Study Paper

On

Trustworthy Artificial Intelligence (AI)

FN Division, TEC

K.L. Bhawan, Janpath, New Delhi

June 2020

Table of Contents

ABSTRACT:-

Artificial Intelligence (AI) is one of the most important innovations for the society, which has the potential for improving the quality of life. This paper discusses the need for a trustworthy AI.

AI has many benefits when it comes to societal, individual or cultural development. However, any mistake either in the development or in the working phase of the AI system can be disastrous, especially when human lives are involved. AI systems should be developed and used in a way that adheres to all the ethical principles, thus providing fairness, prevention of harm and explicability. While Standardization Development Organizations like ITU, ISO/IEC, IEEE have already begun work on certain AI related areas, however, it is imperative to lay down ethical foundation for using AI technology safely/ effectively, creating open process for codifying rights/ regulations around issues such as privacy, security, trustworthiness, robustness, transparency and above all, ethics.

According to report on Independent High-Level Expert Group on Artificial Intelligence set up by The European Commission, trustworthy AI[1] need to be: Lawful, ethical and robust. These three attributes are necessary but not sufficient for the achievement of Trustworthy AI. Creating trustworthy and ethical artificial intelligence requires an understanding not only of the technology itself, but also the societal and ethical conditions present, and how to appropriately account for and assess their impact on the way AI is designed, built, and tested, and the way it interface with human being. It must be sustainable and environment friendly. The traceability of AI system must be ensured, where individuals can have full control over their own data. The main goal of this paper is to understand what really makes an Artificial Intelligence system trustworthy.

Keywords: -Artificial Intelligence, ethical, lawful, robust, trustworthy, fundamental rights, democracy.

[1] https://ai.bsa.org/wp-content/uploads/2019/09/AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf

1. INTRODUCTION

The world, today, is witnessing a technological revolution in form of AI which has potential to change almost every sector may it be industry, government or non-government, education or health. The key capability that separate AI from many of the past technical or scientific breakthroughs in terms of the value it can create is machine- learning capability. Machines have the capability to process, track, and draw insight from millions of data points very quickly. Artificial Intelligence (AI) is actually a set of associated technologies and techniques such as machine learning, deep learning, natural language processing and neural network design that can be used to complement traditional approaches, human intelligence and analytics and/or other techniques. AI is the simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules imbibed in form of algorithms for using the information), reasoning (using rules to reach approximate or definite conclusions) and adaptive learning & self-correction

AI is the most important innovation for the society, which has immense potential of improving the quality of life, which can be utilized in nearly every aspect of life of the people like healthcare services, public sectors, education, electronics, banking etc. AI and increasingly complex algorithms currently influence our lives and our civilization more than ever. For example AI systems can reduce the unwanted needs of resources by accurately monitoring and managing the data of relevant energy needs of the society. This will result in the development of efficient infrastructure and intelligent logistics. Similarly Merging knowledge of human anatomy and AI can offer a new approach to prognosis and preventive health care and development of new drugs that can substantially reduce health care costs. AI is capable of solving complex situations and can be integrated with intelligent automation processes to develop cutting-edge solutions. Industry 4.0 technologies will apply AI for connected machines and processes. The greatest contribution of AI will be to face and resolve the global challenges, given in the UN's Sustainable Development Goals (SDG), a collection of 17 global goals designed to achieve sustainable future for all. To achieve these goals, innovation in the current AI system is of paramount importance for them to encompass a humane perspective and function in society to support and expand human welfare.

Artificial Intelligence is often perceived as a black box technology, with a lingering fear that whether it will be used to manipulate us. How can today's business move beyond these challenges? Answer to such fears lies in building trustworthiness– across business, society and government. Trustworthy AI means AI, which is lawful, ethical and technically robust and reliable. Creating trustworthy and ethical artificial intelligence requires an understanding not only of the technology itself, but also the societal and ethical conditions present, and how to appropriately account for and assess their impact on the way AI is designed, built, and tested, and the way we interact with it. For complete trust between the society and AI systems, both the internal architecture of the AIs and applications and Human Interface utilizing AI needs to be well defined as per principles of trustworthiness.

