redressal mechanism

[211] Committee of Ministers, 'Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems' (April 2020) <https://search.coe.int/cm/pages/result_details.aspx?objectid=09000016809e1154>.

[212] D. Leslie n(58) 23.

[213] Recommendation CM/Rec(2020) n(211), "4.5 Effective remedies: States should ensure equal, accessible, affordable, independent and effective judicial and non-judicial procedures that guarantee an impartial review, in compliance with Articles 6, 13 and 14 of the Convention, of all claims of violations of Convention rights through the use of algorithmic systems, whether stemming from public or private sector actors. Through their legislative frameworks, States should ensure that individuals and groups are provided with access to effective, prompt, transparent and functional and effective remedies with respect to their grievances. Judicial redress should remain available and accessible, when internal and alternative dispute settlement mechanisms prove insufficient or when either of the affected parties opts for judicial review or appeal".

[214] Atomium - European Institute for Science, Media and Democracy, 'AI for people's ethical framework for a good AI in society: Opportunities, risks principles and recommendations' (2019) <https://www.eismd.eu/wp-content/uploads/2019/02/Ethical-Framework-for-a-Good-AI-Society.pdf>.

for AI would have to be widely accessible and include redress for harms inflicted, costs incurred, and other grievances caused by the AI system. It must demarcate a clear system of accountability between organisations and individuals. According to the model suggested by Atomium, redressal can be implemented in two ways:[215]

(i) AI Ombudsperson: This would ensure the auditing of allegedly unfair or inequitable uses of AI reported by users of the public at large through a streamlined judicial process.

(ii) Guided Process for Registering a Complaint: This envisions laying down a simple process, similar to filing a Right to Information request, which can be used to bring to the notice of the authorities, discrepancies, or faults in an AI system.

Similar redressal mechanisms to address the human rights concerns raised specifically by AI systems would have to be designed and implemented in India.

[215] ibid.

APPENDIX: AI PRINCIPLES

This appendix contains a compilation of the various AI principle documents proposed by countries across the world, multinational companies, and civil society organisations and international organisations.

COUNTRIES/ MULTILATERAL BODIES

Asia

- China Electronics Standardisation Institute, “Artificial Intelligence Standardisation White Paper” (2018) (English Translation) <https://cset.georgetown.edu/research/artificial-intelligence-standardization-white-paper/>
- Monetary Authority of Singapore, “Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector” (2018) <http://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf>
- Government of India, NITI Aayog, “National Strategy for Artificial Intelligence: #AI for All (Discussion Paper)” (2018) <https://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf>
- National Governance Committee for the New Generation Artificial Intelligence, China, “Governance Principles for the New Generation AI- Developing Responsible AI” (2019) <http://chinainnovationfunding.eu/dt_testimonials/publication-of-the-new-generation-ai-governance-principles-developing-responsible-ai/>
- Smart Dubai, “Dubai’s AI Principles” (2019) <https://www.smartdubai.ae/initiatives/ai-principles>

- Government of Japan, Cabinet Office, Council for Science, Technology and Innovation, “Social Principles of Human-Centric AI” (2019) <https://ai.bsa.org/wp-content/uploads/2019/09/humancentricai.pdf>
- Personal Data Protection Commission (PDPC), Singapore, “Model Artificial Intelligence Governance Framework” (2020) <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf>

Europe

- House of Lords, Select Committee on Artificial Intelligence, “AI in the UK: ready, willing and able?” (2018) <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>
- European Group on Ethics in Science and New Technologies, European Commission, “Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems” (2018) <http://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf>
- Council of Europe, European Commission for the Efficiency of Justice, “European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment” (2018) <https://rm.coe.int/ethical-charter-en-for-publication-4-december2018/16808f699c>
- Mission assigned by the French Prime Minister, “For a Meaningful Artificial Intelligence: Toward a French and European Strategy” (2018) <https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf>
- Office of the President of the Russian Federation, Decree of the President of the Russian Federation on the Development of Artificial Intelligence in the Russian Federation, “National Strategy for the Development of Artificial Intelligence Over the Period Extending up to the Year 2030” (2019) <https://cset.georgetown.edu/wp-content/uploads/Decree-of-the-President-of-the-Russian-Federation-on-the-Development-of-Artificial-Intelligence-in-the-Russian-Federation-.pdf>

- High Level Expert Group on Artificial Intelligence set up by the European Commission, “Ethics Guidelines for Trustworthy AI” (2019) <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- European Commission, “Building Trust in Human-Centric Artificial Intelligence”, (2019) <https://ec.europa.eu/transparency/regdoc/rep/1/2019/EN/COM-2019-168-F1-EN-MAIN-PART-1.PDF>
- European Commission, “Liability for Artificial Intelligence and other emerging digital technologies” (2019) <https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608>
- European Commission, “White Paper: On Artificial Intelligence - A European approach to excellence and trust” (2020) <https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf>
- Data Ethics Commission, Germany “Opinion of the Data Ethics Commission” (2020) <https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3>
- German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labour and Social Affairs, “Artificial Intelligence Strategy” <https://www.ki-strategie-deutschland.de/home.html>

North America

- Executive Office of the President, National Science and Technology Council Committee on Technology, United States, “Preparing for the Future of Artificial Intelligence” (2016) <https://publicintelligence.net/white-house-preparing-artificial-intelligence/>

