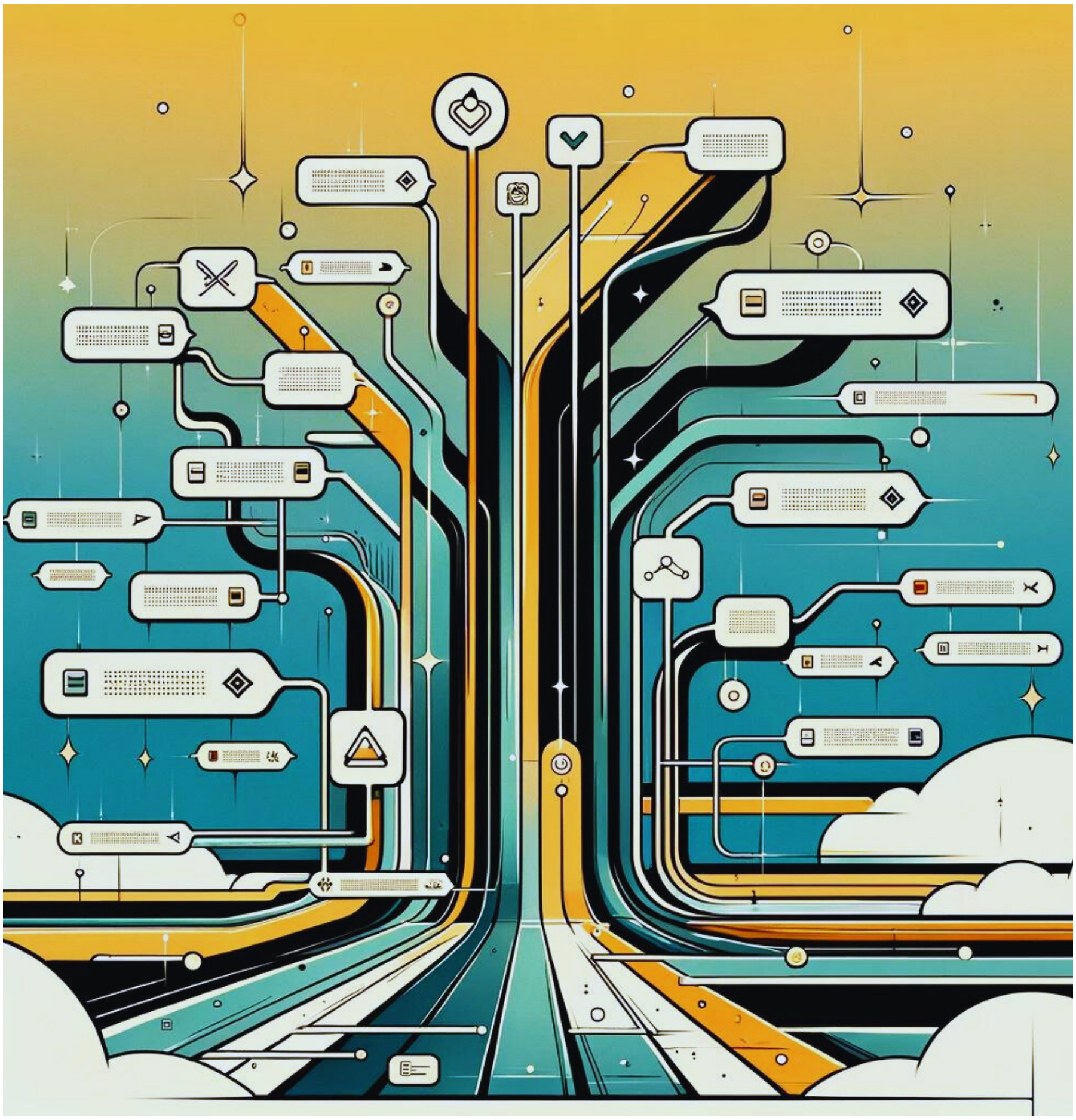
EXPLAINABILITY by DESIGN

THE BEACON INITIATIVE



EXPLAINABILITY AS A SERVICE

WHAT IS AND WHAT SHOULD BE

AN OVERVIEW | A Proposal

Shoshana Rosenberg

A.D. STATEMENTS

PRIVACY STATEMENTS

WHAT IS AND WHAT SHOULD BE

Having spent the last six months focused on understanding and mapping the path to Explainability by Design, I aim to provide you both a comprehensive overview of the current spectrum of technologies aimed at facilitating explainability and interpretability in AI systems, and a proposal of what I deem to be the missing piece.

