ALGORITHMIC DECISION-MAKING IN PAKISTAN

A CHALLENGE TO RIGHT TO EQUALITY & NON-DISCRIMINATION

This is an independent report prepared by the Institute for Responsible Artificial Intelligence & Human Rights, a new initiative of the Centre for Human Rights.

About the Centre for Human Rights

The Centre for Human Rights, housed in Universal College Lahore, is an independent legal research institute that actively researches issues of human rights, and works on legal policy, due process, rule of law and criminal justice reforms in Pakistan. The Centre aims to provide legal analysis which are rights-based, human centric and aimed at enhancing the constitutional freedoms of equality and non-discrimination.

Author

Uzma Nazir Chaudhry

Cover Page By

Adina Azmat Khan

Design Layout By

Sevim Saadat

Published by

Centre for Human Rights
University College Lahore
1.5 km from Thokar Niaz Baig,
Raiwind Road,
Lahore, Pakistan

For more information, please contact uzma.nazir@cfhr.com.pk

Acknowledgements

The Centre for Human Rights greatly appreciates the assistance and support of all the institutions and individuals who have helped make this report possible. In particular, we would like to thank the administration of Universal College Lahore, Dr. Sayyed Asad Hussain and Ms. Urmana Chaudhry for their constant support to the Centre for Human Rights.

We would like to express our sincere gratitude to Umar Mahmood Khan, the Executive Director of the Centre for Human Rights, for his immeasurable commitment to making and supporting this institute for the study and promotion of human rights. A special thanks is extended towards Sevim Saadat, Co-Founder of the Centre for Human Rights, and Fatima Mehmood, Research Fellow at the Centre for Human Rights, for the persistent and generous research assistance, analysis, and editorial support they rendered throughout the duration of this project. Their expert guidance and support has been instrumental in the realisation of this report.

We also extend our gratitude to all the industry experts who contributed to the report through interviews. In particular, the author notes the contributions of the technology experts at Deep Learning Lab, Medical Imaging & Diagnostics Lab, and Intelligent Criminology Lab at the National Centre of Artificial Intelligence, which is the leading hub of innovation and scientific research in Artificial Intelligence & Robotics in Pakistan. A special thanks is extended towards the team of experts at MeVitae, a U.K. based company that uses technology to debias the recruitment process. We appreciate not only their expert contribution through interviews, but also the detailed conversations they had with the author on other occasions about the technical risks and opportunities of Artificial Intelligence. Finally, the author also thanks Cecil Abungu, a legal scholar of Longtermism, AI Risks & Law, based in Kenya, for enhancing the author's understanding of legal problems associated with Artificial Intelligence which are surfacing across the globe, and in developing countries in particular.

Table of Contents

List of Acronyms

AI - Artificial Intelligence

AIS - Artificial Intelligence System

CEDAW - Convention on the Elimination of Discrimination against Women

CJS - Criminal Justice System

EC - European Commission

EU - European Union

GDPR - General Data Protection Regulation

G20 - Group of Twenty

HLEG - High Level Expert Group

ICCPR - International Convention on Civil and Political Rights

ICESCR - International Convention on Economic, Social and Cultural Rights

ICT - Information and Communications Technology

IHRL - International Human Rights Law

IoT - Internet of Things

KSI - Key Stakeholder Interview

NAP - National Action Plan

NCAI - National Center of Artificial Intelligence

OECD - Organisation for Economic Co-operation and Development

PDPB 2020 - Personal Data Protection Bill 2020

PECA 2016 - Prevention of Electronic Crimes Act 2016

QSO 1984 - Qanun-e-Shahadat Order 1984

R&D - Research and Development

UDHR - Universal Declaration of Human Rights

UK - United Kingdom

UN - United Nations

UNESCO - The United Nations Educational, Scientific and Cultural Organization

Executive Summary

Executive Summary

Artificial intelligence is slowly but surely finding its way into Pakistan. This gradual reliance on AI will only pick up pace as we reach the first stage of "knowledge revolution" in support of Pakistan's Vision 2025.[1] With huge amounts being invested in scientific research and development of AI technology in Pakistan, we ought to expect a future where intelligent machines make decisions about Pakistani citizens.

Even though a lot of AI technology is still in its preliminary stages in Pakistan, the Centre for Human Rights has prepared this mapping report to understand the looming threat posed to the right to equality and non-discrimination by decisions made through intelligent and data-driven algorithmic systems. In doing so, we have first explained what algorithmic bias is, and then mapped the following:

(a) Solutions offered to prevent algorithmic bias by intergovernmental and non-governmental organisations.
(b) Use of automated decision-making via intelligent date-driven algorithms in various sectors of Pakistan.
(c) Pakistan's current constitutional, policy, and legal landscape in light of algorithmic bias.

Through this report, we intend to show how AI technology is different from the revolutionary technologies in the past and one of its immediate risks is the manifestation of pre-existing social inequalities through technologically advanced means. In this pursuit, we recognise the immense need for a multi stakeholder dialogue in Pakistan to acknowledge the issue concretely and come up with solutions together as a nation. To this end, we hope to initiate a conversation in Pakistan on the need for algorithmic fairness in design and development of AI systems for algorithmic decision-making in institutions and sectors that have historically shown to not be rights-neutral. This poses a threat not only to Sustainable Development Goals 10 and 16,[2] but also to Pakistan's Vision 2025 of "People First". We also hope that this could lead to a more comprehensive and larger discourse on how to balance the full range of human rights with AI development through a targeted national AI strategy.

Moreover, the entire report is set in the belief that the universal and binding framework of human rights law offers flexibility to deal with the socio-economic challenges that may be brought about through technological development. Pakistan's human rights obligations arise first and foremost from the

[1] Ministry of Planning, Development and Reform, 'Pakistan Vision 2025', (Government of Pakistan, 2014) <https://www.pc.gov.pk/uploads/vision2025/Pakistan-Vision-2025.pdf> accessed 8 June 2021
[2] Goal 10: Reduced Inequalities; Goal 16: Peace, Justice and Strong Institutions. See: 'THE 17 GOALS | Sustainable Development' <https://sdgs.un.org/goals> accessed 8 June 2021

Constitution of the Islamic Republic of Pakistan 1973,[3] and then from its international obligations under the Universal Declaration of Human Rights,[4] International Covenant on Civil & Political Rights[5], and International Covenant on Economic, Social and Cultural Rights.[6] Pakistan is also currently developing a National Action Plan to give effect to the UN Guiding Principles on Business & Human Rights.[7] Therefore, any reference to human rights throughout the report would mean the collective *legal* rights that are enshrined in these documents.

In the recognition that AI systems for automated decision-making are still in the development or commercialisation stage and are to be deployed in the near future, we have offered targeted recommendations with the goal that this can lead to more research and funding in this important area that is meant to revolutionise the Pakistani life in an unprecedented manner across every socio-economic sector. We hope that more Pakistani research in this area can bridge the gap between the computer science and legal community at home, and also add a Pakistani perspective to the international discourse on this subject. Along with our aim of initiating a human rights-based approach to AI in Pakistan, we also hope that this report can offer one more human rights perspective to the international debate. This is so that the global discourse can be nudged forward from the ethics-based approach to a *legal* rights-based approach, at least to the extent of the issues raised by algorithmic bias.

[3] The Constitution of the Islamic Republic of Pakistan 1973 <http://www.pakistani.org/pakistan/constitution/>
[4] 'Universal Declaration of Human Rights | United Nations' <https://www.un.org/en/about-us/universal-declaration-of-human-rights> accessed 8 June 2021.
[5] 'OHCHR | International Covenant on Civil and Political Rights' <https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx> accessed 8 June 2021
[6] 'OHCHR | International Covenant on Economic, Social and Cultural Rights' <https://www.ohchr.org/en/professionalinterest/pages/cescr.aspx> accessed 8 June 2021
Apart from the International Bill of Rights, Pakistan is also party to four other key international human rights conventions, namely, UNCAT, CEDAW, CERD, CRC, and CRPD. For more information on Pakistan's international human rights obligations, see: <https://www.pc.gov.pk/uploads/report/Domestic.pdf>
[7] RSIL, 'Developing a National Action Plan for Business & Human Rights in Pakistan' (*Research Society of International Law | RSIL*) <https://rsilpak.org/project/developing-a-national-action-plan-for-business-human-rights-in-pakistan/> accessed 8 June 2021

Background & Context

Background & Context

In 2018, Pakistan updated its Digital Policy where it officially expressed its intention to embrace new technologies including Artificial Intelligence,[8] Robotics[9] and Internet of Things,[10] among others.[11] The underlying rationale of this Policy is to achieve Pakistan's Vision of "knowledge revolution" by 2025.[12] In order to bring about sustainable digital transformation in the Pakistani society across all socio-economic sectors, the Digital Vision aims to make the youth (which makes up about 60% of the Pakistani population of 225 million) more tech-literate and prepare them to be market ready. Just as any other country, Pakistan's Digital Vision is also aimed at transforming Pakistan into a reliable competitor and service provider internationally. To achieve this, Pakistan has established 9 laboratories in top universities of engineering and technology under a grant of Rs. 1.1 bn to carry out research in the areas of Artificial Intelligence (AI), Robotics, and Internet of Things (IoT).[13] There is also a Presidential Initiative for Artificial Intelligence & Computing which aims to revolutionise education, research, and business by adopting the latest technologies.[14] Pakistan is also investing in building the right ecosystem for adoption of new technologies through start-up incubators and accelerators and by promoting the activities of existing tech-firms so as to put in place foundational factors to enable technological advancement.[15]

Although technologies such as AI, IoT and Robotics are quite promising and offer a number of opportunities which will enable both economic development and human rights, such as right to education, right to health, among others. However, AI also carries risks which may violate human rights whilst simultaneously enabling them, and a unique challenge is now posed to the right to equality and non-discrimination. In fact, the risks of AI carry the potential to violate the full range of human rights guaranteed both under international and domestic laws.[16] This paradox has led to an international

[8] A discipline devoted to the development of systems that demonstrate human-like intelligence.
[9] An interdisciplinary field of study which integrates computer science and engineering.
[10] A giant network of physical objects that are connected to and exchange data with devices and systems over the internet. Examples include smart watches, smart home appliances, among others.
[11] MOITT, 'Digital Pakistan Policy' (Ministry of IT & Telecom, 2018) <DIGITAL_PAKISTAN_POLICY(22-05-2018).pdf> accessed 8 June 2021
[12] In its Vision 2025, Pakistan has recognised that, in comparison to its neighbours, it lags behind when it comes to technological advancements. Therefore, the "knowledge revolution" refers to embracing technology in order to utilise knowledge as a major asset for future development.
[13] Fakhar Durrani, 'Govt Allocates Rs1.1 Billion for Artificial Intelligence Projects' <https://www.thenews.com.pk/print/306187-govt-allocates-rs1-1-billion-for-artificial-intelligence-projects> accessed 8 June 2021
[14] 'PIAIC' <https://www.piaic.org/> accessed 8 June 2021
[15] Digital Pakistan, 'Top Startup Incubators and Accelerators in Pakistan - Digital Pakistan' <https://digitalpakistan.pk/blog/top-startup-incubators-and-accelerators-in-pakistan/> accessed 8 June 2021
[16] Lorna McGregor and others, 'The Universal Declaration of Human Rights at 70: Putting Human Rights at the Heart of the Design, Development and Deployment of Artificial Intelligence' (p.9, *HRBDT*, 20 December 2018). <https://www.hrbdt.ac.uk/the-universal-declaration-of-human-rights-at-70-putting-human-rights-at-the-heart-of-the-design-development-and-deployment-of-artificial-intelligence/> accessed 8 June 2021

discourse on the development of technology which is *trustworthy*, *ethical*, and *responsible* so that the risks of AI technology are minimised and its benefits are maximised. Such an approach also leads to careful and gradual changes, instead of abrupt ones across all socio-economic sectors as we undergo a revolution. This discourse has become highly sophisticated of late in the developed nations such as the EU and the UK where it is no longer a debate about the risks of AI, but a robust discussion of its implications and mitigation strategies at the policy and legal level. Unfortunately, this discourse has not fully reached Pakistan yet, and this is also reflected in Pakistan's Digital Policy which focuses largely on the opportunities of technology and pays little attention to minimising its risks.[17]

Through this report, we adopt a legal rights-based approach and hope to initiate the conversation on Responsible & Trustworthy AI in Pakistan which can lead to a larger multi stakeholder effort to balance out the opportunities of AI with the human rights risks it imposes. The risks of AI are numerous, and include, among others, increased surveillance, issues of privacy and data confidentiality, personal profiling, digital security, unequal access to AI technology, and threat to democracy. Acknowledging the vast array of existing, recognized factors that can create more responsible and trustworthy AI, this report is solely focused on one of these factors and its offshoots; ***algorithmic fairness***. This is because algorithmic *unfairness* imposes a real and immediate risk to the right to equality and non-discrimination. As Pakistan is only just beginning its digital transition and slowly embracing the ongoing technological revolution, we have a window of opportunity to build a future that promotes the common good, human rights and also averts the dangers of technology both from the outset and as it evolves. Sole focus on the opportunities of AI and total ignorance of the risks it presents is a threat to constitutional freedoms. It is also an impediment to Pakistan's ultimate goal of becoming a reliable technology competitor internationally, as technologically advanced countries such as those part of the Group of Twenty (G20), EU, OECD, are all focusing on responsible and trustworthy technology. Therefore, the need to develop an AI strategy which focuses on minimising the risks is no longer optional, but obligatory.

[17] Digital Pakistan Policy 2018 only addresses the lack of female participation, risks of unequal distribution, and lack of accessibility to disabled persons. For more information, see Policy Objectives IV, VIII and XI.

Methodology & Scope of the Report

Methodology & Scope of the Report

With rapid advancement in computing power in recent years, humans are continuously unleashing the potential of technology to tap into innovative areas. We see intelligent technology all around us; we carry it in our pockets, we wear it on our wrists, it helps us write emails, and is now even diagnosing our diseases. The era of artificial intelligence is the Fourth Industrial Revolution,[18] which in Pakistan, at long last, is upon us.[19]

The scope of the report is limited to the use of AI (which could include any AI technique) in social institutions of Pakistan where human decision-making is transitioning into intelligent automated decision-making, such as in the law enforcement or the health sector. The report recognizes that different types of AI techniques may be used in different institutions depending on the type of task the AI has to perform in the particular context of automated decision-making that it has been deployed for. Given the limited information available on deployment and use of AI in Pakistan, this report focuses on the social and legal institutions in Pakistan that use automated decision-making to understand the immediate and potential risk to the right to equality and non-discrimination.

Methodology:

This report relies primarily on qualitative data comprising both primary and secondary data sources. A detailed desk review of existing literature relevant to artificial intelligence, specifically ethics and bias was conducted. This included a review of available and accessible AI related policies, plans and legislative frameworks in Pakistan. Furthermore, an in-depth review of international literature was also conducted to identify best practices and better understand the AI and ethics framework.

In addition to the desk review, semi-structured interviews were conducted with international experts, relevant government and non-governmental stakeholders in Pakistan. The key stakeholder interviews (KSIs) were conducted online with as many stakeholders as accessible in the time frame of the project. The aim of the KSIs was to get a detailed perspective on (i) the use of AI in Pakistan in practice and (ii) the debate on ethics and AI at the global level. The table below lists the stakeholders that were interviewed for this report.

[18] Klaus Schwab, 'The Fourth Industrial Revolution: What It Means and How to Respond' (*World Economic Forum*) <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/> accessed 8 June 2021

[19] 'President Urges Youth to Prepare for Fourth Industrial Revolution' <https://tribune.com.pk/story/1892948/1-president-urges-youth-prepare-fourth-industrial-revolution> accessed 8 June 2021

Stakeholder	Type of stakeholder	# of KSI's
MeVitae	Private company / International expert	2
Medical Imaging & Diagnostics Lab, National Center of Artificial Intelligence (NCAI)	Government Initiative	1
Deep Learning Lab, National Center of Artificial Intelligence (NCAI)	Government Initiative	1
Intelligent Criminology Lab, National Center of Artificial Intelligence (NCAI)	Government Initiative	1
Cecil Abungu	International expert	1

Limitations:

The data collection relied on the stakeholder's willingness and honesty in providing information about their work on AI. In addition to this, some stakeholders were not able to participate in the KSIs due to unavailability. The report focuses only on technologies that have either already been deployed or are planning to be deployed in the near future. We also acknowledge that the development, deployment, and use of AI is still in its infant stages globally, and in Pakistan especially so, but it must be noted that a number of research institutes have been established in different universities of Pakistan with numerous research projects that are underway for commercialisation.[20] Therefore, a lot of innovation in AI is expected in the near future.

