Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

ISSN NO. 2320-5407

Article DOI: 10.21474/IJAR01/11418
DOI URL: http://dx.doi.org/10.21474/IJAR01/11418

*RESEARCH ARTICLE*

# IDENTIFICATION AND MITIGATION OF ALGORITHMIC BIAS THROUGH POLICY INSTRUMENTS

**Rahul Sethi[1], Vedang Ratan Vatsa[2] and Parth Chhaparwal[3]**

1. Indian Institute of Technology Kanpur, India.
2. IT & Management Consultant.
3. Indian Institute of Technology Kanpur, India.

## Manuscript Info

*Manuscript History*
Received: 28 May 2020
Final Accepted: 30 June 2020
Published: July 2020

*Keywords:-*
Algorithmic Bias, Artificial Intelligence, Machine Learning, Policy, Data Governance, Data Analysis

## Abstract

With the increasing implementation of technologies like Artificial Intelligence, Machine Learning and sophisticated data analysis algorithms, the process of decision-making and recommendations is being automated with more emphasis on a trained system to generate valuable required outcomes. Taking into consideration the aspect of how the mechanism of any computer-based decision-making process takes the order, this paper aims at understanding the possibility of bias in any output, which may, further, become a cause for bias in the real world. This paper tries to address a stepwise approach to discovering the factors of how such a bias can be quantified and presents a view of the impact parameters for various stakeholders while trying to process an approach for the overall fairness of the system.

## Introduction:-

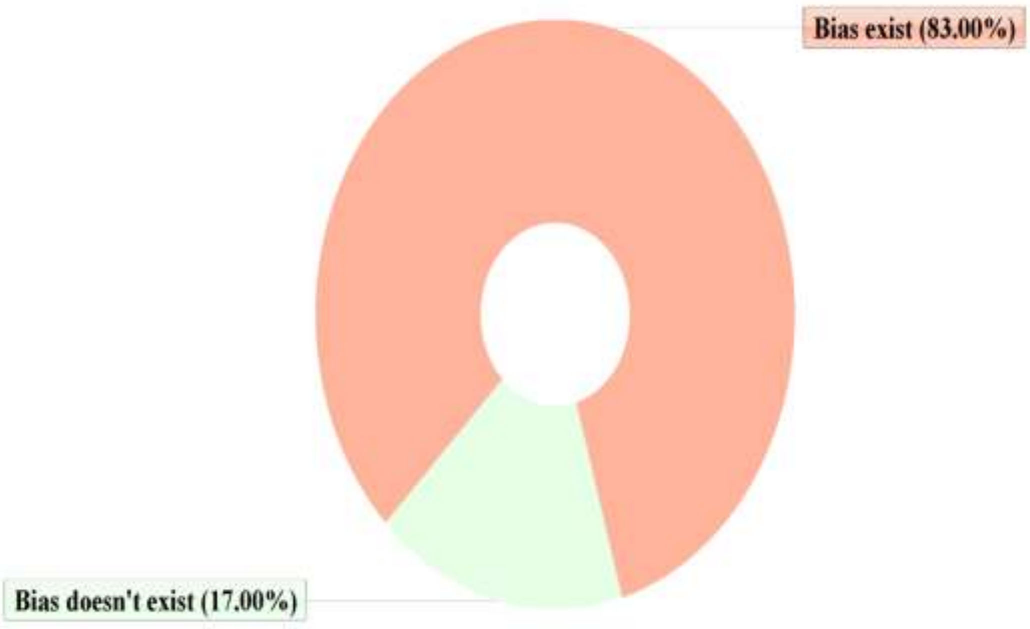
Technological advancements in the last few decades have made us more dependent on computer systems now than ever before, and this has been possible with the help of Artificial Intelligence, Deep Learning, and Machine Learning algorithms[1]. These technologies help to automate decision making at multiple levels - be it on social media, online shopping, search engines, voting behavior and other things too numerous to mention. While this makes our life simpler in many of these fronts and has proved to be very useful, but, in hindsight, computer scientists have noticed that there are some negative impacts linked to it as well.

Algorithmic bias describes systematic and repeatable errors from the computational outcomes that may create a lack of fairness in the process of coming out to a result via a computational process. Algorithmic bias is not new to us and experts have been warning about it for quite some time[2]. In today's scenario, algorithms are ubiquitous and play an important role in our daily life, and many of our decisions are influenced by it, making algorithmic bias a critical issue to be looked upon. The algorithms used for developing the software are trained on data sets, and many a time these data sets aren't properly labeled. An algorithm will become better at the task it is supposed to perform if it is trained with more data. But this training data is often produced with less accuracy, creating a fundamental base for the bias.

