AryaXAI

The Evolving Landscape of AI Regulations in the US

Challenges, best practices and implementing effective AI Governance strategies

Whitepaper by:

Arya.ai

2023

Contents

Foreword

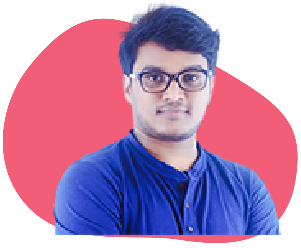
As AI and automation technology mature, the need for inherently interpretable, explainable and responsible models has become the critical focus. While this development is being encouraged, there has been an increased emphasis on managing associated risks with these technologies. The AI/ ML regulatory landscape in the US is changing rapidly; it has become imperative for organizations to make requisite tweaks in their business processes and explain to regulators how their system works to demonstrate compliance with applicable regulations.

The US government has geared up its ongoing efforts on 'Responsible AI' and emphasise the importance of driving responsible, trustworthy, and ethical innovation with safeguards that mitigate risks and potential harms to individuals and society. In May 2023, the Biden-Harris Administration announced new actions that will further promote responsible American innovation in artificial intelligence (AI) and protect people's rights and safety.

While there has been palpable excitement around Responsible AI and AI Governance, it is all still in the conceptual phase. Achieving AI governance will help organizations manage AI risk and scale while complying with the growing AI regulations.

In this whitepaper, we summarize the emerging regulatory framework for AI in the US, discuss challenges and propose concrete steps companies can take to comply with such regulations.

Author



Vinay Kumar

Founder & CEO, Arya.ai

Vinay Kumar Sankarapu is the Founder and CEO of Arya.ai. He did his Bachelor's and Masters in Mechanical Engineering at IIT Bombay. His research was in Deep Learning and published thesis on CNNs in manufacturing. He started Arya.ai in 2013, along with Deekshith to democratize Deep Learning. Since then, Vinay had been working with multiple enterprise in regulated industries like Banks, Insurers and Financial Services.

Vinay Kumar also leads the R&D of AryaXAI product. He wrote multiple guest articles on ‘Responsible AI’, ‘AI usage risks in BFSIs’ and ‘AI Governance framework’. He presented multiple technical and industry presentations globally - Nvidia GTC (SF & Mumbai), ReWork (SF & London), Cypher (Bangalore), Nasscom(Bangalore), TEDx (Mumbai) etc. He was the youngest member of ‘AI task force’ set up by the Indian Commerce and Ministry in 2017 to provide inputs on policy and to support AI adoption as part of Industry 4.0. He was listed in Forbes Asia 30-Under-30 under the technology section. He is part of Forbes Technology Council and Analytics India Magazine Leaders Council.

Introduction

The Biden administration recently announced its intention to gather feedback from the public on potential accountability measures for artificial intelligence (AI) systems as concerns about its effect on national security and education rise. The National Telecommunications and Information Administration (NTIA), a Commerce Department agency responsible for advising the White House on telecommunications and information policy issues, is seeking input from the public because regulators have a growing interest in establishing an "accountability mechanism" for AI systems.

As regulations around AI continue to gain importance, industries that are already highly regulated, such as the financial sector, are leading in developing guidelines to address the potential risks of Artificial Intelligence/Machine Learning (AI/ML). This trend is being mirrored in the US, as lawmakers are looking for measures to be put in place to provide assurance "that AI systems are legal, effective, ethical, safe, and otherwise trustworthy." NTIA plans to draft a report looking at "efforts to ensure AI systems work as claimed – and without causing harm" and said the effort will inform the Biden Administration's ongoing work to "ensure a cohesive and comprehensive federal government approach to AI-related risks and opportunities."

In the US, 2022 saw an initial approach to AI regulation emerge, focused on specific AI-use cases. Regulations like the Equal Opportunity Employment Commission (EEOC) on "algorithmic fairness" in employment or automated employment decision tools (AEDTs) that leverage AI to make or substantially assist candidate screening or employment decisions.

For Banks, the current regulatory guidance in the United States reflects both general and specific concerns related to artificial intelligence. The Office of the Comptroller of the Currency (the "OCC"), responsible for chartering, regulating, and supervising all national banks and federal savings associations, follows a risk-based supervision approach that focuses on key issues such as explainability, data management, privacy and data security, and third-party risk.