2. ETHICS: AN IDENTIFIED AI BENCHMARK

Ethics Guideline for Trustworthy AI by European Commission: It has been observed while offering great opportunities, AI systems also give rise to certain risks that must be handled appropriately and proportionately. Acknowledging the fact that AI is to support human requirements in a human-centric way, European commission formed a high-level committee; the High-Level Expert Group on Artificial Intelligence (AI HLEG), an independent group mandated with the drafting of two deliverables: (i) AI Ethics Guidelines and (ii) Policy and Investment Recommendations.

The National Strategy on AI for India: National Institution for Transforming India (NITI) Aayog (India's policy think tank) has come out with Strategy paper on #AIForAll as it is focused on leveraging AI for inclusive growth. This detailed paper on AI has delineated: technology framework, adoption of technology for benefit of larger good of people of India, how to develop the research ecosystem, promote adoption and address skilling challenges, funding mechanism etc. The strategy also flags important issues like ethics, bias and privacy issues relating to AI and envisions Government promoting research in technology to address these concerns. The national strategy specifically mentions about Fairness / tackling the biases AI / Transparency/ opening the "Black Box" and explainability of AI. It recognises that the data set used for training a machine or creating AI solution may have biases, which may have been reinforced over time. Therefore, there is need to build mechanism to create bias neutrality. This also identifies with discussions taking place in academic, research and policy fora, and definitely merits a combined dialogue and sustained research to come to an acceptable resolution of this problem. Strategy also refers to "Black Box Phenomenon", associated with AI. Opening the Black Box, assuming it is possible and useful at this stage by aiming at explainability of AI systems.

The Artificial Intelligence Task Force Report[2]: Ministry of Commerce and Industry, Government of India created a Task Force on Artificial Intelligence for economic transformation of India, which discussed and mentioned in its report about ethical and responsible AI. The report delineates how in many aspects Al disrupt current social norms and ways of thinking. Therefore, in general, legal and social constructs need to evolve to deal with autonomous systems. It is important for AI systems to show: explainable behaviour, demonstrable either explicitly or statistically and are engineered for safety and security and are rigorously audited to ensure non-contamination by human biases and prejudices.

IEEE Standards Association: AI standardization processes are part of a larger IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. IEEE has released a number of documents regarding the ethical development of AI through their Global Initiative on Ethics of Autonomous and Intelligent Systems, where they consulted across some areas of industry, academia, and government. The IEEE sets out five core principles to consider in the design and implementation of AI and ethics. These include adherence to existing human rights frameworks,

[2] https://dipp.gov.in/whats-new/report-task-force-artificial-intelligence

improving human wellbeing, ostensibly to ensure accountable and responsible design, transparent technology and the ability to track misuse. IEEE SA recently launched the development of an Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS). ECPAIS seeks to develop three separate processes for certifications related to transparency, accountability, and algorithmic bias.

ISO/IEC: Recognising the importance of international standards harmonisation in addressing, managing and regulating new areas of technology, the ISO and the IEC Joint Technical Committee 1 (JTC 1) created Subcommittee 42 – Artificial Intelligence (SC42), in 2017 that has nine standards under development, focused variously on terminology, reference architecture and, more recently, trustworthiness. The Trustworthiness working group is currently drafting three technical reports on robustness of neural networks, bias in AI systems, and an overview of trustworthiness in AI. SC 42 committee is also driving work on the governance of AI within organizational settings, to ensure the responsible use of AI.

ITU Focus Group on AI for Health (FG-AI4H): The ITU/WHO Focus Group on artificial intelligence for health (FG-AI4H) works in partnership with the World Health Organization (WHO) to establish a standardized assessment framework for the evaluation of AI-based methods for health, diagnosis, triage or treatment decisions In the latest meeting of ITU Focus Group on AI for Health held in November 2019 in New Delhi India, a contribution was submitted by TEC on Ethical Issues on AI for Health. After discussion, it emerged that a working group on ethics should be created on the topic.

