ACCOUNTABILITY and RESPONSIBILITY of ARTIFICIAL INTELLIGENCE DECISION-MAKING MODELS in INDIAN POLICY LANDSCAPE

Palak Malhotra[1], Amita Misra [2]

[1]St. Xavier's College (Autonomous), University of Mumbai, Maharashtra, India
[2]Amazon

Abstract

Artificial Intelligence (AI) is the drive behind the fourth industrial revolution (Industries 4.0) that has swept the 21st-century world. Every tech giant, national government, and international institute only has one thing in their mind; how to harness and utilize AI capabilities for economic prosperity and human flourishing. AI, especially autonomous machine learning decision making models have been propagated as a solution to every major challenge faced by entrepreneurs, governments, and social sectors. Completely autonomous machine learning decision-making systems are increasingly becoming devoid of human judgment. Today, algorithms have a definitive say in life-altering circumstances. Identification, security systems, public distribution systems, criminal justice systems, and job opportunities are dependent on algorithmic decisions, whether there is human judgment involved or not. So, how does one bestow the principles of responsibility and accountability on a non-living amoral entity? Here we focus on conceptualizing the functioning of AI and putting them in the wider socio-economic context of society rather than isolated models. We recommend a comprehensive national policy in Indian landscape on ethical AI decision-making models which prescribe a responsible and accountable framework.

Keywords
Artificial Intelligence, Responsibility, Accountability, Bias, Policy

1. Introduction

Artificial Intelligence is the drive behind the fourth industrial revolution (Industries 4.0) that has swept the 21st-century world. Every tech giant, national government, and international institute only has one thing in their mind; how to harness and utilize AI capabilities for economic prosperity and human flourishing. AI, especially autonomous machine learning decision making models has been propagated as a solution to every major challenge faced by entrepreneurs, governments, and social sectors. It is becoming increasingly assertive in the domains of self-driving cars, legal and jurisdiction systems, and automated weapon systems[1]. Today, algorithms have a definitive say in life-altering circumstances. Identification, security systems, public distribution systems, criminal justice systems, and job opportunities are dependent on algorithmic decisions, whether there is human judgment involved or not. So, how does one bestow the principles of responsibility and accountability on a non-living, amoral entity? In this endeavor, it is important to realize that artificial intelligence learning algorithms must be understood in a socio-techno environment rather than as isolated computational systems [2]

In recent years, the AI industry has come under a black cloud. AI rather than being an unbiased and objective mechanism that can contribute to building a fair and equal society has been shown to be doing the contrary. It is perpetuating and even amplifying existing structural biases of society and works favorably in maintaining the power balances in society. In the backdrop of unstable socio-political context of India in which these AI decision making models will soon be implemented require human oversight. Policy intervention and regulation are the need of the hour.

AI is indeed a complex social system that cannot and should not be evaluated on the bases of accuracy and efficiency. This paper will acknowledge the need for the policy sector to intervene and regulate the AI sector to ensure the protection of fundamental rights. Lastly, recommendations will be provided for a comprehensive national policy on ethical AI decision-making models which prescribe a responsible and accountable framework.

2. Bias in AI decision making Models: An Ethical Conundrum

2.1. Will AI lead to the 'Good for All'?

Numerous policy documents, international conventions, and tech giants envision AI under the mantra of 'AI For Good' or 'AI for All'. They focus on emphasizing the positive applications of Artificial Intelligence and the economic prosperity that it will bring about for the country and how it will lead to human flourishing. Algorithms are presented by computer scientists and tech companies as 'purely formal beings of reason' [1]. They are understood to be strictly rational concerns rooted in the disciplines of mathematics and technology. At the face value, an objective, and rational algorithm deciding critical and life-changing things, such as resume screening, parole determination, medical diagnosis, is indeed considered a breakthrough. It is considered to be an objective, equality-based decision-making system that will get rid society of all pre-existing structural biases which are otherwise reflected in human judgment and power relations in society. It will be a catalyst to human flourishing. However, emphasis should be put on the fact that AI's decision-making system while lead to greater good and improve quality of life, it also has the potential to maintain the status quo of society. The existing literature that exists regarding machine learning applications has highlighted the ability of these learning algorithms to exacerbate existing inequalities in societies and even go to the extent of reinforcing them. Thus, it is a rather utilitarian perspective to believe that AI will enable human flourishing [3].