[20] Supra note 13.

Understanding Artificial Intelligence

Understanding Artificial Intelligence

"Artificial" is a fairly simple word meaning "made by *people*, often as a *copy* of something *natural* [emphasis added]."[21] "Intelligence" is more complex to define. For years we have attempted to understand what it means for us to be intelligent, to understand how we think or to make this a bit more wondrous, "*how a mere handful of matter can perceive, understand, predict, and manipulate a world far larger and more complicated than itself*."[22] Philosophical endeavours and continuous advancements in scientific inquiry have helped us better understand our cognitive abilities and behaviour, enabling us to create computers that exhibit human-like intelligence. As a starting point then, the field of artificial intelligence is simply a study of *people* creating systems that are a *copy* of the *naturally* occurring human intelligence.

Moving beyond our simple understanding of artificial intelligence, it is useful to point out that there is no single agreed-upon definition of artificial intelligence[23] just as there is no single agreed-upon definition of human intelligence, which only goes on to reflect the complexity of the notion of intelligence itself (an inquiry beyond the scope of this report). One reason why AI has no universal definition is because with rapidly evolving technology, a technique that constitutes artificial intelligence today may not do so tomorrow as the technique becomes more common. This is known as the odd paradox; new innovations in technology soon become repetitive and routine, and the computer scientists go on to find new computerised solutions which are more impressive and worthy of being labelled as AI.[24]

Regardless of what is or is not AI, what we are certain of is that the scientific discipline of AI is an attempt to replicate in computers what we humans consider to be our most important and identifying property—our intelligence.[25] It is best understood as an umbrella term that uses various techniques to automate the thinking abilities of human beings and enables computers to perform intelligent tasks. This view is captured by a helpful definition provided by Frank Chen which categorises AI capabilities into the following: logical reasoning (eg: playing chess or diagnosing diseases), knowledge representation (this is done by using programming languages through which computers understand and interact with

[21] 'ARTIFICIAL | Meaning in the Cambridge English Dictionary' <https://dictionary.cambridge.org/dictionary/english/artificial> accessed 21 June 2021.
[22] Peter Norvig and Stuart Russell, 'Artificial Intelligence: A Modern Approach' (page 1, 3rd edn, Pearson 2009).
[23] National Science and Technology Council: Committee on Technology, 'Preparing for the Future of Artificial Intelligence' (page 6, Washington D.C. Executive Office of the President, October 2016) <https://cra.org/ccc/wp-content/uploads/sites/2/2016/11/NSTC_preparing_for_the_future_of_ai.pdf> accessed 21 June 2021
[24] Pamela McCorduck, 'Machines Who Think: A personal inquiry into the history and prospects of Artificial Intelligence', (2nd end, AK Peters ltd., 2004).
[25] Ibid.

the real world), planning and navigation (eg: self-driving cars), natural language processing (eg: Siri), and perception (eg: image analysis).[26]

A more recent definition comes from The High-Level Expert Group set up by the European Commission (EC) which currently recommends the use of the following definition of AI:[27]

"*Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.*

As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber -physical systems)."

Therefore, artificial intelligence is a vast scientific discipline[28] of developing artificially intelligent systems by teaching these systems techniques which are also artificially intelligent. Although there are difficulties in defining AI, recently there is general consensus among global stakeholders to define AI for legal certainty and regulation of artificial intelligence systems.[29]

---

[26] 'AI, Deep Learning, and Machine Learning: A Primer - Andreessen Horowitz' <https://a16z.com/2016/06/10/ai-deep-learning-machines/ > accessed 21 June 2021. Please note, the examples were modified by the author of the report.
A similar, but more elaborated definition is provided by a popular Computer Science textbook which divides artificial intelligence into the following categories: (1) systems that think like humans (e.g., cognitive architectures and neural networks); (2) systems that act like humans (e.g., pass the Turing test via natural language processing; knowledge representation, automated reasoning, and learning), (3) systems that think rationally (e.g., logic solvers, inference, and optimization); and (4) systems that act rationally (e.g., intelligent software agents and embodied robots that achieve goals via perception, planning, reasoning, learning, communicating, decision-making, and acting).

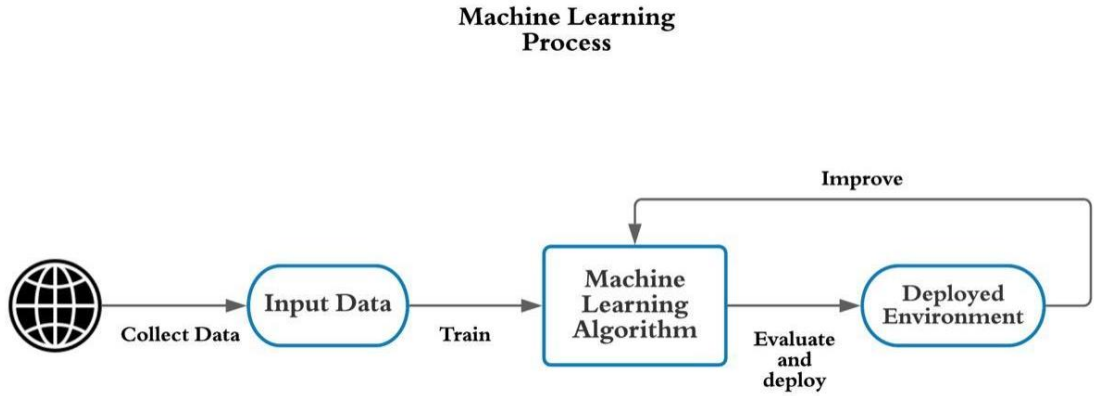
[27] Independent High Level Expert Group on Artificial Intelligence, 'A Definition of AI: Main Capabilities and Disciplines', (page 6, European Commission, April 2019).

[28] Ibid.

[29] European Commission, 'Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS COM(2021) 206 Final' <https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF> accessed 6 July 2021. See also: Jonas Schuett, 'A Legal

Common AI techniques and their mechanics

A fast-growing domain of AI which has led to major technological advancements in recent times is *machine learning*. This statistical process runs on algorithms that have the ability to "learn" and improve their performance over time on the tasks they perform.[30] Simply put, the process starts with a set of data and attempts to map out or derive a procedure on its own that can explain the data or predict future data.[31] This mapping is the *model* which *learns from data*.[32] In this technique of AI, the intelligence of the machine is in the fact that it is not explicitly *programmed* to categorise the patterns in the data in a certain way; it simply does so by figuring out the structure of the data itself.[33]



Practically, programmers begin with a set of data which is divided into *training data* and *test data*. A model is then derived, often from millions of data points, which is a mathematical structure that categorises a range of possible decision-making rules.[34] The model also has adjustable parameters, and an objective function is defined by the programmer to evaluate the desirability of an output based on the adjusted parameters.[35] After the completion of the training of the model, the accuracy and effectiveness of its performance is tested on the test data, i.e. data that the model has not been exposed to before. The aim here is to generalise the model so that its performance is accurate even in cases it has not seen before.[36] The model is then deployed in its applicative environment, and the machine learning algorithm is continuously improved based on what it learns in its deployed environment.

Definition of AI', (September 4, 2019), available at SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3453632> accessed 6 July 2021.

[30] Harry Surden, 'Machine Learning and Law' (Social Science Research Network 2014) SSRN Scholarly Paper ID 2417415 <https://papers.ssrn.com/abstract=2417415> accessed 21 June 2021.

[31] Supra note 23.

[32] Lindsey Andersen, 'Human Rights in the Age of Artificial Intelligence' (Access Now) <https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf> accessed 21 June 2021

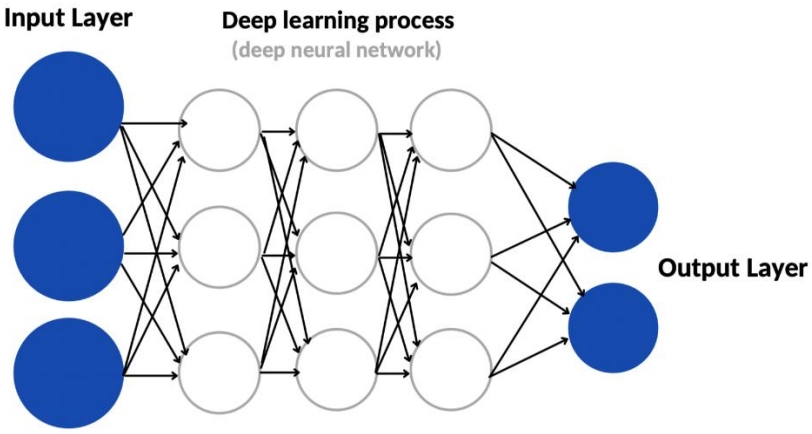
[33] It must be noted, however, that this depends on whether the model was trained through supervised or unsupervised learning.

[34] Supra note 23.

[35] Ibid.

[36] Ibid.

A sub-field of machine learning which has gained prominence in the last decade due to availability of big data[37] is *deep learning* which uses structures called "neural networks" that have been inspired by the human brain,[38] and has been very successful in recognising intricate and complex structures and patterns in large datasets. The learning is done through an artificial neural network to analyse data. This network consists of layered "units" which are modelled after neurons.[39]



The learning procedure is referred to as "deep" because it rapidly discovers or uncovers new layers within the data.[40] In image recognition, for instance, the first layer of units may combine raw data of the image to recognise basic patterns of the image; the second layer of units may combine the results of the first layer to recognise patterns-of-patterns; the third layer may combine results of the second layer[41] and the process goes on till the image is identified. Accuracy increases when the algorithm is fed with more data; to be able to recognise a particular celebrity with better accuracy, for example, the algorithm would need images of the celebrity from different angles, in different hairstyles, etc.[42]

[37] Datasets that are too large and complex to be processed through traditional data processing methods.
[38] KSI with Deep Learning Lab, NCAI.
[39] Supra note 23.
[40] Richa Grover, 'Deep Learning - Overview, Practical Examples, Popular Algorithms | Analytics Steps' <https://www.analyticssteps.com/blogs/deep-learning-overview-practical-examples-popular-algorithms> accessed 21 June 2021
[41] Supra note 23.
[42] KSI with Intelligent Criminology Lab, NCAI.

The Hidden Side of Artificial Intelligence; Algorithmic Bias

The Hidden side of Artificial Intelligence; Algorithmic Bias

The human brain is not only intelligent, it is also biased. In an attempt to encode human intelligence, we are also encoding human biases in machines, leading to what is known as algorithmic bias. To see this in practise, type "greatest leaders of all time" in any search engine of your choosing and you will find a list of prominent personalities which are overwhelmingly male.[43] Now translate "She is a scientist. He is a nurse." from English to Turkish using Google Translate, then press the swap button and watch it change to "He is a scientist. She is a nurse."[44] These are basic examples of gender stereotypes and biases that exist in our world which are being reinforced through algorithms in the virtual world. However, algorithmic bias is not limited to gender, but has also been observed along other protected characteristics, such as race,[45] socioeconomic status,[46] and freedom of expression.[47] Such biases exist not only in online searches, translations and recommendations, but are also present in what is known as automated decision-making.

Automated or *algorithmic* decision-making (the report uses both terms interchangeably) is when decisions are made about people by machines without active human involvement. Although it may seem that automating the decision-making process leads to more efficiency and less prejudice, recent global events[48] have shown that reducing decision-making to algorithms cannot fully encapsulate the complexities of human societies, values, and morals. There is hard evidence which demonstrates that automated decision-making can result in outcomes that are not only biased, but biased in an *unfair* manner, thus *potentially* constituting discrimination at law.[49]

While algorithmic decision-making by way of AI is undoubtedly beneficial in terms of speed, efficiency and accuracy, there is a common misconception that algorithms automatically result in unbiased outcomes and reduce human errors. The misconception exists because algorithms take in apparently

---

[43] 'Artificial Intelligence: Examples of Ethical Dilemmas' (*UNESCO*, 2 July 2020) <https://en.unesco.org/artificial-intelligence/ethics/cases> accessed 21 June 2021

[44] Turkish is a genderless language and the same pronoun "o" is used as an equivalent for the English pronouns "he" "she" and "it". On the other hand, English is largely gender neutral, but has a gendered system of pronouns. For more information, see: Ramazan Göçtü and Muzaffer Kır, 'Gender Studies in English, Turkish and Georgian Languages in Terms of Grammatical, Semantic and Pragmatic Levels' (2014) 158 Procedia - Social and Behavioral Sciences 282 <https://doi.org/10.1016/j.sbspro.2014.12.089>

[45] 'Predictive Policing Poses Discrimination Risk, Thinktank Warns' (*the Guardian*, 15 September 2019) <http://www.theguardian.com/uk-news/2019/sep/16/predictive-policing-poses-discrimination-risk-thinktank-warns> accessed 21 June 2021.

[46] Ibid.

[47] Thomson Reuters Foundation, 'Instagram, Twitter Blame Glitches for Deleting Palestinian Posts' (*DAWN.COM*, 11 May 2021) <https://www.dawn.com/news/1623302> accessed 21 June 2021.

[48] Jeffrey Dastin, 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women' *Reuters* (10 October 2018) <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G> accessed 21 June 2021. See also: Supra note 45 and Supra note 47.

[49] Ibid.

objective points of reference (inputs) and provide standard outcomes, but these inputs and outputs themselves can be problematic.[50] The inputs (such as data) cannot always be objective if prior judgements have been made about those inputs.[51] When this happens, regardless of how well-intentioned the AI solution is, the machine learns the human biases that we thought it would avoid.

The link between human bias and algorithmic bias

The human brain is a system of neurons which helps us to communicate with each other and interact with our environments, constantly gathering information from the large variety of our life experiences and interactions. The stimuli in our brain respond to the information we are gathering every day, and process them. Whilst this natural process enables our cognitive abilities, this is also the process which shapes our cognitive biases.[52]

Cognitive biases are an umbrella of biases that take our past and present experiences and stimuli, categorise them, and make judgements on these experiences over the years. This umbrella of biases can be further categorised into *conscious* and *unconscious* biases. The former is something we are actively aware of, and the latter exists without our awareness; the bias may not be intentional, but the bias is still there, and we do not even know it! Even though our biases are shaped naturally, they can also be *learnt* through our continuous social interactions. However, biases can never fully be eliminated, and can only be reduced. Therefore, as time passes, our cognitive biases may change, mitigate, or even evolve based on the new information that the brain gathers, and the different judgments it makes about that information.[53]

Cognisance of conscious and unconscious biases is very crucial especially when we are making decisions that impact other human beings[54] because historical experiences have shown that human biases may lead to outcomes that are unfair and discriminatory. As human decision-making is gradually being replaced by automated decision-making, there is a risk of pre-existing human biases creeping into the algorithm and potentially exacerbating discrimination. This is because the decisions and choices that our brains have made are now reflected in big data,[55] and as discussed, these decisions and choices may very well be wrapped in our biased judgements.

[50] Robyn Caplan and others, 'Algorithmic Accountability: A Primer' (Data & Society) <https://datasociety.net/wp-content/uploads/2019/09/DandS_Algorithmic_Accountability.pdf>
[51] KSI with MeVitae.
[52] Ibid.
[53] Ibid.
[54] Ibid.
[55] Ibid.

How biases become part of the intelligent data-driven algorithms

Algorithmic bias does not happen through a random chance or machine autonomy. It is a systematic error which occurs due to an inaccuracy in the algorithm itself;[56] the algorithm *systematically*, i.e. *repeatedly* misses out certain groups of people more than others.[57] It is the human or institutional bias which is coded into the algorithm due to biased data or biased parameters of the algorithmic model. In automated decision-making, it ends up impacting some people or groups of people more than others, potentially resulting in unfairness, discrimination, and inequality because the algorithms learn by looking at the world as it is and not what it ought to be. Therefore, biases in intelligent algorithmic decision-making are a reflection of the existence of human biases in the *overall* decision-making processes across all institutions and organisations.[58] The challenge we now face against fairness is not new, but it is being renewed as we make technological advancements. It is an age-old problem of discrimination and inequality which is manifesting itself again through newer and more efficient means which are making it more challenging to deal with an existing issue.

A good example of a discriminatory outcome is Northpointe Inc.'s software COMPAS whose algorithm predicts the risk of recidivism.[59] A landmark research by ProPublica discovered that the algorithms in the USA which predicted the reoffending rates were discriminating on the basis of race.[60] It was reported that black defendants who had not recidivated over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts.[61] On the other hand, white defendants who had re-offended within the next two years were mistakenly labelled low risk almost twice as often as black re-offenders.[62] Therefore, AI risks violating (and in some cases has arguably already violated)[63] the legal right to equality and non-discrimination.

