Transactions on Machine Learning and Artificial Intelligence - Vol. 9, No. 2
Publication Date: April, 25, 2021
DOI:10.14738/tmlai.92.10118.
Ngige, O. C., Awodele, O., & Balogun, O. (2021). Judicial Artificial Intelligence Bias: A Survey and Recommendations. Transactions on Machine Learning and Artificial Intelligence, 9(2). 74-86.

Judicial Artificial Intelligence Bias: A Survey and Recommendations

Ngige, Ogochukwu Constance
Federal Institute of Industrial Research Oshodi, Lagos Nigeria

Awodele, Oludele
Babcock University, Ilishan-Remo, Ogun State Nigeria

Balogun, Oluwatobi
Bells University Ota, Ogun State Nigeria

ABSTRACT

Artificial intelligence (AI) has continued to disrupt the way tasks are being carried out, finding its way into almost all facets of human existence, and advancing the development of human society. The AI revolution has made huge and significant inroad into diverse industries like health, energy, transport, retail, advertising, et cetera. AI has been found to assist in carrying out tasks more quickly and efficiently too. Tasks which were hitherto difficult have been simplified significantly through the use of AI. Slow adoption in judiciary has however been reported, compared to other sectors. A lot of factors have been attributed to this, with AI bias being an issue of concern. Decisions emanating from courts have a significant impact on an individual's private and professional life. It is thus imperative to identify and deal with bias in any judicial AI system in order to avoid delivering a prejudiced and inaccurate decision, thereby possibly intensifying the existing disparities in the society. This paper therefore surveys judicial artificial intelligence bias, paying close attention to types and sources of AI bias in judiciary. The paper also studies the trust-worthy AI, the qualities of a trust-worthy artificial intelligence system and the expectations of users as it is being deployed to the judiciary, and concludes with recommendations in order to mitigate the AI bias in Judiciary.

Keywords: - Judiciary, Artificial Intelligence, Artificial Intelligence Bias, algorithm and Survey.

INTRODUCTION/BACKGROUND

With the advent of the new industrial revolution known and called industry 4.0, businesses are progressively moving the way of AI and machine learning in order to drive automation in simple and complex decision-making processes. Machine learning algorithms are already interwoven into many aspects of our daily lives, influencing in unseen ways as decisions are made for us and about us. The usage of machine learning in different decision-making processes, including in judicial practice, is becoming more and more frequent. The large-scale data digitization and the evolving technologies that use them are positively upsetting most economic sectors, including transportation, retail, advertising, energy, et cetera. Artificial Intelligence is also having an influence on judiciary as digital tools are being used to advance

correctness and drive fairness in judicial functions (Lee, Resnick, &Barton 2019). At its best, AI will assist judges and the judicial system to provide streamlined access to justice, free from human bias. AI can guide court users to navigate many legal issues without the need for a lawyer. So too, AI can provide information to judges that is based on objective factors.

Bias issues in AI decision making systems have become an important area of concern in recent years, as many companies increase the use of AI systems across their operations. Bias is a tendency to lean in a certain direction, either in favor of or against a particular thing. To be truly biased means to lack a neutral viewpoint on a particular topic. In machine learning, bias refers to measurements that lead to a skew that inflicts an unjust outcome upon a population (Lloyd & Hamilton 2018). Bias is not a new problem rather it is as old as human civilization and it is human nature for members of the dominant majority to be oblivious to the experiences of other groups (Adadi & Berrada, 2018). Machine learning bias occurs when an algorithm produces results that are systemically prejudiced due to erroneous assumptions in the machine learning process (Pratt, 2019). Thus, a biased algorithm is one whose decisions are skewed toward a particular group of people. AI bias could result to not just prejudicial decisions, but also imbalanced representations of certain groups. Bias can creep into algorithms in several ways. Artificial intelligent systems make decisions based on training data, which may contain prejudiced human decisions or reflect historical or social inequities. Concerns over racial and gender bias in AI have been seen in various applications including hiring, policing, judicial sentencing, and financial services. If this extraordinary technology is going to reach its full potential, addressing bias needs to be a top priority.

The move towards the adoption of artificial intelligence systems in the criminal justice system have been slow because of increasing concerns regarding the lack of transparency of algorithms and claims that the algorithms are entrenched with biased and racist sentiments. The bias against machine learning is one reason that causes a slow and patchy uptake of computer facilitated decision-making in the criminal justice area, despite the fact that ostensibly this field is a fertile area for the use of algorithms (Bagaric, Hunter, & Stobbs , 2020). This study therefore aims to capture the reasons for AI bias in judicial systems. Specifically, the study seeks to Survey current research efforts and Recommend measures to mitigate AI bias in the judiciary sector.