A number of fascinating explainability tools and methodologies exist at present, each offering unique capabilities and applications that span various stages of the AI lifecycle. These fall into several broad categories, with **Post-hoc (Retroactive) Interpretability Tools** at the heart (and the start) of the evolving toolkit. These tools, such as LIME (*Local Interpretable Model-agnostic Explanations*) and SHAP (*SHapley Additive exPlanations*), analyze AI decisions after a model has been built, providing insights into how different features influence model predictions. They are key for helping to dissect complex models post-deployment, allowing developers and data scientists to gain a deeper understanding of model behaviors and to communicate these insights effectively to stakeholders.

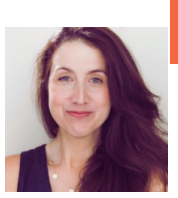
Hybrid Interpretability Solutions merge the capabilities of retroactive analysis with approaches that incorporate interpretability directly into the model design. Tools and platforms in this category, such as InterpretML and AI Explainability 360, offer a blend of techniques, enabling users to not only explore model decisions after they are made but also to utilize inherently interpretable models in development. This hybrid approach facilitates a more integrated strategy for explainability, supporting efforts to balance model performance with transparency.

For a more hands-on approach to understanding model decisions, **Interactive Model Exploration and Analysis Tools** such as the What-If Tool and the Fairlearn Dashboard provide interactive environments for users to experiment with model inputs and observe the impacts on outputs. These tools can support counterfactual analysis and allow for the exploration of "what if" scenarios, making them invaluable for both the development of new models and the refinement of existing ones. The experimentation they encourage and enable can greatly help identify potential biases or weaknesses in models.

Ensuring that AI models remain understandable and behave as expected in live environments is the domain of **Operational Explainability and Transparency Suites** such as Fiddler. These platforms focus on monitoring models in production, providing continuous insights into their performance, detecting biases, and ensuring that the models' decisions are transparent and accountable. These tools are invaluable for maintaining trust in AI applications.

The Missing Piece: Building Explainable by Design AI via a Foundational Explainability Framework
While the current ecosystem of explainability and interpretability tools offers powerful capabilities for analyzing and understanding AI models, a gap remains in the form of comprehensive solutions aimed at building explainable AI from the ground up. Most existing tools (as set out on the following pages) focus on interpreting models that have already been developed or providing insights into operational models, and the ones that aid in development still have gaps that an interoperable and complementary system can help resolve.

There's a glaring need for F**oundational Explainability Frameworks** (and the **Explainability-as-a-Service** tools to deploy them) that not only facilitate more refined analysis and operational monitoring of existing models, but also embed explainability from the initial stages of development. To my mind, the development of comprehensive frameworks and “EaaS” that support the creation of explainable-by-design AI represents a crucial next step in the evolution of responsible AI and in getting us to Explainability by Design. (*See The Beacon Initiative- Page 5.*)



Shoshana Rosenberg

Latin for "After This"

"This", here, is model development and training

POST-HOC INTERPRETABILITY TOOLS

BUILT FOR USE WITH EXISTING BUILT AI MODELS

The primary purpose of these tools is to increase transparency, aid in debugging, enhance trust, and fulfill regulatory requirements by providing insights into the decision-making process of AI models **after training**, but they can also play a key role during the development of AI models. *(Incorporating post-hoc interpretability tools **during the development phase** is a proactive strategy to build more understandable, fair, and robust AI systems- getting us closer to Explainability by Design.)*

Post-Hoc Interpretablity tools can be applied to a wide range of models, making them versatile in explaining different types of AI systems. They offer the flexibility to analyze and understand models without needing access to the model's internal workings, which is ideal for complex or proprietary models. These tools play a crucial role in the broader context of explainable AI (XAI), offering strategies to make machine learning models more understandable and accountable to various stakeholders, including data scientists, non-technical users, and regulatory bodies.

The integration of these traditionally post-hoc tools into the development process could enhance model understanding, guide model improvement, and ensure ethical and fair AI practices from the outset. By applying interpretability tools to different models during the development phase, developers can compare how various models make decisions based on the same data. This insight can guide the selection of models that not only perform well but also whose decisions are easier to interpret and justify. Further, there is value in using combinations of these tools is to cross-validate explanations and gain a more comprehensive understanding of an in-development model's behavior.

