Unit-II Transparent AI

Course Code: AID354

Course: Dependable AI

TY – (AIDS) -Auto

Unit-II Transparent AI

Syllabus

- Defining the nature of transparency,
- The limits of transparency,
- Trust from transparency,
- Secure and Safe AI

AI transparency

- Transparency in AI refers to the ability to peer into the workings of an AI model and understand how it reaches its decisions.
- There are many facets of AI transparency, including the set of tools and practices used to understand the model,
- the data it is trained on, the process of categorizing the types and frequency of errors and biases, and the ways of communicating these issues to developers and users.
- The multiple facets of AI transparency have come to the forefront as machine learning models have evolved. A big concern is that more powerful or efficient models are harder -- if not impossible -- to understand since the inner workings are buried in a so-called black box.
- Basically, humans find it hard to trust a black box -- and understandably so," said Donncha Carroll, partner and chief data scientist at business transformation advisory firm Lotis Blue Consulting. "AI has a spotty record on delivering unbiased decisions or outputs."

What is transparency in AI?

- AI transparency is about clearly explaining the reasoning behind the output, making the decision-making process accessible and comprehensible,"
- said Adnan Masood, chief AI architect at UST, a digital transformation consultancy. "At the end of the day, it's about eliminating the black box mystery of AI and providing insight into the how and why of AI decision-making.

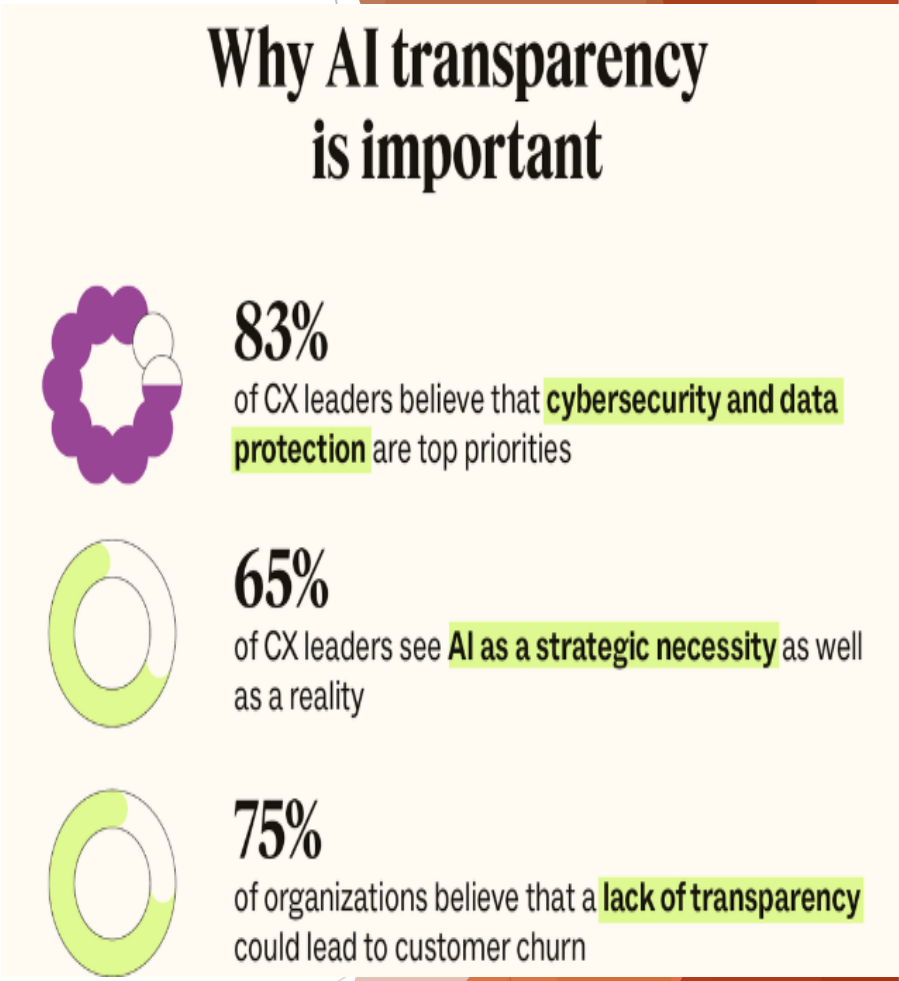
- For the financial services industry, transparency in AI is the ability for lenders to determine how and why an algorithm arrived at its decision about a loan approval or denial.
- While many organizations claim to advocate for the fair and transparent use of AI, their commitment is rarely backed up by real action or business strategies supporting the creation of transparent AI. This leads to a whole host of problems. Without transparent AI, lenders can accidentally sustain discriminatory lending practices from systemic bias, face issues of fairness in their decisions, and build general mistrust from the public — all of which have received increased attention lately when it comes to AI.
- Some algorithmic models do get a bad rap for being black boxes that prone to unfair bias and lack of transparency. However, if you have the right tools to explain your AI's decisions, your institution actually prevents itself from being a black box decision-maker.