3. FRAMEWORK AND FOUNDATION OF A TRUSTWORTHY AI

For successful development of framework for a reliable AI system, three criteria should be met for its development and functioning. According to report on Independent High-Level Expert Group on Artificial Intelligence set up by The European Commission, trustworthy AI[3] must encompass the following three attributes:

a) **Lawful**: The AI system should be compliant with various rules, laws and regulations.
b) **Ethical**: AI system should contain morals and ethics, and adhere to moral values and principles.
c) **Robust**: AI System should be sturdy in both social and technical sense.

Ethical issues of AI are field of applied moral values; it focuses on the various socio-technical discrepancies or issues generated due to the construction function and uses of AI. Ethical field regarding AI has significant value when it deals with problems like safety of individuals, privacy and even unemployment, if created due to AI. The main objective for the developers will be to integrate AI systems with the common life along with disrupting and existing social boundaries for maintaining sustainable law and order in the society.

[3] https://ai.bsa.org/wp-content/uploads/2019/09/AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf

3.1 Lawful AI System:

A lawful AI system should support fundamental rights of the citizen. AI system should comply to all laws of the land protecting human rights and dignity of each and every individual human being. Freedom of individuals means the full autonomous control over their rights that can be- right to education; right to privacy, rights to express etc. An AI system should regard the freedom of individuals by not using any form of coercion, manipulation and deception with them.

3.1.1 Respect for democracy, justice and the rule of law

AI system should not change any current democratic processes, freedom of vote and laws of any country. AI system should also be aware enough for not taking any actions, which can be detrimental to the principles that form the laws.

3.1.2 Equality and inclusivity

AI system should not function in any manner that supports racial issues, religion issues, gender discrimination and any other such unfair criteria. It should be inclusive in nature and be respectful to all, irrespective of their gender, religion and race.

3.1.3 Citizen rights

AI system should be increasing the potential of the ability of various governments to enhance the innovation and efficacy of the public sector as well as the private sector for improvement of quality of life for their citizens.

3.2 Ethical AI Systems:

The way ethics play an important role in our daily lives, similarly, it is necessary to have ethics for AI systems in order to enable the systems to make quick, transparent and responsible decisions. Ethical principles for AI can serve a variety of functions in support of the users. Some of the ethical principles necessary for AI to achieve better outcomes reduce the risk of negative impact and practice the highest standards of ethical business and good governance.

3.2.1 Human centric Approach

The AI systems must not in any case dominate, force, deceive or manipulate human beings, rather, they must be designed in such a way that they support, increase and accompany humans' social and cultural skills as well as their cognitive thinking. The AI systems must follow the design principles that supports the human centric approach and there should always be an upper hand for humans regarding their functionality. The AI systems may also make changes in the working atmosphere aiming for the establishment of meaningful work keeping in mind the proposed limits set by humans.

An AI system must not intend or cause harm to a human being. This involves mental as well physical protection of human beings, while keeping their dignity. The safety and security of the

environment in which the AI systems work must also be kept in mind, so that it is ensured that they are not used maliciously. AI systems should benefit individuals, society and the environment.

3.2.2 Fairness and bias free

AI Programs are made up of algorithms that follow rules. They need to be taught those rules and this occurs by feeding the algorithm with data, which the algorithm then uses to infer hidden patterns and irregularities. If the training data is inaccurately collected, an error or unjust rule can become part of the algorithm- which can lead to biased outcomes. The motive behind using an AI system should be fair and must not include any bias decisions. The ulterior motive behind this principle is to mitigate the results obtained from a discriminate use of data in artificial intelligence.

3.2.3 Explicability

Explicability comes from the word explicable; meaning "capability of being explained". In order to build and maintain trust among users in AI systems, explicability is an important factor. The process through which AI works needs to be transparent and the purpose of the AI system as well as the decisions made by it must be well understood by those affected, directly or indirectly. The extent to which an AI system is explicable is highly based on the context related to which the system is working.