The AI Bias

Despite the numerous benefits of the AI technology and with the technological advancement potentially endless, a risk-averse approach its deployment for court decision-making is essential. Prior to the introduction of any algorithm in a bid to replace or assist a human judge in delivering judgment, we have to ensure that it will make a decision that is at least as fair and defensible as the human arbiter would. The European Commission's High-Level Expert Group on artificial intelligence considers an AI system to be trustworthy when it is lawful, ethical and robust. The major issue for contemplation when considering an ethical artificial intelligence is a **presence of a bias**, either a conscious or an unconscious one, in the algorithm itself or in the data, as it can control and misrepresent the calculation and prediction process. Bias and inaccuracy reduce AI algorithmic criminal justice systems inappropriate for risk assessment when making decisions on either to incarcerate persons or discharge them. Bias can arise as a result of numerous causes, including but not limited to the algorithmic design or the inadvertent or unanticipated use or decisions relating to the way data is coded, collected,

selected or used to train the algorithm. Any organized algorithmic bias in these tools will have a high risk of errors and snowballing disadvantage.

Main Sources of AI Bias

There are three chances for AI to develop bias: data, algorithms and people (Genway, 2021).

Data: This is the first place to look for bias in an AI system: the data used to train the artificial intelligent system. The training data for an AI system should be of adequate size, and should be a good representation of the real-world use. This is not usually the case. It has been shown that even the largest datasets, often gathered from real-world decisions, often show human partiality and fundamental social bias. AI systems don't have a distinctive knowledge of whether their training data is objective or subjective. The output of any learning system is determined by the input data it received. A skewed training set would equally produce a skewed output (Hammond, 2016). If biases find its way into the training data, they will be learnt by the algorithms.

Algorithms: Although algorithms do not introduce bias where there is none, they could magnify biases in the data. Algorithmic bias occurs when the data scientist trains the artificial intelligent system with biased data set. Just as human beings are products of their environments, education and experiences, AI systems are products of its algorithms, and the training data. Bias in this context has nothing to do with data. It's actually a mathematical property of the algorithm that is acting on the data. Managing this kind of bias and its counterpart, variance, is a core data science skill.

People: The final issue is with the people developing the AIs. Those designing artificial intelligent systems are often fixated on achieving a specific objective. All models are made by humans and reflect human biases. The system developers intention is to get the most accurate result with the available data, without necessarily thinking about the broader picture.

Examples of Discriminating Algorithms

Machine learning models can reflect the biases of organizational teams, of the designers in those teams, the data scientists who implement the models, and the data engineers that gather data. Several real-world examples abound where an artificial intelligent system was trained on data which is not comprehensive or diverse enough to be a good representation of the whole population in a just way. An established case is the example of the instrument used by the United States of America courts to for bail decisions making. The software called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), processes the danger of an offender to recommit an additional offence (Mehrabi, Morstatter, Saxwena, Lerman & Galstyan 2019). Juries use COMPAS to determine when to discharge a lawbreaker, or to have him/her remanded in prison. An examination of the software showed unfairness against African-Americans. This software was found more likely to allocate a greater risk score to African-American lawbreakers than to Caucasians with the same crime profile.

Another of such established case of AI bias was found in Amazon's automated recruiting engine. The giant tech company recruiting engine's software was penalizing resumes that included the word 'women' because the artificial intelligent system's training was based on historical data

over a 10-year period of time where male candidates were prevalent, and consequently seen as desirable for some positions (Dastin, 2018).

There is another case of PredPol, an algorithm already being used in many US states to predict when and where crimes would take place. This is with the goal of assisting to reduce human bias in policing and fighting. A 2016 study by the Human Rights Data Analysis Group demonstrated that PredPol was wrongly biased to target certain neighborhoods (Cossins, 2018). When scholars carried out a simulation of PredPol's algorithm to drug crimes in Oakland, California, it continually directed officers to areas with a great percentage of persons from racial minorities, irrespective of the correct situation of crime rate in the areas. While AI can help reduce the impact of human biases in decision making, it can on the other hand make the bias problem worse. AI systems learn to make decisions based on the data and algorithms humans put into them. Often, AI systems inherit human biases because they are trained on data containing human decisions. There are evidences suggesting that AI models can entrench human and societal biases and deploy them at scale (Bockius, 2020). AI models could be used to make important and life-changing decisions in several sensitive situations, therefore it is imperative to ensure that the decisions do not show any form of discriminatory behavior toward certain groups or populations.