2.2. AI reinforces societal bias

The lexicon definition of bias is an inclination or prejudice for or against one person or a group, especially in a way that is considered to be unfair. Bias surfaces when unfair and false judgments are made because the individual making the said judgment is influenced by a pre-existing discriminatory stereotype about members of a particular group and that judgment is in fact not even relevant to the matter at hand. The decisions especially made by ML or AI decision making algorithms are susceptible to be biased against marginalized, and less powerful sections of society. The ethical concern of producing a biased decision by AI learning algorithm is attributed to the unpredictability of the outcome, labeled as the 'Black-Box Problem'. As mentioned above learning algorithms are provided with a historical dataset to learn from to solve a particular problem statement or perform a particular task. But these learning algorithms are not given any correct or incorrect answers to the problem statement. They make their decisions through leanings

of their own by finding patterns and co-relations amongst different variables. For instance, in a resume screening algorithm, the algorithm will be fed a training data set of the resumes of their top-performing employees in the company. From there on, it is the job of the algorithm to find patterns or correlations which exist in all these resumes and indeed makes them the best performing resumes. Those correlations or similarities are then set as an ideal benchmark by the algorithm to pass the screening test for future candidates. This saves on unnecessary human labor and cost. However, human judgment is completely moved out from it. Neither the deployer, nor the user can truly estimate what the algorithm will learn from the dataset fed into it. Therefore, learning from data, interpreting, and making decisions occurs in a 'black box'. This especially becomes worrisome when the developer is feeding into the system a dataset that is riddled with bias. For instance, let's re-consider the example of the resume screening algorithm and how it might end up perpetuating and reinforcing an existing societal bias. If a company has a history of discriminating against hiring less female or black workforce, there will be fewer resumes from these social groups in datasets. The algorithm, unaware of the structural inequalities, might learn from the dataset that resumes of female or black candidates are less than ideal. It interprets a lesser number of black or female candidates' resumes not as a result of structural inequalities that exist in society but due to genuine performative and cognitive abilities (or rather a lack of them). Therefore, even if the minority community is qualified and might indeed be an ideal candidate, the algorithm dismisses the resume or puts it into secondary consideration due to their gender or race. Such unfair and unjust biases have crept up in criminal justice systems, healthcare access, and intelligence-led policing.

2.3. AI exacerbates inequalities in India

From a policy perspective, India's approach to AI is substantially guided by two national policies. Firstly, under Digital India, the union government increased funding towards research, training, and skill building in emerging AI technology. Under the Digital India Mission, AI applications become pertinent in the use of Pradhan Mantri-Jan Dhan Yojana, Smart Cities, E-Pathshala, E-Prison, Farmers Portal, and E-Courts. Second initiative where AI is increasingly being addressed is Make in India where the government is working towards incentivising AI-based domestic investments and innovations. An AI Task Force was also constituted by the Ministry of Commerce and Industry to look at AI as a socio-economic problem solver in the key sectors of national security, agriculture, education, smart cities, finance, and manufacturing. While the Indian legal and policy system should be appraised

for accommodating the fast-growing spread of AI in our everyday lives, the policy initiatives are rather lacking the same enthusiasm in addressing the ethical concerns that arise out of them.

