International Journal on Emerging Research Areas (IJERA) ISSN:2230-9993
DOI: 10.5281/zenodo.10686021

The Integration of Trustworthy AI Values: A Comprehensive Model for Governance, Risk, and Compliance in Audit Architecture Framework context

Manju Susan Thomas
Research Scholar
Marian College Kuttikanam Research center
Lincoln University College (LUC) Malaysia.
manjuthomasmail@gmail.com

Juby Mathew
Faculty of Computer Science and
Multimedia,Lincoln University College,Malaysia
jubymp@gmail.com

Abstract

The Trustworthy AI-compliant Governance, Risk, Compliance Architecture (GRC Architecture) as sources monitor the GRC aspects of an organization in real time and report it to the stakeholders as to where the firm lacks performance/ risk/ compliance measures thereby enabling them to take corrective measures in time. The system is a supervised ML system which has sufficient human intervention and oversight. It is built on features like robustness and safety; the AI system has an in-built model control faculty that would take care of wrong actions by not letting them manifest, by classifying AI activity into multiple classes in which super-critical functions wherein a potential misstep can cause business disruptions; human approval is necessary there. Complying with all the laws and regulations governing the respective domains, the GRC Architecture is ISO-certified and adheres to all the privacy and data governance laws. It is transparent: all its actions are well accounted for. To ensure diversity, non-discrimination, and fairness, a special model is employed. It considers societal and environmental wellbeing in that paperless electronic digital system ensures a reduced carbon footprint contributing to infinite intangible benefits hidden prima facie, as the GRC aspects of many a firm are taken care of. In this research paper, a Governance Risk and Control system architecture is shown to adhere to these parameters of Trustworthy AI

outlined in this paper has the objective of analyzing the performance of a firm vis-à-vis all the three facets. In other words, the system, analyzing the data from multiple

Keywords: Artificial Intelligence, trust in autonomous systems, Business process systems, GRC architecture, Trustworthy AI

1.Introduction

Trust is one factor that keeps the world intact and sound in function. Without human beings trusting other human beings, there are no communities formed. And without communities trusting other communities, nation states are but a work of fiction. While these aspects are entirely humanity-oriented, how do we go about it when several functions of human beings including cognitive, are given away to Artificially Intelligent machines or algorithms to be done? The question thus becomes whether it is advisable to trust a machine recommending you a product or on a different note, sorting your university admission application and deciding on your stream of study which can have life-long consequences for the applicant. What if the AI system is biased? The question of who decides what has now been changed to what decides who. And it needs to be trustworthy. Trustworthy AI has a few parameters. Those AI models or architectures that are trustworthy should adhere to a range of parameters. The EU framework in this regard is substantially robust.

There is little wonder that with the AI systems increasingly taking decisions that affect humanity in multiple fronts like education, healthcare, mobility, security and defense, technology should be trustworthy. The Expert Committee set up by the EU for framing the ethical aspects of AI has made Trustworthy AI its foundational ambition. Without trust in place, humanity will not have confidence in the development of AI technology and its applications and this mandates the existence of a clear and comprehensive framework for achieving trustworthiness.

To facilitate trustworthiness, a range of principles need to be adhered to. The systems, first of all need to be human-centric. This means, the AI technology must be robust enough in its commitment to use itself in the service of the common good. Its primary endeavor should be to improve human welfare and enhance human freedom.

Trustworthy AI has got three main constituent elements to it:

1. Lawful nature: It should be compliant with all the regulations and laws of the time.
2. Ethicality: Must adhere to codes of ethics, values, and principles
3. Robustness: Socially and technically robust and should not be causing unintentional harm

AI technology is deemed to be Trustworthy if it follows the following parameters.

1. It has human agency and oversight.
2. Has got technical robustness and safety.
3. Has got proper privacy and data governance compliance in place.
4. Is transparent.
5. Is diverse, free of non-discrimination and fair.
6. Takes into account societal and environmental wellbeing.
7. Is Accountable.

The following framework captures this concept in its totality.