LIME (Local Interpretable Model-agnostic Explanations)

LIME generates explanations for individual predictions by approximating the model locally with an interpretable model, and is particularly effective for **providing interpretable explanations for individual predictions of complex, black-box models** across different types of data (text, tabular, or images). *LIME's local explanation approach isn't built to offer reliable insights into the model's overall behavior.*

SHAP (SHapley Additive exPlanations)

SHAP values explain the output of any model by computing the contribution of each feature to the prediction. It's applied to built models to break down a prediction into contributions from each input feature. Calculating Shapley values is computationally intensive, especially for models with a large number of features and complex interactions, which can make SHAP impractical for real-time explanations or for use with very large datasets without significant computational resources. SHAP assumes that features are independent when calculating contributions, which might not hold true for all datasets. Interdependent or correlated features can affect the accuracy of SHAP values, leading to potentially misleading interpretations.

Alibi Explain

This library includes various explanation techniques like Anchor Explanations and Counterfactual Explanations, which are used to understand model behavior on specific instances after the model has already made decisions. **Alibi Explain is versatile and suitable for a wide range of applications, from debugging models to providing user-friendly explanations.** *The complexity and computational demand of some of Alibi's explanation methods can be a limitation, particularly in real-time applications. Generating explanations, especially complex ones, can be time-consuming and computationally expensive, potentially limiting its use in low-latency environments. (This is true for most explainability tools at present.)*

Skater and ELI5

An open-source Python library designed to enable model interpretation for machine learning models, with a focus on global interpretations for model understanding.

Integrated Gradients

Integrated Gradients attributes the prediction of a differentiated model (deep neural networks and others that can accurately and efficiently calculate gradients with respect to model variables) to its input features in a fine-grained manner, **providing insights into how each feature contributed to each prediction** - making it particularly useful for models where understanding the exact contribution of each input feature is critical. A key limitation of Integrated Gradients is its requirement for differentiable models, which means it cannot be applied to all types of machine learning models and- like others- its computational cost can be significant for models with a large number of features or complex architectures.

Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM's applicability is primarily limited to convolutional neural networks and tasks involving visual data. It highlights the important regions in the input image that influenced the model's decision, applied after the prediction is made.

Feature Importance Techniques

Various models come with built-in methods to assess the importance of features in predictions, and feature importance techniques are best used for gaining a high-level overview of which features are most influential in a model's predictions. While feature importance provides a useful snapshot of which features are influential, it does not offer detailed insights into how or why these features impact the model's predictions. The techniques can sometimes be misleading, especially in the presence of correlated features, as they do not account for interaction effects or the directionality of the influence (positive or negative).

ELI5

A Python library that offers tools to visualize and debug machine learning classifiers and provided simplified explanations of their predictions, supporting a wide range of models and frameworks.

HYBRID INTERPRETABILITY SOLUTIONS

FOR DEVELOPED MODELS AND THOSE IN DEVELOPMENT

These solutions offer a mix of capabilities, including both inherent model interpretability and post-hoc explanation features, and they are designed to support the development of models that are both high-performing and understandable. While they significantly enhance the accessibility and implementation of interpretability in machine learning projects, awareness of their limitations and optimal uses is the key for effective tooling selection and use.

Microsoft's InterpretML

InterpretML is an open-source package by Microsoft that provides a unified framework for machine learning interpretability. It integrates various interpretability techniques, including glass-box models and post-hoc explanation methods, to make the explanation of machine learning models easier and more accessible.

InterpretML is ideal when a project requires the integration of both transparent (glass-box) models and black-box model explanations, providing a comprehensive interpretability approach.

The wide range of functionalities and advanced features may be overwhelming for beginners in machine learning or those new to interpretability concepts and some complex interpretability computations may be resource-intensive when applied to very large datasets or highly complex models.

H2O Driverless AI

H2O Driverless AI Overview: H2O's Driverless AI is an automated machine learning platform that emphasizes ease of use, efficiency, and scalability. It includes automated feature engineering, model validation, and deployment capabilities, along with built-in interpretability features. For organizations looking to automate the entire end-to-end machine learning lifecycle, from data preprocessing to model deployment, without sacrificing interpretability, Driverless AI Overview is ideal. Its automated feature engineering and model tuning capabilities make it suitable for projects where rapid prototyping and validation are critical.

The automation of many steps in the model development process may leave users with less understanding of the intermediate steps, potentially making it harder to fully grasp the nuances of the model's decision-making process and- access to the full range of Driverless AI's capabilities may be cost-prohibitive for individual users or small organizations.

