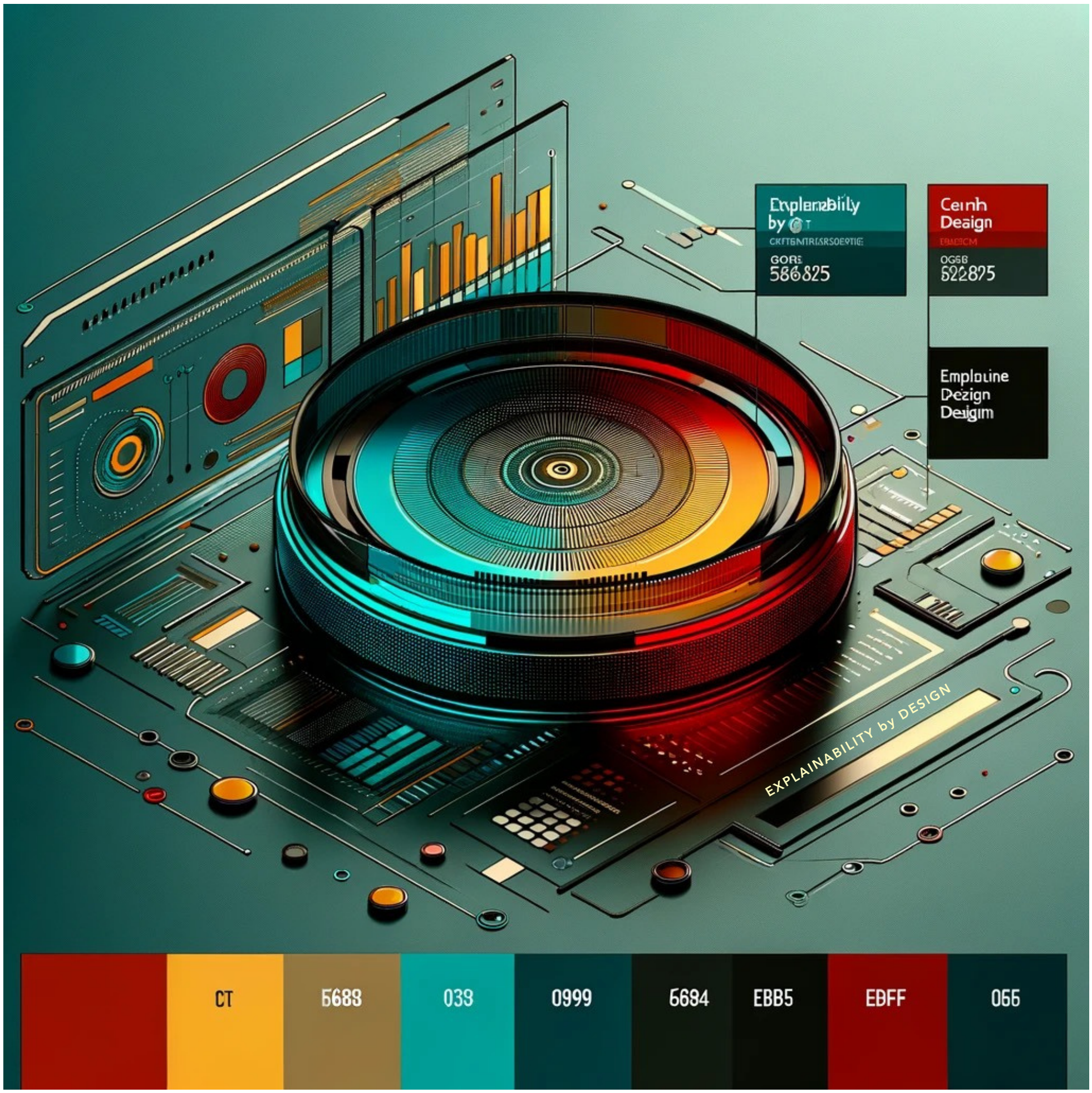
EXPLAINABILITY BY DESIGN:

Minimum Viable Explainability



Hello : Dall-E (The AI model that created this image.)

Shoshana Rosenberg

EXPLAINABILITY BY DESIGN

Minimum Viable Explainability

What does **explainability** mean, from a functional standpoint, knowing that not every aspect of a complex AI system and the underlying processes can be surfaced and truly explained?

I have set out a proposed set of MVE requirements over these few pages, but it is important to note that the concept of AI explainability is directly tied to understanding, and therefore to accessibility.

Akin to drawing a map of a country or a major city, the level of detail and must be shifted to accommodate and make accessible what the individual needs to understand and navigate to make informed decisions. (Imagine a tourist map of NYC or Paris or Dubai compared with a map intended to analyze traffic flows or one depicting elevation levels that would help anticipate flooding.)

AI explanations must vary in depth and complexity based on the needs of the audience or stakeholder, and explainable-by-design AI must be built to allow for more granular explanations that can roll up into more succinct overview.

A key component of EbD has to be the layering of explanations.

With the current glaring issues and the regulations to come that will take aim to address them, not only will **Explainability by Design** need to be part of built AI technologies going forward, but we will also need an arsenal of tools and applications that can read the granular and make it accessible on multiple levels for different groups and myriad audit, business, and regulatory use cases.