Every decisions of the court has a great effect on the individual's private and professional life as well as on the society at large. It therefore becomes imperative to be able to recognize and ideally resolve the bias in the artificial intelligent system so as to avoid a situation where the model renders a partial or erroneous decisions, potentially magnifying the prevailing inequalities in the society.

LITERATURE REVIEW

Bolukbasi, Chang, Zou, SAligrama, & Kalai, (2016) recognizing that the applying machine learning blindly has the tendency of amplifying biases present in data, as seen with word embedding, focused on removing gender stereotypes from texts during word embedding. (ie. associating certain words to a particular gender). The study provided a system of adjusting an embedding to remove some kind of stereotypes, in this case gender (e.g. receptionist and female). The study demonstrates that word-embedding contain biases in their geometry that reflect gender stereotypes present in broader society. Owing to their wide-spread practice as basic features, they reflect such stereotypes as well as amplify them. This becomes a considerable risk and challenge for machine learning and its applications.

In order to reduce the bias in an embedding, the researchers changed the embedding to gender neutral words, by removing their gender associations. For instance, 'nurse 'is moved to be equally male and female in the direction. This significantly reduced the gender bias both directly and indirectly, while preserving the utility of the embedding.

Liang & Acuna (2019) adopting the principle of Psychophysics in experimental psychology proposed an intellectually coherent and generalizable framework to detect biases in AI. This was done in order to relate quantities from the real world into subjective measures in the mind. The authors specially embraced two-alternative forced choice task (2AFC) to extract biases in word embeddings and sentiment analysis predictions to assess possible biases, and the magnitude of those biases in black-box models. 2AFC is perhaps one of the most widely used

and simplest methods in Psychophysics. It entails asking the subjects to repetitively respond to one of two questions depending on stimuli cautiously selected by the experimenter. Based on the set of answers, then, Psychophysics uses curve fitting and interpretation to extract perception biases.

Aysolmaz, Dau and Iren (2020) used Delphi method to propose the Algorithmic Decision Making (ADM) systems development process, and identify sources of algorithmic bias at each step of the process, as well as the remedies. The Delphi method is a process used to arrive at a group opinion or decision by surveying a panel of experts. Experts respond to several rounds of questionnaires, and the responses are aggregated and shared with the group after each round. The researchers used the Delphi method to propose ADM systems development process and identify sources of algorithmic bias at each step of this process together with remedies.

Salmanowitz, (2016) discusses the potentials of adopting virtual reality technology to ease the biases between judges and jurors, and to decrease the presence of inherent racial biases in United States courts. The author was poised to suggest and study an unusual method to reduce inherent racial biases in the courtroom, precisely using the emerging technology of virtual reality. Because virtual reality models have not been clearly intended for use in the courtroom setting, he opined that his work should be seen as a thought experiment. The study establishes the benefit of integrating virtual reality into the courtroom, and not to outline plans for the actual implementation. By encouraging additional research on virtual reality paradigms, and suggesting possible policies for integrating the technology to the courts, the author opined that at the very least, the research would lead future debates on innovative and avant-garde approaches to ease iracial bias in the courtroom.

Guihot, Mathew, & Suzor (2017) argue that artificial intelligence requires supervisors to be involved in the largely unregulated field of AI, and proposed solutions to regulating the development of AI ex ante. The study suggested a two-step process, involving governments firstly setting anticipations and sending signals to influence participations in artificial intelligence development. The study adopts the term ‘‘nudging’ to denote to this form of persuading. Secondly, public regulators must participate in and interact with the relevant industries. This way, information and knowledge about the industries can be gathered risks assessments can be done and consequently regulation of the areas that may pose the most risk. The researchers called on the government of the United States of America to champion this cause, since most of the big AI companies are based in the US.

Pasquale (2019) also agrees with Guihot et. al (2017), and advocates the adoption of regulatory standards for data collection, analysis, use, and stewardship can inform and complement generalist judges.

Erdelyi & Goldsmith (2020) corroborates the findings of Guihot et. al (2017) and Pasuale (2019) posits that suitable regulation is important to exploit the benefits and minimize the risks coming from artitificial intelligence systems. Erdelyi and Goldsmith (2020) argued that artificial intelligence-related issues cannot be embarked upon successfully without honest international coordination supported by robust, consistent domestic and international governance arrangements. With this in mind, the study proposed the institution of an international AI governance framework structured round a new artificial intelligence

regulatory agency, that drawing on interdisciplinary expertise could help creating uniform standards for the regulation of AI technologies and inform the development of AI policies around the world.