IBM's AI Explainability 360

This is IBM’s diverse toolkit that includes a range of algorithms and techniques for model interpretability and explanation, suitable for both developers creating new models and those seeking to understand existing models better. It covers a broad spectrum of explainability techniques, targeting different stakeholders, including data scientists, model validators, and business users and its extensive toolkit is invaluable for ensuring transparency and fairness.

The comprehensive nature of the toolkit means there can be a significant learning curve to effectively utilize its full capabilities, and integrating the toolkit into existing machine learning pipelines and workflows may require additional effort, especially in complex enterprise environments.

INTERACTIVE MODEL EXPLANATION AND ANALYSIS TOOLS

These tools focus on allowing users to dynamically explore and analyze how different inputs affect model predictions, often in a visual and user-friendly manner, and are invaluable for understanding complex model behaviors and testing hypothetical scenarios.

GOOGLE'S WHAT-IF TOOL

Google's What-If Tool Provides an easy-to-use interface for visually exploring model predictions under various conditions, making it possible to analyze model behavior across different inputs without writing code. *(This tool also can integrate into TensorBoard, allowing users to analyze TensorFlow models interactively.)*

MICROSOFT'S INTERPRETML DASHBOARD

Microsoft's InterpretML Dashboard InterpretML, this tool offers interactive visualizations to explore model predictions and explanations, particularly useful when working with Explainable Boosting Machines (EBMs).

FAIRLEARN DASHBOARD

An interactive visualization tool for assessing model fairness and understanding prediction disparities across groups, useful for exploring and mitigating bias in AI models.

OPERATIONAL EXPLAINABILITY AND TRANSPARENCY SUITES

These platforms exemplify the diverse approaches to integrating explainability and transparency into AI operations, each offering a blend of features to support the lifecycle of AI models from deployment to ongoing management.

Fiddler

A platform that provides tools to monitor, explain, and analyze AI models in production. Fiddler's Explainable AI features help users understand model predictions and biases, making it easier to diagnose and improve models.

Seldon Deploy

Part of the broader Seldon ecosystem, Seldon Deploy offers functionalities for model monitoring and explainability, integrating with Alibi Explain for in-depth explanations, offering users a robust set of tools for maintaining the transparency and accountability of AI models in production.

DataRobot

MLOps DataRobot's platform includes comprehensive MLOps capabilities that extend beyond deployment to include monitoring model health, understanding predictions, and ensuring models are fair and unbiased, with an emphasis on explainability and transparency.

IBM Watson OpenScale

Watson OpenScale provides detailed insights into AI model operations, including fairness, explainability, and performance metrics. Its focus on transparency and the ability to work with various machine learning frameworks and platforms makes it a valuable operational explainability and transparency suite.

TIBCO ModelOps

TIBCO's approach to model operations includes features for monitoring and explaining AI model decisions, supporting transparency and governance requirements in operational settings. This makes it a fit for the operational explainability and transparency suite category, though its emphasis may vary compared to more specialized tools.

WhyLabs

WhyLabs focuses on data and model observability, aiming to maintain model reliability and performance. While its primary strength lies in monitoring, the insights provided can contribute to operational transparency and help explain model behavior.

PROPOSAL FOR THE BEACON INITIATIVE

FOUNDATIONAL EXPLAINABILITY FRAMEWORK

UNIVERSAL TAGGING SYSTEM

BUILDING EXPLAINABILITY INTO EVERY STAGE OF DEVELOPMENT

With the Beacon Initiative, I am proposing both the collaborative development and adoption of universal tagging and a three part system that creates a **Foundational Explainability Framework** designed to shift the landscape and timing of transparency and explainability in AI model development. This framework aims to ensure **Minimum Viable Explainability ("MVE")** in AI models that will be built going forward by implementing a systematic Insight Tagging System (ITS) throughout the AI model development lifecycle. The Beacon Initiative and the ITS are designed to seamlessly integrate with and enhance the effectiveness of existing explainability tools and will facilitate **Explainability by Design**, embedding and enabling detailed logging, explainability, and compliance checks across all phases of development.

By making key components, decision points, factor importance, data types and data flow traceable to enable Minimum Viable Explainability, the Beacon Initiative seeks to establish a means to get to **Explainability by Design**, and a new paradigm in AI development, ensuring models are not only effective but also built to be fully interpretable, explainable, and compliant with regulatory standards.