STATEMENTS

Shoshana Rosenberg

PROPOSED BASELINE REQUIREMENTS FOR MINIMUM VIABLE EXPLAINABILITY →

EXPLAINABILITY BY DESIGN

Minimum Viable Explainability Baseline Requirements

System Design and Structure

BLUEPRINTS TELL THE STORY

STATEMENTS

Shoshana Rosenberg

1. Model Architecture Overview

Understanding the basic structure of an AI model is foundational for comprehending and explaining how it processes information and makes decisions.

Understanding the model's architecture can help identify potential biases or privacy risks inherent in the design. (For instance, certain model types may inadvertently amplify biases or be more susceptible to leaking personal data.)

2. Feature Importance

Knowing which inputs significantly impact the model's decisions (and what is given priority or more importance by design or default) helps in understanding what the AI is focusing on and whether these factors are appropriate and fair.

Identifying key features and weighting allows for the assessment of whether the model disproportionately relies on features that could lead to biased outcomes against certain groups.

3. Performance Metrics

Clear metrics for evaluating the AI's performance are essential for gauging its effectiveness and reliability, ensuring it meets the desired standards.

By including fairness and equity metrics, you can evaluate whether the AI has the needed checks and balances to treat all groups fairly and equitably.

EXPLAINABILITY BY DESIGN

Minimum Viable Explainability Baseline Requirements

Data and Training

1. Training Data Overview

The quality, sourcing, and characteristics of the data used to train AI models directly influence their behavior and decisions, which makes transparency about this data crucial to explainability. Understanding the data used to train the model allows the assessment of whether it is representative and diverse and to evaluate whether the model may be predisposed to neglect or misrepresent certain groups. Knowing the data sources and types helps in identifying potential privacy concerns, such as the inclusion of sensitive or personally identifiable information.

2. Fairness and Bias Analysis

Identifying and mitigating biases in AI systems is key to ensuring fair and ethical outcomes, particularly where it could be used in decisions that impact people's lives.

3. Versioning of Models

Keeping track of different versions of AI models is important to allow for understanding how changes affect outcomes, and will help ensure accountability over time. Keeping track of model versions also allows for monitoring changes that might impact DEI or privacy. This historical record helps in understanding and rectifying any emerging issues.

Shoshana Rosenberg

EXPLAINABILITY BY DESIGN

Minimum Viable Explainability Baseline Requirements

Explanations and Interpretability

READING BETWEEN THE LINES.

1. Decision Rationale

Providing reasons for AI decisions is crucial for transparency and trust, especially in scenarios where these decisions have significant consequences.

This also helps ensure that these decisions are fair and unbiased, promoting transparency in how different groups are treated and to highlight if personal data is being used or factored in.

2. Uncertainty Estimation

Knowing the confidence level of AI predictions helps users understand how much they can rely on these decisions.

Understanding a model's certainty in its decisions can be crucial in scenarios where uncertainty might affect minority or traditionally underrepresented groups more severely.

3.User Feedback Mechanism

Allowing users to give feedback on AI decisions is an important and requisite component that helps in refining and improving the system, making it more effective.

Diverse user groups need to be able to voice concerns about potential biases or unfair treatment, fostering inclusive improvement of the AI system.

STATEMENTS

Shoshana Rosenberg

EXPLAINABILITY BY DESIGN

Minimum Viable Explainability
Baseline Requirements

Ethical Standards and Regulatory Adherence

1. Compliance Documentation

Ensuring and evidencing that AI systems adhere to laws and regulations is essential for legal and ethical operations, protecting both users and developers.

2. Model Update and Monitoring Plan

Regularly updating and continuous monitoring of AI systems is key to ensuring they remain effective, fair, and relevant as the world and data change.

3. Ethics and Fairness Controls

Building in measures to address and cure the risks of harm and ethical implications of AI use is crucial for maintaining public trust and avoiding such harms, ensuring the technology benefits society.

This is key to preventing the AI from perpetuating historical injustices or biases and adds a needed layer of re-evaluation of privacy and personal data considerations.

MEASURE TWICE, CUT ONCE.

STATEMENTS

Shoshana Rosenberg

A.I., **Privacy** and **DEI** have a high level of interdependence and interconnectedness and are continuously evolving.

All three are tied directly to ethics, fundamental human rights, the future of work, and decision making and bias, which means:

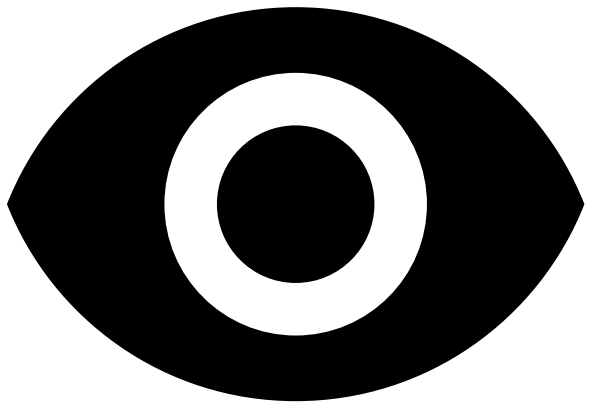
YOU AREN'T ON THE SIDELINES OF THESE THINGS.

YOU ARE CRUCIAL TO THEM BEING WHAT THEY SHOULD.

Do you want to know more?

STATEMENTS

Shoshana Rosenberg



A.I. DE.TEMENTS

STATE OF PRIVACY

Shoshana Rosenberg