Gavoor, and Teperdjian (2021) also advocates the appropriate regulation of AI in administrative agencies in order to balance technological innovation with legal compliance and fidelity to well-tread limiting principles. The researchers proposed practical solutions that the Executive Branch should deploy to harness the most meritorious components of AI in their work while minimizing avoidable pitfalls. The study posits that the technical expertise of Office Science and Technology Policy (OSTP) coupled with the regulatory and implementation expertise of Office of Information and Regulatory Affairs (OIRA) will enable the efficacious implementation and oversight of administrative AI in the United States of America. These reforms can readily be accomplished by the issuance of an Executive Order or Presidential Memorandum.

Adadi & Berrada, (2018) inspired by the earlier concerns and annotations that artificial intelligence algorithms have opacity issues, meaning that it is challenging to have an understanding of their internal mechanism of work, especially Machine Learning (ML) algorithm, proposed Explainable Artificial Intelligence (XAI) in other to work towards a more transparent system. The aim of XAI is to create a set of procedures that yield a more explainable model at the same time upholding high performance levels. Explainability systems could assist to identify if the features considered in a decision reflect bias and can allow more responsibility compared to human decision making. (McKinsey, 2019).

Silberg and Manyika (2019), agreeing with Adadi & Berrada (2018) argued the inclusion of explainability techniques in AI system , the difficulty when using neural networks of explaining how a particular prediction or decision was reached and which features in the data or elsewhere led to the result can also play a role in identifying and mitigating bias. These techniques include local interpretable model-agnostic explanations (LIME), integrated gradients, and testing with concept activation vectors. The authors argued that explainability systems could help identify if the factors considered in a decision reflect bias and could enable more accountability than in human decision making, which typically cannot be subjected to such rigorous probing.

Ferrer , Nuenen, Such, Cot´e, &Criado, (2020) propose a synergistic approach that allows us to explore bias and discrimination in AI by supplementing technical literature with social, legal and ethical perspectives.

Pena, Serna, Morales & Fierrez, (2020) presented a new public experimental framework around automated recruitment aimed to study how multimodal machine learning is influenced by biases present in the training datasets: FairCVtes. The study assessed the capability of standard neural network to acquire biased target functions from multimodal information sources, comprising images and structured data from curriculum vitae, and developed a discrimination-aware learning technique centered on the removal of delicate information such as gender or ethnicity from the learning process of multimodal approaches, and applying it to automatic recruitment model for improving fairness. The researchers established the capability of generally used learning approaches to uncover sensitive information (e.g. gender and

ethnicity) and the need to implement appropriate techniques to guarantee discrimination-free decision making processes. Finally, the study focused on debated modern approaches to preclude undesired effects of these biases, and then examined one of these methods (SensitiveNets) to advance fairness in the artificial intelligence-based recruitment framework. As an alternative, SensitiveNets removes sensitive information at the learning process instead of doing so at the input level, which may not be possible or practical.

Bellamy, Dey, Hind, Hoffman, Houde, Kannan, Lohia, Martino, Mehta, Mojsilovic, Nagar, Ramamurthy, Richards, Saha, Sattigeri, Singh, Varshney, Zhang (2018) introduces a novel open source Python toolkit for algorithmic fairness, AI Fairness 360 (AIF360). The toolkit helps to facilitate the transition of fairness research algorithms to use industrially and to offer a common basis for fairness researchers to share and appraise algorithms. The study is the first system to bring together in one open source toolkit: bias metrics, bias mitigation algorithms, bias metric explanations, and industrial usability. By incorporating these aspects, AIF360 could support robust partnership among AI fairness researchers and practitioners, assisting the translation of our collective research results to practicing data scientists, data engineers, and developers deploying solutions in diverse industries.

Lloyd & Hamilton (2018) is of the opinion that both the government and public sector establishments have a obligation to citizens to begin a discussion with system developers and issue considerate policy round data standards to guarantee varied representation in datasets to avert bias magnification, while ensuring that artificial intelligent systems are put up with inclusion in mind.

