



TECHNOLOGY BRIEFING ON TRANSPARENCY, ACCOUNTABILITY, & EXPLAINABILITY IN AI

AI, especially deep learning, has shown remarkable capabilities in solving complex problems, but its 'black-box' nature has raised questions about its transparency and accountability. As AI systems increasingly influence various sectors from healthcare to finance, *there's a pressing need to ensure these systems are transparent, accountable, and explainable. This demand is not just a theoretical concern but has practical implications in ensuring fair, unbiased, and trustworthy AI.*

Accountability: The *responsibility AI developers and users have regarding the outcomes produced by AI systems.* As AI influences critical decisions in healthcare, finance, and more, ensuring that these systems are trustworthy and that there's a mechanism to address errors or biases becomes paramount.

Transparency: It's about *understanding how an AI model works architecture and design-wise.* Many AI models are seen as "black boxes" due to their complexity, raising concerns about their transparency.

Explainability: Closely tied to transparency, explainability involves detailing in understandable terms *how AI models reach their conclusions.* With the rise of AI, there's been a push for "Explainable AI" or XAI, which aims to make even the most intricate models understandable to humans.

Introduction to AI Algorithms

Basic AI Mechanism

- **Data Input**
 - *Source:* Data is the foundation of AI. It comes from diverse sources, such as user inputs, sensors, databases, and online interactions.
 - *Data Preprocessing:* Before use, raw data is often cleaned and transformed to ensure it's suitable for analysis. Missing values are addressed, and outliers are often removed or adjusted.
- **Processing**
 - *Feature Extraction:* Not all data is useful. Algorithms will pinpoint the most relevant parts of or trends in the data, called "features", to analyze.
 - *Algorithm Selection:* Depending on the task (e.g., classification—trying to identify something, prediction, clustering—grouping similar things together, generation—creating something new based on existing samples), a specific algorithm or combination of algorithms is chosen.
 - *Model Training:* With training data, the model adjusts its internal parameters to best predict the desired outcome. Think of this as the algorithm "learning" from examples.
- **Output**
 - *Prediction/Decision Making:* The trained model, when presented with new data, will make predictions or decisions based on patterns it learned during training.
 - *Validation and Testing:* It's crucial to test the model's predictions against a set of data it hasn't seen before to measure its accuracy and robustness.

Enshrined within AI is *machine learning (ML)*, which intuitively, is the *process of training an algorithm on reference data with features to optimize for so that it can observe trends and form or create*

conclusions without the need to be very explicitly programmed. *Deep learning*, a more intricate form of machine learning, has emerged as the flagship method to solve difficult problems, as *these algorithms identify and learn features from data on their own*.

Deep Learning and Neural Networks

- **Neural Networks**

- *Inspiration*: These algorithms are inspired by the human brain. Just as our brain consists of interconnected neurons, a neural network has "nodes" interconnected in layers.
- *Layers*: There are typically three types of layers within these algorithmic networks:
 - **Input Layer**: Represents the features of the data.
 - **Hidden Layers**: Multiple layers where the processing happens. Data is transformed as it moves through these layers. The more layers, the more complexity and compute power (generally).
 - **Output Layer**: Produces the final prediction or classification.

- **Deep Learning**

- *Definition*: When neural networks have a large number of hidden layers, they're termed "deep" networks. This depth allows for more intricate pattern recognition.
- *Backpropagation*: A crucial technique in deep learning. As the network makes predictions, errors are calculated and sent back through the network to adjust the parameters. This iterative process refines the model's accuracy.

- **Applications:**

- *Image Recognition*: Deep learning models can identify objects in images, sometimes surpassing human accuracy. Platforms like Google Photos employ this for image categorization.
- *Speech Recognition*: Tools like Apple's Siri or Amazon's Alexa use deep learning to understand and process spoken language.
- *Natural Language Processing*: Chatbots, translation tools, and even content generation platforms (like OpenAI's GPT series) use deep learning to understand and generate human-like text.

...and much more, which is detailed further into the memo. Broadly speaking, these models are applicable in any generative, predictive, or instructional task. However, they do have **limitations**.

- *Complexity*: Deep learning models, especially when they have millions of parameters, become "black boxes". It's challenging to pinpoint exactly how they make decisions.
- *Data Dependency*: These models require vast amounts of data to train effectively. The quality and diversity of this data significantly influence outcomes.
- *Computational Intensity*: Deep learning demands powerful computational resources, often specialized hardware like GPUs (processing systems developed by big tech companies like NVIDIA).