Need of AI transparency

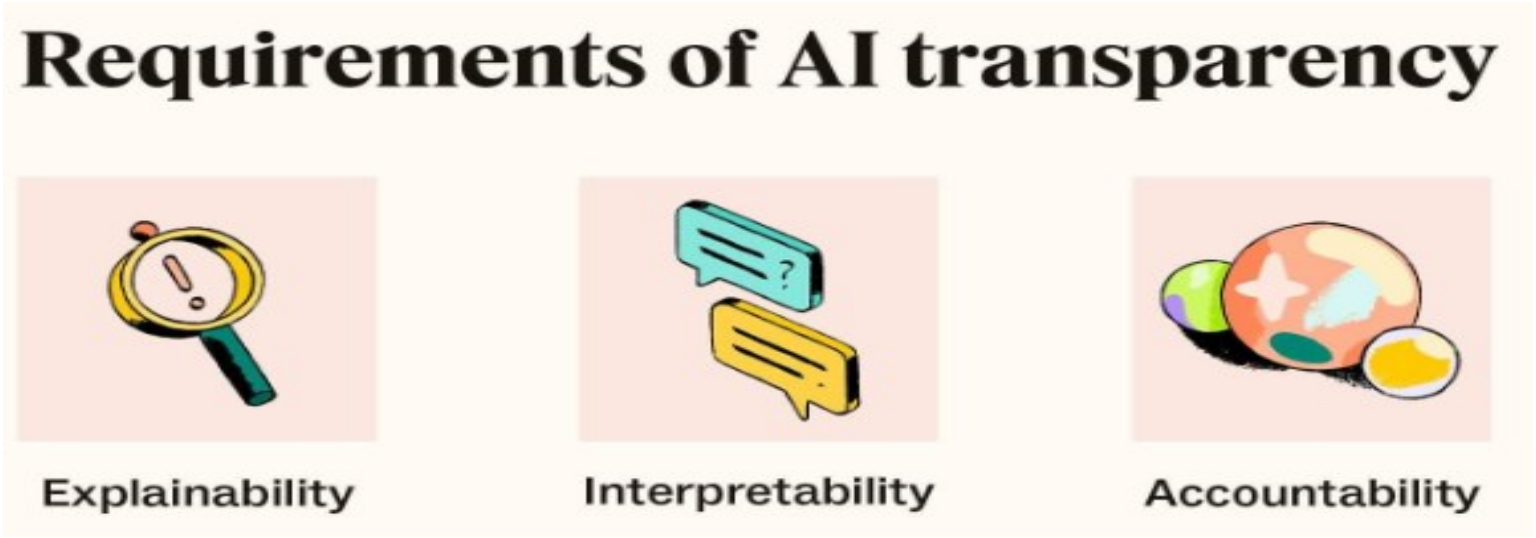
- ✓ Builds trust with customers and employees.
- ✓ Ensures fair and ethical AI systems.
- ✓ Detects and addresses potential data biases.
- ✓ Enhances the accuracy and performance of AI systems.
- ✓ Ensures compliance with new AI regulations like the EU AI Act.
- ✓ To scrub systemic bias from the algorithms.
- ✓ Algorithms must be made explainable and transparent not like a black box.
- ✓ to reduce risk
- ✓ Document and validate your AI model.
- ✓ Enable user control and customization.
- ✓ Provide feedback and error handling.

Why is AI transparency important?

- AI transparency involves understanding its ethical, legal, and societal implications
- According to our CX Trends Report, 75 percent of businesses believe that a lack of transparency could lead to increased customer churn in the future.
- Because AI as a service (AIaaS) providers make AI technology more accessible to businesses, ensuring AI transparency is more important than ever.
- The ethical implications of AI means making sure AI behaves fairly and responsibly. For example, using AI in the workplace can help with the hiring process, but it may inadvertently favor certain groups over others based on irrelevant factors like gender or race.
- The legal Implications of AI involve ensuring that AI systems follow the rules and laws set by governments. For instance, if an AI-powered software collects personal information without proper consent, it can violate privacy laws.
- The societal implications of AI entail understanding how AI affects the daily lives of individuals and society as a whole. For example, using AI in healthcare can help doctors make accurate diagnoses faster or suggest personalized treatments.