---

[56] Luke Jew, 'Algorithmic Bias Explained' (*Bias free hiring within your ATS*) <https://www.mevitae.com/resource-blogs/algorithmic-bias-explained> accessed 21 June 2021

[57] KSI with MeVitae.

[58] Centre for Data Ethics and Innovation, 'Review into Bias in Algorithmic Decision Making' (CDEI, November 2020) <https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making> accessed 21 June 2021

[59] This term refers to the tendency of a criminal to reoffend.

[60] Jeff Larson and others, 'How We Analyzed the COMPAS Recidivism Algorithm' (*ProPublica*)<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=nD-X136_tDm0nh1l4Xtv0LbpjY_BSO3u> accessed 21 June 202. A similar example of racial profiling via algorithms is the increased use of predictive policing. However, the risk of discrimnation is not only against race, but also sexual preference and age. See: 'Predictive Policing Poses Discrimination Risk, Thinktank Warns' (*the Guardian*, 2019) <http://www.theguardian.com/uk-news/2019/sep/16/predictive-policing-poses-discrimination-risk-thinktank-warns> accessed 21 June 2021.

[61] 45% v 23%.

[62] 48% vs. 28%.

[63] Supra note 48.

Bias can enter the algorithm in various ways. Listed below are some non-exhaustive ways that biases become part of the algorithm:

(a) Bias at the Data or Input level

- **Historical data**: It is possible that the data the model is built, trained, tested and operated on could introduce bias if the data represents previously made human decisions which are wrapped up in historical or societal inequalities.[64]

- **Data selection bias**: This occurs when data is not representative of the target population. For example, by over or under recording of particular groups, the algorithm may be less accurate for some people or provide a skewed picture of particular groups.[65]

- **Incomplete or outdated data**: If data is insufficient or if the data no longer reflects current realities, then decisions of the algorithm may very well be inaccurate.[66] This is why ethics guidelines such as those by the European Commission or the Montreal Declaration recommend that algorithmic models have to be continually updated with new and more accurate data.

(b) System Design

Biases can also be introduced by the human designers of the artificial intelligence system (AIS). This is not to say human designers carry malintent. As discussed in the previous section, the nature of the human brain is such that biases are inevitable, even in the most well-intentioned people. Bias through system design can occur in two ways:

1. The bias can exist in the algorithm itself.[67] If a human being has built the algorithmic model, but the model is not a well-motivated one, i.e. it is not based on insights of our natural world,[68] then it may be the case that biases of the programmer are coded into the algorithm. For example, if the algorithm has to decide whether or not to give a loan, it is possible that it may contain a gender bias if it bumps up scores for men by a higher amount than for women.[69] This is an intrinsic bias in the model itself, and may even be

[64] Supra note 58.
[65] Ibid.
[66] Supra note 32.
[67] Supra note 56.
[68] Ibid.
[69] KSI with MeVitae.

easy to spot. However, models built through machine learning or deep learning are far more complex which make it more difficult to spot intrinsic biases in the model.[70]

2. Biases can also enter the algorithm through the variables that the human designers have prioritised for the AIS to optimise on and factor in while making decisions.[71] If an obvious variable or parameter, such as gender, leads to a decision which is biased in an unfair manner, it may be discriminatory. At law, this is better understood as *direct discrimination*. Discriminatory results or biased decisions may also be made as a result of proxies. With increased use of technological devices and the internet, data reflecting people's behaviours, choices, activities, movements, etc. is abundantly available. Globally, as people leave behind traces of their lives in the form of data, what has become available is "big data" which is full of correlations.[72] There are patterns in data which exist within a single person's data, as well as across people.[73] For example, even if a variable which is known to cause unfair outcomes, such as race, is eliminated, the algorithm may still give a racially biased result through common proxies for race such as income, education, or post code.[74] As a result, AI is "*inevitably seeking out proxies for directly predictive characteristics when data on these characteristics is not made available to the AI due to legal prohibitions.*"[75] In legal parlance, this is better known as *indirect discrimination*.

(c) Human Oversight

A recurrent recommendation in various AI ethics guidelines across the globe is that human oversight be exercised throughout the lifecycle of the AIS so that humans remain in control of the technology including in the determination of the final outcome, especially in situations where algorithms cannot apply human judgement in unfamiliar situations.[76] However, in a recent report the Centre for Data Ethics & Innovation has cautioned that there is a risk of bias re-entering the decision-making process depending on how humans interpret and apply their

[70] Ibid.
[71] Fillipo A. Raso and others, 'Artificial Intelligence & Human Rights: Opportunities & Risks' (Social Science Research Network 2018) SSRN Scholarly Paper ID 3259344 <https://papers.ssrn.com/abstract=3259344> accessed 21 June 2021.
[72] Betsy Anne Williams and others, 'How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications' (2018) 8 Journal of Information Policy 78.
[73] Ibid.
[74] Cathy O'Neil, '*Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*', (Penguin Books 2016). See also: Supra note 32.
[75] Anya E.R. Prince and Daniel Schwarcz, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data' (2020) 105 Iowa L. Rev. 1257. <https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data/>
[76] Supra note 58.

own conscious or unconscious biases to the final outcome.[77] Therefore, whilst ensuring that AI technology remains in human control, it is necessary that there is diversity and inclusivity in the teams involved in system design, development, and deployment of AI so that unfair human biases can be mitigated as much as possible.

Rights-based issues emerging from technology

Intelligent algorithms have raised some issues that carry legal ramifications. In this part, we highlight both the technology and legal issues to show the precise gap which exists between the two disciplines, and stress on the need to bridge this gap so that the development of intelligent automated systems can be promoted in a rights-respecting manner.

(a) The Technology Issue

Some of the core issues that intelligent algorithms raise boil down to lack of transparency, explainability, and traceability. Even though these are the same issues that we face with human decision-making, the added danger is that fewer individuals are aware of how the decisions are being made.[78] Barocas and Nissenbaum have discussed this as the transparency box problem.[79] As machine learning is based on large datasets, it is not entirely transparent, even to the creators of the machine, what data was used by the machine to reach a particular decision. This goes hand-in-hand with Pasquale's black-box.[80] While the input and the outputs are known, what the algorithm does inside the opaque black-box remains a mystery.[81] Therefore, tracing how and why a decision was made and on the basis of what data is problematic with AI algorithms which create the models based on complex datasets which are not fully understood by human beings.[82]All of this makes it difficult to explain or pinpoint why an algorithm resulted in a potentially discriminatory outcome. Opening up the opaque black-box and dissecting the algorithm is also met with problems of intellectual property. This then creates a proprietary black box.[83]

---

[77] Ibid.
[78] KSI with MeVitae
[79] Solon Barocas and Helen Nissenbaum, 'Big Data's End Run around Anonymity and Consent', *Privacy, Big Data, and the Public Good Frameworks for Engagement* (Cambridge University Press, 2014)
[80] Frank Pasquale, 'The Black Box Society', (Harvard University Press, 2016), <https://www.hup.harvard.edu/catalog.php?isbn=9780674970847> accessed 21 June 2021
[81] "The Time Has Come for International Regulation on Artificial Intelligence" – An Interview with Andrew Murray - Opinio Juris' <http://opiniojuris.org/2020/11/25/the-time-has-come-for-international-regulation-on-artificial-intelligence-an-interview-with-andrew-murray/> accessed 21 June 2021
[82] Supra note 32.
[83] Megan Stevenson, 'Assessing Risk Assessment in Action', (Minnesota Law Review, 2018) <https://www.minnesotalawreview.org/wp-content/uploads/2019/01/13Stevenson_MLR.pdf> access 21 June 2021

As a result, the precise mechanics of the algorithm remain hidden, and mitigating unfair bias becomes increasingly difficult. However, if bias is not eliminated from the algorithm, and if this bias is potentially discriminatory in nature, then discrimination will take place at a much larger scale in comparison to human decision-making; despite criticisms that human decision-making is more problematic than algorithmic decision making,[84] AI algorithms carry the "*potential to discriminate more consistently, systematically, and at a larger scale than traditional non-digital discriminatory practises*."[85] This is because AI may not only replicate existing biases, but can also significantly scale discrimination and discriminate in unforeseen ways.[86] Where human decision-making is limited to fewer factors and decisions, decision-making through digital means is more consistent and applies to a larger number of people in a more systematic way as it is based on large amounts of datasets representing hundreds of different factors, thus leading to a situation where discrimination can potentially be more widespread.[87]

Therefore, owing to the sophistication and scale of the AI technology, the challenge that this technology raises is a little different from the challenge we confronted with other revolutionary technologies. The core difference between the AI technology and technologies that have come before, such as radio or television, is one of autonomy; we are giving up our autonomy for something we do not fully understand.[88] Previous life-changing technologies were easily understood by both the creators and the beneficiaries of the technology, whereas AI is complex even for its creators.[89]

(b) The Legal Issue

As is clear from the technology issue, intelligent algorithms can be unexplainable by the creators and are hidden from the people about whom the algorithms make decisions. With these issues left unaddressed, we may find ourselves in a scenario where citizens will find it increasingly difficult to question or protest against something that is hidden and AI developers

[84] Alex P. Miller, 'Want Less-Biased Decisions? Use Algorithms', (Harvard Business Review, 2018) <https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms?utm_source=linkedin&utm_medium=social&utm_campaign=hbr> accessed 21 June 2021

[85] Natalia Criado and Jose M Such, 'Digital Discrimination' (Oxford University Press) <https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198838494.001.0001/oso-9780198838494-chapter-4> accessed 21 June 2021. See also: Cecil Abungu, 'Algorithmic Decision-Making and Discrimination in Developing Countries', forthcoming (2022) 13 Case Western Reserve Journal of Law, Technology & the Internet, p. 9.

[86] Jessica Fjeld and others, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI' (Social Science Research Network 2020) SSRN Scholarly Paper ID 3518482 <https://papers.ssrn.com/abstract=3518482> accessed 21 June 2021

[87] KSI with Cecil Abungu.

[88] Supra note 81.

[89] Ibid.

may not be able to justify evaluating people on the basis of algorithms that they themselves cannot provide concrete explanations for.[90] This raises a unique challenge for the rule of law and implementation of constitutional freedoms because anti-discrimination law, as it currently stands in most parts of the world, places the burden on the claimant to prove the co-relation between the protected characteristic and the alleged discriminatory conduct.[91] As such, the combined issues of opacity, lack of explainability, and intellectual property not only make it difficult for litigants to detect that they have been discriminated against, but also to go to court and prove how the algorithmic decision-making process was discriminatory against them.[92] The problem of detection of discrimination is further amplified by the fact that the disciplines of AI and data science are not just about the individual, but about the huge datasets which represent many individuals.[93]

In addition to this, specific legal challenges exist in the country context, such as those related to indirect discrimination under the existing Pakistani legal framework. Under Pakistani law, there is no extensive legislation on anti-discrimination. Rather, "equality" is protected under Article 25 of the Constitution of Pakistan 1973. This constitutional freedom, despite being open-ended, provides protection mainly against direct discrimination, and not as much against indirect discrimination.[94] Indirect discrimination is a legal concept that was largely developed and advanced through the international human rights framework, especially through the **Convention on Elimination of All Forms of Discrimination against Women (CEDAW)**. Therefore, the legal concept of indirect discrimination in Pakistan is largely based on these international commitments. The lack of development of this legal concept in the domestic framework makes it difficult to establish both as a legal concept and as part of existing laws. With deployment of algorithmic decision-making this may prove to be extremely problematic as such decision making carries high potential to discriminate indirectly by way of proxy, thus warranting an amendment in the existing interpretation of discrimination under Pakistani law.[95]

[90] Supra note 74, p. 8.
[91] KSI with Cecil Abungu.
[92] Ibid.
[93] KSI with MeVitae.
[94] Theoretically, it would be reasonable to interpret Article 25 to include both direct and indirect discrimination. However, in practise no such jurisprudence exists to substantiate the claim to indirect discrimination.
[95] The development and enforcement of the concept of indirect discrimination remains murky in Pakistan. Arguably different provisions in the legal and policy framework, including quotas for women, reserved seats for minorities, etc, are indicative of the State's recognition of indirect discrimination at the least. However, substantive strides in localising the legal concept of indirect discrimination are lacking in the Pakistani framework.

International Best Practises to Mitigate Algorithmic Bias

So far!

International Best Practises to Mitigate Algorithmic Bias - *So far!*

Globally, many policy and ethics initiatives have surfaced offering guidelines for minimising bias throughout the lifecycle of the AI by various stakeholders in the private sector, professional organisations, civil society, intergovernmental and non-governmental organisations, research institutes, etc. Such initiatives have emerged from the "*something must be done"* line of reasoning because the possibility of discrimination via intelligent and data-driven algorithms is a hard fact that is largely acknowledged by the different stakeholders. However, there is still no definitive stance on "*what must be done*";[96] some guidelines have explicitly stated that strategies for new technologies must be reviewed and updated over time to keep up with the evolution of technology and knowledge thereof.[97] But there is good news. The numerous guidelines which have attempted to answer "*what must be done*" offer the same or considerably similar solutions, igniting hope that this is not a cause without solutions.

Methodology of Analysis

For our analysis, we have focused only on the documents that take a human rights-based approach to the AI challenges or consider human rights more than just in passing. It must be noted that at the intergovernmental level, even though direct reference is made to the protection of fundamental rights, the human rights approach is actually an ethics-based approach[98] and not a *legal* rights-based approach.[99] This means the standards that are being set globally at the moment are not legally binding, but carry only an ethical force and exist for the purpose of guidance only,[100] and private technology companies have been entrusted with the task of self-regulation.[101] However, at the same time, the guidelines also recognise that many "ethical" principles mentioned for guidance purposes are already

[96] This was also highlighted during our KSI with MeVitae.
[97] Independent High Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI', (p. 3, European Commission, 2019)
[98] An approach that considers principles such as fairness, transparency etc. in light of ethical theories such as deontological, teleological, and virtue-based ethics. Such a framework is not legally binding and lacks legal accountability.
[99] An approach that is aligned with the binding framework of human rights as enshrined under the Constitution of Pakistan 1973 and the International Human Rights Law.
[100] Categorically noted in the Ethics Guidelines for Trustworthy AI by AI HLEG, European Commission. See: Supra note 97.
[101] European Commission, 'COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE EUROPEAN COUNCIL, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS, Artificial Intelligence for Europe (SWD(2018) 137, 25 April 2018) <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&rid=1> accessed 24 June 202.

recognised in existing laws.[102] Not surprisingly then, this approach has been criticised as a means for tech-firms to escape regulation.[103]

At the intergovernmental level, we have reviewed and analysed the recommendations and guidelines set forth in the First Draft of the Recommendation on the Ethics of Artificial Intelligence by UNESCO,[104] Recommendation of the Council on Artificial Intelligence by OECD,[105] G20 AI Principles,[106] Ethics Guidelines for Trustworthy AI by the High-Level Expert Group set up by the European Commission.[107]

For non-governmental documents, the Toronto Declaration prepared by Access Now and Amnesty International[108] and the Montreal Declaration for a Responsible Development of Artificial Intelligence by the Université de Montréal[109] were reviewed. The Toronto Declaration is the most relevant document reviewed for this section as it is the most definitive human rights document thus far on machine learning technology which places a central focus on the right to equality and non-discrimination. The Toronto Declaration also supports a legal rights-based approach over an ethics-based approach due to the universal, flexible, and legally binding framework of international human rights law (IHRL). This approach is in alignment with the purpose and approach of this report.

In the following analysis, we have identified common themes that exist in all of the aforementioned documents. All of the documents followed a similar pattern, categorizing solution-based themes that contained subcategories of challenges. Each subcategory had its own principles which explained how to counter the challenge posed by the particular subcategory. For the purposes of this report, we only focused on those solution-based themes that identified the subcategory of algorithmic bias as a challenge. We analysed the theme in the context of algorithmic bias and discovered various principles that guided how to circumvent the problem raised by the same and achieve the purpose of the theme. In our analysis, we have not commented extensively on the viability of any of the themes as we believe that such a determination demands multi-stakeholder participation. However, we do acknowledge the

[102] Ad Hoc Expert Group (AHEG) for the Preparation of a Draft text of a Recommendation the Ethics of Artificial Intelligence, 'First draft of the Recommendation on the Ethics of Artificial Intelligence', (UNESCO, 2020), <https://unesdoc.unesco.org/ark:/48223/pf0000373434> accessed 21 June 2021. See also: Supra note 97.
[103] Ben Wagner, 'Ethics as an Escape from Regulation.: From "Ethics-Washing" To Ethics-Shopping?' in EMRE BAYAMLIOĞLU and others (eds), *BEING PROFILED* (Amsterdam University Press 2018) <https://www.jstor.org/stable/j.ctvhrd092.18> accessed 21 June 2021
[104] Supra note 102.
[105] OECD, 'Recommendation of the Council on Artificial Intelligence', (OECD/LEGAL/0449).
[106] G20, 'Ministerial Statement on Trade and Digital Economy', (2019).
[107] Supra note 97.
[108] Access Now and Amnesty International, 'The Toronto Declaration', (16 May 2018).
[109] Université de Montréal, 'Montreal Declaration for a Responsible Development of Artificial Intelligence', (2018).

