FEBRUARY 27, 2021
INDIAN SOCIETY OF ARTIFICIAL INTELLIGENCE AND LAW

Preliminary Review Report on Explainable Artificial Intelligence & its Opaque Characteristics

Indian Strategy for AI & Law Programme
isail.in/strategy

Indian Strategy for AI & Law, 2020

Approved by
Abhivardhan, Chairperson

INDIAN SOCIETY OF ARTIFICIAL INTELLIGENCE & LAW

About Us

Indian Society of Artificial Intelligence and Law wants to spread our message of hope and compassion. We believe that a single action can make a difference in the community, and that collective action can greatly impact the world. The Indian Society of Artificial Intelligence and Law is the only society in India, which conceives of the incorporation of Artificial Intelligence into the field of Law, wherein it is a multidisciplinary collaboration between Law and Information Technology (particularly Data Science and AI).

Our Philosophy

Artificial Intelligence has a primordial importance in human society. It renders a suitable relativity with mankind, and reflects a cultural backstop to human nature and its bliss. It dates back to the possibilities of what a machine could realize under Alan Turing, in his paper, 'The Imitation Game'. Nevertheless, the advent of Law, from its traditional positivist approach to a generic synthetic yet positive approach, has affected the directional utility of a globalized world. The use of machine learning as a component of data dimensionality has sincerely affected and led to generic consequences and is deemed fit enough not be ignored by international and domestic legal systems.The Society, which we term as ISAIL as well, focuses on its 3-fold approach; Academic Innovation, Learning Opportunities and Social Entrepreneurship.As AI is a deemed possibility to exist, it is an imperative ground that its legal modalities empower an anthropomorphic future for the natural ecosystem, to regard its decisions and relevance. This is not a question, which is about an Artificial General Intelligence, above every stigma of intelligence. It renders a possibility of major reality where international law and its domains are capable to understand, disseminate, secure and recognize the present and future of AI.

Preliminary Review Report on Explainable Artificial Intelligence & its Opaque Characteristics

Sameer Samal

Junior Research Analyst, Indian Society of Artificial Intelligence and Law

Abstract. Artificial Intelligence and its subsets have developed into advanced systems over the last few years. AI models such as Deep Neural Networks under Deep Learning have penetrated certain critical fields that require absolute certainty regarding their outcomes and predictions. Critical fields such as healthcare, finance and security require flawless predictions with explanatory justifications and sufficient reasoning. However, complex Machine Learning and Deep Learning models, that are widely used in the aforementioned fields, are opaque to a large extent which render them untrustworthy. Therefore, it is deemed necessary to add an interpretability or explanatory layer over the existing outcome. A prediction with reasonable justification regarding any factors that might have influenced its outcome can play an important role in building trust among the users.

This report aims to outline the opaque nature of conventional Artificial Intelligence systems by probing its ethical constraints. The significance of transparency and its facets are also observed. Subsequently, the concept of Explainable Artificial Intelligence (XAI) is explored and the role of XAI in Intellectual Property laws as well as its impact on the Medical field is analysed.

The Opaque System of AI Models

Researchers throughout various academic disciplines have witnessed the penetration of artificially intelligent technologies in their respective fields. However, with this, researchers have also observed the hesitation of users. The premise for the advent of any new technology can be traced to two systems; first, the technology being born out of certain needs, and second, the technology being initially developed and subsequently matched with any requirement. In both these methods, the users initially portray their reluctance in adopting the new technology, but their trust can be gained eventually. This is usually done by providing sufficient proof and evidence in favour of the technology's reliability and safety.

An appropriate example tracing the historical evidence in this context is the introduction of automatic elevators and the public's response towards it. Automatic elevators were introduced in the United States of America during the early 1900s. However,

Available on isail.in/civilized.
The Civilized AI Project – Discussion Paper

these machines, that have shaped the architectural landscape of the modern city, were surrounded with skepticism when first introduced. It was not until the 1950s, after a New York City strike in 1945 that cost the administration over a million dollar in lost taxes and a series of advertisements in favour of the elevators' safety, that trust was built among the public (HennGray 2015). With little deliberation one can establish similarities between the aforementioned events and the present-day mindset of users towards AI powered autonomous vehicles and robots.

Artificial Intelligence and its subsets have the power to transform the human civilisation by improving our autonomy and wellbeing. However, to effectively interact with this technology, it is essential for users to trust it. As stated earlier, trust can be built by sharing adequate information about the technology, its process and all other associated details. While the early models of artificial intelligence were easily understandable and interpretable by humans, the last few years have witnessed the significant growth of opaque or black-box models.