The Consumer Financial Protection Bureau (CFPB) in May 2022, published a circular on Adverse action notification requirements in connection with credit decisions based on complex algorithms. CFPB stated that the Equal Credit Opportunity Act (ECOA) and it's implementing rules apply to all credit decisions, including those made using complex algorithms. The ECOA prohibits discrimination against credit applicants and prohibits creditors from using algorithms that prevent them from providing specific and accurate reasons for adverse actions, such as declined applications.

In October 2022, the White House Office of Science and Technology Policy (OSTP) unveiled its Blueprint for an AI Bill of Rights. The blueprint is a non-binding set of guidelines for designing, developing, and deploying artificial intelligence (AI) systems and consists of a set of five principles and associated practices, namely:

1. **Transparency:** AI systems should be transparent and accountable, and their decision-making processes should be open to scrutiny and public examination.
2. **Fairness:** AI systems should be designed to be fair, unbiased, and inclusive and should avoid reinforcing existing social and economic inequalities.
3. **Privacy:** AI systems should respect the privacy rights of individuals and comply with relevant privacy laws and regulations.
4. **Security:** AI systems should be secure, resilient, and able to withstand cyber attacks or other malicious activities that could compromise their integrity or functionality.
5. **Accountability:** Those who design, develop, and deploy AI systems should be accountable for the impact of their systems on society, and should take measures to mitigate any negative impacts that arise.

Designing an AI Governance framework

Although the regulations vary from country to country, there are common themes found throughout all approaches that can guide compliance efforts. In order to follow the conduct, organizations can consider adopting the following 'AI Governance' components:

- **Reliability:** The recommendations discuss using accurate and exhaustive training data such that algorithms have been provided with enough training data to perform and are reliable in production. Consistency of these algorithms is critical as the algorithms are used as decision automation/augmentation tools that can directly influence business performance and risk.

- **Auditability:** Having algorithms are part of the process; it is required to ensure there is auditability of these algorithms for regulatory requirements and be able to provide functional information when needed.

- **Transparency:** Algorithms should be explainable and traceable such that there is transparency and trust in the process. Such transparency should be provided to all stakeholders in the process, like Underwriters, Data Scientists, Product Managers, Business Owners, Risk Managers, Regulators and customers.

- **Responsible AI:** Algorithms should be built responsibly and used responsibly! Responsible AI is not just limited to the usage of 'AI' but also how it has been built. Training data should be procured correctly, and no restricted data should be used for training purposes.

- **Fairness in underwriting:** Organizations should ensure that there is no algorithmic bias and ensure that underwriting is fair underwriting practices, impartial and provides an inclusive opportunity to everyone.

- **Data Privacy:** Institutions should ensure that the data is protected and secure.

So far, we've outlined various components required for 'AI Governance'. Now, let's discuss the challenges, threats and opportunities while building such 'AI Governance' in your organization.

Reliability: How do you promise reliability when 'AI' can fail?

There is a heavy assumption that algorithms, once built, they'll continue to deliver similar or better performance over time. In reality, models can fail quite often, like any other software.

- Credit score provider, 'Equifax' issued wrong credit scores for thousands of customers resulting in erroneous decisions by lenders. Data drift may have caused the issues, resulting in model failures. (Ackerman and Andriotis 2022)
- Zillow (Metz 2021), a real estate startup, had to shut down the home purchase business after the AI debacle. They were using AI to predict prices, but when the models failed, they incurred tremendous losses and had to shut down that entire business unit.

The reasons for model failure are many - data drift, concept drift, data sanity issues etc. They need to be monitored and maintained to ensure that the model is safe and that the right data is going for inferencing. Vela et al. 2022 tested multiple models to understand the temporal effects on model performance. They showed that more than 91% of models failed over time.