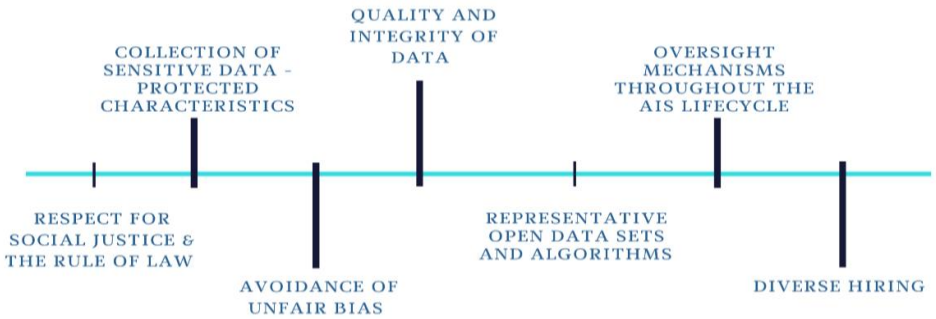
commentary which is developing on the practicality of the globally offered solutions, especially in the context of developing countries.[110]

Whilst reading the common themes, one must be cognisant of two things; (a) the themes and principles are indivisible and interdependent; (b) the context in which intelligent algorithmic decision-making is deployed is important as different situations raise different challenges.

In addition to reviewing non-binding guidelines, we have also reviewed recent landmark legal developments that have been made in the EU that highlight the EU's efforts to bring AI within legal regulation, such as the General Data Protection Regulation (GDPR) and the recent *Proposal* for a Regulation of The European Parliament and The Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.

Intergovernmental & Non-governmental Recommendations and Guidelines

1. Fairness & Non-Discrimination



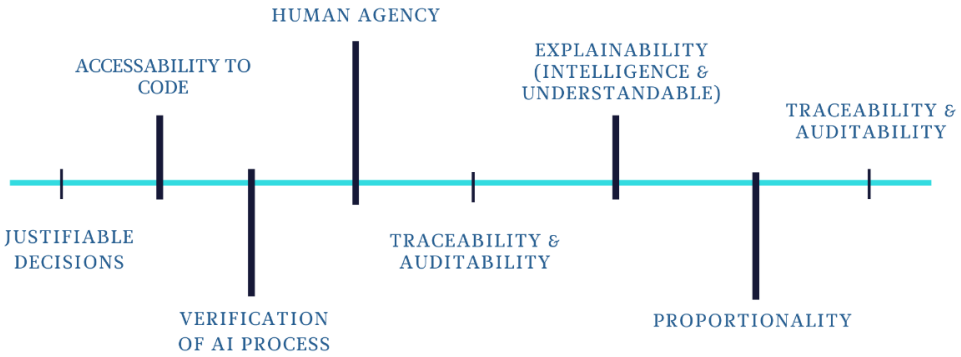
This principle is the basic foundation that enables all the other solutions to bring about algorithmic fairness and to remove social inequalities and discrimination.[111] The cornerstone of this principle is to uplift the rule of law and promote social justice. This can be done by ensuring that AIS do not reinforce historical or socially constructed biases that can lead to the discrimination of protected groups and characteristics.[112] This is especially important as AIS tends to impact such groups and characteristics more than others.[113] This principle places AI developers under a soft obligation to consider ways to avoid unfair biases in the entire lifecycle of the AIS.[114]

[110] Cecil Abungu, 'Algorithmic Decision-Making and Discrimination in Developing Countries', forthcoming (2022) 13 Case Western Reserve Journal of Law, Technology & the Internet.
[111] Supra note 102, p.7; Supra note 105, p. 7; Supra note 106, p. 11; Supra note 109, p.13.
[112] Such as racial and ethnic groups, gender, age, religion etc.
[113] Supra note 108, p.5.
[114] Supra note 97, p.6; Supra note 105, p. 11; Supra note 106, p. 4.

There are a number of solutions that have been put forth in the intergovernmental and non-governmental documents that can help achieve fairness and non-discrimination. One of the ways is by placing legal safeguards on the collection and use of sensitive data under data protection legislations.[115] Not only does this lead to more trustworthy AI, but the right to privacy also overlaps with right to equality and non-discrimination here because AI developers are guided to remove identifiable and discriminatory biases in the data collection phase.[116] Therefore, the quality and integrity of the data collected is made an essential requirement as it can greatly impact the performance of the AIS.[117] As such, the data must be representative, and should not be ridden with socially constructed biases before the algorithm is even trained on the dataset. However, avoidance of bias has to be tested and documented at every stage of the AI development such as planning, training, testing and deployment.[118] In order to collect more representative data, the guidelines support the development of common algorithms and accessibility to open datasets to expand their use as a socially equitable objective.[119]

The guidelines also recognise that bias cannot just enter the AIS through data, but also during the programming phase whereby people developing the AIS may bring in their own biases while designing the algorithm.[120] In order to minimise bias in the modelling phase, the guidelines recommend against any homogenisation of opinions or practices,[121] and encourage AI developers to hire people from diverse backgrounds and make their opinions more inclusive in the AI development process.[122] Another solution is to put in place oversight mechanisms that can evaluate the AIS's purpose, limitations, requirements, and decisions in a clear and transparent manner.[123]

2. Transparency & Explainability



[115] Supra note 97, p. 17; Supra note 106, p. 3; Supra note 108, p. 7.
[116] Supra note 97, p. 17.
[117] Ibid.
[118] Ibid.
[119] Supra note 105, p. 8; Supra note 106, p. 13; Supra note 109, p. 13.
[120] Supra note 97, p. 36; Supra note 108 p. 6.
[121] Supra note 109, p. 14.
[122] Supra note 97, p.18; Supra note 109, p.14.
[123] Supra note 97, p.15.

Under automated decision-making through AI, decisions made can impact a person's life, quality of life, reputation etc. As there is a very real possibility that the decision can unfairly discriminate against the person or the sensitive aspects of their life, such decision-making should be transparent.

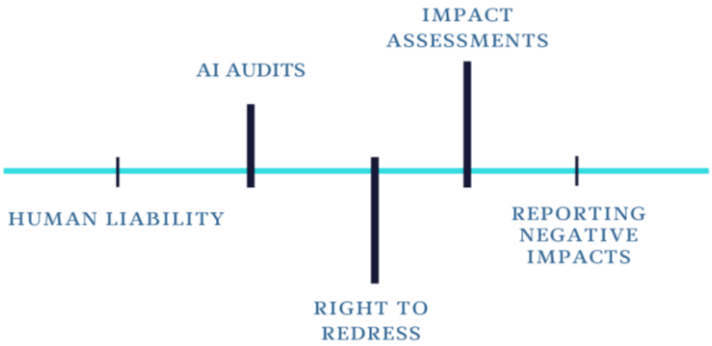
Transparency is an umbrella term that contains sub-meanings in the context of algorithmic decision-making through AIs, and all meanings are interdependent. Firstly, transparency means that the decision-making should be intelligible to the creators of the AIS as well as those who have been impacted by the automated-decision. More often than not, computer scientists and mathematicians are the ones who can comprehend and understand the codes, the data, and the factors that determine an outcome. However, under self-learning systems, it can be difficult even for computer scientists and mathematicians to understand why an AIS made the decision that it made. Intelligibility also means that the AI decision-making process is to be made transparent to the users in a language that they can comprehend and understand. Therefore, the guidance for overcoming this issue is to make the decision-making process more intelligible, which overlaps with the next recommendation; traceability.

Intelligibility is necessary to foster traceability of the automated decision that is made in the event that the decision is challenged. For this reason, it is recommended that AI developers make the AIS more transparent in a manner that is proportionate to the context in which the AIS is deployed so that its decisions can be explained and justified to the impacted people, such as by explaining the inputs, the outputs, the most important factors and parameters that shaped the decision, and the limitations of the AIS. Justifiability also includes accessibility of algorithms by public authorities and relevant stakeholders for the purposes of verification and control, unless where the algorithm has the potential to be misused or to impose serious danger.

Transparency is achieved by documenting each process of the AI development such as, among other things, data gathering and labelling, and development of the algorithm so that the black box can be opened and the flaw in the decision-making process can be located efficiently.

Moreover, to enable people to challenge the decisions, human agency must be paramount; while interacting with an AIS, people must be informed that they are not interacting with a human being, and they must be given appropriate tools to comprehend and interact with AIS. This can help them self-assess outcomes, and effectively challenge the same. By creating such a mechanism, we can enable both the right to information as well as the right of appeal and redress. Therefore, the need for appropriate transparency is crucial for the successful achievement of accountability.

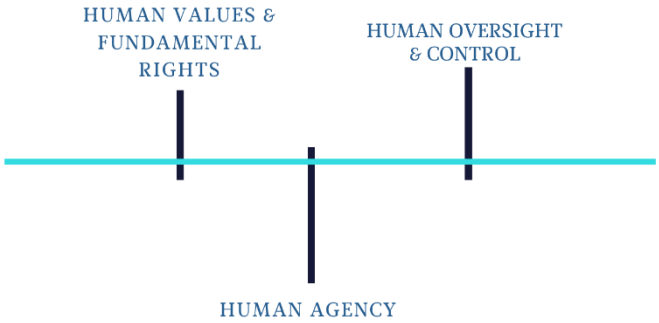
3. Accountability



One of the most important requirements of this theme is that only human beings can be held accountable or liable for the decisions made by an AIS. As such, only the AI actors are responsible and accountable for the proper functioning of the AIS based on their roles and context. This is a fair solution as the threat of liability will place the AI developers under a greater burden to develop a system that does not impact human beings unfairly.

Although it is not entirely clear if the guidelines recommend accountability through judicial mechanisms, they offer guidance on how accountability can be achieved through other non-judicial means. This includes auditability of the AIS (whether internally or externally), human oversight throughout the AIS's lifecycle, identifying, assessing, documenting and minimising the potential negative impacts of AIS. This is necessary especially for those who may be indirectly impacted by the performance of an AIS. Where any negative impacts are identified or if they materialise, they are to be reported to those who have been impacted. In this regard, human rights impact assessments are becoming an important part of the current discourse on trustworthy and responsible AI.

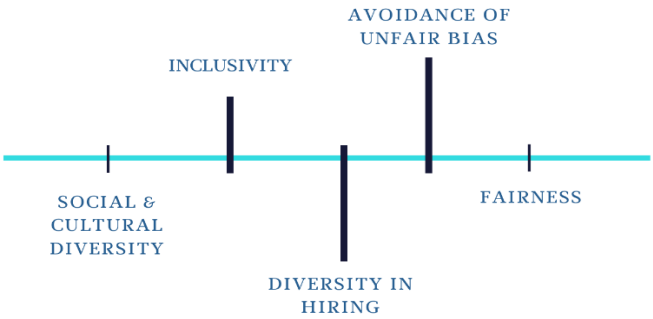
4. Human-centric & human control



There is a large consensus among the many stakeholders of AI that all AI technology must promote and respect human values, human autonomy and agency, democracy, rule of law, fundamental freedoms, and social justice. This theme is made central to all AI guidelines as this is what ought to create AIS

that is more trustworthy and responsible. For AIS to be more human-centric then, we need to increase human involvement throughout the lifecycle of the AI so that each stage of the AIS's lifecycle remains within human control. Governance mechanisms through approaches such as human-in-the-loop, human-on-the-loop, and human-in-command can be established to retain human control.[124] What these approaches mean is a task for the AI experts to determine by demonstrating *how* such approaches can be implemented and to what extent. However, one thing that this definitely means in the context of automated decision-making is that capacity must be created to enable a human being as the final decision maker.[125]

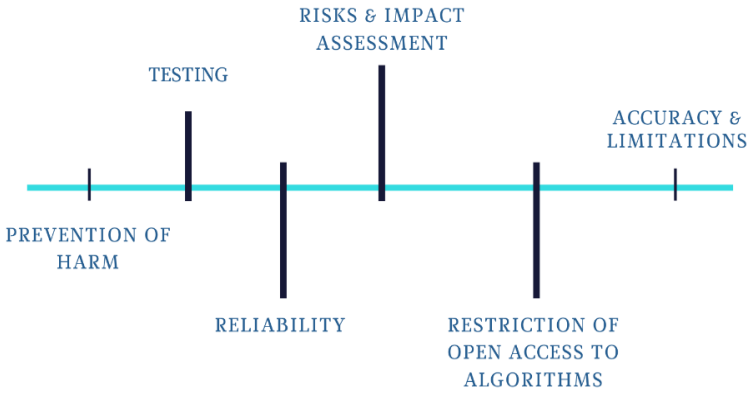
5. Diversity & Inclusion



One of the most important ways to avoid algorithmic bias is by increasing diversity and inclusion throughout the lifecycle of the AIS; diversity in society must be reflected in the AIS through inclusion of diverse opinions and practises from the moment algorithms are conceived till the very deployment of the AIS. This can be achieved by hiring people from diverse backgrounds. As algorithmic decision-making, as well as the broader AI challenge, is at its heart a socio-technological problem,[126] this theme attempts to deal with the issue of bias and discrimination at its root. This is considered to be a key component for protecting and upholding the right to equality and non-discrimination,[127] as it can avoid unfair bias.[128]

[124] Supra note 97, p. 16.
[125] Ibid.
[126] Microsoft 'Responsible AI Principles from Microsoft' (*Microsoft*) <https://www.microsoft.com/en-us/ai/responsible-ai> accessed 22 June 2021.
[127] Supra note 108, p. 6.
[128] Supra note 97, p. 18.

6. Safety & Security



The theme aims to prevent any homogenisation of society through standardisation of behaviour and opinions.[129] Through inclusion and diversity of underrepresented and marginalised groups, and other diverse groups can actively participate and meaningfully contribute during the design, development, deployment and use of the AIS to ensure that the systems are rights-respecting.[130] This theme is closely connected with the theme of fairness as diversity and inclusion also mean that equal treatment must be afforded to all diverse groups, and equal access must be granted through inclusive design.[131]

The underlying aim of safety and security is the principle of *prevention of harm* to develop systems that are more reliable and trustworthy.[132] Any AIS must prevent harm and behave in a manner that it was intended to. Such an approach is meant to minimise unintended and unexpected harms from AIS.

The implementation of this approach is largely dependent on rigorous testing, risk assessments, auditing, and pre-release trials etc. in controlled environments to facilitate an agile transition from the research and development (R&D) phase to the deployment phase.[133] Such testing can assist in the identification of deliberate or inadvertent biases which can then be quantified and mitigated.

Such approaches can also aid in determining the ability of the AIS to accurately make predictions or judgements and decisions over a range of inputs in a range of situations.[134] In the event that occasional inaccurate predictions cannot be avoided, the likelihood of the occurrence of such errors must be made known as limitations of the AIS.[135]

[129] Supra note 109, p. 14.
[130] Supra note 108, p. 6.
[131] Supra note 97, p. 12.
[132] Ibid, p.16.
[133] Supra note 105, p. 9.
[134] Supra note 97, p. 17
[135] Supra note 108, p. 9.

Private tech-companies

As no regulatory framework for AI exists at state or international level yet, the tech-giants have formulated their own internal ethics policies for AI. The nature of these policies is self-regulatory; something that is also supported by the European Commission for the time being.[136] Although there are common themes among the policies of the tech-giants, what a particular theme means to a tech firm is dictated by the personal values of that firm. There is no uniformity in their definitions, however, broadly, all hope to achieve the same goal of creating responsible AI. Some definitions are broad and vague, while others place the firms under strict ethical expectations.

We reviewed the policies of Google,[137] IBM,[138] Microsoft[139] and Tencent[140] in the context of algorithmic bias and discrimination, and the common themes among them in that context were *fairness*,[141] *reliability and safety*,[142] *privacy*,[143] *transparency*,[144] *accountability*,[145] and a *human-centric approach*.[146] These themes are not only common between the internal policies of the tech-giants, but they are also common with the inter-governmental and non-governmental guidelines that were analysed in the previous section.