AI transparency requirements



Explainability

- Explainable AI (XAI) refers to the ability of an AI system to provide easy-to-understand explanations for its decisions and actions. For example, if a customer asks a chatbot for product recommendations, an explainable AI system could provide details such as:
- "We think you'd like this product based on your purchase history and preferences."
- "We're recommending this product based on your positive reviews for similar items."
- Offering clear explanations gives the customer an understanding of the AI's decision-making process. This builds customer trust because consumers understand what's behind the AI's responses. This concept can also be referred to as responsible AI, trustworthy AI, or glass box systems.
- On the flip side, there are black box systems. These AI models are complex and provide results without clearly explaining how they achieved them. This lack of transparency makes it difficult or impossible for users to understand the AI's decision-making processes, leading to a lack of trust in the information provided.

Interpretability

- Interpretability in AI focuses on human understanding of how an AI model operates and behaves. While XAI focuses on providing clear explanations about the results, interpretability focuses on internal processes (like the relationships between inputs and outputs) to understand the system's predictions or decisions.

- Let's use the same scenario from above where a customer asks a chatbot for product suggestions. An interpretable AI system could explain that it uses a decision tree model to decide on a recommendation.

Accountability

- Accountability in AI means ensuring AI systems are held responsible for their actions and decisions. With machine learning (ML), AI should learn from its mistakes and improve over time, while businesses should take suitable corrective actions to prevent similar errors in the future.

- Say an AI chatbot mistakenly recommends an item that's out of stock. The customer attempts to purchase the product because they believe it's available, but they are later informed that the item is temporarily out of stock, leading to frustration. The company apologizes and implements human oversight to review and validate critical product-related information before bots can communicate it to customers.

- This example of accountability in AI for customer service shows how the company took responsibility for the error, outlined steps to correct it, and implemented preventative measures. Businesses should also perform regular audits of AI systems to identify and eliminate biases, ensure fair and nondiscriminatory outcomes, and foster transparency in AI.

The black box problem of AI

- The black box problem was acceptable to some degree in the early days of AI technology,
- but lost its merit when algorithmic bias was spotted.
- For example, AI that was developed to sort resumes disqualified people for certain jobs based on their race,
- and AI used in banking disqualified loan applicants based on their gender.
- The data the AI was trained on was not balanced to include sufficient data of all kinds of people, and the historical bias that lived in the human decisions was passed to the models
- While these are some of the more extreme cases of misclassification -- and include some purposely designed adversarial inputs to fool the AI model -- they still underline the fact that the algorithm has no clue or understanding of what it is doing
- For the same reason, unusual alterations in the pattern make the model vulnerable, and that's why AI transparency is needed
- When using AI for critical decisions, understanding the algorithm's reasoning is commanding. An AI model designed to detect cancer, even if it is only 1% wrong, could threaten a life.

Levels of AI transparency

- There are three levels of AI transparency, starting from within the AI system, then moving to the user, and finishing with a global impact. The levels are as follows:
- Algorithmic transparency
- Interaction transparency
- Social transparency

- Algorithmic transparency focuses on explaining the logic, processes, and algorithms used by AI systems. It provides insights into the types of AI algorithms, like machine learning models, decision trees (flowchart-like models), neural networks (computational models), and more. It also details how systems process data, how they reach decisions, and any factors that influence those decisions. This level of transparency makes the internal workings of AI models more understandable to users and stakeholders.
- Interaction transparency deals with the communication and interactions between users and AI systems. It involves making exchanges more transparent and understandable. Businesses can achieve this by creating interfaces that communicate how the AI system operates and what users can expect from their interactions.
- Social transparency extends beyond the technical aspects and focuses on the broader impact of AI systems on society as a whole. This level of transparency addresses the ethical and societal implications of AI deployment, including potential biases, fairness, and privacy concerns.