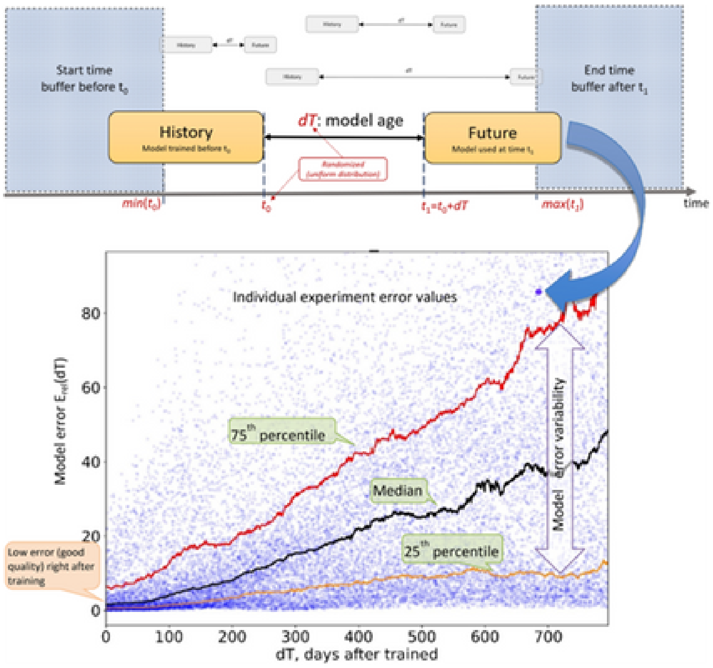


Figure 1: Plot of error distribution with age of model from training time period (Vela et al. 2022)

Data drift is caused when the data distribution changes or untrained data is added in inference samples. It could be caused because of changes in the process or acquiring new profiles, or going into new markets or integrating with new systems etc. Any of these issues can impact model performance.

Concept/model drift occurs when the target is changed because of changes in the processes, business guidelines etc. For example, the definition of a good profile can be changed during COVID, resulting in different underwriting decisions as compared to training data.

Data sanity is another prominent reason for failure. It is caused when the data pipelines are broken or altered without changing the models. Let's say you were using the milliseconds in training data, and now you're sending seconds data during production.

When algorithms fail, it not only affects the business financially but also damages the reputation and can invite huge unnecessary regulator challenges.

The excitement of deploying using AI should be carefully balanced against the caution that is required to use it. Else, institutions would only focus on 'manufacturing the algorithms' without monitoring them.

Fairness and Bias: Can models develop bias over time?

Fair underwriting promises an equal evaluation and allocation of decisioning logic without considering 'sensitive' features defined by societal or geo prejudices. The use of such sensitive features in modelling is prevented as per regulations in geographies like the US. RBI hasn't restricted the usage of these features directly but suggested using fair and non-biased algorithms.

When Apple launched the credit cards with Morgan Stanley in 2019, it was observed that there was bias (Vigdor 2019) in underwriting and preferring males as compared to other genders. This was quickly highlighted and resulted in a PR mess. It was later investigated by New York's Department of Financial Services (BBC News 2019).

To ensure there is no bias, institutions should list out such sensitive classes (these are debatable and changing over time) and track them during testing and production. Simpler bias-sensitive features are 'Gender', 'Caste', 'Location' etc. Eventually, the institutions are expected to deliver the same decision outcome for the same profiles irrespective of whether they are Male or female, caste or where they are coming from.

While the execution looks simpler, it gets complicated because of data and concept drift. You may ensure that there is no bias during training and testing, but bias can be developed over time because of the drifts. (Vela et al. 2022) observed that the bias increases as the models are used for longer periods.

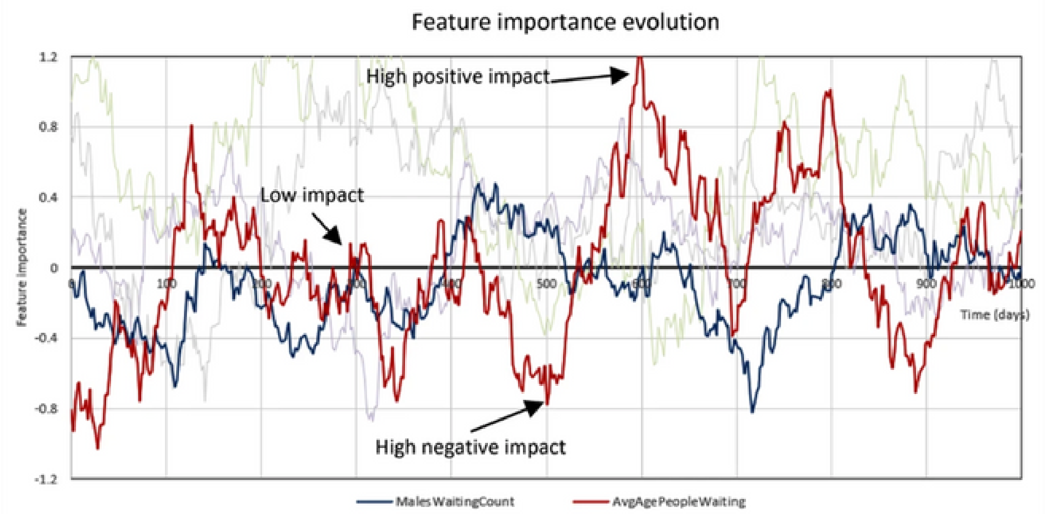


Figure 2: In the experiment, it is observed that feature importance of sensitive features like 'Gender of patient' and 'Age of patient' can vary over time and can influence negatively, positively or null! (Vela et al. 2022).