Some uncommon themes among the tech-giants that are common with the inter-governmental and non-governmental guidelines are *inclusiveness*,[147] *socially beneficial AI*,[148] and *value alignment*.[149]

Legal Efforts to Regulate AI

1. EU's General Data Protection Regulation

As highlighted earlier, an overlap exists between data privacy and data bias when it comes to the *collection* and *use* of data. This overlap has explicitly been dealt with in the GDPR. This is the EU's most decisive effort to legally protect users and their personal data consistently all across the EU, but

[136] Supra note 101.
[137] Google, 'Our Principles' (*Google AI*) <https://ai.google/principles/> accessed 22 June 2021.
[138] IBM, 'AI Ethics' (15 June 2021) <https://www.ibm.com/artificial-intelligence/ethics> accessed 22 June 2021.
[139] Microsoft 'Responsible AI Principles from Microsoft' (*Microsoft*) <https://www.microsoft.com/en-us/ai/responsible-ai> accessed 22 June 2021.
[140] Tencent ,'"ARCC": An Ethical Framework for Artificial Intelligence' <https://www.tisi.org/13747> accessed 22 June 2021.
[141] Supra note 137; Supra note 138; Supra note 139; Supra note 140.
[142] Supra note 137; Supra note 139; Supra note 140.
[143] Supra note 141.
[144] Ibid.
[145] Supra note 137; Supra note 138; Supra note 139.
[146] Supra note 141.
[147] Supra note 139.
[148] Supra note 137.
[149] Supra note 138.

its impact has also crossed borders.[150] The GDPR is largely open-ended as it aims to keep up with the ever-changing nature of technology, and the legislation has been promulgated to respect and protect fundamental freedoms, especially the right to privacy.[151]

The GDPR has made it to the list of the international best practises in this report as it has brought some of the aforementioned "*ethical*" principles to a legal footing. First and foremost, it has legally dealt with automated decision-making under Article 22, Section 4 [Rights of the Data Subject], GDPR:

Automated Individual Decision Making, including profiling

"The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."

Specific guidelines on "automated individual decision-making and profiling" have been developed and adopted by the Article 29 Data Protection Working Party on the Protection of Individuals with regard to Processing of Personal Data.[152] According to the guidelines, this provision is interpreted as there being a general prohibition on automated decision making instead of a right that has to be invoked by individuals. This is so that individuals can automatically be protected from the potential effects of such processing.[153] However, if the processing falls under the exceptions listed in Article 22(2), then Article 22(3) is triggered which requires that there be safeguards where processing is done in accordance with Article 22(2). The safeguards include the *right to be informed* (such as meaningful information about the logic involved, as well as the significance and envisaged consequences for the data subject), *right to obtain human intervention* and the *right to challenge the decision.*[154] The guidelines have also defined the criteria for ensuring how these rights are to be given effect. The three safeguards are equivalent to the *ethical* principles of transparency and explainability, human control, and accountability discussed in an earlier section.

[150] The legislation applies to all businesses that interact or do business with EU Citizens. Moreover, there is also a persuasive transnational impact of the legislation. For instance, in Pakistan, the Data & Privacy Protection Procedures (DP3) for the Punjab Safe Cities Authority claim to be in line with the GDPR 2016/679 of the European Union.
[151] GDPR Article 1(2); Recitals 1, 2 & 4.
[152] WP251/2017 last revised 2018-19: <https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053>
[153] Ibid, p.20.
[154] Ibid.

Apart from this, the Working Party has also identified other general provisions that also apply in the context of automated decision-making, such as *lawful, fair and transparent*,[155] *accuracy*,[156] *right to object*,[157] which are also very reminiscent of the *ethical* principles discussed earlier.

Although there is some criticism about both the GDPR as a whole and Article 22 in specific[158] (an analysis beyond the scope of this report), we consider the legal recognition of automated decision-making and profiling and general prohibition of the same as a step in the right direction as it has legally concretised principles that, according to global discourse, reside in the ethics realm. Moreover, it also adds some certainty to the broad discussion on "*what must be done*" by showing that *something* can definitely be done!

2. *Proposal* for a Regulation of The European Parliament and The Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.

In April 2021, the European Commission put forward a draft proposal for the legal regulation of AI.[159] The proposal lays down a legal and regulatory framework to enable development of AIS that Europeans can trust. It addresses the concerns of EU stakeholders that legislative gaps exist in the EU to regulate AI and the need to either amend existing laws or create new legislation. In addition, the proposal is also a response to the consistent calls from the European Parliament and European Council to create a legal and regulatory ecosystem for AI so that risks and benefits of AI can be addressed at the Union level, and the well-functioning of the internal market can be ensured.

This recent development addresses numerous risks of AI, including algorithmic discrimination. One of the key objectives of the proposal is to address opacity, bias, unpredictability and partially autonomous behaviour of certain AIS.[160] In response to the consensus among the stakeholders to define AI narrowly and precisely, the proposal has provided a "future-proof" definition of AI, and a proportionate risk-

[155] Article 5(1) (a) GDPR.
[156] Article 5(1) (d) GDPR.
[157] Article 21 GDPR.
[158] For a thorough impact of the GDPR on AI, see: Giovanni Sartor and others, 'The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence: Study' (European Parliament, 2020) <http://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf> accessed 21 June 2021
[159] European Commission, 'Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS COM(2021) 206 Final' <https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF> accessed 6 July 2021.
[160] Ibid, p.2.

based approach has been put forward to categorise AIS based on the risk they pose. Risk categories include:

(a) Unacceptable Risk: AIS that clearly threaten the livelihood of people have been proposed to be banned. This includes AIS that manipulate human behaviour or circumvent user's free will, such as toys with voice assistance encouraging dangerous behaviour among minors, live biometric identification systems in publicly accessible spaces and social scoring by governments.[161]

(b) High Risk: AIS that pose a high risk are proposed to be subject to strict obligations before they can be put on the market. Obligations include risk assessments and mitigation systems, high quality data sets, logging of activity to ensure traceability, clear and adequate information, human oversight, high level robustness, security and accuracy. A key development in this group is the classification of all remote biometric identification as high risk AIS due to their potential to violate fundamental rights, particularly human dignity, privacy, protection of personal data and non-discrimination. Therefore, their use is prohibited, with a few narrow exceptions. Examples of high risk AIS include scoring of exams, CV-sorting softwares for recruitment, credit scoring for loans, among others.[162]

(c) Limited Risk: Due to the limited risk they pose, the draft regulation proposes that such AIS be subject to specific transparency obligations. For instance, if a user is interacting with a chat bot, then they should be made aware that they are not interacting with a human so that they have the opportunity to decide whether to continue or step back.[163]

(d) Minimal Risk: As the risk posed by such AIS is only minimal, such as spam filters, they are proposed to not be regulated as they do not threaten the rights and safety of citizens.[164]

The draft regulation is pending review and approval of the European Parliament. Once approved, the regulation will be adopted by the European Parliament and the Member States, and will become directly applicable across the EU.[165]

---

[161] 'New Rules for Artificial Intelligence – Q&As' (*European Commission*) <https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683> accessed 6 July 2021.
[162] 'Europe Fit for the Digital Age: Artificial Intelligence' (*European Commission - European Commission*) <https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682> accessed 6 July 2021.
[163] Ibid.
[164] Ibid.
[165] Ibid.

Automated decision-making across various sectors in Pakistan

Automated decision-making across various sectors in Pakistan

This part of the report maps out the development, deployment, and use of intelligent data-driven algorithms to make decisions across various state-run institutions in Pakistan. This section also highlights some of the problems associated with deployment of AI in these state-run institutions as well as some of the pre-existing biases and discriminatory practises which can be reinforced through algorithms in each sector. A lot of AI designing, development and deployment in Pakistan is being supported through targeted federal funding and through public-private partnerships so as to promote homegrown players in AI, IoT, and Robotics.[166] The mapping of AI use in the Health Sector, Judiciary, and the Criminal Justice System (CJS) has been done through expert input provided by computer science researchers at the National Center of Artificial Intelligence; an AI research institute established by the Government of Pakistan in support of the Digital Policy which is quickly becoming the hub for scientific research in Pakistan. NCAI has nine operational labs in some of the top engineering and technology universities in the country. It is working under a grant of the Higher Education Commission and aims to commercialise AI technology that NCAI is developing.

A. Health Sector

This is one of the sectors where research and development of AI driven decision-making is rapidly emerging in Pakistan, especially after the widespread transmissions of Covid-19. Development of AI in the health sector of Pakistan is being led by the Medical Imaging & Diagnostics Lab at NCAI and current federal research in this area is targeted towards developing automated systems for diagnosing tuberculosis, breast cancer, brain tumor, and the Covid-19.[167] The direct beneficiaries of this technology are hospitals, radiologists, and citizens.[168]

These intelligent automated systems for medical diagnosis work in much the same way a medical practitioner would, such as a radiologist, as the systems are modelled on the same.[169] For example, a doctor would first receive supervised training in becoming a radiologist and would be exposed to different cases of tuberculosis, breast cancer, and brain tumor and would learn to make a diagnosis. After the training is complete, the radiologist would then be exposed to new cases and would be able to make a diagnosis without any supervision. Similarly, algorithms are also provided with a lot of training data where they *learn* to make the diagnosis through supervised learning so that errors in diagnosis can be fixed by the computer scientists

[166] Policy Guideline 17, Section III, 'Enabling the Digitization of key socio-economic sectors', Digital Pakistan Policy 2018, p. 16.
[167] KSI with Medical Imaging & Diagnostics Lab, NCAI.
[168] Ibid.
[169] Ibid.

while the algorithm is being trained. Then the algorithm is provided with test data where the intelligent automated system makes decisions based on what it learned during the training phase.[170] This technology is currently in the commercialisation phase, however, hesitancy at the end of the hospitals has been observed due to misplaced fears that the machines would replace humans completely.[171]

Potential Risks

Data related issues that can impact algorithmic fairness are quite a few in this area. Currently, due to a lack of standard national policy on data privacy or data sharing in Pakistan, there is a general hesitation among hospitals in sharing data for development of intelligent automated systems. As a result, hospitals in bigger cities are unwilling to share data, and villages have no data to provide due to unavailability of machines there. Unavailability of machines, however, is not an issue unique only to villages, but is also faced by some hospitals in provincial capitals. Data is directly available only from smaller cities that are forthcoming in data sharing, however, the data can still not be said to be representative of the target population.[172]

Moreover, the lab is not involved in data collection;[173]collected data is provided to the lab. As we do not know who collects the data, we cannot comment on whether or not data is sampled and disaggregated according to what it represents. For instance, in the health sector there is a possibility for pre-existing gender biases to creep into the algorithms through biased data. Due to historical patriarchal structures, historical data has gender biases. There is a large data gap, with availability of predominantly male data, even in the health sector.[174] Medical research has shown that diseases impact men and women differently as "*there are biological differences between male and female immune systems at every biological level, from cell to organ-to-organ system to individual as a whole*",[175] and this may lead to different symptoms for men and women for the same disease. In tuberculosis, for example, some research has shown that TB lung lesions may not appear as severely in women as in men.[176] In a heart attack, women may experience atypical symptoms such as pain in the jaw, neck, shoulder blades, and also fatigue,

[170] Ibid.
[171] Ibid.
[172] Ibid.
[173] Ibid.
[174] Caroline Criado Perez, 'Chapters 11 & 12', *Invisible Women: Exposing Data Bias in a World Designed for Men* (pp. 193-235, Vintage 2020)
[175] 'Taking Sex and Gender into Account', (World Health Organisation, Western Pacific Region) <https://iris.wpro.who.int/bitstream/handle/10665.1/7977/9789290615323_eng.pdf>; See also: Supra note 174.
[176] 'Women and Tuberculosis, Taking Action against a Neglected Issue' <https://www.action.org/uploads/documents/2010_womenTB.pdf>. See also: Supra note 174.

dizziness and palpitations instead of the more common chest and left-arm pains.[177] As a result, women are frequently misdiagnosed if their symptoms do not conform to that of men.[178] To this end, the WHO cautions against making the frequent mistake of not accounting for the importance of symptoms that may occur in just one sex, and of using symptom profiles for one sex to form as a basis for diagnosis of all sexes.[179]

Therefore, if the data is not sampled and sex-disaggregated to determine whether or not it is equally representative of all genders, the algorithm may be biased towards one gender more than the other. This may result in misdiagnosis even where its performance is accurate because it may not matter that the algorithm carries a minor inaccuracy so long as it performs well on the majority, but the result of this minor accuracy may just be an unfair bias towards a particular gender. Similarly, there can also be other factors that may lead to biased data, such as age. If these biases are perpetuated through the use of algorithms, then this can potentially violate not only the right to equality and non-discrimination, but also the right to life.

In an attempt to minimise bias, the Medical Imaging and Diagnostics Lab follows the international best practice of anonymising the data. All private and sensitive information of patients that is not relevant or necessary for the working of the algorithm is anonymised for better performance of the algorithms. However, even so, the Lab cautioned that if data is not representative of the population from all regions, there is a possibility of biases creeping in. While it is commendable that the Lab is following some international best practises, we recommend that greater efforts be made in mitigation of bias in every component of the algorithm and at every stage of the lifecycle of the AIS that are deployed in this sector, and also carry out extensive testing to prevent unintended outcomes.

B. Criminal Justice System (CJS)

The purpose of CJS is to detect, prevent, and mitigate crimes, and in the achievement of this purpose, the police, lawyers, and judiciaries work together to control crime and guard the safety of the citizens. Around the world, countries are investing in building “Safe Cities" by using technology to assist the actors involved in the CJS to achieve the purposes of the CJS more

[177] ‘Sex Differences in Heart Disease: A Closer Look’ (*Harvard Health*, 1 October 2020) <https://www.health.harvard.edu/heart-health/sex-differences-in-heart-disease-a-closer-look> accessed 21 June 2021

[178] Caroline Criado Perez, ‘Chapter 11: Yentl Syndrome’, *Invisible Women, Exposing Data Bias in a World Designed for Men* (p. 221, Vintage 2020)

[179] Supra note 174, p. 55.

efficiently.[180] For instance, AI is being used by the Police for 'predictive policing',[181] by the Prosecutors to predict outcomes of cases,[182] and by the Judiciary to carry out risk assessments of offenders.[183] Despite being well-intentioned, shortcomings of these AIS have been widely reported.[184] Pakistan is one of the many countries that has taken the initiative to build Safe and Smart Cities as a response to growing ubranisation, globalisation, terrorism and natural disasters.[185] The Safe City project aims to transform the police culture in Pakistan through use of state-of-art technologies.[186] The Intelligent Criminology Lab at the NCAI is developing the technology to advance the purpose of the Safe Cities.[187]

Below are examples of the AI which is being developed by the Lab to serve the three-fold purpose of the CJS:

(a) Crime Detection: To serve this purpose, AI is being developed for facial recognition, detection of forgery, suspicious activity, suspicious objects, anomalous crowd behaviours, and graphic surveillance such as keeping an automated record of the number of people entering particular area or vehicle analytics, among others.[188]

(b) Crime Prevention: AI is being deployed for crowd and social media analytics. The former is related to detection of anomalous crowd behaviours and the prevention of the same, whereas the latter addresses the issues of hate speech and social media monitoring through the use of AI algorithms.[189]

(c) Crime Mitigation: AI solutions are being developed to mitigate the crimes by studying and analysing virtual crime environments.[190]

---

[180] 'Safe Cities Index 2019' (*Safe Cities, The Economist*) <https://safecities.economist.com/, https://https://safecities.economist.com/> accessed 22 June 2021.
[181] Supra note 45.
[182] Irma Isabel Rivera, 'The Implementation of New Technologies under Colombian Law and Incorporation of Artificial Intelligence in Judicial Proceedings' <https://www.ibanet.org/article/14AF564F-080C-4CA2-8DDB-7FA909E5C1F4> accessed 22 June 2021.
[183] 'Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System' (*The Partnership on AI*, 23 April 2019) <https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/> accessed 22 June 2021.See also: Marion Oswald and others, 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and "Experimental" Proportionality' (2018) <https://doi.org/10.1080/13600834.2018.1458455>.
[184] For the shortcomings of Predictive Policing, see: Supra note 45. For the risks posed by risk assessments of offenders in the USA and the UK, see: Supra note 183.
[185] 'Concept Paper: The Future of Punjab Police', (pp. 3-7, *Punjab Safe Cities Authority*, August 2015)
[186] <https://psca.gop.pk>
[187] KSI with Intelligent Criminology Lab, NCAI.
[188] Ibid.
[189] Ibid.
[190] Ibid.