Regulations and standards of transparency in AI

- Here are a few key regulations and standards to help govern artificial intelligence:

- General Data Protection Regulation (GDPR): established by the European Union (EU) and includes provisions surrounding data protection, privacy, consent, and transparency
- Organisation for Economic Co-operation and Development (OECD) AI Principles: a set of value-based principles that promotes the trustworthy, transparent, explainable, accountable, and secure use of AI
- U.S. Government Accountability Office (GAO) AI accountability framework: a framework that outlines responsibilities and liabilities in AI systems, ensuring accountability and transparency for AI-generated results
- EU Artificial Intelligence Act: an act proposed by the European Commission that aims to regulate the development of AI systems in the EU
- These regulations can standardize the use and development of AI, locally and globally. AI systems can be consistently more clear and trustworthy by emphasizing transparency, ethical considerations, and accountability.

The benefits of AI transparency

- Transparent AI offers many benefits for businesses across ethical, operational, and societal realms. Here are a few advantages of transparency in AI:
- Builds trust with users, customers, and stakeholders: Users, customers, and stakeholders are more likely to engage with AI technologies or businesses that utilize an AI help desk when they understand how these systems function and trust that they operate fairly and ethically.
- Promotes accountability and responsible use of AI: Clear documentation and explanations of AI processes make the responsible use of AI easier and hold businesses accountable in case of errors or biases.
- Detects and mitigates data biases and discrimination: Visibility into the data sources and algorithms allows developers and data scientists to identify biases and discriminatory patterns. This allows businesses to take proactive steps to eliminate biases and ensure fair, equitable outcomes.
- Improves AI performance: Developers who clearly understand how models operate can fine-tune algorithms and processes more effectively. Feedback collected from users and insights from performance data allow for continuous improvements to enhance the accuracy and efficiency of AI systems over time, especially with AI for the employee experience.
- Addresses ethical issues and concerns: Transparency in AI enables stakeholders to evaluate the ethical implications of AI-powered decisions and actions and ensure that AI systems operate within ethical guidelines.
- Embracing transparency in AI not only enhances the reliability of AI systems but also contributes to responsible and ethical usage.

Challenges of transparency in AI (and ways to address them)

1. Keeping data secure

- Ensuring customer data privacy while maintaining transparency can be a balancing act. Transparency may require sharing details about the data used in AI software, raising concerns about data privacy. According to our CX Trends Report, 83 percent of CX leaders say data protection and cybersecurity are top priorities in their customer service strategies.

- **How to handle this challenge:**
- Appoint at least one person on the team whose primary responsibility is data protection. Brandon Tidd, the lead Zendesk architect at 729 Solutions, says that “CX leaders must critically think about their entry and exit points and actively workshop scenarios wherein a bad actor may attempt to compromise your systems.”

2. Explaining complex AI models

- Some AI models, especially those utilizing deep learning or neural networks, can be challenging to explain in simple terms. This makes it difficult for users to grasp complex AI models' decision-making and intelligent automation processes.
- How to handle this challenge:
  Develop visuals or simplified diagrams to illustrate how complex AI models function. Choose an AI-powered software with a user-friendly interface that provides easy-to-follow explanations without the technical stuff.

3. Maintaining transparency with evolving AI models

- As AI models change and adapt over time, maintaining transparency becomes increasingly more difficult. Making updates or modifications to AI systems or retraining them on new datasets can alter their decision-making processes, which can make it challenging to maintain transparency consistently.
- How to handle this challenge:
  Establish a comprehensive documentation process that tracks the changes made to an AI ecosystem, like its algorithms and data. Provide regular and updated transparency reports that note these changes in the AI system so stakeholders are informed about these updates and any implications.

AI transparency best practices

- **Be clear with customers about how their data is collected, stored, and used**
- Provide transparent and understandable explanations to customers about the collection, storage, and utilization of their data by AI systems. Clearly outline privacy policies detailing the type of data collected, the purpose of collection, storage methods, and data usage in AI systems. Protecting customer privacy starts with obtaining explicit consent from users before collecting or using their data for AI purposes.

- **Detail how you're preventing inherent biases**
- Conduct regular assessments to identify and eliminate biases within your AI software. Communicate the methods used to prevent and address biases in AI models so users understand the steps being taken to enhance fairness and prevent discrimination. Maintain records of bias detection, evaluation, and processes to show a commitment to customer transparency and bias prevention.

- **Explain what data is included and excluded in AI models**
- Clearly define and communicate the types of data included and excluded from AI models. Provide reasoning behind the selection of data used in AI training to help users understand the model's limitations and capabilities. Avoid including sensitive or discriminatory data that could result in biases or infringe on privacy rights.