ML Explainability: how do you build trust in the algorithms when they are not understandable to all stakeholders?

ML Explainability is one of the most interesting and challenging components of AI Governance. ML Explainability delivers not only AI trust and acceptance but also better auditability. First, let's understand what is 'ML Explainability'

Problem 1: The definition of acceptable explanation changes by whom you ask.

When asked about explainability, many institutions disclosed that they use feature importance as part of explainability. For a data scientist, a feature importance map could be an acceptable explainability, whereas an underwriter cannot understand these graphs. In such cases, how can one ensure the algorithm output is understandable by all stakeholders? The data science teams might have spent a lot of time building these models and started producing these accurate predictions. But the prediction itself is not sufficient to build confidence in the ecosystem.

In the underwriting process, data scientists are least involved in the processes once the model is pushed into production. Using poorly explainable models can limit the usage of the model predictions and high chances of slowly becoming obsolete. And for critical decisions, stakeholders can not continuously inquire about the reasoning and evidence with data science teams on case to case basis needing huge time investment by all stakeholders. Poor explainable systems for regulators can result in poor auditing and evidence with data science teams on case to case basis needing huge time investment by all stakeholders.

Poor explainable systems for regulators can result in poor auditing and incomplete risk assessment. Explainability is not a default output of the models, and it needs additional layers to deliver acceptable explainability for all stakeholders.

As algorithms are slowly underwriting the majority of the book, how safe is it to use black box models, provide wrong explanations, or use opaque models that are not understandable by all stakeholders?

The design of explainability should actually start with understanding the requirements of these stakeholders. For example, decision experts would be interested in understanding the cause/decision trail behind the decision and the correlation of their decision logic with the algorithm logic. At the same time, regulators or business owners would go one-more step to find the source of the learning to evaluate the reliability and authenticity of the decision/data.

Problem 2: No explanation is better than a wrong explanation.

For example, let's say you are using some generic XAI method to explain your model, which is not optimized for your model; you may be sharing wrong explanations to all the stakeholders, including regulators, about the model's functionality. This can be riskier than using a black box model.

Inconsistent explanation also cause severe damage to the model confidence. If you are using techniques like SHAP or LIME, the explainability is not consistent and could vary with changing number of pertubation or base sample size.

Problem 3: As the technique is more complex, explainability becomes harder.

Simple rules-based underwriting engines are easier to interpret and understand than classic machine learning models. And classic machine learning models are easier to explain than deep learning models. New techniques are suggested to build intrinsically explainable models, even using deep learning, but the explainability is not heavily different from using the deep learning models as-is.

Algorithmic Auditing: A Curious Case of Evidence Formulation

Auditing is a systematic method of capturing artifacts designed to risk aversion. Traditional auditing only focuses on human-dependent processes with key SPOCs like Underwriting, Business Owners and Finance teams. But many institutions are not following even a bare minimum algorithmic auditing practises today. Any algorithm auditing today is simply limited to the practice of attaching a very high-level report that has very limited information about the aforementioned AI governance components recommended by regulators. The issues related to lack of skills, the urgency to deploy models, poor explainability, not capturing the required model artefacts, and lack of model evaluation skillsets. Algorithmic auditing requires multidisciplinary skillsets in both Machine Learning and Business risk evaluation. And this is missing from both regulators and institutions.

While there is time for regulators to catch up, but for institutions, this means they could be using risky models and do not even know about them. The case study of Zillow, Equifax or Apple, while they are known for their technical sophistication, financial institutions are largely process heavy. If they face business threats of using risky models, financial institutions are carrying heavy risks and do not even know about them.

Institutions should at least start following basic algorithmic auditing habits to start capturing the minimum artefacts, to begin with, and expand with maturity. Else, these can cause heavy reputational and financial damage to the business.

Data Privacy: Can algorithms leak data?

Yes, algorithms can leak sensitive data. (Carlini et al. 2022) Suggested 'Inference attacks' can disclose the data used in training the model. While it talked about deep learning models, a modified method can be used to predict whether a 'sample' was used in training. Such membership inference attacks (Shokri et al. 2016) have been becoming a challenging security concern for the use of 'AI'. (Jegorova et al. 2021) Presented a survey outlining the research literature on studying this kind of attack.