AI Model Evolution Over Time

The de-facto AI models used continuously developed and changed over time in response to issues and malfunctions that warranted an increase in accountability, explainability, and transparency. *This ethical consideration is not new and will persist as the technology is deployed.*

Late 1990s-2000s: Emphasis on Transparency

- **Ensemble Models**

- *Basics*: Ensemble models are a blend of multiple machine learning models that come together to make predictions. This synthesis aims to produce better results than any single model could on its own.

- *How They Work*: These models operate by pooling together the predictions of each individual model (like decision trees or neural networks) and making a final decision based on aggregated results. This can be viewed as seeking the "wisdom of the crowds."
- *Advantages*:
 - Balanced Biases: Because they integrate various models, ensemble models can neutralize individual model biases.
 - Improved Accuracy: By combining models, ensemble techniques often enhance predictive accuracy.
- *Challenges*:
 - Overfitting: These models can become too attuned to their training data, hindering their generalization to new data.
 - Complexity & Interpretability: Aggregating results from multiple models can make it harder to understand how the final decision was reached.
- **Decision Trees**
 - *Basics*: A graphical representation of decisions and their potential consequences. It's like navigating a series of "if-then" conditions to reach a conclusion.
 - *Transparency*: Decision Trees are easily visualized and understood, making them a preferred choice for applications where interpretability is crucial.
 - *Challenges*:
 - Data Sensitivity: Small changes in the data can result in very different trees.
 - Complexity Limits: They might not capture intricate relationships in highly complex datasets.

2010s: Deep Learning Dominance

- **ImageNet Competition**
 - *Background*: A pivotal competition in the machine learning community. It challenged algorithms to classify images into thousands of categories.
 - *Significance*: The competition saw deep learning models, specifically convolutional neural networks, significantly outperform traditional techniques, marking deep learning's ascendancy in computer vision.
- **Healthcare & Protein Folding**
 - *Basics*: Predicting how proteins fold is a monumental task in biology. The structures proteins take on influence how they function and are crucial for understanding diseases and creating medicines.
 - *DeepMind's AlphaFold*: A deep learning system that predicts protein structures with remarkable accuracy. This was a significant leap, potentially transforming drug discovery and disease understanding. (this was further developed in the 2020s and said to be a viable solution to the age-old biology problem).
 - *Relevance to Transparency*: While the results are groundbreaking, deep learning models like AlphaFold are intricate, making it difficult to grasp how they arrive at their conclusions. This brings in the challenge of explainability in AI.

The Alignment Problem

The "*Alignment Problem*" in AI refers to the *challenge of ensuring that AI systems, especially as they grow more capable and autonomous, behave in ways that are consistent with human values and intentions*. Properly aligned AI would not just do what it's told but would also act in the best interests of humanity. A misaligned AI, on the other hand, might optimize for unintended and potentially harmful outcomes. Aligning AI with our values becomes paramount as it intertwines with concepts of accountability (who's responsible when AI goes awry?), transparency (can we see how AI makes its

decisions?), and explainability (can we understand those decisions?). Key areas within algorithmically-fueled injustices are highlighted below.

- **Facial Recognition Technology (FRT)**
 - *Basics*: FRT identifies or verifies a person based on their facial features.
 - *Issues*:
 - Misidentification: FRT has been shown to have accuracy gaps, especially misidentifying people of color, women, and age extremes, leading to civil liberties concerns.
 - Civil Liberties: Erroneous identifications can lead to wrongful arrests, as seen in cases like Ousmane Bah suing Apple and Robert Williams' wrongful arrest in Detroit.
- **Healthcare Algorithms**
 - *Basics*: Tools assisting in diagnosis, prognosis, and treatment.
 - *Challenges*:
 - Racial Biases: Algorithms can perpetuate racial biases present in training data. A notable example is the racial bias in a widely-used health risk prediction algorithm that favored white patients over Black patients with similar health needs.
 - Data Collection: Biases often originate from the data collection stage, with datasets not representing the broader population, leading to biased outcomes.
- **Hiring Algorithms**
 - *Basics*: Tools for automating recruitment processes.
 - *Issues*:
 - Gender & Racial Biases: Algorithms can inadvertently favor or disfavor certain groups. Amazon, for instance, had to scrap a resume screening tool that was biased against female candidates.
 - Opacity & Discrimination: Platforms like HireVue, which assess candidates via video interviews, have been critiqued for potential biases and a lack of transparency in their assessments.
- **Criminal Justice Algorithms**
 - *Basics*: Tools aiding decisions in bail, sentencing, and parole.
 - *Challenges*:
 - Racial & Socioeconomic Biases: Algorithms like COMPAS have been criticized for biases against Black defendants and for potentially exacerbating existing disparities in the criminal justice system.
- **Education Algorithms**
 - *Basics*: Tools for personalized learning and performance assessment.
 - *Issues*:
 - Quality & Equity: Algorithms can introduce biases that impact the quality of education. For instance, an algorithm used for grading standardized tests favored verbose language over clarity.
- **Social Media Algorithms:**
 - *Basics*: Algorithms that curate and recommend content based on user behavior.
 - *Challenges*:
 - Bias & Extremism: Platforms like YouTube have been critiqued for promoting extremist content, while algorithms on Twitter and Instagram have faced backlash for racial and body image biases.

