Algorithmic Transparency:

The Need for Sophisticated Regulation

BLOCKSUITS

August, 2020

Authored by:

Samaksh Khanna
Shivani Agarwal

www.blocksuits.com
blocksuits@gmail.com

Disclaimer:
The views expressed in this White Paper are those of the author(s) and do not necessarily represent the views of the BlockSuits or its members. It should not in any capacity be construed as legal advice. White Papers are submitted to the BlockSuits as contributions to its insight areas and interactions. White Papers describe research in progress by the author(s) and are published to elicit comments and further debate.

August 20, 2020

Executive Summary

The ever increasing digitisation has led to domination of AI in almost every sphere of life, from recruitment to making judicial decisions. Successful implementation of many AI systems remains still remains a challenge due to various issues and algorithm bias is one of them.

Organisations need to fundamentally rethink their AI operating models and the government interference is of utmost importance. Organisations are more likely to think of their own commercial benefit, therefore, the government interference can ensure the three basic principles advocated by the authors:

(i) Transparency

(ii) Auditability

(iii) Accountability

The European Union has already has a governance framework in place which regulates AI and has subsequently released many guidance for a fair deployment of AI. The United States also has an Algorithmic Accountability Act of 2019 in discussion which is first ever attempt to consolidate principles of algorithmic decision making and regulate AI systems across industries.

The first step towards ensuring a fair deployment of AI is to have a data protection regime in place. In India, data protection is currently governed by the SPDI Rules (as defined hereunder), however, a Personal Data Protection Bill, 2019 (PDP Bill) is also tabled in the parliament and is likely to be passed soon. The PDP Bill does not provide any restriction for decisions solely based on automated processes rather provides the data subject (data principal in the context of the PDP, 2019) the right to receive the personal data which has been processed in a structured and machine-readable manner.

BlockSuits recommends an (i) ex-ante approach; and (ii) ex-post approach, to ensure transparency, auditability and accountability.

Introduction

Algorithms have been an essential part of organisational decision making for some time now. One of the most prominent aspects of data analytics through algorithms is the basis of the 'profiling' of individuals. Profiling is a combination of analysing personally identifiable information or 'personal data' and grouping them into segments for a case by case usage.Algorithmic accountability by regulators has often perceived to be within the scope of antitrust issues. A recent example of the same could be found with the release of India's 'e-commerce policy' wherein the Government of India (**'GOI'**) has preserved the right to seek disclosure of e-commerce businesses' algorithms to ensure that foreign e-commerce businesses do not undermine domestic e-commerce and there is no 'bias' or discrimination.However, even with the introduction of legislative instruments, India substantially lags behind other jurisdictions in regulation of technologies. This paper is focused on a potential regulatory regime for algorithmic accountability in India.

Many jurisdictions have directly related the use of algorithms with that of artificial intelligence (**'AI'**) systems as AI systems are the primary source of decision making through algorithmic analysis. Modern AI systems are developed to not only procure and follow instructions but also provide a personalised experience to the consumer by learning the behavioural patterns of consumers. This in turn means that personal data, which is being collected through online and social activities, shall be processed at multitudes of levels for the AI machines to understand and predict patterns or in simple words, learn. Such decision making by AI systems or any organisation using algorithmic patterns for the purpose of profiling bears the risk of discrimination and breach of privacy, hence, the need for adequate accountability. One could argue that the current data protection regime in India when compared with technology-heavy jurisdictions such as EU, United Kingdom (**'UK'**), Australia, Canada, and the United States of America (**'US'**), is lacking and the upcoming Personal Data Protection Bill, 2019 (**'PDP, 2019'**) does not provide an adequate solution to technological developments. Moreover, an important thing to note is that the privacy regime in any jurisdiction is ever-changing with the development of juridical standards with regards to modern technology. For example, with the increasing use of blockchain technology, the European Parliament in the EU released a study on 'Blockchain and the General Data Protection Regulation (**'GDPR'**)'. However, Indian regulators have not been remarkably regulating disruptive technologies and have not provided any guidance as to the usage/regulation/guidance towards such technologies. The increased usage of disruptive technologies such as blockchain, open banking, automated decision making, etc requires a technologically adept country like India to formulate regulations for the industry to grow within a space of automation and decentralisation. For this purpose, the authors have applied a multi-jurisdictional approach to evaluate the nature of laws that could be adopted in India.