Sharmar (2020) surveys various kinds of bots and discourses the consequences and bases of biases in chatbots. The study embarked on the review of literature, and uses real-world instances to underscore that the performance of chatbots imitates our societal biases, which cannot be treated in isolation. The study proposes the importance of working on both the design deficiencies in the algorithm itself as well as on the manner data is chosen for training the algorithms. The study opines that the challenge can be alleviated by adopting a tripartite strategy: (1) careful selection of data for training purposes, (2) due diligence in algorithm selection and design, and (3) incentivizing and rewarding responsible behavior by businesses. The study also adopting the work of (Schlesinger et. al., 2018) concludes that objectionable talks is difficult for chatbots because their understanding of the different contexts that their human counterparts use is often difficult to capture, even by the most sophisticated database.

Sperrle & Keim (2019) are of the view that human involvement in machine-learning processes is often necessary due to the limitations of current models and the non-availability of domain knowledge to those systems. This involvement introduces the problems of bias and model manipulation. Systems should model their trust-level towards their users. Drastic model changes should not be allowed until users have proven that they are consistent in their interactions.

Amini, Soleimany, Schwarting, Bhatia, & Rus, (2019) set to tackle the challenge of integrating debiasing capabilities directly into a model training process that adapts automatically and without supervision to the shortcomings of the training data, and developed a novel, tunable algorithm for mitigating the hidden, and potentially unknown, biases within training data. The

approach adopted by the study involves an end to-end deep learning algorithm that learns the desired task (e.g., facial detection) as well as the fundamental latent structure of the training data simultaneously. The study showed that given a biased training dataset, our de-biased models show improved classification correctness and reduced categorical bias through race and gender as against standard classifiers.

Greene, Shmueli, Fell, Lin, Shope, & Liu, (2020) describe four types of inconsistencies introduced by risk prediction algorithms which threaten to violate the principle of treating similar cases similarly and often arise from the need to operationalize legal concepts and human behavior into specific measures that enable the building and evaluation of predictive algorithms. The authors argued that several of these inconsistencies arise not from the final application of the algorithms themselves, but rather from the human elements involved earlier in the data gathering, cleaning, and variable selection phases. A second source of inconsistency arises due to selecting predictors—the measures used as inputs into the predictive algorithm, which are assumed to be predictive of the outcome. Finally, data scientists may add inconsistency when attempting to improve model performance. For example, an outcome measure including rare crimes might create an unbalanced dataset that is difficult to model; the data scientists might thus group together the set of rare crimes. The authors recommend that it is especially important to assess predictive performance in a context as close as possible to the deployment context. The algorithm fuses the original learning task with a variational auto encoder to learn the latent structure within the dataset and then adaptively uses the learned latent distributions to re-weight the importance of certain data points while training. The algorithm has the capability of recognizing instances of under-representations in the training dataset and consequently improves the likelihood with which the learning algorithm samples these data points. The study concludes that this difficulty can be averted by modifying our alignment to human-machine conversation. While it may be difficult/impossible for one chatbot or database to accomplish, bots with specialized areas of expertise could be used, and then trained in understanding context.

Bagaric, Hunter, & Stobbs (2020) carried out an examination on the desirability of using algorithms to predict future offending and in the process analyze the innate resistance that human have towards deferring decisions of this nature to computers. It happens that many individuals have illogical misgivings about the use of computers for decision-making, known as “algorithmic aversion”, resulting from the confidence that humans would most likely make more precise decisions than machines, particularly when the issue encompasses a large quantity of multifaceted and nuanced variables. The study provided recommendations vis-à-vis the steps that are required to overcome algorithmic aversion and provide the basis for the advancement of fairer and more effective sentencing, parole, and probation systems. The study found that the key is to identify precise and indirect causes of biases in the integers that drive the algorithms. An additional and significant condition is for the coding to be transparent, incorporating new systems into the existing systems methodically and slowly in order to guarantee suitable approval and not to weaken trust in the criminal justice process.

THE TRUST-WORTHY AI

The continued and rapid expansion of AI have left many businesses with the challenges that are more human than machine. As artificial intelligence continues to disrupt and enable businesses in nearly every industry, these set of challenges ironically slows its widespread deployment.