With an unprecedented increase in sharing and accessing information on the internet in combination with the opportunity to remain anonymous, hateful and offensive content against people of colour, Muslims, women, and other vulnerable groups has become widespread.[191] Therefore, on virtual platforms that have millions of users, it is now more urgent than ever to prevent the spread of hateful content, and the Intelligent Criminology Lab is working towards such prevention by using intelligent algorithms to automatically police hateful and offensive content by detecting the same in text, image and video data.[192]

Potential Risks

"Hate" is a sensitive area; what constitutes hate cannot be defined and definitions that are present at law are wide and vague. Hence, whether or not something amounts to hate is best determined on a case-by-case basis especially because what is considered offensive is subjective and depends largely on the social context and has to be constantly balanced with the freedom of expression. A recurrent issue with free speech is the *judicial* determination of whether or not it has reached a breaking point beyond which it is not protected constitutionally and constitutes a hate crime.[193] Due to the complexities involved in determining what the breaking point is, it is not possible to have a standard definition for hate speech for use in all circumstances.

For this reason, we must be careful in deployment of intelligent algorithms for detection of hate speech as they carry a high risk of perpetuating pre-existing biases that could negatively impact the same groups that such systems are designed to protect, thus potentially leading to further marginalisation of minority groups. In the Pakistani context, this may mean a proliferation of wrongful blasphemy accusations through digital means. At the same time, extreme care must be taken in ensuring that a proper balance is struck with the freedom of expression. For instance, in a recent study it was discovered that black-aligned tweets were classified as racism, sexism, hate speech, harassment, and abuse at higher rates than white-aligned tweets.[194]

---

[191] György Kovács, Pedro Alonso and Rajkumar Saini, 'Challenges of Hate Speech Detection in Social Media' (2021) 2 SN Computer Science 95.
[192] KSI with Intelligent Criminology Lab, NCAI.
[193] 'The Ongoing Challenge to Define Free Speech' Vol 43, No. 4 (American Bar Association Human Rights Magazine)<https://www.americanbar.org/groups/crsj/publications/human_rights_magazine_home/the-ongoing-challenge-to-define-free-speech/the-ongoing-challenge-to-define-free-speech/> accessed 21 June 2021.
[194] Thomas Davidson, Debasmita Bhattacharya and Ingmar Weber, 'Racial Bias in Hate Speech and Abusive Language Detection Datasets' [2019] arXiv:1905.12516 [cs] <http://arxiv.org/abs/1905.12516> accessed 21 June 2021.

One reason for this was because some terms which are considered racial slurs when used by outsiders are not considered racial slurs when used by members of the black community themselves. As such terms were present more often in black-aligned texts, they were more likely to be flagged. This shows the importance of social context and highlights the challenge in detecting hateful content. By mistakenly considering speech by a targeted minority group as abusive we run the risk of unfairly penalising the same, but failure to identify abuse against them will render us unable to take action against the perpetrator.[195] Researchers have cautioned that even though no model can entirely avoid such problems, we still need to be cognisant of the fact that such models also carry the potential to be systematically biased against certain groups.[196] With this in mind, we must consider as a society whether detection of hate speech is a suitable task for AI which does not carry the cognitive power to balance out conflicting interests and rights, such as freedom of expression versus freedom to profess religion, as algorithms lack the ability to understand the nuances of language in different "social contexts".

C. Judiciary

In 2019, former Chief Justice of Pakistan Asif Saeed Khosa announced that the Supreme Court of Pakistan was going to make use of Artificial Intelligence to aid decision-making.[197] This project is being developed by the Deep Learning Lab of NCAI. The project is still at its preliminary stages and there is still a long way to go before it can be deployed due to the complexity of the project.[198] In any case, the initiative is very well-intentioned and can greatly assist in the revival of speedy justice in Pakistan.

The project involves building an intelligent knowledge-based system that will have all cases that have been decided in the history of Pakistan. Judges from anywhere in the country will be able to input the facts, and the system will generate similar precedents and their judgements, and this similarity will be determined based on data points such as citations, case names, court names, among others. The intelligent system will also be able to generate summaries of the judgement, and assist in judgement writing as well. Based on the knowledge it has from the facts and previous precedents and the analysis it has made, it will also be able to *recommend* to the judge what a decision *could be*, but the judge will not be bound by any such recommendations.[199] For this latter part, the development of this technology does not entail replacement of human judges with technology because legal judgements require a higher level

[195] Ibid.
[196] Ibid.
[197] Dawn.com, 'Artificial Intelligence to Help Judiciary's Performance: CJP' (*DAWN.COM*, 19 June 2019) <https://www.dawn.com/news/1489143> accessed 21 June 2021.
[198] KSI with Deep Learning Lab, NCAI.
[199] Ibid.

of cognition and technology is not yet sophisticated enough to exercise the kind of cognition required of judges.[200] Rather, the point is simply to aid judges, and make the entire process from research to judgement more consistent.

At the moment, the Deep Learning Lab has taken a top-down approach in the development of the intelligent database. As such, currently the database is being developed with decided cases from the apex and high courts as their data is more easily available, and later, cases from lower courts will also be added so that an end-to-end system can be created. After the completion of the first phase, the system will continue to be fine-tuned as new data becomes available.[201] This intelligent system, once developed, will minimise the laborious human effort required in legal research, and help the legal industry to become more efficient.[202]

Potential Risks

Although the development of this AIS for the judiciary represents a good-use-case to the extent of the development of an intelligent database of all precedents, the second phase of the project where the system can also provide recommendations on what the judgement ought to be is where a number of problems can arise.

An associated risk with intelligent recommendations on decisions is that of value lock-ins and stagnation of law.[203] This problem refers to a state where a particular value perpetuates a particular state of existence which may become permanent and "lock-in" the society to a particular status-quo.[204] The human rights development in Pakistan has been very slow and Pakistan's social challenges have been reflected in its laws and judicial judgements. A great example comes from Pakistan's rape laws. Previously, the Qanun-e-Shahadat Order 1984 (QSO) contained sub-clause 4 under section 15, which permitted the evidence of the victim's *immoral character* during the rape prosecution of a man.[205] Moreover, evidence of bad character went hand-in-hand with the two-finger test which would determine the virginity and sexual history of the survivor. This evidential requirement has historically served as a basis for acquittal of rapists in Pakistan, and has given rise to the notion that rape can only be found

[200] Ibid.
[201] Ibid.
[202] Ibid.
[203] Ameen Jauhar and others, 'Responsible Artificial Intelligence for the Indian Justice System' (Vidhi Centre for Legal Policy, April 2021) <https://vidhilegalpolicy.in/research/responsible-ai-for-the-indian-justice-system-a-strategy-paper/>
[204] Ibid.
[205] QSO 1984, Section 151 (4): *"when a man is prosecuted for rape or an attempt to ravish, it may be shown that the prosecutrix was of generally immoral character."*

where a survivor is a virgin prior to rape.[206] In 2016, the QSO 1984 was amended to omit sub-clause 4 of section 151 by the Criminal Law Amendment (Offences Relating to Rape) Act 2016, and only as recently as January 2021, the Supreme Court of Pakistan declared sexual history of the survivor and virginity testing as unconstitutional.[207]

However, despite the amendment in 2016, in practise, judges have continued to rely on the survivor's "bad character". The Centre for Human Rights, in a case study conducted in 2018 of rape cases in the lower judiciary of the city of Lodhran, discovered that character of the survivor continued to play an integral part in acquittal of rapists whereby in 10 out 63 cases (16%), judges explicitly relied on the survivor's character and in the remaining 53 cases (84%), judges made indirect references to the survivor's character.[208] For example, in a 2017 case, the judge noted that the victim was "*a lady of easy virtue and hence corroboration of her statements was needed…*"[209] and also "*I do agree with the learned counsel for prosecution that simple penetration is sufficient and positive reports of swabs and DNA are not a requirement of law but I would say that when the victim is enjoying bad character, then these things are important to believe her version*."[210]

Even though the law has been amended, the amendments are quite recent, and the cases which have made evidential requirements more survivor-centric are very much in the minority. The better part of Pakistan's judicial history is flooded with judgements which blame the survivor and use the sexual history of the survivor as a justification for acquittals. They make up the majority of the cases, and there is a possibility that the algorithm may "lock-in" sexism and reinforce it through its recommendations on what a judgement ought to be. Therefore, as the AIS is still being developed and will take some time before it can be deployed, we believe that we have an opportunity to minimise the human biases that may be perpetuated in this area, and make our best attempt to balance the design and development with human rights.

Another risk associated with AI's assistance with judicial decision-making which has been highlighted in literature and also in a recent thought-provoking report by the Vidhi Centre for Legal Policy is related to the constitutional role of judges and separation of powers. They have aptly pointed out that AI cannot carry out careful and reasoned decision-making that experienced judges can. This is even more so in cases of judicial review where the judiciary

[206] Fatima Y. Bokhari and Sevim Saadat, 'Accountability for Rape: A Case Study of Lodhran' (Centre for Human Rights, 2019).
[207] Criminal Appeal No. 251 / 2020 & Criminal Petition No. 667 / 2020.
[208] Supra note 206.
[209] Case FIR No. 195/2017, Sessions Case No. 124/OSC of 2017 (Para. 26, 21)
[210] Ibid, para. 23.

has to balance out competing interests by exercising innovative thinking. The cognitive ability of human judges comes from experience on the bench and continuous exposure and engagement with the law.[211] This enables judges to appreciate and approach facts and issues of different cases more holistically. Landmark precedents such as those set in **Shehla Zia vs. WAPDA**[212] which, through extraordinary legal reasoning, held that the entitlement of citizens to a non-hazardous environment was a part of their constitutional freedom of right to life, is a testament to the cognitive abilities of judges that are learned through experience. Moreover, it was due to this experience on the bench that US judges in **Brown vs. Board of Education** were able to determine, through judicial reasoning, the intrinsic inequality of the "separate but equal" doctrine and declared racial segregation in schools as unconstitutional. As AI lacks such cognition and human experience, its recommendations may not be very out of the box.

Moreover, every now and then the law is confronted with *novel* situations, i.e., cases representing unique facts that have not come to the court before. For novel situations, data can never be available to train the algorithms on, which means the algorithm will not have seen every possible combination of facts.[213] In such situations, reliance on the algorithm's recommendation may lead to injustice. Even though it has been clarified that the judge will exercise the ultimate discretion in whether or not to follow the recommendation, there is danger that the judge may be psychologically compelled to rely on a recommendation provided by sophisticated technology exhibiting human-like intelligence.[214]

Therefore, any technological development in this area ought to research other biases that may exist in other laws, apart from the bad rape law discussed above, which can potentially be reinforced through intelligent systems so that people can trust these systems. We must make our best efforts to ensure that design, development, deployment and use of intelligent systems in the judicial sector must function on the idea of natural justice and uplift the rule of law, and not undermine it. To this end, the Deep Learning Lab has stressed the necessity for a multi stakeholder approach in development of intelligent technologies for the promotion of the common good.

[211] Supra note 203.
[212] PLD 1994 SC 693.
[213] KSI with MeVitae.
[214] Ibid.

The Constitutional, Policy & Legal Landscape of Digital Pakistan

The Constitutional, Policy & Legal Landscape of Digital Pakistan

In this section, we attempt to contextualise the global themes to Pakistan's landscape. The reason for carrying out a legal analysis instead of an ethics one is because a number of ethical principles that are proposed in the ethics guidance documents are already well-recognised legal principles. We acknowledge that tensions can exist between laws and ethical norms, however, we believe that such tensions must be resolved by human rights law, and not just ethics, as human rights law is binding and grounded in ethics in any case.

We first analyse the Constitution of the Islamic Republic of Pakistan 1973 and IHRL to demonstrate the nature of Pakistan's human rights obligations to understand why Pakistan needs to take cognisance of algorithmic unfairness. We shall then contextualise the global themes by carrying out an analysis of Pakistan's Digital Policy to determine whether or not algorithmic bias is a key consideration as we make technological advancements as a nation. This shall be followed by an analysis of Pakistan's current cyber laws to determine to what extent, if at all, they are reflective of (a) algorithmic bias, and (b) if they have pre-existing elements to support legal recognition of algorithmic justice. The rationale behind this determination is to know whether or not we can rely on the provisions and objectives of current legislation to deal with algorithmic bias *or* re-interpret the current provisions to make room for algorithmic fairness *or* are we in need of new laws and policies altogether.

A. Understanding Pakistan's Human Rights Obligations

"The Constitution is not merely an imprisonment of the past but is also alive to the unfolding of the future."

Chief Justice Muhammad Haleem [215]

This report adopts a *legal* conception of human rights, and supports a legal rights-based approach to the development of intelligent data-driven technologies in Pakistan. This chapter is set in the belief that the standards of obligation imposed by human rights law along with its broad principles which are not only grounded in morality and ethics, but also carry legal force,[216] offer the right type of flexibility to formulate the pathway to accommodate the 4th Industrial Revolution in Pakistan. As new technology

[215] Benazir Bhutto v Federation of Pakistan [PLD 1988 SC 416]; A similar view was also endorsed by the Supreme Court of Pakistan in Muhammad Nawaz Sharif v Federation of Pakistan [PLD 1993 SC] where Justice Frankfurter in Sweezy v Hampshire was quoted, "*While the language of the Constitution does not change, the changing circumstances of a progressive society for which it was designed yield new and fuller import to its meaning."*
[216] Aryeh Neier, *International Human Rights Movement: A History*, (p. 57, Princeton University Press 2012).

has the potential to reinforce pre-existing risks to human rights, and also create new ones, there should be a reconceptualisation of existing human rights as well as a willingness to recognise new ones, when necessary, in order to deal with the challenges that intelligent data-driven technology is raising and will continue to raise. As we globally undergo a technological revolution, we must allow domestic Constitutions and IHRL to actively assist "*the unfolding of the future*" through a long established, flexible, and binding framework. As algorithmic bias is the focus of this report, the following discussion is based on the legal right that algorithmic bias has most potential to violate; the right to equality and non-discrimination.

Constitution of the Islamic Republic of Pakistan

The nature of the fundamental rights as enshrined in the Constitution is both restrictive and preventive. The fundamental rights not only impose a restriction on the arbitrary exercise of power in relation to an activity that an individual can engage in,[217] but also impose upon the State a positive obligation to protect fundamental rights.[218] This means that the Constitution should not only be interpreted with reference to what *shall not* be done, but also with reference to what *should* be done. The Constitution obliges the State to take effective measures to *prevent* any violations of the fundamental rights. In this regard, the obligation is to safeguard fundamental rights by creating laws which are in line with the Principles of Policy, and uphold fundamental rights.

Article 25: Equality of citizens

(1) All citizens are equal before law and are entitled to equal protection of law.

(2) There shall be no discrimination on the basis of sex.

This fundamental right in the Constitution is derived from the American Constitution,[219] and judicial interpretation of the right combines both American and English concepts, therefore, jurisprudence of both is relevant.[220] The American jurisprudence has been given effect under *Brig F.B. Ali vs. The State* by *Hamoodur Rehman CJ* to the extent that no rule can be formulated to cover every case that involves this fundamental right. It would be both impractical and unwise to do so because it would generalise

[217] Muhammad Nawaz Sharif vs. Federation of Pakistan [PLD 1993 SC]
[218] Shehla Zia vs. Wapda [PLD 1994 SC 693]; Suo Motu No. 16 of 2011 [PLD 2011 SC 979].
[219] Chief Justice Hamoodur Rahman in Brig F.B. Ali vs. The State [PLD 1975 SC 506].
[220] Justice Fazal Karim, 'Article 25: Right to Equality', *Judicial Review of Public Actions*, vol II (p.1330, Second Edition, Pakistan Law House).

the right.[221] Therefore, each case is to be decided as it arises.[222] Moreover, this right is a fundamental value of every democratic society as it is based on the considerations of justice and fairness.[223]

Even though Article 25 is framed without reference to specific characteristics such as colour, race, age, caste, ethnicity, etc., it still protects the same because the Article is phrased in general and universal terms, and its universality is to be given effect under the law.[224] This also ties in well with the constitutional requirement to interpret constitutional freedoms liberally so that the Constitution is able to continue to embolden freedom, equality, tolerance and social justice.[225] In any case, a lack of reference to these characteristics is substituted by Pakistan's obligations under the Universal Declaration of Human Rights (UDHR), International Covenant on Civil & Political Rights (ICCPR) and International Covenant on Economic, Social & Cultural Rights (ICESCR) which specifically protect certain characteristics.