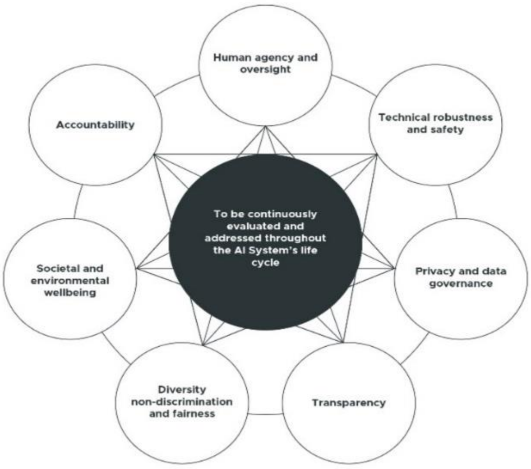

Human agency and oversight
Accountability
Technical robustness and safety
To be continuously evaluated and addressed throughout the AI System's life cycle
Societal and environmental wellbeing
Privacy and data governance
Diversity non-discrimination and fairness
Transparency

Figure 2: Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle

The GRC Architecture as represented in this paper adheres to all these parameters and embodies it.

2.Literature Review

Prince Chacko Johnson mentions in detail how AI is presently being utilised to enhance actions of humans (less than 60%) within R&D rather than to pre-set different matters in motion (just above 10%). [1]

With AI-integrated technologies getting more and more mainstream, there has arisen a need to identify and examine the degree of trust that users have when it comes to such technologies. The paper by Hyesun Choung *et al* argues that with the progress happening in the development of AI, an equal measure of understanding of trust in AI technology is also needed. The paper studies this aspect by elaborating the role of trust in one's intent to utilize AI related technologies.[2]

To conceive, create and develop AI-based architecture that the users trust and the society trusts, there is a specific requirement to understand how ML technologies have an impact on trust. The paper written by Ehsan Toreini et al has a concerted approach to relate considerations on trust from the segment of social sciences to technologies that are worthy of trust in the sectors of products and services. [3]

And in this paper by Frank Marcinkowski *et al*, they are of the view that Algorithmic decision-making (ADM) is assuming a greater degree of prominence in all walks of social life. In tertiary education, ML systems, since they have the capacity to process huge chunks of data, are utilised to arrive at effective decisions. But there are also queries about fairness being raised. [4]

In another paper by Drew Roselli et al, the bias in AI algorithms is studied and solutions suggested. The authors are aware that the propensity of bias raises the risk of deployment of AI algorithms, and they feel that removal of unwanted bias may not be feasible because of encoding of biases of a historical nature by AI systems. [5]

In a paper by Ninareh Mehrabi et al, the team of authors studied multiple real-world applications of AI that have biases in them in a multitude of ways, and a list was drawn up that portrayed a range of fountainheads of biases that can influence the applications. Taxonomy for definitions pertaining to fairness was also defined. [6]

The study penned by Eirini Ntoutsi et al provides a multidisciplinary bird's eye view of the regions of bias in AI systems. It focuses on the multitude of challenges and solutions of technical nature and provides for a brand-new direction for research well-heeled in a frame of law. Big Data AI is the topic here. [7]

Arun Rai in one of his papers, differentiate between AI models thar are interpretable and black-box models that are into deep learning and mentions about how to turn black-box models into the ones that are glass-box models. [8]

Pei Wang in his paper mentions 5 typical ways AI definitions stand and how the definitions are mutually compared. The research direction for each AI definition, even as each may appear legitimate, may take a different pathway than intended, struggling to provide AI a proper identity. A solution follows afterwards.