Multi-jurisdictional analysis with comparison to India

EU-GDPR

The GDPR is based on the principle of protecting the consumer from any breach of privacy and the creation of a rigid compliance requirement. For this purpose, Article 5(5) and Article 22 of the GDPR require that controllers (organisations determining the purpose for processing of data) shall portray that they are in compliance with the principles of transparency, fairness, and lawfulness. This means that all algorithms which are being utilised for decision making shall portray no biasness/ non-discriminatory principles with a legitimate justification for processing data through algorithms. Interestingly, the GDPR under Article 4(4) has clearly defined 'profiling' to state "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular, to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements" under Article 4(4). The definition includes behavioural aspects of a data subject based on socio-economic backgrounds which shall help organisations to understand and evaluate the compliance of the data that is being collected with the provisions of the GDPR. Moreover, the data subject also has the power to object to any profiling as per Article 21 of the GDPR.

The contention mentioned above is explained through the following hypothetical scenario:

Most data sets are utilised to provide a score or a ratingto people based on various patterns such as e-commerce shopping, social browsing, etc. Such information may be utilised by credit agencies to determine eligibility forloans, economic benefits, employment, and other social benefits. Such data may be prone to bias which shall be determined by a case to case basis. The GDPR, although comprehensive, provides for a wide range of member states with flexibility. Hence, the proving of bias is a hefty task when proving a claim of breach of rights under GDPR. Credit scoring can actually be better formulated by algorithms rather than human intervention since the decision shall be based on payment data on not individual characteristics in some instances. However, such decisions shall be evaluated by the AI machine based on the limited'training data' that was used initiallyand fed in which may or may not be considered accurate or 'applicable to all' by courts in the EU. This is applicable to scenarios where the AI system may 'reject' a certain gender or race if limited fed in the training data shows that certain such genders or races have had specific criteria in the past. This could relate to assumption of defaulting in loans based on internet searches. However, such an assumption on the AI machine's part may lead to bias as the scheme will be based on a limited and specific training data which may not amount to a full proof mechanism. While it may be a possibility that a certain community or gender is more likely to be prone to certain issues in certain areas, subjecting the determination on such a possibility may involve bias.

The GDPR also appears to be divided on the 'right to explanation/right to access information' aspect provided by Articles 13-15. Through these articles, the data subject has the right to access information on the logic behind the decision made during profiling. In this context, an important question arises as to what requirements should be there in proving the logic behind algorithmic decision making, and how does one prove the legitimacy of such a decision? This would require vast amounts of data interpretability in explaining the correlation between the input data and the outcome of the data towards a decision. In this scenario, there is no availability of one case fits all status. The GDPR could simplify this approach by providing principles of proving the legitimacy of algorithmic decision making.

Moreover, it is a practice amongst data controllers to interfere with the outcome of algorithmic decision making to forward the business approaches of their organisations. The EU in this regards introduced a Governance Framework for Algorithmic Accountability and Transparency (**'Governance Framework'**), which provides an example for the above point,"even if a bank can explain which data and variables have been used to make a decision (e.g. banking records, income, postcode), the decisions turn on inferences drawn from these sources. Thus the actual risks posed by big data analytics and AI are the underpinning inferences that determine how we, as data subjects, are being viewed and evaluated by third parties". In this regard, the Governance Framework provides for a proposal within the data protection regime in the EU. The proposal states that there should be an existence of a 'right to reasonable interferences'. The right to reasonable interferences shall focus on not just how data is collected but also evaluate the method and justification prior to deployment of the data analytics so that data subjects are able to identify and understand how their data is being collected and what is the rationale/ logic behind the algorithmic decision making.

United States

The US is one of the biggest data analytics states globally. The public and investigative authorities in the US utilise large chunks of data for the purpose of profiling. Hence, it has become increasingly important to regulate data analytics and decision making by following the notions of transparency and accountability. For this purpose, US Senators in the 1st session of the 116th Congress had proposed an Algorithmic Accountability Act of 2019 (**'AA Act'**) in April 2019 which is still under discussion. The AA Act is the first-ever effort in the US to consolidate principles of algorithmic decision making and regulate AI systems across industries. The AA Act provides the Federal Trade Commission (**'FTC'**) the authorisation to enforce regulations on corporations and persons who are storing, using, processing, and sharing consumer's personal information. The FTC shall also direct such organisations/persons to conduct impact assessments and address issues regarding biases and discrimination. The AA Act shall be applicable to 'covered entities' which the meaning of Section 5(a)(2) of the Federal Trade Commission Act meaning that it will have a wider impact on not just technology companies but also banks, credit rating agencies, etc who process the data of more than 1,000,000 (one million) consumers or consumer devices as per Section 5(b) of the AA Act. Since the US does not have any federal or consolidated data protection regime, the AA Act shall also help in establishing the grounds for personal information which has been described under Section 10 of the AA Act as "any information, regardless of how the information is collected, inferred, or obtained that is reasonably linkable to a specific consumer or consumer device". An 'automated decision system' in the AA Act is provided by Section 2(1) to mean "a computational process, including one derived from machine learning, statistics, or other data processing or artificial intelligence techniques, that makes a decision or facilitates human decision making, that impacts consumers". Such a definition is wide enough to cover a variety of user ambits such as product/buying recommendations based on a consumer's search history. An essential problem with the AA Act that the authors feel is that it only places liability and targets big firms having more than 1,000,000 consumers. However, in the growing age of AI, small firms have equal ground to provide interference which may facilitate discrimination and breach of privacy. Moreover, under the AA Act, organisations are required to undergo 'impact assessments' on 'high risk automated decision systems', but nowhere in the text of the AA Act has it specified that organisations shall disclose the findings of such impact assessments. If the specific framing of the AA Act is taken into account, then it shall be noted that the AA Act mostly targets 'automated high-risk decision making' and not all 'high-risk decision making'. This scope of liability is very restrictive in nature as many organisations may be having human intervention in decision making while using algorithms as the underlying principle/source.