Unfair and unjust biases have crept up in Indian criminal justice systems, healthcare access, intelligence-led policing. The main reason attributed to such biased outcomes has been historical bias in the dataset [4]. For instance, in a data-driven algorithm deployed by law enforcement agencies to identify criminal hotspots on the basis of a data set regarding neighborhoods where most arrests for crimes have been taken. This becomes extremely problematic in countries, like India, where police often practice arbitrary arrest. In India, police is known to profile neighborhoods and suspects on the basis of religion and caste. Such biased data will only flag neighborhoods that are pre-dominated by vulnerable caste and religious groups. This will lead to an unnecessarily increased deployment of police in the area leading to unjust arrests. These arrests are only utilized further as additional training data. Hence, not only an AI algorithm perpetuates bias but exacerbates it. This phenomenon is already being witnessed in Delhi's CMPAS which is reinforcing caste and religious prejudices of the police [1]. Not only is this infringement on the right to freedom against discrimination and rights of minorities, it has fatal consequences for the life and liberty of the said socially vulnerable groups. However, these trends have yet to be confirmed by the government.

Even if the dataset is not riddled with historical biases, it can still be biased due to inefficiencies caused by incomplete datasets. For instance, it has been proved that facial recognition algorithms that are employed by law enforcement agencies and employers could intensify systemic biases against color and gender. This is because not enough individuals from different social groups are used in datasets. While face recognition algorithms boast a 90 per cent accuracy rate, it is not universal [5]. Their error rate still continues to be higher for vulnerable social groups. There have been numerous case studies on how facial recognition software, which are, employed in surveillance, airport passenger screening, and employment and housing decisions falsely or incorrectly identify non-white, non-male individuals, and Asian populations 10 to 100 times more often than they did with white faces [6]. Such software also has a higher error rate in falsely identifying women as compared to men. Also, automatic gender classification implicitly assumes that gender is a static concept that does not change frequently across time and culture [7]. While biases emerging from AI have already been reported in Western countries, it has yet to be investigated in India. However, given how un-

equal and patriarchal the current Indian state is, such AI decision making models will continue to reinforce existing power structures in India. It is only a matter of time that these AI decision-making systems that are used to automate decisions regarding an Indian citizen's eligibility and entitlement to opportunities and social benefits. Therefore, it can potentially interfere with the Indian constitutional rights of due process, and the right to freedom from discrimination. Currently, no policy or legal regulation in the current Indian landscape discusses the biases that AI decision-making systems might produce, and how to prevent or mitigate them. The only work in progress is NITI Aayog's document [8] on National Strategy on Artificial Intelligence (NSAI) which attempts to establish a framework for responsible and accountable AI in order to prevent and mitigate any harm that might arise from the AI decision-making model.

3. Conceptualising Responsibility and Accountability

The lexicon definition of responsibility is the state of having a duty to have control over something or someone. When one talks about responsibility in Artificial Intelligence, especially Machine Learning algorithms, responsibility refers to the role of individuals in their relation to the AI systems.

Responsibility does not pertain to the fact that the computer software of any kind itself should be responsible. Rather, the organization and the employee within the organization that composes the socio-technical environment of AI algorithms should be responsible. On the other hand, accountability flows from having responsibility. In terms of ethics and governance, it is equated with answer-ability, blameworthiness, liability, and the responsibility to justify actions to a forum that holds the said party accountable. It refers to the requirement for the system to be able to explain and justify its decisions to users and other relevant actors [4]. The idea of accountability especially arises to connect an agent in case of an occurrence of harm or an injury. A key element of answer-ability is that of explanation and justification [4]. It is providing information and explanation that allows the accountability forum to assess and judge whether the actions taken were ethical or not. With respect to the Indian policy landscape, as aforementioned, it is only NITI Aayog's National Strategy on Artificial Intelligence (NSAI) that recognizes the Principle of Accountability and Responsibility in AI Decision Making systems. Furthermore, India's recent signatory to the United Nations Education, Scientific and Cultural Organization's first global agreement on Ethics of Artificial

---

[1]https://blogs.lse.ac.uk/humanrights/2021/04/16/predictive-policing-in-india-deterring-crime-or-discriminating-minorities/

Intelligence [2]. However, due to these principles not being binding on countries and policies and national strategies being in nascent stages, organizations have managed to shirk away responsibility and accountability from their shoulders through numerous ways.