A black-box machine learning (ML) model contains thousands of parameters and hundreds of layers that render these models practically impossible to understand. These models are being increasingly used in making important predictions in critical context (Arrieta, Díaz-Rodríguez, Del Ser, Bennetot, Tabik, Barbado, Garcia, Gil-Lopez, Molina, Benjamins, ChatilaHerrera 2020). Therefore, all the affected stakeholders have begun demanding more transparency to understand these models better. Explanation about the output from the model is crucial in some fields, such as healthcare, where a life is at stake and the slightest error in understanding the prediction from the model can have serious repercussions. Therefore, it is crucial for these models to be understandable, so that the decisions based on the output can be justifiable.

The Concept of Explainable Artificial Intelligence

The nature and process of Explainable Artificial Intelligence (XAI) can be better explained with the help of visual illustrations.

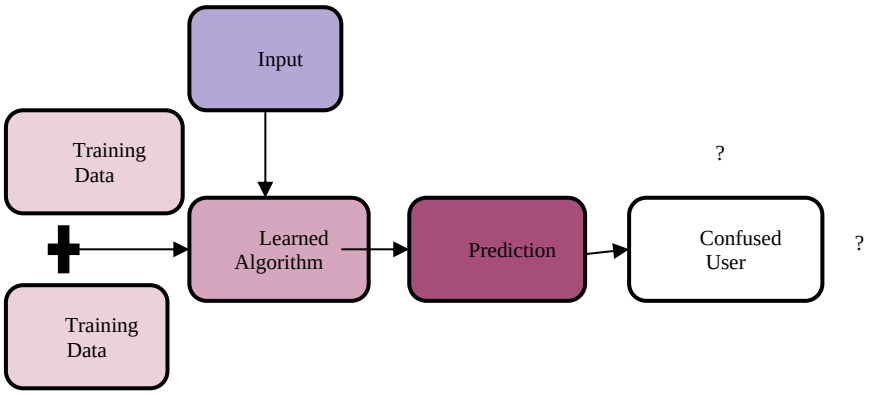

Fig. 1. ML Workflow (Image Structure Credit- Jamie Zornoza. Published at- towardsdatascience)

A Machine Learning model is initially fed some training data that goes through a specific learning process to result into a learned function. The learned function can then be fed input to derive predictions, which are labeled as output in the above figure. The biggest flaw of this model is its lack of transparency that leaves the end-user confused and skeptical. It is this opaque structure of a Machine Learning system that has awarded itself the title of 'black-box' model. Considering the fact that the predictions from this model, i.e., the outcome, does not come with any justification, a confused and an ill-informed user will have difficulty in trusting and relying on the prediction of the Machine Learning model. Adequate justification and reasoning behind a specific prediction will help the user trust and make informed decisions. This can be achieved by adding a layer of interpretation or explanation to the aforementioned process. Thus, transforming the traditional Machine Learning model into a Explainable Artificial Intelligence model.

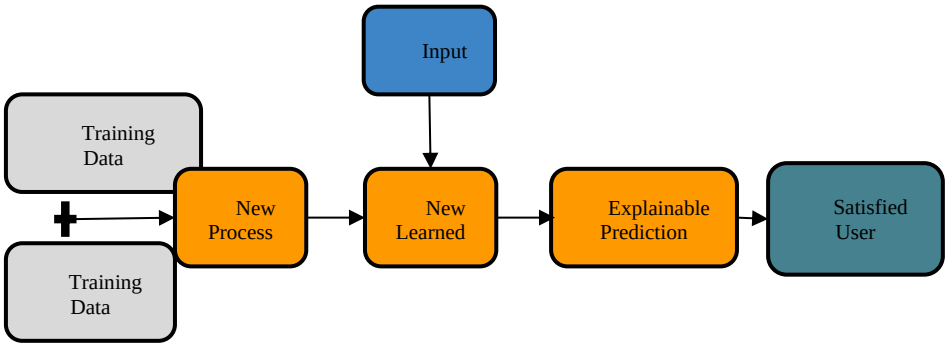

Fig. 2. XAI Incorporated ML Workflow (Image Structure Credit- Jamie Zornoza. Published at towardsdatascienc)

In this model, a new learning process has been used that gives us an 'Explainable Model' as a learned function. This Explainable Model can be fed input to derive predictions that provide reasoning and justification. The 'explanation interface', i.e., the output of the Explainable Artificial Intelligence incorporated Machine Learning workflow provides adequate justification with every prediction that enables the end-user to make an informed decision based on the prediction. This aspect can be very crucial in some fields that mandate precise and accurate decisions, such as healthcare, finance, and security.