3.2.4 Data privacy, protection and security

AI system should respect and uphold privacy rights and data protection, and ensure the security of data. This includes ensuring proper data governance and management for all data used and generated by the AI systems. Data used for training the system need to be anonymised for the sake of data privacy at the same time it need to be protected from misuse and pilferages and to be secured against any kind of security violations.

3.3 Robust AI System

3.3.1 Reliability

Every AI system is deployed by a human organization. In high-risk applications, the combined system of human plus AI must function as a high-reliability association in order to avoid catastrophic errors. Such AI systems should perform in a safe, secure and reliable manner, and safeguards should be foreseen to prevent any unintended adverse impacts.

3.3.2 Safety by Design

The AI system should work in a safe and secure manner. It should be designed by foreseeing all adverse impacts it can create and such unintended impacts should be averted by robust designs. This is needed both from, technical perspective (ensuring the system's technical robustness as appropriate in a given context, such as the application domain or life cycle phase), and social perspective (in due consideration of the context and environment in which the system operates).

4. REQUIREMENTS OF TRUSTWORTHY AI SYSTEM

Responsibility of creating trustworthy AI lies with all stakeholders being part of an AI system. Therefore, different set of stakeholders: developers, deployers and end-users have to play their part for designing a trustworthy AI system. Developers who design and develop the system need to be aware and vigilant while doing research and designing the system. Deployers who deploy AI to create products, services, facilities etc. need to be aware of their responsibility of using AI in a manner which is according to law, ethics of the society. End users should insist upon making AI compliant to three basic requirements of trustworthy AI.

AI systems should meet the following requirements in order to be deemed trustworthy.

4.1 Support to Human Rights

Fundamental Rights: AI systems have the capacity to equally support or hamper fundamental rights. For instance, they can balloon in the field of education, thus supporting someone's right to education. However, the same AI system can negatively affect someone's fundamental rights. In such situations, proper fundamental rights violation assessment must be performed. This must be done before the development of the AI system.

Human Agency: There should be a flexible system between the user and the AI system. The user should have the necessary knowledge and tools in order to comprehend and make changes in the AI system according to their needs and goals. However, this must be limited to a certain degree.

Human Oversight: Human oversight can be beneficial. Proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches.

4.2 Technical robustness and safety

Resilience to attack: Just like any software, AI systems also have the vulnerability of being attacked by adversaries (e.g. hacking). In case an adversary attacks an AI system, there are chances that the AI system may respond differently and produce an unwanted output. It may even shut down. Hence, in order to mitigate such unforeseen results, the AI's security must be taken into account while designing and developing the AI system.

Fall Back Plan: Every AI system must have a fall-back plan in case a problem occurs. It must be ensured that the AI acts according to the proposed regulations towards its goal without harming any human being or the environment. The fall-back may include moving from a statistical approach to a rule based approach. The system may even take permission from the human operator before performing further tasks.

Accuracy: An AI system must be accurate enough to make correct judgements. This is very crucial at times and situations where human lives are at risk. Inaccurate predictions may lead to damage to property and even loss of human lives.

Reliability and Reproducibility: An AI system must be reliable as it work with a variety of input in order to obtain different outputs. In addition, an AI task must produce the same output when repeatedly performed under the same conditions so as to ensure consistency of its results with similar conditions.

4.3 Data Safety, Security and Protection

Delegation of decision making to algorithms, may require built in mechanism for reduced bias, discrimination and improved privacy protection. However, even if a technological method helps user delegate that responsibility of decision making to an AI system with improved outcomes, he cannot get away from his core responsibility of assuring data privacy and protection. Therefore, it is important that higher standards to be set up for:

Privacy, Data protection: The information provided by the user and the personal information of the user must be kept safe by the AI system at all times. The AI system must not misuse it for any reason whatsoever.

Quality and Integrity of Data: Whenever any data is gathered by the AI system, there are chances that the data may be full of errors and mistakes. Feeding such type of data may change the system behaviour. The system must also reject any malicious data.