M. S. Thomas and J. Mathew describe in detail the architecture of a Supervised ML Model for automating an IA workflow [9]

Jean-Marie John-Mathews in the paper has come up with two different scenarios for the progress of ethical AI: a greater degree of external regulation and liberalization of descriptions for AI. [10]

Emily Shearer et al prescribes a system that can look deep into algorithmic quality to enhance trust in the tools, those related to healthcare sector. Process regulation of AI development rather than product regulation is the key the authors suggest. [11]

Embedding of ethics, assessment, ways of incorporation and construction should be applied within the lifecycle of the technology along with role formulation and role execution. A set of actors should go by predefined roles including but not limited to giving ethical and factual information, their knowledge prowess, analysis and the like. The paper by Lan Xue and Zhenjing Pang draw from examples of Autonomous vehicles. [12]

A contrast study is often a work of insights and a plethora of issues dissected. Bernd Carsten Stahl et al comes out with the suggestion of a Europe-wide agency and discusses the essential form, structure and nature of it so that it is in a position to deal well with ethical and rights issues for the development, deployment and utilisation of AI. The study then contrasts its conclusions with the suggested European Artificial Intelligence Board as mentioned in the draft AI Act of Europe. [13]

There are also studies, for instance the one written by Bo Li et al that offers solid guidance for specialists and stakeholders in society to enhance the trustworthiness of AI. [14]

The European Commission recently came up with suggestions from its Expert Group or AI. Even as it is shown to have deficits by an author like Michael VEALE, the centrality of the document vis-à-vis TAI discussions cannot be ruled out [15].

Artificial Intelligence and its Elements

In the year 1956, Dartmouth College in a place called New Hampshire in the US witnessed an agglomeration of researchers in the science of computing to discuss ideas on AI or Artificial Intelligence. At that time, the field of AI was at its emerging stage. The researchers and scientists were eager and bold enough to imagine a realm where machines use language just like human beings do, form abstract constructs and concepts and solve the problems of the type reserved for humans, and enhance themselves. The meeting was undoubtedly historic. It set the stage for decades after decades of government and industry research in the field of AI. A few technologies that resulted from these research measures would be:

- Mapping technology
- Smartphones with voice-assistance feature
- Hand-writing recognition aiding the conventional (snail) mail delivery.
- Trading algorithms in finance
- Smart logistics
- Spam filtering systems
- Language translation systems

AI denotes a system which is machine-based that can make the following functions work and that too for a set of human-defined objectives influencing real or virtual environs:

- Make predictions.
- Give recommendations.
- Take decisions.

The term AI encompasses a range of factors. The AI enabled machines are supposed to have the power of perception, be it real or virtual environments. The perception is made an abstract of to arrive at models by virtue of analysis through automation. The models are utilised to formulate options or responses. Apart from this, Artificial Intelligence is expected to perform not rudimentary computational tasks but should be able to perform several ‘advanced functions’ like those of its ability to see, understand to a level, spoken and written language and translate the same. It should also be able to carry out problem solving and pattern recognition. A level of human-mimicking or imitation is expected out of machines when it comes to cognitive functions in AI. The machines with AI capabilities are capable enough to perform a range of advanced functions including visualizing, understanding language, translating spoken and written language, analyzing data and making recommendations amongst many other things. Largely, they should be capable enough to solve cognitive issues generally associated with human intelligence in the lines of learning, solving of problems and recognition of patterns.

Types of AI

Having stated the above facets of Artificial Intelligence, it should be noted that all forms of AI do not come painted on with the same brush. There are different types of AI systems based on their 'cognitive maturity'.

Reactive AI emulates the human beings' capability to respond to stimuli. It can field a limited response to a combo of inputs. But it has also got a set of limitations in that it has an absence of memory-based functionality. Experience of the past cannot be utilised in the present and there is nil scope of improvement on its own. Limited Memory AI meanwhile responds to stimuli and learning from responses. Used in chatbots, virtual assistants and self-driving cars. It stops short of the Theory of Mind AI. The latter class of AI responds to human feelings and also executes the actions of Limited Memory AI with added efficacy. It is used in Autonomous cars and other applications and is yet to take off in a truly practical sense. And robot armies, mechanical overlords, sentient humanoids, all being hypothetical form the part of AI that is aware of itself, Self-aware AI.

The maturity of AI can be classified into three categories:

1. Artificial Narrow Intelligence: A good example of this category of AI would be that of natural language processing AI or NLP AI. But its capabilities are limited. For instance, this type of Intelligence can respond to voice commands and perhaps execute a few actions related to the same. But it cannot go beyond this stretch.