Another technique used to preserve data was anonymization. But deanonymization is possible, and feature correlation can also be exposed, leading to data attacks. (Carlini et al. 2022) Were able to find the customers from anonymized data of Netflix review data by correlating them with Amazon reviews.

If you are generative AI models trained on sensitive data, there is a high chance of making these models disclose such information. There were instances where generative AI models like large language models (LLMs like ChatGPT) or diffusion models (like Midjourney) disclosed sensitive information in the output (Carlini et al. 2023), (Duan et al. 2023). Another type of attack is 'induced poisoning of training data'. (Tramèr et al. 2022) demonstrated how the training data sets could be poisoned to reveal sensitive information, including credit card numbers.

Responsible training Data: Responsibility is not just limited to usage but also how it is built

We've already talked about the responsible use of AI in our earlier section pertaining to Fair underwriting models, explainable AI that is understandable to all stakeholders and auditable AI to ensure accountability can be traced. Many institutions miss out on another critical aspect of responsible AI: how these algorithms are built. There have been many examples of misuse of customers' data and privacy to achieve better credit scoring. Responsible 'AI' ensures an ethical way of procuring training data and allowing flexible controls to customers on skipping their data in the models.

Regulators or risk managers, when auditing these algorithms, the source of training data is also part of the audit. There are challenges for regulators to check the ethical boundaries of the institutions as algorithmic auditing is challenging as it requires interdisciplinary skills, which are lacking today.

Because of these reasons, institutions can find ways to fool the auditing systems. For example, SHAP and LIME are very traditional explainable approaches. It is found that these can be fooled by using scaffolding attacks. (Slack et al. 2020) showcased how these methods could be spoofed and hide biased features from auditing.

AI usage risk - what happens when the models fail?

So far, we have discussed the recommendation given by the regulators. Another component that needs attention is 'AI usage risks'. Algorithms are complex software that can fail or go rogue for various reasons. As they are a critical part of the business, they can create unmanageable losses that can harm the business if they fail. So, how do institutions ensure that they deploy 'AI' safely?

To understand the risk associated with the algorithms, institutions must stress test them to find the gaps and identify failure areas. After identifying the issues, institutions can resolve them by giving necessary feedback to the models or creating another layer called 'Policies' that can provide directional inputs to the model to avoid these pitfalls. 'Policies' can also be used to define risk boundaries that optimize the value between risk and returns. Institutions should ensure that this layer is well evaluated before productionising any algorithm. Such a layer can also provide confidence to regulators about the institution's ability to use algorithms at scale without risking the business viability. 'Prompting' layer in LLMs is the 'Policies' layers. While prompting is now well known because of LLMs, but such layers in other ML techniqies are very underlooked.

But the challenge lies in finding enough data to stress testing these models, as it is critical to identify sufficient data to have a valid and satisfactory test outcome. Institutions can use 'Synthetic AI' to produce enough testing samples to stress test models or validate the policies.

Conclusion

The Biden-Harris Administration's announcement has come at a time when there is a global heightened attention to initiatives against potential risks from using Artificial Intelligence/ Machine Learning (AI/ML). While there is a broad category of regulations, there is still much confusion regarding the degree and manner in which these regulations apply to the use of AI.

While the regulations are still catching up with the rapidly evolving scope and nature of AI technology, it is likely that we will soon see specific regulations and guidelines tailored to different industries and use cases of AI. Combining government oversight with industry self-regulation through activities such as industry consultations and educational programs to train regulatory personnel on technology knowledge and skills will be crucial in staying informed and adapting to the evolving AI regulatory landscape.

Like any other software, 'AI' is not explainable, can fail, be biased, leak data, and have risks in using it, but the potential outgrows all these characteristics. And these can be managed by deploying comprehensive 'AI' governance tools. Regardless of any further actions, industries - primarily highly regulated- like financial services, healthcare, and fintech need to start taking incremental steps considering current regulations and use cases.