A quick survey with 115 entries, asking 'What do you think about the fairness of computational decision-making and recommendation outcomes?', reveals that the majority thinks that bias exists in the outcomes of computational

**Corresponding Author:- Vedang Ratan Vatsa**
Email Address:- vedangvats@gmail.com

decision-making and recommendation process, which acts as a basis of our study to understand more about how a framework can be developed to study the impact of such a bias.



**Existence of Algorithmic Biases:**

1. The criminal justice algorithm COMPAS (Correctional Offender Management Profiling for Alternate Sections), which is used by courts for law enforcement - to predict whether the accused should be released or detained, was found biased against African-Americans according to ProPublica in 2016[1].
2. Gender based discrimination in the process of displaying STEM career ads[3].
3. According to research by Princeton University, it was found that the keywords ‘women’ or ‘girls’ were more likely to be linked to arts and humanities, whereas ‘men’ or ‘boys’ were more linked to mathematics and sciences. Similarly, African-Americans’ names were considered less ‘pleasant’ as compared to Europeans[1]. This was based on Machine Learning software used to link different words and ignorantly reinforced the existing gender, caste, and regional bias by humans.

**Evaluating Types of Algorithmic Governance:**

The key idea behind algorithmic governance is that digital technologies structure our society in multiple ways. Transparency matters a lot as it is a founding pillar of government, but since the algorithms are too complicated in their structure and aren’t easy to understand, so it’s not the main underlying issue[4]. There are many factors that need to be looked upon to comment on whether algorithmic governance is beneficial. Some of the controversies and concerns arising out of algorithmic governance include:

**Datafication and Surveillance**

Tons of data are being generated in almost every sector in the present times, and processing this overwhelmingly large set of data is becoming harder for scientists across the world. This situation is known as ‘data deluge’ and leaves us with no option with the existing technologies to process this entire set of data. Hence the surveillance of this data is not easily possible, and it is mostly left under-processed.

**Agency and Autonomy**

Algorithms are all-over today in our day to day lives, and this has led to questions of its possible impact on human autonomy and agency. Humans can act in accordance with objective morality, but machines and programs can’t. There is also a fear of AI takeover, with machines and computers taking away control of the planet from humans and becoming the most intelligent species on Earth.

### Transparency and Opacity

Data Transparency becomes important when automated decision-making systems are questioned about their fairness of the decisions affecting an individual[5]. But too much transparency can backfire for the companies, as modern AI is making other factors apart from the source code more important, which includes their training data sets. These machine learning algorithms - especially deep learning methods, only have code of about a few hundred lines. Even with this transparency, one would have to look up this massive data and understand its algorithms. This would mean that responsible companies would not hesitate to have transparency, rather they could educate their users of the key factors driving their algorithm's decision.

### De-Politicization and Re-Politicization:

It has always been a common belief that algorithms are de-politicizing, because of their 'objectivity and truth'. But the notion of algorithmic bias defies it, not completely but to a certain extent. This is however due to the prevalent social inequality and discrimination that gets reflected in our data sets and hence algorithms. It leaves people doubting if these algorithms were made biased to discriminate against a set of people and may even be used for re-politicization if there are vested interests.

### Bias and Fairness:

The biases in algorithms reflect the discrimination and social inequality already prevalent in society and the algorithm is not to be blamed about it. There are biases in our data sets and models. Although, once these biases are reflected in algorithms during operation, it only amplifies the social discrimination and prejudices. Hence, it needs to be removed from the very basic level and we need to look for a way to end these social, gender, caste and racial biases from our society if we want to eventually make our algorithms and automated decision-making process fair.