2. Artificial General Intelligence: The objective of this AI, of which research is ongoing and is largely at a theoretical level, is to create lifelike intelligent assistants to humans.

3. Artificial Superintelligence: Intelligence that would be the brightest on planet earth and superior to human intelligence: nowadays a part of science fiction literature.

Ethical issues/ challenges posed by AI.

AI is deemed to be the portal that would usher in human progress and innovation by enabling the flourishing of humanity in terms of its achievement of individual and societal wellbeing. The technology is expected to work for the greater common good. At a practical level, AI systems are professed to aid in the achievement of the UN's SDGs or Sustainable Development Goals, especially for the promotion of gender parity, balance and tackling climate change. It may also help us rationalize the use of our natural resources, healthcare enhancement, aiding in production processes and mobility enablement. It can even help with monitoring progress against social cohesion and sustainability parameters. But the advancement of AI may give rise to or has given rise to, in some cases, a range of issues or challenges, including those that related to bias. A few of the challenges are provided below:

Challenges related to life and living.

While the Theory of Mind AI and Self-aware AI are still on paper and belong to the realms of hypotheses, AI in its present form itself presents a bevy of ethical challenges and dilemmas. The best example is that of the trucking industry. It employs many millions of people in the United States and Canada alone. However, the future of those truck drivers comes under question as self-driving trucks and autonomous vehicles take to roads in a big way. Many of them are sure to lose their jobs. And conventionally human beings generally sell their time to earn the necessary monetary resources to sustain themselves and their immediate family. While a number of people who may lose their jobs may be hypothetically able to upskill or reskill or reskill or reskill or reskill or reskill or reskill, others may not be so privileged. Thus, the ethical challenge becomes the following: the challenges of sustaining the lives and living standard of those who lose their jobs to AI technology in the face of inability to upskill or reskill making them devoid of the capability to sell their time and skills.

Challenges related to inequality.

There is a widening wealth-gap in the world economy. And we can see that entrepreneurs rake in millions and billions in monetary value even as laborers are left in the lurch. With AI technology getting more and more matured, the human workforce will be cutdown and revenue and income will increasingly go to the personnel at the top. Thus, the ethical challenge becomes the following: the challenge of distributing wealth generated by the aid of AI enabled machines in an equitable fashion and ensuring fairness.

Challenges related to artificial stupidity.

There are a couple of challenges with regard to Artificial Stupidity. There are generally two phases to ML. The training phase and the testing phase. The former enables the machine with training in patterns followed by testing wherein it is given more examples and data from the real world. The systems can be fooled in multiple ways. For instance, an image recognition system may be supplied with a random dot pattern and asked to come up with recommendations or action points. The machine may simply fail to identify the random dot pattern and “see” things that are not there and recommend ‘stupid’ actions. Thus, the challenge becomes the following: the challenge of stupidity in AI ruining the objectives for which it was created in the first place.

3.Proposed System: TAI-embodied GRC Architecture

Risk assessment, detection of fraud, classifying images, filtering of spam and a host of other processes are made possible due in part to supervised learning technique. Machines are trained and tested in a system of learning in the AI field utilising which predictions are made regarding the potential output based on both the data and training. The labelled data or the training data in the model is a significance that some of the input data which has been fed to the system has been tagged with the appropriate output.