Access to Data: Not everyone must have access to the data collected by the AI system. Certain rules and regularities must be maintained regarding who will have the access to such data and under what circumstances this data can be extracted.

4.4 Transparency and Accountability

Traceability: All of the information that the AI system gathers stores or communicates between other systems or users must be open to tracking for security purposes. This should be done under proper guidelines documented under the best possible standard. Traceability helps remove any errors in the decisions made by the AI system, and prevents any future mistakes.

Explainability and Transparency: There must always be an explanation of why an AI system made a particular decision. There are some situations in which analysing a particular decision made by the AI system is necessary.

Communication and Accountability: Every user has the right to know that they are interacting or communicating with an AI system. A user can knowingly choose to have a human based interaction with its AI system, but that too under certain conditions. In addition, this must not violate any fundamental rights under any condition.

4.5 Non-discrimination and Inclusiveness

Bias Aversion: The information that goes through an AI system (whether that data is used to interact with the user or is used while developing the AI system) may contain some historical

events that are related with biases in the past. This piece of information may continue to create various cultural, racial or sexual bias and prejudice in the future as well. In order to alleviate the problem, people from a diverse background may be hired while developing the AI system. The teams developing, designing, testing, deploying, maintaining, and procurement of AI systems takes into consideration the diversity of users and society in general. This ensures objectivity and contributions of various perspectives and needs. Generally, team diversity is in terms of not only gender, age, social group, culture but also includes skill sets, professional skills and background.

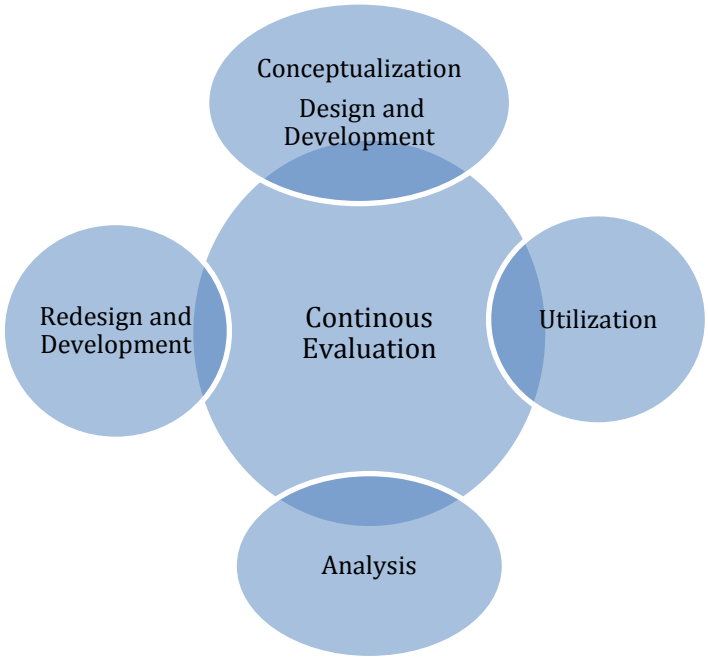
Sustainable and environment friendly AI: An AI system's design, development and usage processes must be performed in an environment friendly way, e.g. energy consumption during the AI's usage process must be tracked and kept under certain limits. Ideally the goal should be to enhance environmental efficiency.

Social impact: AI systems have the ability to alter our social lives, be it in areas of entertainment, work life or social life. They cannot only make our social lives better, but can deteriorate it too. When it comes to AI's negative impact on our social life, they include both physical as well as mental effects. In order to mitigate this, the AI systems must be kept under observation and monitored regularly.

Society and Democracy: Apart from using AI systems to improve an individual's life, they must be used to affect the society, for e.g. analysing the flaws of a democracy and suggesting decisions to improve its structure.

5. REALIZATION OF A TRUSTWORTHY AI

For realizing a trustworthy AI, it is necessary to create an end to end life cycle design of AI system, which means measuring performance of the system from initial conceptualization to final end product and making a close loop feedback system to avoid any diversions.