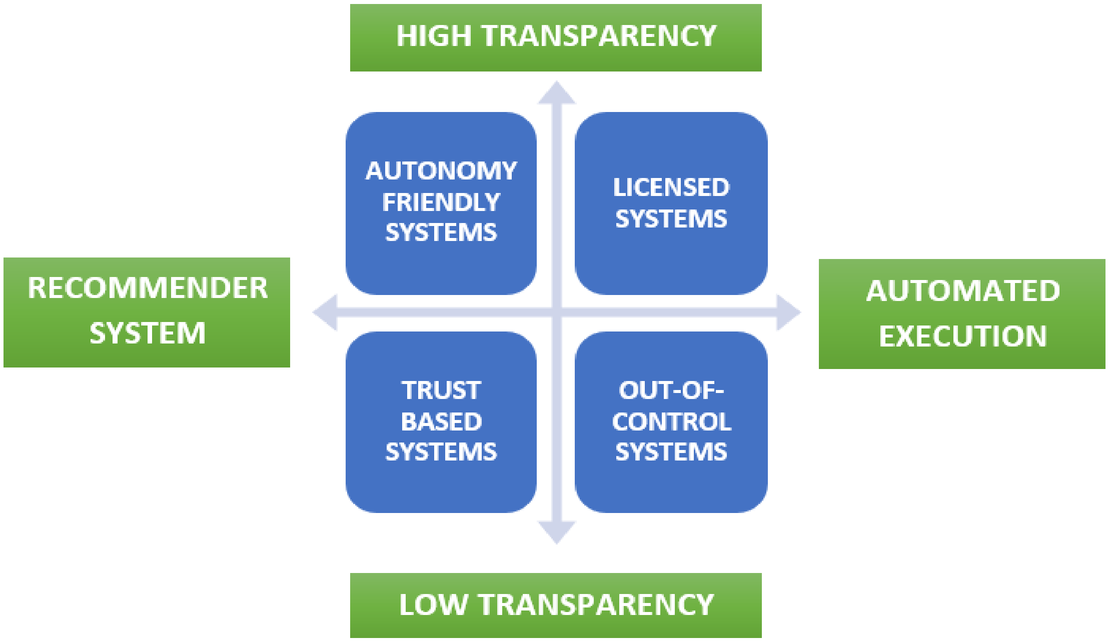


**Figure:-** Types of algorithmic governance systems.

### Algorithmic Impact Assessment:

The automated algorithmic decision making, making use of certain parameters of evaluation structure the format of how various public systems work, like in case of the criminal justice system, criminal activity prediction, predictive policing, energy usage optimization, personalized education, performance evaluation systems, profile matching algorithms, etc. Often such systems operate as black boxes with little or no transparency since it generally stands out of the perspective of external scrutiny, monitoring and accountability[6]. Some of the key elements to consider while considering the inadequacies of algorithms include:

1. Self-assessment of the automated decision systems by the implementing organizations to understand its impact on the factors of biasing within the community members.

2. Periodic external technological auditing of the processes to consider the diverse impacts and evaluate the techniques.
3. A transparent discloser of the algorithmic flow and input data for the data-based decisions.
4. A provision to enforce a due process for the beneficiaries to question the architecture base on the impact of inadequate, inefficient or biased assessments of any technological process-in-place.
5. The direct or indirect tradeoff between profitability and decisions of the concerned entity.
6. An open documentation to learn about the algorithmic data flow within a digital process.
7. Study of long sets of input and output data or information to conclude an error or ineffectiveness of the core process.

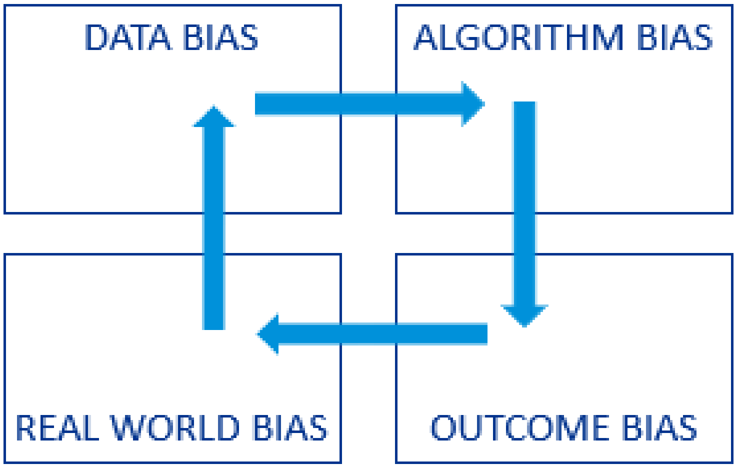
**Estimating the Impact of Bias:**

For any estimation process to conclude the impact of possible bias or error, it is important to ponder upon the following question to deeply understand and estimate the impact of underlined bias:

1. Who are the beneficiaries of the final decisions and how can they be affected by any biased result?
2. How is the training data being chosen and what were the factors of diversity being taken into consideration while collecting the data?
3. Is the input data reliable and include the diversity factors?
4. How can an algorithm be tested?
5. What are the methods to ensure the inclusion of diversity in the design and implementation phase?
6. What is the threshold for bias?
7. What are the ways in which bias may offer any form gain to the concerned stakeholders?
8. What interventions may be required to restrict the impact of bias in any system?
9. What can be the level of openness that the design for an algorithm may have?