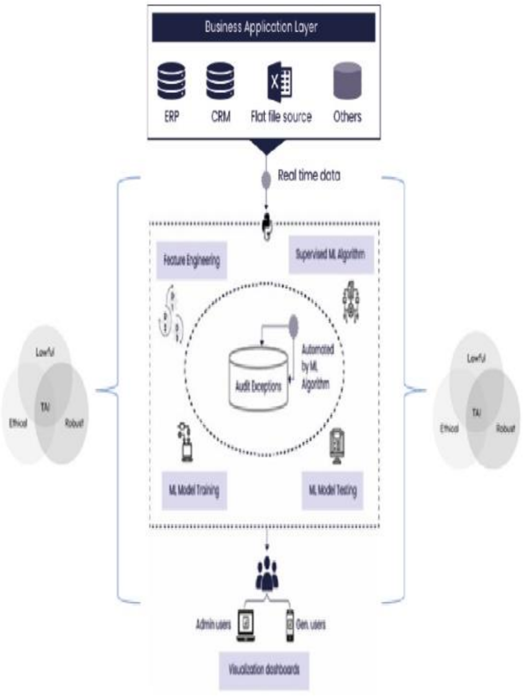


Figure 2: Audit architecture with TAI enabled.

Architecture consists of two segments:

1. Organization Business Applications Layer/ Business Applications Layer: This layer or segment has a range of applications feeding data into it. ERP or Enterprise Resource Planning systems, CRM (Customer Relationship Management) systems, flat file sources and a host of other apps can connect to this.
2. Audit Analytics Layer: This is a layer constituted by the Audit Data repository and an Audit Exceptions engine which is automated by a Supervised Machine Learning Model.

The implementation of Architecture

1. The data from the Business Applications layer will be redirected to the Audit Analytics layer intermediating the Business Application Layer and the Supervised ML layer. The data from the former layer is extracted, transformed and added/ loaded by an ETL process and stored in the intermediary level of Audit Data Repository or Warehouse and the data is processed in accordance with the Audit Rule Scripts. In other words, data from the audit data repository moves into the audit rules engine wherein all the processed rules are stored apart from pre-processed Exceptions. It is pertinent to note that GRC Architecture follows human agency and oversight as it is supervised ML. The AI model is guided at appropriate times using well-ordained procedures to facilitate intervention of a human stakeholder as and when necessary. The AI Model guidance is planned for and implemented herein.

The potential for business disruption is primarily pre-empted due to the following set of reasons. The AI system has an in-built model control faculty that would take care of wrong actions, preventing them from manifesting by dividing or classifying AI activity into the following classes:

Tier 1 Events: These are super-critical functions wherein a potential misstep can cause business disruptions. All the Tier 1 Events to manifest, human approval is necessary. Thus, human intervention is deemed high for these types of events.

Tier 2 Events: There are sub-super-critical functions/ events that warrant only moderate or occasional human intervention.

Tier 3 Events: There are certain events that can progress on autopilot with low to no human intervention.

Thus, the hybrid model of decision-making depending on the severity of the events ensures that the GRC Architecture is TAI-compliant.

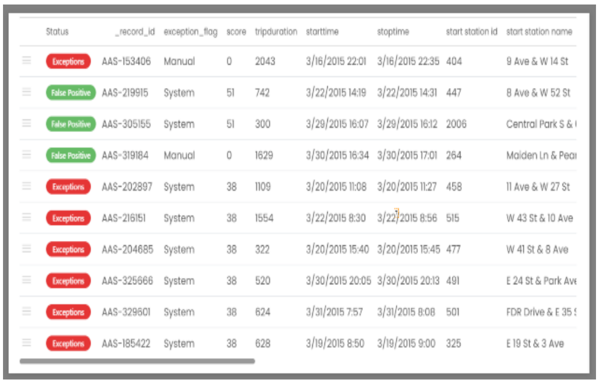
2. The engine has a proper feature engineering process and is automated by supervised ML model. The feature engineering carried out by Auditors by flagging of the data as whether a false positives or exceptions. This data will be used to train the ML model further to perfection. It is a continuous process with the entire population and not based on samples, there is provided a section that brings out Analytical Insights and rationale behind event classification or parameter selection. Every decision taken is accounted for and explained. The key driver behind each decision is well-represented, self-explainable and transparent to each stakeholder.

The data which is used in the entire process is subjected to data cleaning as well as pre-processing of data. The process in fact cleans the complete data set from null values, corrects it and if needed removes inaccurate and improper records. This step ensures that the data is standardized and normalized. Additionally, in the preprocessing of data, values from the JDOC column are extracted. This feature will be employed for the model training. To contribute to automating the audit exception engine, a supervised classification algorithm known as the random forest model is used. The residual data is categorised and is used in training the model and model testing.

