ikigai
AI Council

Explainability for AI Predictions

Quarterly Report – Q1 2024

Explainability for AI Predictions

The first AI Council meeting centered around the concept of explainability: how should AI predictions be explained so decision makers can trust and use them meaningfully?

Why Explainability

To set the context, AI Council Member, MIT Professor, and Ikigai CEO Devavrat Shah shared why explainability is critical to AI use in business:

- Impact on the world (e.g., ethics):
  - Businesses should be accountable for using AI in a positive way for the world
  - Regulators need mechanisms to establish rules for lawful business use
  - Businesses and regulators need mechanisms to measure compliance
  - People require options for recourse if AI will affect their daily lives
- Impact on enterprise (e.g., business performance):
  - Businesses should understand AI-generated insights before using them to make critical decisions
  - Explainability leads to trust, and without trust, AI insights will not be used when they could improve business outcomes
  - Exceptions will be common; businesses need ways to address them

During the discussion, AI Council Member and GW Law scholar Aram Gavoor noted the importance of explainability, and ethical AI use more broadly, given the still-developing regulatory landscape. Currently, U.S. regulation on AI is spearheaded by the Executive Branch with "existing laws being applied in new ways."[1] States are becoming "labs of democracy" with government bodies opting for "regulation through enforcements." In this environment, AI companies have both an opportunity and an obligation to lead by example.

Key Insights from the Session

AI Council Member and Penn Computer Science Professor Michael Kearns kicked off the discussion with a provocative remark: "Explainability research is deeply broken...what's missing in the science of explainability is the behavioral component."

1. Executive Order on AI: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

When explaining a prediction, companies need to think about not only *what* they're trying to communicate, but also *how* and to *whom;* feature analysis may be useful to the data analyst, but it is less helpful when explaining an AI-driven decision to users less numerically oriented.

AI Council Member and UC Berkeley Computer Science Professor Michael I. Jordan concurred and added an illustrative example: "Let's say I go to a bank and am denied a loan because an AI system took in my covariates and told me no." Here, a potential solution to explain the result could be **case-based reasoning,** which is how humans often explain non-AI predictions as well. The bank should be able to explain their result in the following forms while maintaining data privacy:

- The application was similar to X other types of applications that were also rejected (This is "prediction similarity.")
- The application was similar to X other types of applications that were not rejected, but ultimately defaulted on their loans (This is "ground-truth similarity.")

Dr. Kearns noted that if the bank can't provide information like the above, it could be that the model wasn't trained on a diverse dataset and could be inaccurate. For example, an AI system for local college admission may be very good at predicting applicants' success if they come from in-state schools, but it lacks the historical data to correctly predict success for out-of-state applicants. Case-based reasoning would expose this deficiency.

AI Council Member and MIT Professor Munther Dahleh agreed that case-based reasoning is critical for engagement between the people deploying AI systems and the people using or receiving the AI predictions. He then posed a question to the group: can statistical thinking be used to both explain predictions and enable users to evaluate trustworthiness?

Statistical thinking includes **cross-validation, confidence intervals, and calibration**. Dr. Dahleh pointed out that "confidence intervals depend on sets of assumptions that need to be validated," and "cross validation can allow us to test our assumptions" by testing models on different data. Dr. Kearns added that once the model is cross-validated, confidence intervals can be calibrated by observing their performance over time.

If an AI system provides a rigorously defined confidence interval[2] that is updated based on real-world model performance or testing, humans can trust that the system tracks reality. However, from Dr. Kearns' experience in industry, "even when models output confidence intervals, they're not always [correctly defined]; [companies] don't have the semantics" to make them rigorous. Dr. Jordan agreed. Software companies could "start using [technically rigorous] language [and confidence intervals] more" if they want to demonstrate trustworthiness for technically-minded enterprise users.

Dr. Dahleh agreed that calibration can build trust and noted the importance of rigorous but easy-to-understand explanations. Humans "usually understand things if [they] have some form of a simplified model." Is there a way for AI companies to provide users with simple explanations for predictions, or the tools to uncover them?

Dr. Jordan posed **provenance** as a potential solution. Historically, provenance was utilized in database systems to capture data origination and lineage. Provenance is critical to understanding how AI systems intake and transform data, which is in turn critical to explainability and trust. For example, provenance could tell you that certain data are much older and therefore predictions using that data should be treated with more room for error, which is exactly what confidence intervals aim to show. Dr. Jordan pointed out that in the modern world, "data will very often be stale, and confidence intervals should automatically adjust, or typically grow larger, as the collection date of data recedes backward in time." If you want to build "trustable infrastructure," understanding the quality of input data and tracing the path of transformations is key, Dr. Jordan noted. Not all data is the same; humans recognize it, and provenance can ensure algorithmic systems recognize it as well.

Towards the end of the session, Dr. Kearns pointed out that explainability is far from perfect: "if you really want to know why [you received] a prediction, the science doesn't point us to good answers." However, the council members agreed that AI companies can still provide the "tools to uncover" potential rationale, empowering people to either build trust in the systems or find options for recourse. ■

See next page for Recommended Actions for AI companies to implement prediction explainability and build trust.

2. Dr. Jordan noted that the frequentist definition of confidence intervals works well for software companies: a confidence interval at X% represents an interval that contains the value being estimated X% of the time it is measured.