Moreover, through such a structuring, it also appears that the US regulators hold automated decision making in a more conflicted regard than human decision making, meaning that automated decision making is more susceptible to failure and less trustworthy, which may not be the case in every scenario. The impact assessment scope of the AA Act is much more restrictive when compared with Articles 35, 36, and 57 of the GDPR which formulate the basis and principles of a Data Protection Impact Assessment (**'DPIA'**).

Hence, at this point considering the very limiting nature of the AA Act, it may be said that several revisions and amendments are required before the actual enactment, especially considering the monetisation level of US organisations through data analytics.

The Indian Perspective

While most data-driven nations have adopted a certain kind of guidance towards ensuring non-discrimination and transparency in algorithmic decision making, India does not have any particular approach towards algorithmic transparency. One of the major reasons for this is the uncertainty in the data protection regime in India. While the PDP, 2019 is still under legislative debates, the government has also proposeda Governance Framework for Non-Personal Data (**'NPD Framework'**). Both the PDP, 2019 and the GPDR consider the impact of automated decision making. The GDPR under Article 22 has restricted the capability of organisations to make decisions solely on automated means, meaning that data subjects have been provided with the right not to be subjected to a decision solely based on automated processing. The PDP, 2019, in contrast to the GDPR, provides a 'right to portability' under Clause 19. PDP, 2019 does not specifically provide any restriction for decisions solely based on automated processes rather provides the data subject (data principal in the context of the PDP, 2019) the right to receive the personal data which has been processed in a structured and machine-readable manner. The NPD Framework has classified insights involving the application of algorithms under 'private non-personal data' and has provided that such algorithms may not be considered for data sharing, hence, enshrining an antitrust sense to algorithmic data. While there is still no comprehensive regulation for algorithmic accountability, the NPD framework does introduce 'data sandboxes' where algorithms can be deployed, with only output being shared. This means that organisation using automated decision making will be able to test their systems using training data and may be able to deploy such systems in the data sandbox to test their compliance in a controlled environment.

The Information Technology (Reasonable security practices and procedures and sensitive personal data or information) Rules, 2011 (**'SPDI Rules'**), the current framework for data protection regime in India, also does not provide for algorithmic accountability and non-discrimination practices. Hence, it has become necessary for India to adopt a separate bill for algorithmic accountability and transparency which provides for separate audits for algorithmic data and processing and such data to be made publicly available wherever required. However, it is contemplated that the initial step towards culminating any transparency provisions for automated decision through AI systems and algorithms should be the passing of the PDP, 2019 to ensure an adequate data protection regime to be in place.

BLOCKSUITS COMMENTS

Transparency lies at the key outset for ensuring algorithmic accountability, and for this purpose, academicians suggest two sets of explanations for deploying algorithms for decision making. As mentioned above, the right to explanation as provided by laws like the GDPR may undermine transparency if it does not clearly support the two sets of explanations, mainly, ex-ante and ex-post explanations.

Ex-ante explanation

This concept is based on the premise that data subjects shall have sufficient information to consent to data processing. For this purpose, organisations designing various models utilising algorithms shall consider the interpretability of the model by data subjects so that the data subjects may be aware of their rights while analysing what kind of data shall be processed by such models. In simple words, from the initial design of the AI or machine learning system itself, organisations/individuals shall observe the impact of such models. The consideration for the ex-ante explanation would include all kinds of inputs for training data before deployment. Ex-ante explanations are based on tests and forecasts of the initial design rather than actual results. Hence, in order for the data subject to make informed consent, organisations are also required to provide post-ante explanations.

Ex-post explanation

The ex-post explanation is based more on the specificity of models and provides for more features that are utilised in automated decisions. These are based on actual results and include more examples of how the data is being processed in a practical format such as language, visualisations, etc. The ex-post explanation system is more modified or customised to a specific data subject in order to provide an explanation of how on how algorithmic data processing has created an impact on data subjects. Since the data subjects will be able to comprehend the specific aspects and features used by algorithms in their context, they may be able to challenge the decision in an account of harmor breach of rights of data subjects. The concept of ex-post explanation is based on the fact that algorithms should be able to demonstrate the actual factors that are undergone in automated decision making. This is especially important for self-learning AI systems, where the AI system may adapt to individual choices of data subjects and go beyond the training data that was initially utilised, leaving data controllers open to risks and uncertainties.