Diversity, non-discrimination, and fairness are characterised by a special model. Decisions taken by AI is the first line of action. Stakeholder-involved decision is the second line of action. Independent external/ internal auditing validates/ authenticates all the layers.

Figure 2: This figure portrays the state wherein the machine has marked Exceptions and False Positives on its own.

The said GRC Architecture is ISO 27001:2013 compliant. Complies with all the laws, regulations and norms governing the respective domains. Paperless electronic digital system ensures a reduced carbon footprint for the GRC Architecture making it TAI-compliant. Proactive Continuous Controls Monitoring or CCM coupled with real time reporting to the higher-ups ensure TAI compliance.



Status	_record_id	exception_flag	score	tripduration	starttime	stoptime	start station id	start station name
Exceptions	AAS-153406	Manual	0	2043	3/16/2015 22:01	3/16/2015 22:35	404	9 Ave & W 14 St
False Positive	AAS-219915	System	51	742	3/22/2015 14:19	3/22/2015 14:31	447	8 Ave & W 52 St
False Positive	AAS-305155	System	51	300	3/29/2015 16:07	3/29/2015 16:12	2006	Central Park S &
False Positive	AAS-319184	Manual	0	1629	3/30/2015 16:34	3/30/2015 17:01	264	Maiden Ln & Pear
Exceptions	AAS-202897	System	38	1109	3/20/2015 11:08	3/20/2015 11:27	458	11 Ave & W 27 St
Exceptions	AAS-216151	System	38	1554	3/22/2015 8:30	3/22/2015 8:56	515	W 43 St & 10 Ave
Exceptions	AAS-204685	System	38	322	3/20/2015 15:40	3/20/2015 15:45	477	W 41 St & 8 Ave
Exceptions	AAS-325666	System	38	520	3/30/2015 20:05	3/30/2015 20:13	491	E 24 St & Park Ave
Exceptions	AAS-329601	System	38	624	3/31/2015 7:57	3/31/2015 8:08	501	FDR Drive & E 35
Exceptions	AAS-185422	System	38	628	3/19/2015 8:50	3/19/2015 9:00	325	E 19 St & 3 Ave

Results and Discussion

Discussion based on the results of the TAI embodiment of the GRC System Architecture

1) Human Agency and Oversight

The GRC System does augment and add to human capabilities. For instance, the system can mark exceptions on its own enabling humans to save time, energy, and money. However, there are built-in safeguards as well. There is a system of well-conceived practices followed to mark false-positives wherever necessary by virtue of human intervention. The Architecture, thus, in action follows human agency and oversight as it is into Supervised ML. An AI Model Guidance Plan is followed.

The series of screen shots provided below is.

an excellent example in this regard:

Figure 1: This figure (given above) portrays the status of Exceptions as provided in the system.

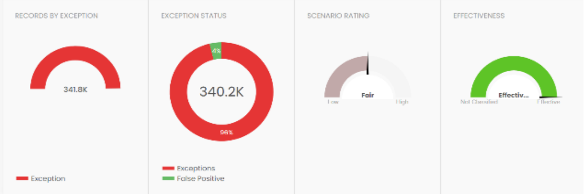
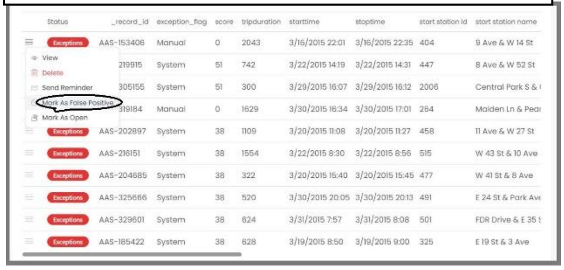
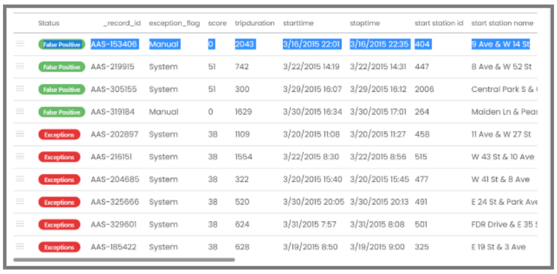


Figure 3: This figure portrays the state wherein human intervention takes place and an Exception is marked as False Positive by a human being in guidance to the system.