The risks of a misbehaving AI tool are also increasing exponentially to the point where it literally becomes a matter of life and death. Real-world instances of these systems gone askew exist, where AI systems are biased against people based on sex, race, age, etc. As these intelligent systems are deployed on large scale, the attendant risks also increases, hypothetically having serious consequences for the society at large. Sentiments and understandings about what makes artificial intelligence trustworthy differ, preconditions and (ethical and regulatory) requirements that have to be satisfied are unequally prioritized across the world. Considering that 'trust' in general is a complex phenomenon that has sparked many scholarly debates in recent decades, it is not surprising that the conceptualization of trust in AI and what makes AI trustworthy-as of today-remains inconclusive and highly discussed in research and practice. According to the Independent High –Level Expert Group on AI, 2019, trustworthy AI is founded on the knowledge that trust shapes the foundation of societies, economies, and sustainable development. It follows therefore that the society can only be able to actualize the full benefits of AI if trust can be established well in it. Artificial intelligence is seen as trustworthy by its users if its development, deployment and use are in a manner that not only guarantees its obedience to all applicable laws but also specially adheres to general ethical values (Thiebes, Lins, & Sunyaev, 2020.

Qualities of a Trustworthy Artificial Intelligence

Fairness: An AI system must be designed and trained to be fair and consistent in making decisions. It must have in place internal and external checks to decrease bias. Artificial intelligence systems learn from the data used for its training. If the training data contains real-world bias, the system acquires it, magnifies and proliferates the bias at a high speed and scale. The scaling of such bias may lead to inadvertent and indirect prejudice against certain groups or people, thus worsening partiality and marginalization. Involving experts from diverse background to be part of the AI development process might help in ensuring that the developed system is free from bias. This move advances the course of militating against AI bias by ensuring that diversity of opinions is factored in in the system development. Any detected bias should be understood, and then eased through the recognized procedures for determining the issue, and rebuilding client trust in the system.

Transparent and Explainable: A trustworthy AI system allows all participants the right to understanding how their data is being deployed, and how the system makes decisions. The aim of the algorithm must be transparent and clear in such a manner that non-experts in data analytics can clearly understand it. The system algorithms, characteristics and relationships must be made open to all, and the decisions fully explainable. Artificial intelligence should no longer be seen as a "black box" , that generates output after receiving an input, without a clear knowledge of what goes on inside (Saif &Ammanath, 2020) the United states of America court system that uses AI tools are facing the argument over the use of opaque artificial intelligent tool to inform criminal sentencing decisions. The explainability traits of an AI tool concern the ability to explain both the process of the system, and the decisions taken by the system. It should be traceable and understood by persons concerned. Wherever the system's decision has a significant impact on people's lives, appropriate explanations of the AI's decision making process becomes imperative. These explanations should be timely, and adapted to the level of the persons concerned.

Responsible and Accountable: This requirement necessitates that checks should be in place to ensure responsibility and accountability of artificial intelligent systems, their outcomes before and after their development, deployment and use. This is a key requirement that ensures that the technology itself is not blamed in cases of poor decisions and miscalculations. As AI is increasingly being deployed in taking critical decisions for instance in criminal sentencing, disease diagnosis, autonomous driving, etc., policies should be established to show who should be held accountable and responsible for their output. If an autonomous vehicle causes a collision for instance, who becomes responsible and accountable for the damage? Will it be the vehicle, the owner, the vehicle manufacturer or the artificial intelligence agents?

Robust and Reliable: an AI system should be technically robust and reliable just as the traditional system or people it is trying to augment or totally replace. It must be readily available when needed, generating consistent and reliable outputs at all times. It should be scalable, maintaining robustly and reliability as the impact grows and increases. An artificial intelligent system should be built with a pre-emptive approach to danger, in such a manner that they behave reliably as intended while reducing unintended and unanticipated risk. This also applies to possible changes in their operational settings or the presence of other agents that may relate with the system in an antagonistic way. They should be protected from vulnerabilities that may allow them to be abused by enemies. Security threats can lead to erroneous decisions or harm.

Privacy: This is one of the fundamental human rights, principally affected by the AI systems. It is important for all type of data system, and particularly important for artificial intelligence since the data used in the development of AI systems are always very personal and well detailed. A responsible system should comply with data regulations, using data only for the purpose for which it is stated.

Safe and Secure:an AI system should be protected against security risks and threats that could lead to danger of both physical and digital harm. Safety and security is important for any AI system because of the huge and increasing role and effect on real-world events.