**Framework for Evaluating Algorithmic Policy Instruments:**

The existence of algorithmic biases in technical products has persuaded cultural authorities, such as the United Nations, to establish proper legislation on algorithmic biases. Although efforts made to determine how algorithmic policies should be made have been meager, significant qualitative frameworks have been established to evaluate algorithmic policy instruments.



Before establishing a framework to evaluate algorithmic policy instruments, it is important to identify that algorithms are just instruments that enumerate how to process social and cultural content. However, these algorithms are optimized based on the data they entail, and therefore, the data becomes the source of the bias. Hence, policymakers need to establish policies on optimizations that regulate these algorithms. Every algorithm has 3 components, which are fed with data for optimization. These components include the input, the code, and the context. Bias may arise in each of the three components.

The first bias includes bias in input, which arises when the algorithm is fed with massive datasets. These datasets may be culturally biased themselves and hence reflect cultural inequity in the algorithm[7]. An example to explain this bias would be machine learning algorithms, which when fed with English texts, associate the word “genius” with men and “model” with women. The source of this bias is the dataset provided, which in turn is biased because of the procedure humans use to source this dataset.

The second bias includes bias in the code, which perpetuates when a certain technology is constructed in such a way that it works better for a certain group of people as compared to another group. An example of this is the facial recognition technology, which is wired to identify faces of color and hence, creates a racial bias[8].

The third bias includes bias in the context where the input and code meet the purpose of the algorithm. The purpose is the real-world application which comprises a business-model and biases catering that business model paves way for bias to creep into the algorithm. An example of this would be news platforms, whose aim is to maximize their profits and focus on the major section, knowingly or unknowingly providing less focus on minorities. This bias initiates the other 2 biases.

The above 3 biases in an algorithm are the primary barriers to algorithmic policy instruments. The following includes a basic framework to remove these barriers and evaluate algorithmic policy instruments:

**Barriers to unbiased inputs:**

1. Proprietary data: The datasets that are input in algorithms consist of information, generally inaccessible to the general public and are hence confidential. This data is hence an asset for companies and can be exchanged at high prices. Technical product firms don’t have direct access to unbiased data and hence, the lack of data accessible to these firms tends to initiate a bias.
2. Prejudice in data: Data sets fed into the algorithm might be knowingly or unknowingly prejudiced and providing value to certain parameters over the other. For example, in Boston, individuals use the Street Bump app which uses smartphone data to examine and mark road conditions. The improper penetration of this app in different locations has caused the inbuilt algorithm to discriminately allocate repair resources to different locations.
3. Personalization: Algorithms personalize processes and results as per the users. Hence, the previous information of the user can be, sometimes, used to decide what should be shown and hence presents a single perspective to the user.
4. Localization: Algorithms tend to customize processes and results based on location. For example, a news algorithm may suggest trending news as per the location of an individual

**Barriers to unbiased code:**

1. Existence of black boxes: Most software doesn’t reveal its underlying logic due to security reasons and to maintain a competitive high ground. This causes consumers to be unaware of the methodology of processes and results being offered. With the advent of sophisticated techniques like machine learning, it can be difficult for both developers and users to identify the underlying logic.
2. Lack of traceability: An existing bias can be eliminated if it may be traced. However, due to the presence of a plethora of sources of bias, it becomes extremely difficult to trace it. For example, children’s Youtube recommendation system works on machine learning, the bias which can be attributed to either of the developer, the training data or malicious users.
3. Unpredictability of code: The presence of millions of lines of code for an industry-ready software makes it prone to errors and more difficult to find bugs. Complex systems can be difficult to manage, and biases may creep-in while fixing bugs.
4. Instability: Software firms nowadays release multiple updates in a short duration of time, owing to cloud computing and other such sophisticated technologies. The algorithm of the software is hence sensitized to a constantly changing operation of a dataset.
5. Lack of diversity in development teams: While drafting the code for a software, the lack of perspectives among developers makes them miss out on important aspects in avoiding bias. The lack of perspectives might be due to a lack of diversity in the development teams. For example, the tech industry still has more male than female members.