Technologies & Methodologies

Below are methodologies and technologies that aim to shed light on the opaque nature of advanced AI models in this three-fold context.

Techniques

- **LIME (Local Interpretable Model-agnostic Explanations)**
 - *Basics:* LIME is a technique designed to explain the predictions of any machine learning classifier in a way humans can understand.
 - *How It Works:* It perturbs the data input and observes the changes in predictions. By doing so, it approximates the complex model with a simpler, local model for individual predictions, offering insights into how features influence outcomes.
 - *Importance:* Provides insights on a per-instance basis, highlighting how different features played roles in specific decisions made by the algorithm.
- **SHAP (SHapley Additive exPlanations)**
 - *Basics:* SHAP values are a unified measure of feature importance or weight on an algorithmic decision.
 - *How It Works:* It distributes the "credit" for a prediction among its features based on their contributions. It is built on the concept of Shapley values from cooperative game theory, ensuring a fair distribution of contributions.
 - *Relevance:* SHAP provides a consistent way to interpret model outputs, ensuring that the interpretation is the same regardless of the context.
- **Attention Mechanisms in Deep Learning**
 - *Basics:* Attention mechanisms enable models, especially in tasks like language processing, to focus on specific parts of the input data.
 - *How It Works:* Much like human attention, this mechanism allows the model to weigh different parts of the input differently, giving more "attention" to the most relevant parts when making a prediction.
 - *Significance:* Provides insights into what the model deems important in the data, allowing for a level of interpretability in neural network decisions.

Accountability Measures

Here's how government, corporations, academia, and civil society can establish guardrails to assign trust and accountability to potentially adversarial outcomes of automated technologies (and prevent them from occurring in the first place).

- **Audits:**
 - *Basics:* Regular examinations of AI systems to ensure they align with set standards.
 - *Purpose:* Audits ensure that AI systems remain transparent, ethical, and operate as intended. They can help in uncovering biases, vulnerabilities, or other issues.
 - *Relevance:* Regular audits can build trust with the public, ensuring that AI systems remain under scrutiny and adhere to evolving standards.
- **Red Teaming:**
 - *Basics:* An adversarial approach where independent teams challenge an AI system.
 - *How It Works:* These teams simulate real-world attacks or adversarial inputs to test the system's robustness and identify vulnerabilities.
 - *Significance:* Red teaming offers a proactive approach to uncovering potential flaws before they become actual issues, ensuring that AI systems are robust against adversarial attacks.

- **Ethical Guidelines & Frameworks:**

- *Basics:* Sets of principles guiding the responsible development and deployment of AI.
- *Examples:* The Montreal Declaration for Responsible AI, which outlines principles such as well-being, autonomy, and justice, setting a moral compass for AI development.
- *Relevance:* Such frameworks provide AI developers and stakeholders with a roadmap, ensuring that AI technologies are developed in a manner consistent with societal values and norms.

Transparency, accountability, and explainability are fundamental pillars in the responsible development and deployment of AI systems. As AI becomes more integrated into various sectors, understanding its decision-making processes becomes vital to ensure fairness, mitigate biases, and foster public trust. AI developers and stakeholders must prioritize these principles, ensuring systems are transparent in their operations, accountable for their actions, and can be explained in understandable terms. By doing so, society can harness the benefits of AI while safeguarding against unintended consequences and ethical pitfalls.