UNIVERSAL TAGGING STANDARDS

The foundation of the Beacon Initiative is laid with the creation and adoption of a comprehensive set of universal tagging standards and protocols and standardized, universally accepted tags designed to label every critical component, decision point, and data pathway within an AI model. These standards and tags are meant to be developed through a collaborative, cross-industry effort, engaging international standards organizations organisations to ensure its relevance, applicability, and adoption across diverse AI development contexts. *(I do have a draft set of tags for you, of course. Stay tuned.)* This standardized tagging schema is essential for achieving a baseline of **Explainability by Design**, setting the stage for advanced insights into AI model functioning and decision-making processes.

These tags provide the groundwork for developing the Insight Tagging System; the sophisticated AI interpretation and explanation tools suite that can accurately convey the complexities of AI decision-making to a diverse audience. The implementation of these tools would represent a significant leap forward in building Explainable-by-Design AI that is more transparent and understandable and aligns with the deluge of additional regulations to come and the growing need for responsible AI development, ensuring that as AI systems grow in complexity, their explainability does not diminish but rather evolves alongside them.

UNIVERSAL TAGGING SYSTEM COMPONENTS

Deployable as "Explainability as a Service" (EaaS) software, these tools will ensure that any AI model built with the system, regardless of its complexity, maintains a robust level of interpretability, explainability, and compliance with privacy and regulatory standards.

Components of the Insight Tagging System:

Insight Tagging Engine (ITE)

The ITE is automated, AI-powered software designed to apply insight tags throughout the AI development process. It will identify and tag key aspects such as data sources, data types, decision points, algorithmic features, and compliance-related elements. The ITE will be hosted on a developer-controlled server to maintain privacy and data integrity and functioning independently of third-party cloud services. This tool also offers developers the ability to introduce additional, specific tags to meet unique industry or regulatory requirements, enhancing versatility and relevance.

The Insight Tagging System provides for Minimum Viable Explainability ("MVE") by systematically tagging key components of the AI development process, including data sources, data types, decision points, features, feature importance, fairness metrics, and regulatory aspects.

Granular Tag Reader (TR)

The outputs of the ITE require the GTR, a sophisticated tool capable of reading and interpreting the tagged components' metrics, tracing the model's development and decision-making process with granular detail. The GTR ensures that every tagged aspect contributes to a baseline level of explainability, sufficient for understanding by another AI system or advanced tool.

Explainability Reader Toolkit (ERT)

This third level in the tools suite would be an open platform to allow all manner of businesses and industries to develop explainability readers, designed to interpret and present the information from the Granular Tag Reader in a tiered and user-friendly manner. Provides foundational elements to ensure accurate and meaningful interpretation of tagged information, supporting the development of diverse tools catering to different industry needs.

STATEMENTS

Shoshana Rosenberg

It will slow development! What is the ROI?! AI is too complex for this!

I've got you. More to come. A great deal more.

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

THE WAY I SEE IT

DIVERSITY, EQUITY, AND AI ARE THE FUTURE

AI will accelerate the interconnectivity of the world and will be deeply ingrained in all kinds of decision-making processes.

DEI is essential for sustainable progress, innovation, and harmony.

Embracing DEI in the AI era is the only viable path to ensure that AI does not exacerbate or perpetuate existing biases and inequity.

INCLUSION IS THE KEY TO DIVERSITY AND EQUITY

Diversity will not be sustainable and equity not possible unless a full spectrum of the community is represented, integrated, accepted, respected and valued.

PRIVACY IS THE KEY TO INCLUSION

Inclusion cannot be measured, refined or fostered effectively without candid feedback and diversity data, both of which put individuals **at risk** without true privacy and anonymity, and both of which are too often collected without the preservation of privacy rights.

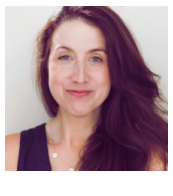
DATA IS THE KEY TO AI

To ensure ethical AI, the data it is trained on must be diverse, representative, and gathered properly (with authorization and/or consent) and used responsibly.

AI CAN BE AN UNPARALLELED KEY TO EQUITY

AI that is properly and thoughtfully designed with DEI principles can identify and help flag and rectify systemic disparities across any number of sectors and disciplines and processes, as well as helping to identifying gaps in the policies or tools that support them.

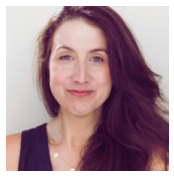
STATEMENTS



Shoshana Rosenberg

TRANSPARENCY, ACCOUNTABILITY, FAIRNESS, AND TRUST ARE KEY TO ALL THREE

A.I. DEVELOPMENTS IN STATE PRIVACY



Shoshana Rosenberg