RECOMMENDATION

An investigation of bias in artificial intelligence should recognize that bias and prejudice mainly come from biases inherent in humans. The AI models and tools developed are usually a reflection of who we are. Machine learning systems are a product of what they eat. The training data for these natural language processing problems is human language, thus the models have the propensity to spread human bias. In a bid to develop a system that is fair, efforts must be channeled by the various stakeholders involved in the development of the system to correct the gender imbalance. This is already argued by Bolukbasi, Chang, Zou, SAligrama, & Kalai, (2016), who advocates the elimination of gender and racial sensitive words, while using general neutral words at the word embedding phase. The human understanding g of how input affects the reliability of an AI system is a crucial factor in the development of any AI system. To this end, it is pertinent to ensure that the right people are providing the input data, paying particular attention to their training with respect to bias and ethics.

Because of the sensitive nature of the application of AI systems in judiciary, and the need for parties to know the reason for any judgments given, the works of Adadi & Berrada, (2018) and

Silberg & Manyika (2019), that includes the explainability techniques in AI systems in their model comes to the fore. It is important to identify the use cases in the development of an AI system, for which explainability is necessary. This is to foster fairness, and offer grounds for the party to appeal a judgment if need be.

The Supreme Court which is the highest court in the country should be part of AI model building, since judgments emanating from SC are considered both 'final' and 'supreme'. The justices should be involved in the data preprocessing stage of the system development. To tackle bias, they need to understand its sources. To reduce bias in data they need a true understanding of the underlying data and the hidden values and human nuances it represents. They should be ready to question bias at every stage and identify where there is need for external expert input.

CONCLUSION

Massive datasets availability has made it easy to learn new insights through computers. Algorithms therefore have become a more revealing and pervasive tools for computerized decision-making. Though algorithms are used in many contexts, the focus should be on intelligent tools that make decisions using data from people, containing their identities, demographic attributes, preferences, and their likely future behaviors, as well as the objects related to them, and ensure that such a system is well designed and developed, and free from any form of bias and prejudice. It is tempting to assume that, once trained; a machine-learning model will continue to perform without oversight. In reality, the environment in which the model is operating is constantly changing, and managers need to periodically retrain models using new data sets.

With Artificial intelligence systems being deployed in diverse applications, operators and every concerned stakeholder must be thorough and proactive in addressing issues that may contribute to bias. Creators of the machine-learning models that will drive the future must consider how bias might negatively impact the effectiveness of the decisions the machines make. Confronting and reacting to algorithmic bias would possibly forestall damaging impacts to users and weighty responsibilities against the operators and creators of algorithms, including computer programmers, government, and industry leaders. Failure to do this could lead to emasculating AI's potentially positive benefits by building models with a biased mind of their own.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6 , 52138–52160.

Amini, A; Soleimany, A; Schwarting, W; Bhatia, S& Rus, D (2019). Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. In Proceedings of the 2019 AAAI/ ACM Conference on Artificial Intelligence, Ethics, and Society (AIES), 27-28 January, 2019, Honolulu, Hawaii, United States, AAAI/ACM, 2019.

Angileri, J; Brown, M; DiPalma, J; Ma, Z & Dancy, C () Ethical considerations of facial classification: Reducing racial bias in AI. https://www.researchgate.net/profile/John-Angileri/publication/338225415_Ethical_Considerations_of_Facial_Classification_Reducing_Racial_Bias_in_AI/links/5e0916c6299bf10bc382bb88/Ethical-Considerations-of-Facial-Classification-Reducing-Racial-Bias-in-AI.pdf.

Aysolmaz, B; Dau, N & Iren, D (2020) Preventing Algorithmic bias in the development of algorithmic decision-making systems: A Delphi study. In proceedings of the 53rd Hawaii International Conference on System Sciences 2020S

Bagaric, M; Hunter, D; & Stobbs, N (2020) Erasing the Bias Against Using Artificial Intelligence to Predict Future Criminality: Algorithms are Color Blind and Never Tire, 88 U. Cin. L. Rev. 1037 (2020) Available at: https://scholarship.law.uc.edu/uclr/vol88/iss4/3

Bellamy, R.K; Dey, K; Hind, M; Hoffman,S.C; Houde, S; Kannan, K; Lohia, P; Martino, J; Mehta, S; Mojsilovic, A; Nagar, S; Ramamurthy, K.N; Richards, J; Saha, D; Sattigeri, Singh, M; Varshney, K.R & Zhang, Y (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:1810.01943v1 [cs.AI] 3 Oct 2018

Bolukbasi, T; Chang, K; Zou, J; SAligrama, V & Kalai, A (2016) Man is to Computer program as woman is to homemaker? Debiasing word embeddings. arXiv:1607.06520v1 [cs.CL] 21 Jul 2016

DeBrusk, C (2018) The risk of machine learning bias and how to prevent it. Risk Journal/ Rethinking Tactics 2018. https://www.oliverwyman.com/content/dam/oliver-wyman/v2/publications/2018/november/Risks-of-machine-learning.pdf

Erdelyi, O.J and Goldsmith,J (2020) Regulating artificial intelligence proposal for a global solution. https://arxiv.org/pdf/2005.11072.pdf.