Although there are no reports of discrimination as a result of algorithmic bias in Pakistan at the time of writing this report,[226] we ought not to wait for a violation of the right to equality before we take cognisance of this problem (both at law and otherwise). This is because our Constitution places the State under a positive obligation to *prevent* violation of this fundamental right. The case by case approach and the universal phrasing of Article 25 both offer the right type of flexibility to deal with algorithmic bias at law because the word "*bias*" is of a much wider import than the word "*discrimination*".[227] While not all biases may constitute discrimination in a legal sense, algorithmic bias does have the potential to introduce new biases that can skew the outcomes in an unfair manner.[228] An example of this is where a non-protected characteristic, such as a postcode, may impact people from poorer socio-economic backgrounds unfairly.[229] Therefore, the flexibility which exists in the

[221] According to the precedent set in F.B. Ali vs. The State [PLD 1975 SC 506], the only generalisation which can be made about this right is that equal laws apply to all in the same circumstances and differences can lead to different applications of the law so long as it is reasonable to do so.
[222] Justice Fazal Karim, 'Article 25: Right to Equality', *Judicial Review of Public Actions*, vol II (Second Edition, Pakistan Law House)
[223] Aharon Barak, foreword to "The Role of a Supreme Court in a Democracy", Harvard Law Review (116 Har. L.R. 16, 2002).
[224] Supra note 220, p.1328.
[225] Justice Mansoor Ali Shah in Jurists Foundation through Chairman vs. Federal Government through Secretary, Ministry of Defence [2020 PLD 1], paragraph 7.
[226] At the time of writing this report, no information was available on whether incidents of discrimination as a result of algorithmic bias have been reported till date in Pakistan. It is essential to not treat the lack of reporting as being equivalent to the inexistence of algorithmic bias. While use of AI is becoming more common in Pakistan (both in the public and private sector), monitoring and data collection lag behind. Hence a number of factors including, but not limited to, lack of access to data and information on algorithmic bias, lack of awareness raising on algorithmic bias, etc. could explain the lack of reported cases of discrimination in this area.
[227] Supra note 58.
[228] Ibid, p. 21.
[229] Sam Shead, 'How a Computer Algorithm Caused a Grading Crisis in British Schools' (*CNBC*, 21 August 2020) <https://www.cnbc.com/2020/08/21/computer-algorithm-caused-a-grading-crisis-in-british-schools.html> accessed 25 June 2021.

interpretation of the right to equality enables us to expand our understanding of what is "fair" in the technologically advanced 21st century.

International Bill of Rights

Pakistan's obligation to uphold equality also comes from the UDHR, ICCPR and ICESCR.

Article 2, UDHR
"Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. Furthermore, no distinction shall be made on the basis of the political, jurisdictional or international status of the country or territory to which a person belongs, whether it be independent, trust, non-self-governing or under any other limitation of sovereignty."
Article 26, ICCPR
"All persons are equal before the law and are entitled without any discrimination to the equal protection of the law. In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status."
Article 2 (2), ICESCR
"The States Parties to the present Covenant undertake to guarantee that the rights enunciated in the present Covenant will be exercised without discrimination of any kind as to race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status."

The protected characteristics identified in IHRL recognize that historically people have been treated unfairly on the basis of these protected characteristics, and that we have universally conceded that such unfairness is unacceptable.[230] To be able to prevent such unfair treatment on the basis of certain characteristics, the same has been granted legal enforceability universally as our disapproval of unfair treatment on the said grounds. In addition, the protected characteristics listed in IHRL are also non-exhaustive and offer room for flexibility of interpretation of what they encapsulate.[231]

[230] Supra note 58, p.27.
[231] The United Nations High Commissioner for Human Rights has recognised the necessity of preventing discrimination against additional cases. See: <https://www.unfe.org/standards/>

At the global level, the UDHR is the leading document of the rights that every human enjoys by virtue of being born as human. This document was adopted through a UN General Assembly resolution in 1948,[232] and even though the UDHR is non-binding, a number of its provisions are now part of customary international law and jus cogens norms, which are binding.

Additionally, the ICCPR and the ICESCR elaborate on the rights that were first expressed in the UDHR. These sister covenants are both binding and categorize the rights of the UDHR into civil and political rights, and economic, social and cultural rights. However, whilst the commitments under ICCPR become obligatory immediately upon ratification,[233] the ICESCR requires states to take progressive measures to enable the realisation of the economic, social and cultural rights, having due regard to the state's economic conditions and resources.[234]

This broadly establishes that Pakistan shoulders a double obligation (one under domestic law and the other under international law) to protect the right to equality and non-discrimination through a *respectful*, *preventive* and *remedial* framework of human rights. This includes a duty to respect human rights in their own conduct, and to also prevent natural and juridical persons from violating human rights.[235] Therefore, taking a preventive approach to algorithmic bias from the very-get go is necessary for Pakistan to uphold its human rights obligations especially because we have only recently tapped into algorithmic decision-making in various institutions. We have a window of opportunity to get this right, and we must seize it.

UN Guiding Principles on Business and Human Rights

Human rights prevention is not only an obligation of the State, but also of third-parties, such as businesses.[236] The UN Guiding Principles are grounded in three principles; *protect*, *respect*, *remedy*.[237] They are not open for signatures or ratification, but exist to provide an authoritative global standard for preventing and addressing the risk of adverse human rights impacts linked to business activity. They

[232] U.N.G.A. Res. 217 A (III).
[233] International Covenant on Civil and Political Rights (New York, 16 Dec. 1966) 999 U.N.T.S. 171 and 1057 U.N.T.S. 407.
[234] 'Progressive Realisation and Non-Regression' (*ESCR-Net*) <https://www.escr-net.org/resources/progressive-realisation-and-non-regression> accessed 21 June 2021; 'OHCHR | What Are the Obligations of States on Economic, Social and Cultural Rights?' <https://www.ohchr.org/en/issues/escr/pages/whataretheobligationsofstatesonescr.aspx> accessed 21 June 2021.
[235] Supra note 71.
[236] General Principles of the Guiding Principles on Business and Human Rights identify three areas as the focus of the Guidelines: (a) The State (link State responsibility with regulation of Business Activity to create a mode of attribution of liability on States if enterprises abuse human rights); (b) Business Enterprises (impose individual responsibility on enterprises to respect human rights); (c) Remedies (to redress human rights violations).
[237] Ministry of Human Rights, Government of Pakistan, 'Business and Human Rights in Pakistan' (BHR Pakistan, 2019) <https://bhr.com.pk/wp-content/uploads/2019/10/BHR-Brochure-1.pdf>

have been created to police corporate conduct and ensure that the third parties operating within a State, i.e. the business enterprises, also respect human rights as they extensively contribute to the development of a State.

Although the Guiding Principles do not carry legal force, they apply to all States and all business enterprises (both transnational and others)[238] and may be argued to be binding as they do not create any new obligations. Rather, the Guiding Principles have been derived from pre-existing international human rights instruments[239] which provide useful guidance on how businesses can operate in a rights respecting manner.[240] They have created a globally expected standard of conduct for enterprises, and although they are arguably soft law, their widespread application "socially bind" corporations to respect human rights.[241] The Guiding Principles place business enterprises under high burdens, and rightfully so, given their ability to violate human rights in favor of profit.

The impact of these guidelines have been that many countries have either developed and implemented a National Action Plan[242] or are in the process of developing the same.[243] In Pakistan, a National Action Plan (NAP) is currently under development to create a legitimate framework to promote rights-respecting business activity.[244] This is being done through multi-stakeholder involvement to review human rights instruments, both domestic and international as well as domestic laws and policies that currently exist for corporations.[245] Pakistan's aim is to improve its reputation[246] and prove internationally that it is a trusted business partner. A State-led initiative to incorporate the Guiding Principles would bestow the same with legal force domestically. This initiative, once completed, can create a strong obligation upon private technology companies to be more mindful of algorithmic bias which can be a breeding ground for unintended human rights violations, as the sense of accountability would then be greater.

[238] Ibid.
[239] Principle 12, UN Guiding Principles on Business and Human Rights.
[240] Justine Nolan, 'The Corporate Responsibility to Respect Rights: Soft Law or Not Law?' (Social Science Research Network 2013) SSRN Scholarly Paper ID 2338356 <https://papers.ssrn.com/abstract=2338356> accessed 21 June 2021.
[241] Ibid.
[242] Such as the Netherlands and Germany, see: <https://bhr.com.pk/wp-content/uploads/2019/10/BHR-Brochure-1.pdf>
The list also includes the UK and USA, see: <https://bhr.com.pk/resources/#1571019484527-781d8510-d3af>
[243] Ibid. See also: <https://globalnaps.org/country/>
[244] Supra note 7.
[245] <https://bhr.com.pk/wp-content/uploads/2019/11/Lahore-Consultation.pdf >
[246] To read the commentary on Principle 5, see: Supra note 7.

B. Pakistan's existing policy and legal framework

Digital Policy of Pakistan 2018

With strong state support for mass adoption of new and emerging technologies, the holistic strategy aims to make Information and Communications Technology (ICT) a broad enabler of every sector of socio-economic development through sectoral digitalisation, with economic development being a central focus of the policy vision. The policy envisions improving the lives of Pakistani citizens by making ICT services accessible, affordable, reliable, and universal.[247] To achieve this, Pakistan aims to use the potential of its youth, especially women and girls, to promote equality and achieve its sustainable development goals through education, skills training, and encouraging entrepreneurship and freelancing.[248] Overall, the policy is well-intentioned and hopes to bring about positive change across all socio-economic sectors by making most of the opportunities presented by technology.

As mentioned before, the discourse on *responsible, ethical,* and *trustworthy* AI is quite broad and a number of sub-categories exist under every particular theme. Although Pakistan's Digital Policy is reflective of various themes already, however, one major distinguishing factor between Pakistan's Digital Policy and global policies discussed in a previous section is that Pakistan's policy deals with a range of digital technologies,[249] instead of having a narrow focus on AI technology. As such, it is a "Digital" policy and not an "AI" policy. Identified below are some of the broad themes of trustworthy/responsible AI that the Digital Policy already reflects to some extent, however, it must be noted at the outset that none of them reflect a recognition of algorithmic bias as an issue. As a result, none of the themes have been discussed in the context of algorithmic bias. Rather, the themes are reflected in the context of the broad range of "digital" technologies and not just AI.

(a) Well-being of all sentient beings and Accessibility:[250]

Under the intergovernmental policies for AI, this principle requires that AIS must improve the living conditions such as health and working conditions. This is reflected in Pakistan's Policy Vision which states the Government of Pakistan aims to improve the living conditions and *economic* well-being of its citizens through provision of high- quality ICT services. We

[247] Ministry of IT & Telecom, 'Policy Vision & Goals, Digital Pakistan Policy 2018' (p.5, Ministry of IT & Telecom 2018).
[248] Ministry of IT & Telecom, 'Policy Objectives, Digital Pakistan Policy', (pp. 5-6, Ministry of IT & Telecom 2018).
[249] In addition to focusing on AI, IoT, and Robotics, the focus is also on technology not powered by AI, such as e-Clinics, increased broadband connectivity and provision of 3G/4G services, online dispute resolution, digital forms, smartphone applications etc.
[250] These principles are also reflected in the policies discussed under international best practises.

recommend the specific addition of mental, physical, social, cultural and political well-being to the policy vision.[251]

The Policy also considers improved access to health through ICT services for women and girls,[252] and also considers increasing online access for Persons with Disabilities.[253] Increasing access is also reflective of the fairness principle, but this principle is not discussed explicitly in the Policy.

(b) Inclusivity:[254]

The Policy aims for digital inclusion to reduce rural-urban divide, gender disparity, unserved and underserved areas, inequality for the person with disabilities, by connecting the unconnected with *broadband*. The intergovernmental policies recognise the potential for AI to create such a divide and their solutions are therefore focused on how to increase inclusivity through AI.[255] Pakistan's policy focuses on "inclusivity" through broadband technology. However, the need for inclusivity, regardless through which type of technology, is recognised in the Policy. Therefore, we recommend that the aim of inclusivity should also be discussed in the context of automated decision-making.

(c) Human-centric approach:[256]

A human-centric approach is a very broad and holistic concept which cannot be expressed in a single definition. The intergovernmental policies that were reviewed for this report held the broad aim of a human-centric approach as a central focus of the policy which was reflected in nearly every theme. Pakistan's Digital Policy primarily focuses on economic development by utilising the human capital of the country, and the term "human-centric" has not been used in the Policy. However, human-centrism can be read into various parts of the policy. For instance, one of the human-centric focuses of the Montreal Declaration is to empower citizens regarding digital technologies, increasing their knowledge base and teaching them fundamental skills to

[251] We acknowledge that economic, social, and political well-being is considered in the policy, but that is done only in the context of women empowerment.
[252] Ministry of IT & Telecom, 'Policy Objective IV, Digital Policy of Pakistan 2018', (Ministry of IT & Telecom 2018)
[253] Section 7 of the Digital Pakistan Policy 2018.
[254] Theme is also reflected in the policies discussed under international best practises.
[255] Supra note 97, p.12.
[256] Theme is also reflected in the policies discussed under international best practises.

foster critical thinking.[257] This is reflected in section 16 of the Digital Policy which focuses on modernising ICT education and making it more mainstream.[258]

(d) Transparency and Accountability:[259]

It must be acknowledged at the very outset that this theme is not discussed in the context of algorithmic bias in the Digital Policy, unlike the intergovernmental policies. Rather, the Digital Policy discusses it under the policy objective of e-Governance,[260] and *transparency*, *efficiency* and *accountability* should be achieved through integrated governmental databases. However, as far as accountability is concerned, it is not clear *who* would be made accountable; would government agents be made accountable for data collection (in which case we can make a case for collection of biased data) or is this meant to increase the accountability of wanted criminals and civil offenders? Since the objective of e-Governance is good governance, it may be presumed that an integrated database would be made transparent for the government and that criminals and civil offenders could be made accountable more efficiently. If this is the case, then once again, the digital policy only favors the opportunities technology promotes, and risks of technology have not been accounted for in the digital policy.

Additionally, transparency is also discussed in the context of data privacy. Section 1 of the Policy calls for a legislative framework for data protection that provides transparency and security of sensitive information. Even though sensitive data has not been discussed in the explicit context of algorithmic bias in the Policy, the recognition of sensitive data and the need to protect it provide sufficient impetus to encapsulate algorithmic bias within this protection. There is a need to recognise algorithmic fairness at a state level, and one of the basic elements that can lead to this recognition (such as protection of sensitive data) is present in the current consultation draft of the Personal Data Protection Bill 2020.[261]

Recommendations:

The Digital Policy discusses a broad range of technologies, in addition to Artificial Intelligence and IoT. However, as the challenge posed by Artificial Intelligence is different from the technologies that have come before, we recommend that Pakistan should also develop an AI Strategy through a robust

[257] Supra note 109, p.9.
[258] Ministry of IT & Telecom, 'Section 16, ICT Education, Digital Pakistan Policy 2018', (p.16, Ministry of IT & Telecom).
[259] Theme also reflected in the policies discussed under international best practises.
[260] Ministry of IT & Telecom, 'Policy Objective IX, Digital Pakistan Policy 2018', (p.7, Ministry of IT & Telecom 2018).
[261] Section 28 of the Personal Data Protection Bill 2020.

multistakeholder effort the way other States have.[262] This is necessary as it will not only recognise AI as different from other more straightforward technologies, but also enable us to identify AI specific issues and come up with solutions together as a nation.

The Digital Policy, as it currently stands, only discusses how to encourage the development and use of AI, but does not discuss its risks.[263] A separate policy shall explicitly codify the various themes, such as transparency, accountability, and fairness, in the context of AI technology, removing the need to implicitly read these themes into the existing Digital Policy.

We also recommend a more active involvement of the Ministry of Human Rights in the implementation of the current Digital Policy,[264] and in the development of any human-centric AI Strategy in the future as the challenges posed by AI are 21st century challenges to human rights, therefore, resolutions must come directly from the human rights law.

Cyber Laws of Pakistan

Historically, as new technologies have come to Pakistan, the law has played a key role in creating a framework to both support and regulate these technologies. Examples of this range all the way from the **Telegraph Act 1885** to **Citizen Protection (Against Online Harm) Rules 2020**. In an analysis of Pakistan's cyber laws, we have attempted to discover whether or not some elements required for the recognition of new technological challenges are already present in current laws. In reading the following, one must be mindful that there are no laws, policies or guidelines to deal specifically with algorithmic bias in Pakistan yet. Therefore, the subsequent analysis only aims to show the legal recognition of some ethical guidelines analysed earlier which can serve as foundations to steer a dialogue towards enabling the recognition of algorithmic bias and challenges associated therewith in Pakistan.