Status	_record_id	exception_flag	score	tripduration	starttime	stoptime	start station id	start station name
Exceptions	AAS-153406	Manual	0	2043	3/16/2015 22:01	3/16/2015 22:35	404	9 Ave & W 14 St
	219915	System	51	742	3/22/2015 14:19	3/22/2015 14:31	447	8 Ave & W 52 St
	305155	System	51	300	3/29/2015 16:07	3/29/2015 16:12	2006	Central Park S &
	319184	Manual	0	1629	3/30/2015 16:34	3/30/2015 17:01	264	Maiden Ln & Pea
Exceptions	AAS-202897	System	38	1109	3/20/2015 11:08	3/20/2015 11:27	458	11 Ave & W 27 St
Exceptions	AAS-216151	System	38	1554	3/22/2015 8:30	3/22/2015 8:56	515	W 43 St & 10 Ave
Exceptions	AAS-204685	System	38	322	3/20/2015 15:40	3/20/2015 15:45	477	W 41 St & 8 Ave
Exceptions	AAS-325666	System	38	520	3/30/2015 20:05	3/30/2015 20:13	491	E 24 St & Park Av
Exceptions	AAS-329601	System	38	624	3/31/2015 7:57	3/31/2015 8:08	501	FDR Drive & E 35
Exceptions	AAS-185422	System	38	628	3/19/2015 8:50	3/19/2015 9:00	325	E 19 St & 3 Ave



Status	_record_id	exception_flag	score	tripduration	starttime	stoptime	start station id	start station name
False Positive	AAS-153406	Manual	0	2043	3/16/2015 22:01	3/16/2015 22:35	404	9 Ave & W 14 St
False Positive	AAS-219915	System	51	742	3/22/2015 14:19	3/22/2015 14:31	447	8 Ave & W 52 St
False Positive	AAS-305155	System	51	300	3/29/2015 16:07	3/29/2015 16:12	2006	Central Park S &
False Positive	AAS-319184	Manual	0	1629	3/30/2015 16:34	3/30/2015 17:01	264	Maiden Ln & Pea
Exceptions	AAS-202897	System	38	1109	3/20/2015 11:08	3/20/2015 11:27	458	11 Ave & W 27 St
Exceptions	AAS-216151	System	38	1554	3/22/2015 8:30	3/22/2015 8:56	515	W 43 St & 10 Ave
Exceptions	AAS-204685	System	38	322	3/20/2015 15:40	3/20/2015 15:45	477	W 41 St & 8 Ave
Exceptions	AAS-325666	System	38	520	3/30/2015 20:05	3/30/2015 20:13	491	E 24 St & Park Av
Exceptions	AAS-329601	System	38	624	3/31/2015 7:57	3/31/2015 8:08	501	FDR Drive & E 35
Exceptions	AAS-185422	System	38	628	3/19/2015 8:50	3/19/2015 9:00	325	E 19 St & 3 Ave

Figure5: This figure portrays the state wherein the machine imbibes the new learning and performs accordingly.

2) Technical Robustness and Safety along with Data Governance and Privacy.

As mentioned above, all the super-critical functions wherein a potential misstep can cause business disruptions need mandatory human approval. All the sub-super-critical functions/ events may have only moderate or occasional human intervention. These measures ensure technical robustness and safety. The Architecture, above and beyond this is ISO 27001:2013 compliant. Thus, it complies with all the laws, regulations and norms governing the respective domains.