Refrences

- Ackerman, Andrew, and Annamaria Andriotis. 2022. "Equifax Sent Lenders Inaccurate Credit Scores on Millions of Consumers." The Wall Street Journal, August 2, 2022. https://www.wsj.com/articles/equifax-sent-lenders-inaccurate-credit-scores-on-millions-of-consumers-11659467483.
- Artificial Intelligence Committee Reports: https://www.meity.gov.in/artificial-intelligence-committees-reports
- BBC News. 2019. "Apple's 'Sexist' Credit Card Investigated by US Regulator." BBC News, November 10, 2019. https://www.bbc.com/news/business-50365609.
- Carlini, Nicholas, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. "Membership Inference Attacks From First Principles." 2022 IEEE Symposium on Security and Privacy (SP). https://doi.org/10.1109/sp46214.2022.9833649.
- Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. "Extracting Training Data from Diffusion Models," January. https://doi.org/10.48550/arXiv.2301.13188.
- Duan, Jinhao, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. 2023. "Are Diffusion Models Vulnerable to Membership Inference Attacks?," February. https://doi.org/10.48550/arXiv.2302.01316.
- Jegorova, Marija, Chaitanya Kaul, Charlie Mayor, Alison Q. O'Neil, Alexander Weir, Roderick Murray-Smith, and Sotirios A. Tsaftaris. 2021. "Survey: Leakage and Privacy at Inference Time," July. https://doi.org/10.48550/arXiv.2107.01614.
- Metz, Rachel. 2021. "Zillow's Home-Buying Debacle Shows How Hard It Is to Use AI to Value Real Estate." CNN. November 9, 2021. https://www.cnn.com/2021/11/09/tech/zillow-ibuying-home-zestimate/index.html.
- Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2016. "Membership Inference Attacks against Machine Learning Models," October. https://doi.org/10.48550/arXiv.1610.05820.
- Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. "Fooling LIME and SHAP." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3375627.3375830.
- Tramèr, Florian, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. 2022. "Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets," March. https://doi.org/10.48550/arXiv.2204.00032.
- Vela, Daniel, Andrew Sharp, Richard Zhang, Trang Nguyen, An Hoang, and Oleg S. Pianykh. 2022. "Temporal Quality Degradation in AI Models." Scientific Reports 12 (1): 1–12.
- Vigdor, Neil. 2019. "Apple Card Investigated After Gender Discrimination Complaints." The New York Times, November 10, 2019. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.
- Blueprint for AI Bill of Rights, https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf
- Statement from Office of the Comptroller of the Currency, May 13, 2022: https://www.occ.gov/news-issuances/congressional-testimony/2022/ct-occ-2022-52-written.pdf
- Consumer Financial Protection Circular 2022-03 by CFPB, MAY 26, 2022: https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/#15
- AI Accountability Policy Request for Comment by National Telecommunications and Information Administration, U.S. Department of Commerce: https://ntia.gov/sites/default/files/publications/ntia_rfc_on_ai_accountability_final_0.pdf
- Proposed regulations on Automated Employment Decision Tools by The Department of Consumer and Worker Protection: https://rules.cityofnewyork.us/rule/automated-employment-decision-tools-2/
- Artificial Intelligence and Algorithmic Fairness Initiative by U.S. Equal Employment Opportunity Commission (EEOC): https://www.eeoc.gov/ai

Company Profile

About AryaXAI:

AryaXAI is the ML Observability tool for mission-critical 'AI'. It aims to find the gaps in AI solutions and fix them in auto-mode or provide insights for manual intervention. It addresses key issues in the ML solution life cycle like Explainability, Model Performance Degradation, Monitoring, Auditability and AI usage risks. AryaXAI uses patent-pending algorithms for explainability in Deep Learning along with leveraging multiple open source methods for ML Monitoring. AryaXAI is already used in Arya.ai's products like Automated Claims Processing, Automated Underwriting & Fraud Monitoring in Insurance. It is currently in closed beta.

About Arya.ai:

Arya.ai is one of the first deep learning startups from India to deploy AI & Deep Learning for mission-critical functions in Banks, Insurers and Financial Services. Arya.ai is solving the 'AI' acceptability challenges by offering a 'Vertical AI cloud' for the BFSI industry and providing plug-and-use or easy-to-customize product layers specific to the BFSI industry. This allows BFSIs to expedite the roll-out of 'AI' safely and responsibly. Key products offered are - AryaXAI - ML Observability tool for mission-critical AI & Arya APIs - plug and use pre-trained models and solutions for BFSIs.

Arya.ai has been working with partners like Microsoft Azure, Nvidia, Intel etc., to collaborate and contribute to the responsible 'AI' ecosystem. Arya.ai has also received multiple accolades, mentioned from different technical and industry forums. It is named as one the top 61 AI startups globally by CB Insights, Arya.ai founders are listed in Forbes 30 under 30, received an AI innovator award from Nasscom etc. Arya.ai is mentioned in multiple research reports globally by EY, BCG, Gartner, Apis etc.

Looking to start similar initiatives in your organization?
Schedule a call with the AryaXAI team today.

BOOK NOW