Black-box, Meet Transparency

As explained in the aforementioned example of automatic elevators, building trust among the public at large requires a multi-faceted approach. There exists technical as well as societal challenges that need to be addressed (Tubella, Theodorou, DignumDignum 2019). The trust in Artificial Intelligence or any relatively new technology, for that matter, is directly linked with transparency. A transparent system or process enables all stakeholders to easily understand it. It is important to verify whether the Artificial Intelligent system complies with the established legal framework and the society's ethical and moral values. Transparency, in this context, means the visibility and understandability of the factors that influence the outcomes and predictions of the AI model. It should be effectively understood by the individuals who employ it, the individuals and organisations that regulate it, as well as by those who are either directly or indirectly affected by it. Therefore, transparency does not merely refer to the clarity of process, but also clarity of the factors that influence the various steps involved in the process.

The contemporary ideals of transparency are not restricted to the traditional notions of privacy considerations, but are spread across the far-reaching horizons of fairness and ethics. Therefore, it is essential for developers and regulators to venture beyond the isolated elements of privacy and ethics, and look into the wide palette of concerns that paint the canvas of AI with bias.

Aesthetic-Geographical Constraints in India

All individuals have a value system that influences their choices. A value system is a hierarchy of moral and ethical values that is unique to every individual. These values are based on one's virtues or vices, and experiences (*value system - Wiktionary* 2021). Therefore, an individual may prioritize a certain value more than others. Nations are held to be the collective construct of the individuals that build it. This 'construct' is not only limited to its corporeal existence but also extends to the ethical and moral principles that its citizens collectively hold to be of significance. Therefore, similar to individuals, even nations have a value system that influences their legal framework, judicial principles, legislative enactments, and executive decisions. A value system or a set of values in hierarchy based on their priority is inferred from a premise or a set of premises. These premises may differ from one nation to another based on their generalised practices, traditions, and norms. This premise will verify an Artificial Intelligence system's adherence to ethical and socio-legal values of that country.

The Role of Explainable Artificial Intelligence in Intellectual Property Policy and Administration

Artificial Intelligence engages with various fields of Intellectual Property at many levels and capacities. The World Intellectual Property Organisation (hereinafter referred to as 'WIPO' for the sake of brevity), has identified the following three aspects of Artificial Intelligence that engage with Intellectual Property (World Intellectual Property Organisation 2020):

1. Intellectual Property Policy
2. Strategic Capabilities: AI Capacity and Regulation
3. Intellectual Property Administration:
   a. classifying patents and goods and services for trademark application;
   b. searching of patent prior art;
   c. identifying elements of trademark;
   d. other trademark formality compliances; and
   e. client services and automated help desks.

WIPO has also identified a list of issues that are associated with Artificial Intelligence and its interactions with Intellectual Property. Conventionally, the debate around Artificial Intelligence and Intellectual Property revolves around the questions of 'authorship' and 'ownership'. However, it is imperative to discuss various other facets of Artificial Intelligence and their impact on Intellectual Property Policy and Administration. WIPO has categorised the issues into separate sections. Relevant sections and some of the pressing issues under them are as follows (World Intellectual Property Organisation 2020):

Patents: questions revolving around the impact of this technology began with the advent of Artificial Intelligence. In the last few years, Artificial Intelligence has transformed into an essential element of the notion that we use to perceive contemporary patent law. The biggest challenge that AI, as an assisting technology, had to face while penetrating the field of Intellectual Property, was the issue of inventorship and ownership. The latter is also generally referred to as proprietary rights. However, it is necessary for the debate to move beyond these conventional issues and shed light on the challenges arising from the introduction of AI in Intellectual Property Policy and Administration. Some of them are:

- incorporation of legal standards and ethical notions into AI systems that exclusively deal with the process of identifying and regulating patentable subject matter, as well as their guidelines;
- to revisit the requirements of inventive step and non-obviousness in patent registration; and
- the necessity of disclosure and its boundaries.