4. Evading the Principles of Accountability and Responsibility

Bestowing accountability has been a concern that can be predated to 1996. [9] warned of the erosion of human responsibility due to increased reliance on computer systems. Something emphasized back then and still remains a critical ethical concern is who is accountable when recommended decisions, and actions taken by computer systems, which are, non-living entities, that harm a particular individual or group of individuals. Bestowing accountability and responsibility in computer systems, especially machine learning algorithms is met with uncertainties, and deliberate evading of the issue at hand through the following ways;

4.1. Many Hands Problem

The concept of many hand problems in computerized systems was brought on by who argued that there exists a machine learning pipeline and there are many individuals or teams who are involved in the design, development, and implementation of a machine learning algorithm. The ML pipeline goes through three stages [10]. The first stage involves formulating a problem statement. Problem formulation can involve the collection, selection, and curation of a dataset and is responsible for operationalizing a concrete task or a problem statement. This is followed by the implementation stage where a certain type of historical dataset is selected which will be fed into the learning algorithm. From that particular historical dataset, the algorithm is supposed to find correlations and patterns. This is followed by model training and evaluation wherein the algorithm runs through multiple procedures and models. Finally, the most satisficing model is adopted. If bias develops in early stages it only has the potential to go unnoticed, further accumulate and magnify as it passes down various stages. Therefore, if the decision-making algorithms produce biased decisions it becomes extremely difficult to isolate one individual or a group to be held accountable.

4.2. Blaming the Algorithm

Tech giants have often hidden under the guises of offering computer systems as scapegoats [10]. As mentioned above, a machine learning algorithm works in a black box, which means that the system hides its internal logic from the user [4]. They are even hidden to the developer as the co-relations recognized by the machine may not be identifiable as valid or recognizable features to the human mind itself. The algorithm is learning on its own. However, corporations have often shrugged off responsibility and accountability under the pretext of **'we could not have predicted this'**. However, the issue is that you are prescribing intent and attributing moral agency to the algorithm. Companies have often made it sound that 'Intelligence' in Artificial Intelligence or 'learning' in Machine Learning algorithms suggests some sort of intention, awareness to social context, and adaptability to patterns which can intensify biases and thus connote moral agency to the algorithm. However, while it is participating in life-altering decision-making, it is not making certain decisions on intentions. Bestowing morality of knowing what is right and what is wrong becomes a scapegoat route for the organizations to get off the hook.

4.3. Ownership Without Liability

Yet another principle propagated by [9] was of 'Ownership Without Liability'. Third-party providers of data-driven algorithmic systems refuse to expose their systems to scrutiny by independent auditors on the ground of trade secrets. Furthermore, manufacturers and owners of hardware materials, such as autonomous vehicles, shift the liability on environmental factors or human users. For instance, a car manufacturing company has the legal ability to evade any liability and responsibility when it comes to autonomous or semi-autonomous cars [11]. This becomes worrisome given how the fully automated cars, which are, currently in testing stages, have been known to show discriminatory practices to minority pedestrians. However, the recent example of Mercedes [3] proves that it is up to organizations whether they want to indulge in mere ethics washing or genuinely adhere to the principles of accountability and responsibility by taking on full liability in case of an accident.

4.4. Secrecy and Proprietary

There is a critical concern regarding the principle of transparency. Even if the algorithm is available to everyone, it might not be comprehensible to an individual due to technical illiteracy. However, it has been observed that

even if the algorithm's features and operations are understandable, they can still be secretive due to proprietary reasons. Algorithms remain a secret to public forums and to harm individuals or bodies. For instance, even after accusations of machine bias in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) deployed by the criminal justice system in predictive policing and sentencing accused, the workings behind the algorithm were not disclosed publicly[4]. However, when an algorithm is not made available to individuals who might be in harm's way or even to independent auditing parties then companies become less likely to report real machine learning biases. If companies are not made to at least audit their algorithms, then it becomes highly unlikely that they will have no threat of financial, punitive measures, or reputation at stake, putting good faith in companies on truthful reporting on their model's malfunctioning which becomes highly unlikely.