- University of Montreal, Canada, “Montreal Declaration for Responsible AI” (2018) <https://www.montrealdeclaration-responsibleai.com/the-declaration>
- The White House Office of Science and Technology Policy (OSTP), United States, “Memorandum for the heads of executive departments and agencies” (2019) <https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf>
- Defense Innovation Board (DIB), Department of Defense (DoD), United States, “AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense” (2019) <https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF>
- Government of Canada, “Responsible use of artificial intelligence (AI)” <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html#toc1>

South America

- C Minds, “Towards an AI Strategy in Mexico: Harnessing the AI Revolution” (2018) <https://www.cminds.co/copy-of-ai>
- IA-Latam, “Declaración de Principios Éticos Para La IA de Latinoamérica” <https://ia-latam.com/etica-ia-latam/>

Africa

- University of Pretoria, “Artificial Intelligence for Africa: An Opportunity for Growth, Development, and Democratisation” <https://www.up.ac.za/media/shared/7/ZP_Files/ai-for-africa.zp165664.pdf>

Australia

- Department of Industry, Innovation and Science, Australian Government, “AI Ethic Principles” (2019) <https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles>

Global

- G20 Ministerial Meeting on Trade and Digital Economy, “G20 AI Principles” (2019) <https://www.g20-insights.org/wp-content/uploads/2019/07/G20-Japan-AI-Principles.pdf>
- The Organisation for Economic Co-operation and Development (OECD), “OECD Principles on AI” (2019) <http://www.oecd.org/going-digital/ai/principles/>

CIVIL SOCIETY, INTERNATIONAL AND ACADEMIC ORGANISATIONS

- Future of Life Institute, “Asilomar AI Principles” (2017) <https://futureoflife.org/ai-principles/?cn-reloaded=1>
- The Future Society, Law & Society Initiative, “Principles for the Governance of AI” (2017) <https://thefuturesociety.org/2017/07/15/principles-law-and-society-initiative/#:~:text=Principles%20for%20the%20Governance%20of%20AI,-Principle%201%3A%20AI&text=Principle%202%3A%20AI%20shall%20be,of%20AI%20shall%20be%20accountable.>
- Internet Society, “Artificial Intelligence and Machine Learning: Policy Paper” (2017) <https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-AI-Policy-Paper_2017-04-27_0.pdf>
- Amnesty International and Access Now, “The Toronto Declaration” (2018) <https://www.accessnow.org/cms/assets/uploads/2018/08/The-Toronto-Declaration_ENG_08-2018.pdf>
- Public Voice Coalition, “AI Universal Guidelines” (2018) <https://thepublicvoice.org/ai-universal-guidelines/>
- Access Now, “Human Rights in the Age of AI” (2018) <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf>
- Alan Turing Institute, “Understanding Artificial Intelligence Ethics and Safety” (2019) <https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf>
- EQUINET, “Regulating for an Equal AI: A New Role for Equality Bodies” (2020) <https://equineteurope.org/wp-content/uploads/2020/06/ai_report_digital.pdf>

- UNICEF, “Artificial Intelligence and Children’s Rights” (2020) <https://www.unicef.org/innovation/media/10726/file/Executive%20Summary:%20Memorandum%20on%20Artificial%20Intelligence%20and%20Child%20Rights.pdf>
- Institute of Electrical and Electronics Engineers, “Ethically Aligned Design: Version 2” <https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf>
- UNI Global Union, “Top 10 Principles for Ethical Artificial Intelligence” <http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf>

CORPORATIONS

- Intel, “AI Public Policy white paper” (2017) <https://blogs.intel.com/policy/files/2017/10/Intel-Artificial-Intelligence-Public-Policy-White-Paper-2017.pdf>
- OpenAI, “OpenAI Charter” (2018) <https://openai.com/charter/>
- Artificial Intelligence Industry Alliance (AIIA), China, “Draft Joint Pledge on Artificial Intelligence Industry Self-Discipline” (2019) <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-ai-alliance-drafts-self-discipline-joint-pledge/?from=timeline&isappinstalled=0>
- Telia Company, “Telia Company Guiding Principles on trusted AI ethics” (2019) <https://www.teliacompany.com/globalassets/telia-company/documents/about-telia-company/public-policy/2018/guiding-principles-on-trusted-ai-ethics.pdf>
- Google, “Artificial Intelligence at Google: Our Principles” <https://ai.google/principles>
- Microsoft, “Microsoft AI principles” <https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6>
- Deutsche Telekom, “Guidelines for Artificial Intelligence” <https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366>

- Sony, “Sony Group AI Ethics Guidelines” <https://www.sony.net/SonyInfo/csr_report/humanrights/hkrfmg0000007rtj-att/AI_Engagement_within_Sony_Group.pdf>
- Telefónica, “AI Principles of Telefónica” <https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles>
- DeepMind, “DeepMind Ethics & Society Principles” <https://deepmind.com/about/ethics-and-society>
- Samsung, “Principles for AI Ethics” <https://research.samsung.com/artificial-intelligence>
- IBM, “Everyday Ethics for Artificial Intelligence: Five Areas of Ethical Focus” <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>
- Philips, “Philips AI Principles” <https://www.philips.com/a-w/about/artificial-intelligence/philips-ai-principles.html>
- Accenture, “Responsible AI and Robotics: An Ethical Framework” <https://www.accenture.com/gb-en/company-responsible-ai-robotics>