Under the **Prevention of Electronic Crimes Act 2016 (PECA 2016)**, section 2 has defined data as including both content and traffic data and the former is defined as "*...information or concept for processing in an information system including* ***source code*** *or* ***a program*** *suitable to cause an information system to perform a function*". Additionally, powers of authorised officers are listed under

[262] Thomas A. Campbell, 'Artificial Intelligence: An Overview of State Initiatives' (Future Grasp, 2019) <http://www.unicri.it/artificial-intelligence-overview-state-initiatives> accessed 21 June 2021.
[263] Section 17 of the Digital Policy focuses on IoT, AI, Fintech, & Robotics. The gist of this section is that innovation centers for each shall be established in major cities, and a highly integrated ecosystem shall be developed to promote home grown players in, inter alia, AI. It also focuses on capacity building, public-private partnerships, and modernising the education curricula.
[264] Currently, the involvement of the Ministry of Human Rights is extremely limited. It is only responsible for making IT more accessible for PWDs. See: Policy Initiative "PWD", Digital Policy of Pakistan (p.23, Ministry of IT & Telecom, 2018).

section 35 of **PECA 2016**. Under subsection (g) of section 35, there is a requirement that the data be decrypted when an authorised officer is granted access to the data so that the data is "intelligible". The provision also includes an explanation for what "decryption information" means; transforming encrypted and ciphered data to a readable and *intelligible* form. A similar requirement also exists under section 6 of the **Citizen Protection (Against Online Harm) Rules, 2020 (Rules of 2020)**.[265] Both these provisions refer to presenting information to an *authorised officer/agency* in an intelligible form. However, with concerns for the right to privacy on the rise, Pakistan's **Personal Data Protection Bill 2020 (PDPB 2020)**, which is still under consultation, requires that when access is provided to a citizen for their personal data, a copy of personal data must be provided to the *citizen* in an intelligible form.[266]

This shows that computer programs are already legally recognised in Pakistan, and the difficulty of understanding computer codes by lay persons is also recognised which is why intelligibility has been made a legal requirement under **PECA 2016**, the **Rules of 2020** and will also be a legal requirement under the **PDPB 2020** when it becomes law. Therefore, what we see here is the legal existence of one component of the ethical requirement of transparency and explainability under the global guidelines analysed earlier.

Not only are computer programs/algorithms legally recognised in Pakistan, but the term "automated" has also been legally defined as "*without active human intervention*" under the **Electronic Transactions Ordinance 2002**. Therefore, this separate legal recognition of algorithms and automation creates scope for the merger of the two legal concepts in a new statute or policy to more accurately represent the technological realities of the 21st century. However, it may be argued that such recognition is currently underway through the **PDPB 2020**, which allows *processing* of personal data, "*whether or not through automated means*", in institutions such as healthcare, finance, judiciary, and law enforcement.[267]

Moreover, various laws in Pakistan that establish regulatory bodies often include the requirement for non-discriminatory decision-making by the human-decision makers of the Authority under the law. This requirement also exists under Pakistan's cyber laws. An example of this is section 6 (b) of the **Pakistan**

[265] **Section 6 - Provision of information by Social Media Company**: The Social Media Company shall provide to the Investigation Agency designated or established under section 29 of the Act, any information or data or content or sub-content contained in any information system owned or managed or run by the respective Social Media Company, in decrypted, readable and comprehensible format or plain version in accordance with the provision of the aforesaid Act. *Explanation.* - The information to be provided may include subscriber information, traffic data, content data and any other information or data.
[266] Section 16.2 (b) of the Personal Data Protection Bill 2020.
[267] See sections 2 (f), 28, and 30 of the Personal Data Protection Bill 2020.

Telecommunications Authority Act 1996.[268] This provision requires that any decisions made by the Authority should be "*made promptly, in an open, equitable, non-discriminatory, consistent and transparent manner*". This provision is followed by two additional provisions which mandate that people affected by the decisions have an opportunity to be heard as well as an opportunity to appeal the decision.

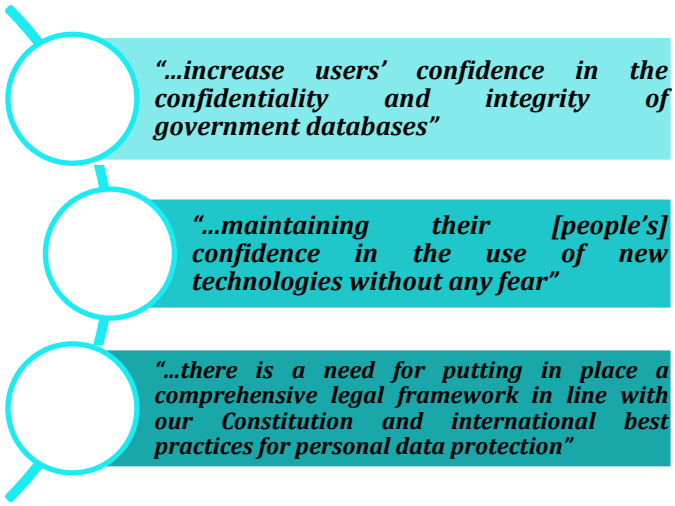
This shows that generally in Pakistan there is a recognition that human decision-making can be discriminatory which is why the need for transparent, equitable and non-discriminatory decision-making has been brought to a legal standard for regulatory bodies that regularly make decisions about citizens. The possibility of challenging the decisions is also included within the legal procedure as the idea of justice is incomplete without the ability to appeal decisions. All of this is very reminiscent of the current global discourse on AI ethics which agrees that automated decision-making through AI suffers from the many of the same defects that human decision-making does. Therefore, the pre-existing legal standards set for human decision-making can be extended or translated to automated decision-making as well along with creating clear mechanisms for accountability.

However, while terminology such as transparency and fairness of algorithmic decision making can definitely be discussed in normative language and eventually even brought to a legal standard, the threshold of transparency and fairness of AIS has to be created by the AI experts themselves as they have requisite technical skills to make such systems more transparent and fairer. Such transparency and fairness would also be based on the context in which the system has been deployed. The role of law is not only to mandate transparency and fairness of algorithmic decision making, but is also to regulate it. But such regulation can only take place when we know what transparency and fairness of an AIS means in technical terms in a particular context, and this is a huge gap in the current research of AIS in Pakistan. Therefore, an active multi stakeholder dialogue between the legal and technology communities should take place in Pakistan to determine how this gap between setting thresholds and regulating the said thresholds can be bridged.

Another pertinent legal reference for the current discussion is the **PDPB 2020**. Although the Bill has been under consultation since 2018 and is not law yet, we shall still refer to it because it serves as an excellent portal into the mind of the legislature. If the Bill is promulgated as an Act, this law will govern the collection, processing, use, and disclosure of personal data in a rights-respecting manner. Perhaps the most important part of the Bill for our purposes is the language used in the statement of objectives. A holistic reading of the objectives shows that the underlying goal of the legislature is to foster

[268] PTA is a governmental authority in Pakistan which regulates the establishment, operation and maintenance of telecommunication systems and provision of telecommunication services in Pakistan. <https://www.pta.gov.pk/en/functions>

confidence of people in big data and new technologies which may lead one to think that Pakistan recognises there are both risks and opportunities associated with new technology, and is willing to work on minimising the risks. This is particularly evident from the following excerpts from the objectives:



Moving now to some particular aspects of the Bill which can be discussed in the light of algorithmic justice, as far as *collection* is concerned, there are no mechanisms to ensure that data bias be minimised in the collection phase through data sampling or segregation to determine what the data represents or whether or not the data is even representative of the target population.

As regards *processing*, the Bill has defined the term "processing" under section 2 (f). The definition is quite broad, but the noteworthy aspect of the definition is that processing of datasets can be both automated and manual. The Bill imposes a general prohibition against processing *sensitive personal data*, which includes among other things, an individual's medical records, ethnicity, religious beliefs, *or any other information*.[269] These touch upon a few of the "protected characteristics" discussed earlier, but "any other information" offers flexibility to *read in* other protected characteristics as part of "sensitive" data. Personal data is also categorised in terms of "*critical personal data*"; there is no definition for this under the Bill, but responsibility for classifying personal data as critical data has been bestowed upon the Personal Data Protection Authority, with approval from the Federal Government. There is nothing in the Bill to suggest what criteria is to be followed to determine how personal data could be classified as critical personal data. Therefore, it is not possible to say whether or not it deals with protected characteristics, but there is a general prohibition against processing critical personal data as well.

[269] Defined under section 2 (k) of the Personal Data Protection Bill 2020.

However, the Bill also creates exceptions to this general prohibition. Sensitive and critical personal data can be processed for medical purposes (such as medical diagnosis),[270] judicial purposes, prevention or detection of crimes or taxation purposes. It is interesting to note that so far these are the institutions where automated decision-making by way of AI has been deployed or is being planned to be deployed, and how the development of such AIS has taken place without any laws in place. Nonetheless, based on context, it is justified to use different types of sensitive data in each of these institutions, but since the decisions of these institutions have the power to seriously impact a citizen's life, they can be considered high risk institutions, thus greatly warranting the need to place strict mechanisms to ensure pre-existing discriminatory biases of these institutions are not reinforced when decision-making in the institutions is automated.

Recommendation:

As is evident from the foregoing, there are no explicitly focused laws, policies or guidelines on dealing with algorithmic bias or any other risks associated with AI technology. It is also not possible to say whether or not algorithmic bias is a consideration in development of AIS in Pakistan because the notion of algorithmic bias either had to be *read into* the cyber laws in the previous analysis or the interpretation of existing provisions had to be extended to include the issue of algorithmic bias as well. Even where we can say that the current interpretations of certain provisions from cyber laws are open-ended and can easily encapsulate issues of algorithmic bias by way of reinterpretation, there is the problem of non-consolidation; such provisions are not under one single cyber law and one has to pick and choose provisions from various cyber laws to make a case for algorithmic fairness legally.

Therefore, there is a dire need to have a separate AI policy which, at a state level, clearly recognises the challenge of algorithmic bias in addition to all other challenges associated with AI technology. Such policy should also be complemented by new additional laws that can recognise such challenges and create ways to regulate AI technology whilst ensuring that human rights remain at the heart of any such laws and policies.

[270] Section 28 of the Personal Data Protection Bill 2020.

Way Forward and Recommendations for Pakistan

Way forward and Recommendations for Pakistan

In order to ensure that Pakistan is prepared for AI development and deployment it is necessary that relevant institutions and stakeholders are empowered to work pre-emptively to identify red flags and potential problems. A cautious, preventive and precautionary human rights-based approach is needed to, in the words of Saleem Akhtar J in the Shehla Zia case, "*avert a catastrophe at its earliest stages*".[271] In addition to this, the State needs to recognize the impact of AI is vast and call for the protection of all socio-economic groups and persons in society. It is essential that Pakistan plans ahead and utilises existing literature and best practices from states that are actively addressing the risks of AI. The switch towards AI is inevitable and hence careful planning is necessary in the current climate as AI is slowly making its way into the Pakistani framework. There is a need to ensure that AI development and its use is in alignment with human values. The recommendations below provide a more holistic approach to moving forward on algorithmic decision-making in Pakistan by focusing on not just the legal and policy framework but also effective monitoring of AIS and implementation of ethical guidelines and practices in Pakistan. The recommendations, although divided, are all complementary across the different groups identified below:

A. Short-term interventions:

(a) State intervention/State Action:

- Support relevant government institutions to start the discussion on AI technology and specifically the associated risks. Gaps between knowledge of different stakeholders need to be bridged. Increasing the understanding and knowledge of relevant government stakeholders needs to be prioritized to ensure a safe way forward for AIS in Pakistan.

- Increase investment and prioritization of policy, legal, and technical research needs on how to build responsible AI technology that citizens can trust. Currently, a big gap exists in research on AI risks in Pakistan. Most of the current research on AI is coming from the ICT industry itself which is relevant to AI experts for the most part.

- Advise existing stakeholders to coordinate and communicate with government stakeholders to ensure that the AIS developed are transparent, explainable, traceable, and accountable. Moreover, existing stakeholders can share challenges and success stories in AIS in Pakistan with the government to support future policy-making.

[271] Shehla Zia v WAPDA. pp. 709-710.

(b) Legal and Policy:

- Support recognition of the potential discrimination and harm that can be caused by use of AI in Pakistan. This can be done through capacity-building of relevant government departments, Parliamentarians and the judiciary on the role of AI and ethics.
- Adopt a human rights lens to the issues of algorithmic bias, enabling protection of the broader right to non-discrimination and equality enshrined in the Constitution of Pakistan.
- Promote planning and conversations on the future of technology and specifically AI in Pakistan, at the policy-level. This can include lobbying and drafting of targeted and human rights compliant guidelines and/or legislation on AI development and deployment.

B. Medium-term interventions:

(a) State intervention:

- Develop a coordination mechanism or a working group for responsible AI that can support communication between AI stakeholders in Pakistan to engage in multi-stakeholder dialogues, investigate the numerous risks of AI including that of algorithmic bias, how the risks can manifest in different contexts and determine the extent to which we can allow automation to meddle with human lives and how and where to draw the line.
- Support and promote a policy environment that enables smooth transition from R&D to deployment and improvement of the AI systems.
- Promote information sharing and create awareness amongst the youth on AI and use of AI in Pakistan. The Ministry of Information & Technology should disseminate information on the legal / policy framework for AI and any other relevant guidelines on how to develop trustworthy AI.

(b) Legal and Policy:

- Develop a national AI policy/strategy for a trustworthy and fair future of AI through multi stakeholder dialogues. Any such strategy should be constantly reviewed and updated as our understanding of AI evolves.
- Revise existing anti-discrimination laws in Pakistan and make them more litigant friendly by coming up with creative solutions to deal with the black-box issue in a way that the unfair

burden of detecting algorithmic bias and its potential unfairness is not disproportionately borne by the litigant.

- Revise **PDPB 2020** in light of international best practises, and then promulgate to guide the collection of necessary and "good" data.
- Develop guidelines / regulatory framework aimed at the ICT industry to support human diversity and inclusivity in AIS.

(c) Capacity-building of stakeholders in Pakistan on AI

- Invest in capacity-building of government institutions (federal and provincial level) to better understand AI technology, its use and potential impacts and consequences. Build the capacity of government stakeholders on AI to support responsible use of AIS.
- Mandate training / seminars for all universities working independently or in partnership with other entities on creation / use of AI in Pakistan.
- To uplift the rule of law, build the capacity of the judiciary to make decisions related to AI systems.

(d) Accountability and Monitoring framework:

- Lobby for and develop regulatory mechanisms to monitor (under the Ministry of Information & Technology) at the federal level to support and monitor AI creation and deployment in Pakistan.
- Develop oversight and monitoring frameworks/protocols/policies to ensure avenues for redress and review of automated-decisions are available to all. The law should provide for a procedure for review, appeal and challenge of an outcome made by an automated system. In addition to this, use of automated systems, especially in criminal justice institutions, should come with increased human accountability to ensure dispensation of justice remains a priority.

C. Long-term interventions:

(a) State intervention

- Support awareness-raising and access to information about the AI technology, how it works, and where it is being used among the general public. Educating the lay-person is necessary for

them to exercise the range of their constitutional freedoms effectively. Awareness campaigns must also focus on busting the myths about AI technology.

- Invest in developing an AI audits industry and train home-based experts to become AI auditors.

(b) Legal and Policy

- Mandate the collection and use of good data which represents the best practise so pre-existing biases can be minimised as much as possible. Legal framework that promotes safe sharing of data can also encourage institutions to give data to AI developers without hesitation. With availability of more data which represents good practises, the performance of AI systems can be improved.

(c) ICT Industry

- Exercise due diligence before deploying AIS by collecting good and exemplary data, examining the type of data collected, what it represents, and the consequences it can have.
- Carry out extensive testing of AI on lots of data so that biases can be detected and quantified.
- Ensure teams designing, developing, testing, and deploying AIS need to be reflective of the diversity of the users of the AIS. AIS should incorporate human diversity and inclusivity to remain within the regulatory framework. This is necessary because the real world in which AIS is deployed is diverse.

(d) Accountability and Monitoring framework

- Develop a watchdog / monitoring body (semi-government or non-governmental) to ensure automated decision-making is not used in high-risk sectors.
- Develop mechanisms for human oversight at all stages of the lifecycle of an AIS. This helps promote a human-centric approach to AI use in Pakistan.
- Develop mechanisms to carry out human rights risks assessments of AI systems.