5. Policy Recommendation for making AI Accountable and Responsible

As discussed above, the increasing and pervasive use of AI decision-making models to facilitate, influence and inform welfare services, criminal justice systems, and other decisions has a high-risk impact on the constitutional fundamental right to equality, freedom, and life liberty. Given how ambiguous the principles of responsibility and accountability can be in the domain of Artificial Intelligence, it becomes a matter of policy and governance to ensure that all the various stakeholders in the sociotechnical environment (that are, researchers, developers, and managers) are aware of their responsibility. The organization as a whole must be willing to take the liability for any harm arising from bias or an inherent flaw in the machine learning algorithm. NITI Aayog's NSAI and UNESCO's principles for AI have also acknowledged that a responsible and accountable policy framework is at the center of designing an ethical AI decision-making system.

5.1. Multidisciplinary Training of Employees

Employees from STEM-related fields must have training in basic social sciences where they are educated regarding the nuances of power imbalances in society, structural social inequalities and how technology can produce a

---

[4]https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

new form of systemic oppression and marginalization of such social groups. This can be achieved through a diverse course curriculum and government-mandated sensitization workshops in the workplace.

5.2. Algorithmic Literacy and Transparency

As mentioned above, the AI algorithms work in a 'black-box construct' and even their codes are held in secrecy and proprietary. Furthermore, there are certain complexities to the principle of transparency, such as the technical illiteracy of the general public. However, a national policy on ethical AI should propagate a sensible way to mandate the provision of transparency to overcome the challenges of technical literacy. Some of the provisions which can be considered are; making algorithms more explainable to the end-user especially if it involves high-risk social costs and attempts to increase general data literacy, and training data and algorithms must be made transparent.

5.3. Informed Consent

The national policy on ethical AI should mandate that organizations have informed consent of end-users where it is explicitly stated that the services which are being availed or the screening processes any individual is going through are automated. It is quite surprising that something as comprehensive as NITI Aayog's NSAI has been quiet on issues of informed consent to end-users.

5.4. Diversity and Inclusion

The road to truly building an ethical AI starts from who is given a seat at the table [7]. Diversity and inclusivity lie at the center of developing an ethical AI. There is a need to adopt diversity-in-design [12]. Accurate representation should not only be in datasets but also in AI-centric research and development teams. Diverse employment leads to better and improved decision making and group thinking as researchers from ethnic and gender minorities bring an objective perspective and possess the abilities to question ideas and norms of a homogenized group. Ideally 'diversity-in-design' through incentives schemes can be considered, however still important groups, such as, children and persons with disability can be left out. Therefore, there is also a need to include full community of stakeholders who will go beyond in having a technical expertise. Such stakeholders will include subject matter experts, and end users especially from those communities who are at high risk of vulnerabilities. National Policy can mandate the establishment of advisory or consultative bodies composed of academia and civil society organizations as advisors will attribute

legitimacy to the models which otherwise are absent and quell public hesitations.

5.5. Organizational Responsibility and Accountability

As propagated by NITI Aayog's NSAI, the national policy for ethical AI ensures accountability and responsibility through proportionate liability. Secondly, the 'human in the loop' principle should be adopted where decisions that are identified as high risk require human confirmation before any sort of action is taken place. This is already being implemented in various industry sectors in India which are using automated AI. 'Human in the loop' is also strongly propagated by the NSAI document[5].

5.6. Independent Auditing and Certification

There should be a provision for independent oversight by an external and technically competent oversight body. They would have a duty to lay down regulations and follow a risk based approach towards it. This would mean that AI systems with high-risk social costs will be subjected to more extensive scrutiny and compliance while the burden for automated algorithms with low-risk social costs can be less demanding. The independent regulatory body will investigate, validate and test AI-based algorithms, applications, and products against the well-defined principles of ethical AI. Furthermore, the regulation requirements can follow a risk-based approach where following that, an algorithm can be certified. Certification will reflect that AI algorithms and products are accountable and trustworthy. Already at an international level, IEEE's Ethics Certification Program for Autonomous and Intelligent Systems exists that certify AI systems that comply and adhere to the principles of transparency, fairness and unbiasedness, accountability, and responsibility.