Examples of companies practicing transparent AI

- **Zendesk**
- At Zendesk, we create customer experience software that enables users to enhance their customer support with AI and machine learning tools, like generative AI and AI chatbots. Zendesk AI emphasizes explainability by providing insights into how its AI-powered tools work and how AI decisions are made.
- Zendesk also offers educational resources and documentation to help users understand AI's integration into customer service software, the ethics of AI in CX, and its impact on customer interactions.
- **Lush**
- Cosmetic retailer Lush is vocal about ethical AI usage in its business operations. The company is transparent about not using social scoring systems or technologies that could infringe on customer privacy or autonomy. Lush engages in public discussions and shares its stance on ethical AI practices through its communications and social media channels.
- **OpenAI**
- OpenAI, an AI research laboratory popular for its generative AI applications ChatGPT and Dall-E, regularly publishes research papers and findings that provide insights into its AI developments and breakthroughs.
- OpenAI is transparent about its goals, ethical guidelines, and the potential societal impact of AI through comprehensive documentation. The company encourages collaboration and engagement with the wider AI community, fostering transparency and sharing knowledge about AI development.

Weaknesses or limits of AI transparency

- **Vulnerable to hacking. :**Transparent AI models are more susceptible to hacks, as threat actors have more information on their inner workings and can locate vulnerabilities. To mitigate these challenges, developers must build their AI models with security top of mind and test their systems.
- **Can expose proprietary algorithms:** Another concern with AI transparency is protection of proprietary algorithms, as researchers have demonstrated that entire algorithms can be stolen simply by looking at their explanations.
- **Difficult to design:** Lastly, transparent algorithms are harder to design, in particular for complex models with millions of parameters. In cases where transparency in AI is a must, it might be necessary to use less sophisticated algorithms.
- **Governance challenges:** Another fundamental weakness is assuming any transparency method out of the box will satisfy all needs from a governance perspective,. But these models might still rely on users to identify biased and inaccurate information. If an AI chatbot cites a source, it's still up to the human to determine whether the source is valid. This takes time and energy and leaves room for error.
- **Lack of standardized methods to assess transparency.** Another issue is that not all transparency methods are reliable. They may generate different results each time they are performed. This lack of reliability and repeatability might reduce trust in the system and hinder transparency efforts.

How to reach a balance in AI

- When implementing AI, an organization must pay attention to the following four factors:

- **Legal needs.** If the work requires explainability from a legal and regulatory perspective, there may be no choice but to provide transparency. To reach that, an organization might have to resort to simpler but explainable algorithms.
- **Severity.** If the AI is used in life-critical missions, transparency is a must. It is most likely that such tasks are not dependent on AI alone, so having a reasoning mechanism improves teamwork with human operators..
- **Exposure.** Depending on who has access to the AI model, an organization might want to protect the algorithm from unwanted reach. Explainability can be good even in the cybersecurity space if it helps experts reach a better conclusion. But if outsiders can gain access to the same source and understand how the algorithm works, it might be better to go with opaque models.
- **Data set**. No matter the circumstances, an organization must always strive to have a diverse and balanced data set, preferably from as many sources as possible. AI is only as smart as the data it is trained on. By cleaning the training data, removing noise and balancing the inputs, we can help to reduce bias and improve the model's accuracy.

- **Why is AI transparency Hard**: Transparent AI can be challenging because of the complexity of AI models, the explainability of datasets, and trade-offs between explainability and AI performance. Some AI systems operate as black boxes and can rely on vast, complex datasets. This lack of transparency makes it difficult to trace how the data influences outcomes and how to explain AI models' decision-making processes in ways that are easy to understand.
- **What does lack of Transparency in AI Mean?:** Lack of transparency in AI refers to situations where users don't have the necessary visibility or understanding of how AI systems work, how they make decisions, or any influencing factors. This can result in people being untrustworthy of AI systems and how they use customer data.
- **Why is transparency a concern in AI?**
- lack of transparency in AI is concerning for several reasons, including:

- **Ethical or biased implications:** Lack of transparency can lead to unethical or biased outcomes, impacting fairness and potentially causing harm or discrimination.
- **Accountability issues:** Without transparency, assigning responsibility or accountability for AI-generated decisions becomes challenging.
- **User trust and understanding:** Lack of transparency diminishes user trust in AI systems, leading to skepticism and reluctance to adopt or rely on AI technologies.
- **Regulatory compliance:** Increasingly, regulations require transparency in AI systems to ensure compliance with ethical, legal, and privacy standards.