Ferrer , X; Nuenen, T; Such, J.M; Cot´e, M &Criado, N (2020) Bias and Discrimination in AI: a cross-disciplinary perspective.
arXiv:2008.07309 **[cs.CY]**

Gavoor, A and Teperdjian, R. A Structural Solution to Mitigating Artificial Intelligence Bias in Administrative Agencies (January 18, 2021). 89 George Washington Law Review Arguendo (2021 Forthcoming), Available at SSRN: https://ssrn.com/abstract=3768214 or http://dx.doi.org/10.2139/ssrn.3768214.

Genway, S (2021) Three causes of AI bias and how to deal with them. Tessella Altrain Group. https://www.tessella.com/insights/three-causes-of-ai-bias-and-how-to-deal-with-them

Greene, T; Shmueli, G;Fell, J; Lin, C; Shope, M.L & Liu, H (2020). The Hidden Inconsistencies Introduced by Predictive Algorithms in Judicial Decision Making. https://arxiv.org/ftp/arxiv/papers/2012/2012.00289.pdf.

Guihot, M; Mathew, A.F & Suzor, N.P (2017) Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence. Vanderbilt Journal of Entertainment and Technology Law, 20(2), pp. 385-456.

Hammond, K (2016) unexpected sources of bias in artificial intelligence. https://techcrunch.com/2016/12/10/5-unexpected-sources-of-bias-in-artificial-intelligence/

Lee, N.T; Resnick, P &Barton, G (2019) Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Lee, N.T; Resnick, P &Barton, G (2019). Algorithmic bias detection and mitigation: best practices and consumer policies to reduce consumer harms. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Liang, L & Acuna, D.E (2019) Artificial mental phenomena: Psychophysics as a framework to detect perception biases in AI models. . ArXiv: 1912.10818v1 [cs.CL] 15 Dec 2019

Lloyd, K & Hamilton,B.O () Bias Amplification in Artificial Intelligence. https://arxiv.org/ftp/arxiv/papers/1809/1809.07842.pdf.

Mehrabi, N; Morstatter, F; Saxena,N; Lerman, K & Galstyan, A (2019). A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635v2 [cs.LG] 17 Sep 2019

Pasquale, F (2019) Data –informed duties in AI development. Columbia Law Review. Vol 119-m1917. https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=2629&context=fac_pubs.

Pena, A; Serna, I; Morales, A and Fierrez, J (2020) Bias in Multimodal AI: Testbed for Fair Automatic Recruitment. https://openaccess.thecvf.com/content_CVPRW_2020/papers/w1/Pena_Bias_in_Multimodal_AI_Testbed_for_Fair_Automatic_Recruitment_CVPRW_2020_paper.pdf.

Pratt, M (2019). Machine Learning bias (AI bias) https://searchenterpriseai.techtarget.com/definition/machine-learning-bias-algorithm-bias-or-AI-bias

Salmanowitz, N (2016) Unconventional Methods for a Traditional Setting: The Use of Virtual Reality to Reduce Implicit Racial Bias in the Courtroom, the University of New Hampshire Law Review, Volume 15, Article 2, November, 2016.

Schlesinger, A., O'Hara, K. P., & Taylor, A. S. (2018, April). Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 315). ACM.

Sharma, A (2020) Bias in AI-Enabled Chatbots: Implications, causes and solutions. https://anukritisharma.com/blog/2020/3/24/bias-in-ai-enabled-chatbots-implications-causes-and-solutions

Silberg, J & Manyika, J (2019) Notes from the AI frontier: Tackling bias in AI (and in humans). Mckinsey Global Institute. https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/MGI-Tackling-bias-in-AI-June-2019.pdf

Sperrle, F & Keim, M.D () Human trust modeling for bias mitigation in artificial intelligence. http://kops.uni-konstanz.de/bitstream/handle/123456789/45547/Sperrle_2-1dk03mfjgsyvb7.pdf?sequence=1&isAllowed=y

Thiebes, S; Lins, S &Sunyaev, A (2020) Trustworthy artificial intelligence. Electronic Markets, Springer https://doi.org/10.1007/s12525-020-00441-4