5.7. Protection for Whistle-blower

Whistle-blower protection should be given utmost priority since only then will employees will be confident in coming forward for malpractices which are being conducted by any tech companies[6] This again has been missing from consideration in NSAI, UNESCO's framework, and even EU's General Data Protection Regulation.

---

[5]https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf
[6]https://www.reuters.com/article/us-alphabet-google-research-idUSKBN2AJ2JA

5.8. Sector-Specific Regulation

As mentioned above, AI is ubiquitous and overlaps with numerous sectors of healthcare, automobile, e-commerce, insurance, and so on. Therefore, in addition to framing a comprehensive national policy on the development and application of ethical AI, governments should engage and encourage various industries to formulate their standards and guidelines regarding the deployment and operationalization of AI. This should be compatible with national policy and international standards. This will allow them to adequately and comprehensively respond to the challenges which AI might present in their respective industry sectors.

5.9. Grievance Redressal and Arbitration Mechanisms

Preventive policies can only go so far given the black-box problem of AI Decision Making models. However, if an unfair bias occurs from the AI decision-making model and it infringes upon the rights of an individual or their safety and security, grievance redressal mechanisms should be set up in every tech firm branch of India. Furthermore, if an organization or enterprise refuses to make its algorithm transparent and takes liability even after proven harm or infringement on rights, then the aggrieved individual must have a right to petition in an arbitration tribunal which is presided by an industry expert.

6. Conclusion

Not only is Artificial Intelligence an inextricable part of our daily lives, but the progress in AI domains will also soon be synonymous with the kind of growth trajectory a country is on. However, one must be cautious and remember that Artificial Intelligence, especially automated decision-making models are a work in progress. This means that there is still time for them to be perfected and errors are unavoidable. At such a critical juncture, organizations cannot be allowed to excuse themselves from responsibility and accountability. AI have the potential to infringe upon some of the most basic universal and fundamental rights of an individual and have a potential threat to their security. Currently, these ethical and social safety concerns are not alarming, especially in the Indian scenario. However, given the speed of Industries 4.0, it will not be long before unjust and biased AI is translated into a new form of institutionalized systemic bias. It becomes pertinent for the Indian government to formulate a national policy for ethical artificial intelligence which bestows accountability, responsibility, and liability upon tech giants.

References

[1] N. Spaulding, Is human judgement necessary?, In The Oxford Handbook on Ethics of AI (2020) 380–401.

[2] V. Dignum, Responsibility and artificial intelligence, The Oxford Handbook of ETHICS OF AI (2020) 215–231.

[3] S. Bringsjord, N. S. Govindarajulu, Artificial Intelligence, in: E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Summer 2020 ed., Metaphysics Research Lab, Stanford University, 2020.

[4] M. Busuioc, Accountable artificial intelligence: Holding algorithms to account, Public Administration Review 81 (2020) 825–836.

[5] A. Najibi, Racial discrimination in face recognition technology, "https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/ (2022).

[6] B. Rauenzahn, J. Chung, A. Kaufman, Facing bias in facial recognition technology. the regulatory review, https://www.theregreview.org/2021/03/20/saturday-seminar-facing-bias-in-facial-recognition-technology/ (2021).

[7] T. Gebru., Race and gender, In The Oxford Handbook of ethics of AI (2020) 253–268.

[8] N. Aayog, Responsible ai aiforall, https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf (2020).

[9] F. Cooper, B. Laufer, E. Moss, H. Nissenbaum, Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning, arXiv (2022) 1–13.

[10] A. F. Cooper, B. Laufer, E. Moss, H. Nissenbaum, Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning, CoRR abs/2202.05338 (2022). URL: https://arxiv.org/abs/2202.05338.

[11] X. Zhu, Z. Gu, Z. Wang, Ethical challenges and countermeasures of autonomous vehicles, E3S Web of Conferences 233 (2021) 04016. doi:10.1051/e3sconf/202123304016.

[12] N. T. Lee, P. Resnick, G. Barton, Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms (2019).