3)Transparency

The rationale behind every action is well recorded and accounted for by the system. In fact, there has been provided a specific location that has Analytical Insights and rationale behind event classification or parameter selection. Thus, every decision taken is accounted for and explained. The key driver behind each decision is well-represented and self-explainable. It is transparent to each stakeholder.

4)Diversity, Non-Discrimination, And Fairness

Independent external/ internal auditing validates/ authenticates all the actions taken by AI and humans which is into an ordained feedback cycle with the GRC system ensuring proper learnings and constant and complete adherence to diversity, non-discrimination and fairness.

5)Societal and Environmental Well-being

Apart from being a paperless electronic digital system ensuring a reduced carbon footprint for the GRC Architecture, the system is used only by trained professionals, it being a B2B application pre-empting the prevention of the development of attachment and empathy by human beings towards the system.

6)Accountability

Proactive Continuous Controls Monitoring or CCM coupled with real time reporting to the higher-ups ensure TAI compliance apart from other safeguards as characterised by supervised ML model and human oversight.

Conclusion

For AI to remain strong and relevant, trustworthy AI is the need of the hour. The parameters of Trustworthy AI have been outlined well in this paper along with how a GRC system architecture adheres to these parameters of Trustworthy AI. The system is seen to be adhering to the basic tenets of Trustworthy AI and performing as lawful, ethical and quite robust.

References:

[1] Prince Chacko Johnson, Christofer Laurell, Mart Ots, Christian Sandström, Digital innovation and the effects of artificial intelligence on firms' research and development – Automation or augmentation, exploration or exploitation? Technological Forecasting and Social Change, Volume 179, 2022, 121636, ISSN 0040-1625,

[2] H. Choung, P. David, and A. Ross, "Trust in AI and its role in the acceptance of AI Technologies," *International Journal of Human–Computer Interaction*, pp. 1–13, 2022.

[3] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel, "The relationship between trust in AI and Trust worthy Machine Learning Technologies," *Proceedings of the 2020 Conference On Fairness, Accountability, and Transparency,* 2020.

[4] F. Marcinkowski, K. Kieslich, C. Starke, and M. Lünich, "Implications of AI (un-)fairness in higher education admissions," *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2020.

[5] D. Roselli, J. Matthews, and N. Talagala, "Managing bias in ai," *Companion Proceedings of the 2019 World Wide Web Conference*, 2019.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[7]E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M. E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, and S. Staab, “Bias in data-driven Artificial Intelligence Systems—an introductory survey,” *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2020.

[8] A. Rai, “Explainable AI: From black box to glass box,” *Journal of the Academy of Marketing Science*, vol. 48, no. 1, pp. 137–141, 2019.

[9] M. S. Thomas and J. Mathew, "Supervised Machine Learning Model for Automating Continuous Internal Audit Workflow," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), 2022, pp. 1200-1206, doi: 10.1109/ICOEI53556.2022.9776888.

[10] Jean-Marie John-Mathews, Some critical and ethical perspectives on the empirical turn of AI interpretability, Technological Forecasting and Social Change, Volume 174, 2022, 121209, ISSN 0040-1625.

[11] Emily Shearer, Mildred Cho, David Magnus, Chapter 23 - Regulatory, social, ethical, and legal issues of artificial intelligence in medicine, Editor(s): Lei Xing, Maryellen L. Giger, James K. Min, Artificial Intelligence in Medicine, Academic Press, 2021, Pages 457-477, ISBN 9780128212592.

[12] Lan Xue, Zhenjing Pang, Ethical governance of artificial intelligence: An integrated analytical framework, Journal of Digital Economy, 2022, ISSN 2773-0670,

[13] Bernd Carsten Stahl, Rowena Rodrigues, Nicole Santiago, Kevin Macnish, A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values, Computer Law & Security Review, Volume 45, 2022, 105661, ISSN 0267-3649.

[14] Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2021). Trustworthy AI: From Principles to Practices. *arXiv preprint arXiv:2110.01167*.

[15] Veale, Michael. "A critical take on the policy recommendations of the EU high-level expert group on artificial intelligence." *European Journal of Risk Regulation* 11.1 (2020): 1-10.