Recommended Actions

Three ways AI companies can implement prediction explainability and build trust:

1. Enable **case-based reasoning**

For input and output pairs (e.g., situation and prediction), **build the AI system so it can provide users the ability to see similar cases**. The AI system should provide explanation that can fit the following form:

- “Your input was similar to X other types of inputs, which had similar predictions {Y-hat}”
- “Your input was similar to X other types of inputs, which resulted in similar ground-truths {Y}”

2. **Cross-validate and calibrate** using rigorously-defined **confidence intervals**

For AI predictions, a user should be provided with a rigorously defined confidence interval. Confidence intervals should be automatically calibrated as new data is recorded.

In addition, users should be able to perturb the AI system (i.e., alter the input or historical data) to see how predictions change.

3. Track data **provenance**

For AI predictions, a user should be able to determine the set of data points that influenced the prediction. The set of data points includes both historical data as well as any human input.

AI Council Members

Devavrat Shah
Viterbi Professor of AI and Data Science and MIT; Ikigai CEO and co-founder

Massachusetts Institute of Technology

Michael I. Jordan
Professor in the Department of Electrical Engineering, Computer Science, and Statistics

Berkeley University of California

Michael Kearns
Professor and National Center Chair, Department of Computer and Information Science

Penn University of Pennsylvania

Munther Dahleh
William Coolidge Professor in Electrical Engineering and Computer Science, Founding Director of the MIT Institute for Data, Systems, and Society

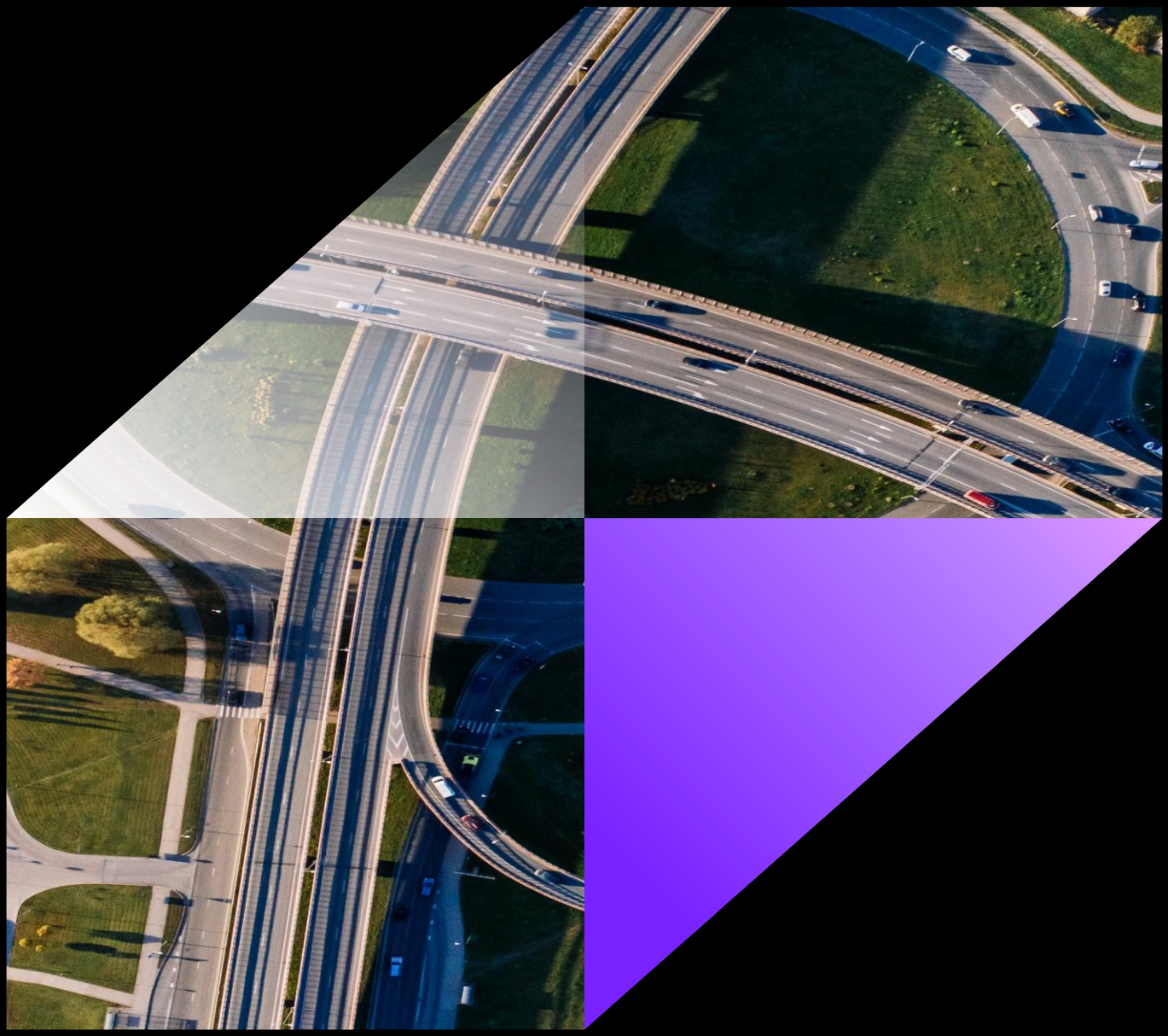
Massachusetts Institute of Technology

Aram Gavoor
Professorial Lecturer in Law; Professor, Trachtenberg School of Public Policy & Public Administration

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Kamal Ahluwalia
President of Ikigai; ex-President of Eightfold AI, ex-CRO Apttus

APTTUS



Visit our website to learn more.

https://www.ikigailabs.io/

Report by Parvathi Narayan; please reach out to parvathi@ikigailabs.io for questions.

ikigai