**Barriers to unbiased context:**

1. Anticipated use and goals: While establishing the optimization technique, firms tend to anticipate the behavior of the users and hence introduce biasing to establish a perfect business model and maximize their profits.

2. Changing media habits: Data has integrated into the media industry and is providing players in the media industry with information such as the demographics of people living in a location and their usual times of browsing. Content optimized as per this information may be biased towards a group of people.
3. Vertical integration: When on a platform, a firm is involved in the production and sales of its own content, the inbuilt algorithm might be conferred in such a way that it provides a bias towards the firm's own product over another firm's product. This is precisely common in E-Commerce firms selling their own products, such that pricing policy is biased towards the contents produced by the firm.
4. Manipulation: The most apparent way of creating bias is data manipulation. Inorganic users to fudge numbers and create a disturbing dataset may create a forced bias in the algorithm.

**Critical Questions for Algorithmic Accountability:**

The above barriers to unbiasedness in algorithms can be abridged into a set of 3 question sets, that provide the tool to deem an algorithmic policy as efficient. The following are the questions:

- Is the data entered in an algorithm culturally exhaustive?

Is the source of this data dependable and does that data differ significantly from similar data sets of other cultures? Is the dataset free from prejudices?

- How effectively is the algorithm regulating the data and deriving bias-free results?

Are users and developers aware of the logic behind the algorithm? How effective is automation while maintaining unbiasedness? Are the individuals who established the logic culturally exhaustive?

- Are algorithms efficient in any given cultural situation?

Does the algorithm behave vaguely when users change their behavior? Can the algorithm account for conflicts in user expectations? Do the makers of the algorithm consider cultural unbiasedness as an important factor while developing the algorithm?

**Fairness and Accuracy Tradeoff:**

The above discussion has led to the addition of an important driver of an ideal algorithm, that is fairness. Hence, fairness and accuracy are two aspects to determining whether an algorithm function ideally or not. Two pool of thoughts exists in the relationship between fairness and accuracy. The first pool of thought states that the fairness and accuracy of an algorithm are mutually proportional to each other. An example to support this pool is that the addition of more culturally exclusive data to an under-represented dataset to deem it as fair might also improve the accuracy of the algorithm which now has a larger data set to draw inferences from. Another proposition supporting this pool of thought is that eliminating bugs in the code to remove cultural bias and deem the algorithm as fair will also ensure that the algorithm functions in a better way, hence making it more accurate. The second pool of thought enforces a mutual inversely proportional relationship between fairness and accuracy. An example supporting this pool is that the use of machine learning on a platform tends to align the algorithm accurately with the goals of the firm owning the platform. However, in that process, machine learning creates a black box which is full of biases. Removing machine learning will reduce accuracy while improving fairness. Also, maintaining high standards of fairness may be expensive and lead to the creation of fairness costs. Both pools of thought are situationally correct. The aim of a technology company should be to optimize the algorithm of the firm while maintaining accuracy and increasing fairness. The first step to this would be to find ways to remove cultural bias among various groups without affecting the accuracy of the algorithm. The second step is to determine if the social costs are justified for the company and if cultural groups are open to the solution provided by the algorithms. If somewhere, human decision-makers are the utmost requirement for fairness, then automation should be avoided.

To cater to this issue, the European Union released "Ethics guidelines for trustworthy AI" which promotes 7 governance principles[10] to ensure cultural fairness through technology:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Environmental and societal wellbeing
7. Accountability

This document proved a breakthrough to define “fairness” in algorithms. However, it doesn’t cover the aspect of fairness from the human-perspective, where people themselves don’t push any human-made biases into an algorithm. A human arbiter to oversee the functioning and decision making in an AI program shall help ensure fairness and follow all seven points mentioned in the European Union document.

## Recommendations:-

Recommendations for having a check on the algorithmic bias include the practice of upgradation of numerous laws to include the digital processes and practices, including the ones that derive data from the data-based processing and analysis. With the advancement of digital service delivery modules and practices by the state and the organizations, it is important to have a regulatory check in place, in the form of an auditing infrastructure for digital solutions.

An implementation of sandboxes to encourage anti-bias implementation of algorithms may further facilitate and detect the sensitivity of bias while making sure that the same can be eliminated via the implementation of detect-and-mitigate frameworks in the form of approvals. It is important to have guidelines in the form of anti-bias self-regulatory practices for the technology organizations. Implementation of sandboxes have already encouraged innovation for the companies in the financial domain that are keen on leveraging the technological aspect of innovation and service delivery[11].