Copyright: similar to patents, majority of the issues addressed about copyright are regarding its authorship and proprietary rights. However, copyright laws must also consider the grey area of copyright infringement and its exceptions in terms of Artificial Intelligence. Machine Learning or Deep Learning models might infringe existing copyright laws by allowing these models to be trained on training data that might be subject to copyright protection. Therefore, it is essential to revisit the provisions of copyright infringement and its exceptions in light of the recent technological developments. Further, the controversial issue of 'Deepfakes' must also be addressed by initiating dialogue on the very nature of the technology in question- the ambit of copyright and its capabilities of bringing 'Deep Fakes' under its purview of regulation.

Designs: similar to inventions, designs are also made with the assistance of Artificial Intelligence and other computer programs. Therefore, the question of authorship and ownership of these designs arises. Additionally, when Artificial Intelligence models work alongside humans, the issue of infringement of copyright-protected data exists.

Trademarks: it is well established that Artificial Intelligence does not affect trademarks in the same way as patents and copyrights, but there still exists a slight interference. Artificial Intelligence can be used in administrative processes associated with trademark registration.

Artificial Intelligence and its subsets already impact Intellectual Property in a number of capacities. The World Intellectual Property Organisation as well as domestic legislations of various countries have identified the aforementioned issues that attention. Explainable Artificial Intelligence can resolve the majority of issues that relate to IP Administration and Regulation. The inherent characteristic of explainability and interpretability in opaque Artificial Intelligence models aim to improve transparency and clarity. With improved transparency, stakeholders can identify the factors that influence an AI model's predictions and base their decisions accordingly.

The Role of Explainable Artificial Intelligence in Healthcare

The role of Explainable Artificial Intelligence in the healthcare field is a highly debatable topic. Undoubtedly, Artificial Intelligence systems have proven to be highly efficient and at times more accurate than their human counterparts, but the lack of transparency in their output and predictions raise genuine concerns. The issue of transparency does not necessarily affect other fields with the same gravity as it affects the healthcare sector and it is absolutely crucial that the factors influencing an AI model's outcome are clearly communicated with the end-users. Considering the healthcare sector's critical nature, the decisions based on predictions from the AI model may directly or indirectly affect an individuals life.

The following aspects in this section can be considered by healthcare enactments:

Regulators may indicate a minimum level of transparency in the Artificial Intelligence powered Clinical Decision Support Systems before their suggestions may be relied on or decisions may be based upon them.

Stringent regulations may be prescribed regarding the consensual collection of clinical data of patients. Additionally, guidelines relating to its storage, processing and dissemination may be formulated.

Conclusions

Artificial Intelligence and its subsets, specifically Machine Learning and Deep Learning models, have firmly embedded their presence in sectors such as Intellectual Property and Healthcare. While their utility and positive impact is widely appreciated, it is necessary to consider their transparency concerns. A multidisciplinary approach towards Explainable Artificial Intelligence might resolve most of the issues faced in these two sectors.

References

1. ARRIETA, ALEJANDRO BARRETO, DÍAZ-RODRÍGUEZ, NATALIA, DEL SER, JAVIER, BENNETOT, ADRIEN, TABIK, SIHAM, BARBADO, ALBERTO, GARCIA, SALVADOR, GIL-LOPEZ, SERGIO, MOLINA, DANIEL, BENJAMINS, RICHARD, CHATILA, RAJA and HERRERA, FRANCISCO, 2020, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* [online]. 2020. Vol. 58, p. 82-115. [Accessed 18 January 2021]. DOI 10.1016/j.inffus.2019.12.012. Available from: http://www.sciencedirect.com/science/article/pii/S1566253519308103
2. HENN, STEVE and GRAY, LEE, 2015, *Remembering When Driverless Elevators Drew Skepticism*. [radio]. 2015.
3. TUBELLA, ANDREA, THEODOROU, ANDREAS, DIGNUM, FRANK and DIGNUM, VIRGINIA, 2019, Governance b Glass-Box: Implementing Transparent Moral Bounds for AI Behaviour. In : *International Joint Conference on Artificial Intelligence Organisations* [online]. 2019. p. 5787-5793. [Accessed 18 January 2021]. Available from: https://doi.org/10.24963/ijcai.2019/802
4. value system - Wiktionary, 2021. *En.wiktionary.org*[online]
5. WORLD INTELLECTUAL PROPERTY ORGANISATION, 2020, *Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence* [online]. World Intellectual Property Organisation. [Accessed 18 January 2021]. Available from: https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=499504