Realization of a trustworthy AI is possible with following methods:

i) Trustworthiness built-in in design;
ii) Policy, normative guidelines, standardisations, testing and certification.

Technical methods should ensure that the trustworthy AI to be employed in the development, designing and is utilized in all phases of an AI system. Besides there can be different not so technical methods, which plays an important role in maintaining and securing the AI. AI for All should aim at enhancing and empowering human capabilities to address the challenges of access, affordability, shortage and inconsistency of skilled expertise.

Design of a trustworthy AI architecture involves:

- *Recognition and Articulation:*
  It involves recognition of factors such as Laws, Policies, normative behaviour of AI system, identification of undesirable system behaviour necessary to follow requirements of trustworthiness. These need to be articulated well so that it can be translated in to system architecture.
- *Planning:* It allows involvement of those plans that adhere to all the requirements.
- *Execution:* This involves actual technical design of the system architecture.

While actually designing the system it is very important to clearly identify system behaviour i.e. deterministic or uncertain. Uncertainty arise when system is perpetually self-learning and very dynamic. In such cases it is necessary that desired and undesirable behaviours of AI system clearly defined in the architecture and it is to be ensured through close loop of identify-measure and change of behaviour pattern by AI system itself.

5.1 Law compliance by design

Law by design provides accurate and explicit links between the abstract principles, which the system should obey, and implementation specific decisions. The norms should be obeyed for implementation of trustworthy AI system. It provides safe shutdown in case of failure and resume the operation after a forced shutdown. Therefore, designers of AI system are responsible for identifying the status of law compliance of their AI system and also impact of system on human interface involved and society at large.

5.2 Explainable AI (XAI)

Explainable AI (XAI) is an idea where behaviour of system must be analysed before interpreting its results for achieving a trustworthy AI system. Therefore, XAI tries to address unpredictability involved with dynamic learning systems by introducing pre-analysis and justification for systems

behaviour pattern. The Explainability quotient should be such that the AI algorithm and associated machine learning is able to[4]:

- Produce more explainable models, while maintaining a high level of learning performance; and
- Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

5.3 Identification of Quality and Service KPIs

Key Performance Indicators needed to adjudge performance of an AI system for its trustworthiness need to be predefined besides parameters related to functionality, performance, reliability, safety and security.

5.4 Testing and validation

Testing and validation of the system must be provided as it ensures the system behaves as desired throughout its life cycle. It must include all components of an AI system, including data, pre-trained models, environments and the behaviour of the system as a whole. The output must be consistent with the results of the preceding processes, while comparing them with the previously defined policies to ensure that they are not violated.

5.5 Regulation, Standardisation and Certification

There has been a lot of debate on how the Artificial Intelligence (AI) should be regulated. Governments around the world have stressed that AI should be governed by a strong set of regulations. Governments and regulators should not only ensure that creativity involved in building AI systems to be nurtured but also to ensure that technology is harnessed for good and available to everyone. To ensure this, it is essential that certain safeguards be created. Therefore, a legislative framework for a trustworthy AI should be in Place to acts as an enabler for adopting trustworthiness as well as a safeguard against deviations from normative guidelines.

Standardisation of designs, business processing and manufacturing services act as a quality management system for AI by providing the users, organisations, research institutions consumers and governments with the ability to identify and encourage ethical code of conduct for their purchasing decisions.

The certifications apply standardised designs; manufacturing services developed for different application domains and align them appropriately in different contexts of industrial and societal standards. Certification cannot replace the responsibility. Therefore, disclaimers as well as review, accountable frameworks and readdressed mechanisms, should complement it.

5.6 Normative Guidelines

An organisation should document its purpose and intentions when working with AI systems. In addition, it should follow standards of some expected values such as transparency, fundamental

[4] Discussion Paper on National AI strategy by NITI Aayog , June 2018

rights, and protection from harm. These guidelines to be in accordance with extant policies and regulation. This can be in form of a charter of responsibility of all stakeholders for assuring design, development, deployment and utilization of a trustworthy AI system.