It is very important for organisations utilising AI systems to divulge on both ex-ante and ex-post explanations. This is something that the GDPR has not considered in depth currently as organisations do not provide ex-post explanations unless specifically asked or provided for. Hence, in order to ensure transparency, both, ex-ante and ex-post explanations shall be considered by organisations.

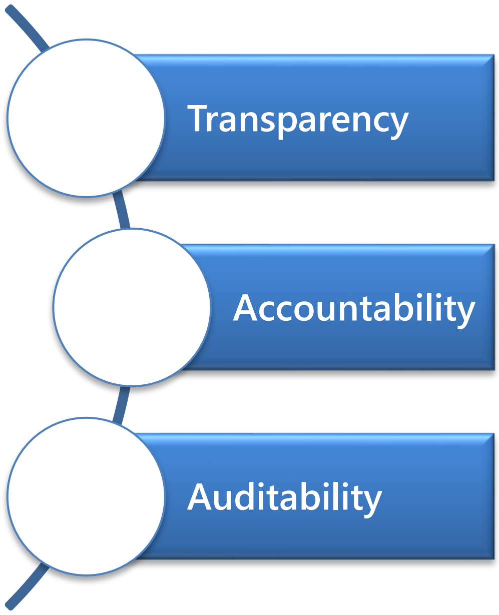
It is observed in many instances that even the usage of AI system is not disclosed to the users. While many regulations around the world govern the data privacy aspect, they fail to address the concerns around disclosure of AI’s existence in the first place. This is also commonly seen with AI chatbots. Some conversations may also be

AI chatbot induced which many users may not feel comfortable interacting with or may not be aware of the same. Therefore, we recommend that the first step will be to ensure that the companies using AI disclose its usage.

The authors also stress the need for government interference. The AI systems, along with their impact on safety, bias, privacy etc shall be disclosed to the government.

The AI systems, overtime analyse, learn, and enhance on certain aspects that the developers may not have intended initially. Therefore, a bias may be developed and enhanced at a larger scale by the AI depending upon the data that is fed in. The developing entities ought to consider the impact of feeding in every information. The data fed to the AI system shall be regularly audited to ensure that there is no element of bias. This can be done through regularly exercising a DPIA. The training data is usually accompanied by data such as compliance with existing rules and regulations. The developers should be careful and update their networks and softwares to ensure that the AI systems do not operate on outdated laws.

In the light of the above, in order to reduce bias, if not completely eliminate it, the authors suggest focus on the following three aspects:



Transparency

Accountability

Auditability

Transparency

One of the main reasons for such algorithm bias is lack of transparency. BlockSuits calls for the deployment of more transparent AI systems. A transparent AI does not essentially have to reveal the source code entirely, but merely ensure that the AI system is transparent enough to explain the automated decisions to the employees and the customers so that it aligns with the core values of any organisation. While AI can have astonishing results at times it can be alarming. Having a transparent AI can particularly be difficult considering the opposite nature an AI. However, approach towards transparency shall be technical in nature. The developer should carry out regular tests and make the report available to those who are impacted by such AI. The role of a developer does not and should not end at completion of such AI models. It should also extend to analysing such automated results statistically and ensure that no algorithmic bias exists.

Auditability

BlockSuits also advocates auditability. Organisations using AI should allow impacted users and employees to look inside the 'blackbox' of AI. The hypothesis of imposing auditability is that all decisions impacting public at large must be backed by a sound reasoning/explanations. If any organisation refuses to submit to such audit of their decisions based on AI, it should be a prima facie evidence of lack of justification and acting arbitrarily.

Accountability

Transparency and auditability will be futile if accountability standards are not framed and adhered to. The harm caused shall be assessed on the nexus causation test, and thus the test will be whether the harm was a foreseeable consequence of deploying blackbox AI. Further, where AI systems are opaque, the burden of justifying its decision making shall be on the proponent of such AI. It is also suggested that where claims are made on AI based decision, the scope of intent shall not be narrowed down or made specific. The test shall be whether the organisation deploying AI was negligent in ensuring a bias free training of its data sets, testing and timely reporting.

Data processing accountability through algorithms can only be made possible when there is a substantive data protection regime in place. This shall provide data subjects and consumers a proper redressal mechanism and grievance procedure in an event of breach of data rights. The initial step for authorities and regulators shall be to provide for a comprehensive and practical regime in data processing while resolving all conflicts and clarifications in the upcoming PDP, 2019, and the NPD Framework.