In the advancement towards Digital inclusion, Digital Exclusion may come out as a new form of vulnerability. On the other hand, some concerns include:

1. A voice assistant processing your voice for the "robustness" of their AI
2. Making a user allow the usage of cookies for "providing a better web experience"
3. Sharing app usage data for "seamless experience"
4. Advertisement pop-ups for something that the user just talked about with someone

Introduction of Corporate Digital Responsibility: Guiding principles in the form of ethical digital practices to ensure that technology implementation may follow a balanced tradeoff between profitability (direct or indirect) and usability (experience, design and privacy)[12].

While it is important to emphasize the practices to curb the stated bias by the state, it is imperative to make the users more literate about the algorithms embedded within the digital products that are being offered, especially for the essential services. It is important to define the ethics of the usage of advanced technologies like machine learning and artificial intelligence in the usage of various processes and implementations like that in Internet of Things[9], so that a holistic growth may be made possible covering all parts of the society. The creation and maintenance of algorithms need to consider the aspect of diversity; starting from the development team, the data used to train the applied models and the qualitative aspect of culture and other such factors while formulating the flow and architecture of a decision-making process. There may be cases when the effect of bias may bring a very small change in the overall process and output, but the aspect of user quantity may distort the overall efficiency of diversity-by-design principle. An independent review of the said facet may further facilitate the effectiveness of the overall system[13]. A formal audit by an independent entity may check for the bias both from the view of input data and output decision and can be considered as a best practice for the same.

## References:-

1. Nicol Turner-Lee, a., 2020. Algorithmic Bias Detection And Mitigation: Best Practices And Policies To Reduce Consumer Harms. [online] Brookings.
2. Dickson, B., 2020. What Is Algorithmic Bias?. [online] TechTalks. Available at: <https://bdtechtalks.com/2018/03/26/racist-sexist-ai-deep-learning-algorithms> [Accessed 21 July 2020].
3. Lambrecht, Anja & Tucker, Catherine. (2016). Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. SSRN Electronic Journal. 10.2139/ssrn.2852260.
4. Katzenbach, C. and Ulbricht, L., 2020. Algorithmic Governance. [online] Internet Policy Review. Available at: <http://policyreview.info/concepts/algorithmic-governance> [Accessed 20 July 2020].
5. Harvard Business Review. 2020. We Need Transparency In Algorithms, But Too Much Can Backfire. [online] Available at: <https://hbr.org/2018/07/we-need-transparency-in-algorithms-but-too-much-can-backfire> [Accessed 22 July 2020].

6. James Vincent, “Google uses DeepMind AI to cut data center energy bills,” The Verge, July 21, 2016, <www.theverge.com/2016/7/21/12246258/google-deepmind-ai-data-center-cooling>
7. Harvard Business Review. 2020. The Hidden Biases In Big Data. [online] Available at: <https://hbr.org/2013/04/the-hidden-biases-in-big-data> [Accessed 14 July 2020].
8. Hunt, Robert, and Fenwick McKelvey. “Algorithmic Regulation in Media and Cultural Policy: A Framework to Evaluate Barriers to Accountability.” Journal of Information Policy, vol. 9, 2019, pp. 307–335. JSTOR, www.jstor.org/stable/10.5325/jinfopoli.9.2019.0307. Accessed 23 July 2020.
9. Vatsa VR, Singh G (2015) A literature review on internet of things (iot). Int J Comput Syst (ISSN: 2394-1065) 2(08).
10. Ec.europa.eu. 2020. Ethics Guidelines For Trustworthy AI. [online] Available at: <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419> [Accessed 2 July 2020].
11. Home.treasury.gov. 2020. [online] Available at: <https://home.treasury.gov/sites/default/files/2018-08/A-Financial-System-that-Creates-Economic-Opportunities---Nonbank-Financials-Fintech-and-Innovation_0.pdf> [Accessed 11 July 2020].
12. Atos. 2020. Corporate Digital Responsibility: Facing The Bigger Picture Of Technology Change - Atos. [online] Available at: <https://atos.net/en/blog/corporate-digital-responsibility-facing-the-bigger-picture-of-technology-change> [Accessed 18 July 2020].
13. Locklear, Mallory. “Facebook Releases an Update on Its Civil Rights Audit.” Engadget (blog), December 18, 2018. Available at www.engadget.com/2018/12/18/facebook-update-civil-rights-audit (last accessed July 19, 2020).