Trustworthy AI encourages the collaborative and instructed participation by all stakeholders. Communication, education and training are important factors for ensuring the potential impact of AI systems and makes people aware as they have a vital part in shaping the society having AI Systems.

5.7 Organizations Responsibility

Some governance frameworks should be established internally and externally by organization to account for the ethical decisions related to deployment, development and usage of AI system. Communication channels should also discuss dilemmas and report emerging issues incorporating ethical concerns.

5.8 Stakeholder Commitment

AI systems may offer huge benefits so it should be guaranteed that they are available to all. This requires discussions and dialogues between various social partners and stakeholders. The process should also include the public for their views.

6. ASSESSING TRUSTWORTHY AI

Development of assessment criteria is a very important step for creation of trustworthy AI systems. These need to be framed in close coordination of multiple interested parties from private and public sectors, stakeholders and government. Open dialogue should be initiated to generate awareness and sufficient efforts to be made for advocacy of such efforts. Assessment criteria should include indicators related to all three main aspects: Lawful, Ethical and Robust AI systems. Such criteria once created need to be tested through small projects. Various small-scale projects are to be executed first for getting the relevant feedback on the limitations of the current AI system. Hard rules and limitations of AI's functions are to be outlined by referencing several factors like safety, advancement of AI and social acceptance of the people.

7. CONCLUSION

AI systems are expected to have very strong influence over business practices, governance structure and society at large. They are presently having numerous positive impacts in various sectors like health, educational, defence etc. However, these systems also accompany equally large risks and negative impacts on the society if are not properly used. For example, AI system can be used with various measuring instruments and life support devices to provide a high level of accuracy and control for aiding the doctors. In case of indirect influence, using measurement recorded by the AI, doctors will be able to determine any potential diseases or problems in the

patients and appropriate preventive measures can be taken. Trust on the measurement of the AI devices and their lack of bias by the doctors can improve the present conditions of treatment exponentially. However, if AI system has inbuilt bias it has equal potential for huge misuse.

Therefore, development of the framework for the system through which they can be regarded as trustworthy is of paramount importance before prolific acclimatisation of AI systems in the daily lives of people and organisations. A human centric ethical approach is required for creating trustworthy AI systems.

ABBREVIATIONS

AI:	Artificial Intelligence
CCA:	Climate Change Adaption
FG-AI4H:	Focus Group on AI for Health
ICT:	Information and communication technology industry
ITU:	International Telecommunication Union
KPI:	Key Performance Indicators
NITI:	National Institution for Transforming India
SDG:	Sustainable Development Goals
SDN:	Software defined networking
TEC:	Telecommunication Engineering Center
XAI:	Explainable AI

References

[1] Asilomar conference on Beneficial AI, 2017

[2] Discussion Paper on National AI strategy by NITI Aayog, June 2018

[3] Floridi, Luciano. "Establishing the rules for building trustworthy AI." *Nature Machine Intelligence* 1, no. 6 (2019): 261.

[4] Livingston, S., & Risse, M. (2019). The Future Impact of Artificial Intelligence on Humans and Human Rights. *Ethics & International Affairs, 33*(2), 141-158.

[5] Marnau, N. (2019). Comments on the "Draft Ethics Guidelines for Trustworthy AI" by the High-Level Expert Group on Artificial Intelligence.

[6] Report of Task force on Artificial Intelligence formed by Ministry of Commerce and Industry Government of India.

[7]Siau, K., Wang, W. (2018), Building Trust in Artificial Intelligence, Machine Learning, and Robotics, *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), S. 47–53.

[8] Vakkuri, Ville, and Pekka Abrahamsson. 2018. "The Key Concepts of Ethics of Artificial Intelligence." Proceedings of the 2018 IEEE International Conference on Engineering, Technology and Innovation, 1–6.

[9] Yu, Han, Zhiqi Shen, Chunyan Miao, Cyril Leung, Voctor R. Lesser, and Qiang Yang. 2018. "Building Ethics into Artificial Intelligence." arXiv, 1–8.