Trust and transparency for AI on the IBM Cloud

Build, run and manage your AI in your enterprise, with trust and transparency

Highlights

- Monitor for accuracy, performance and fairness of AI over time
- Detect and counteract harmful bias in your models
- Enable line-of-business knowledge workers to understand, audit and explain AI decision-making
- Accelerate and optimize how AI is built and used in organizations through trusted, explainable outcomes

Build, run, and manage your AI—with trust and transparency—to drive business value

New trust and transparency capabilities from IBM represent the cornerstone of how we're helping businesses build, run and manage AI models and applications across their organizations.

Roadblocks on the AI journey

Businesses today are increasingly certain that AI will be a driving force for the evolution of their industries in the immediate term. Many are already taking their first steps on this journey, building AI-powered chatbots to augment their call centers or automating back-office tasks by using AI to process documents or recognize images.

Yet for every successful AI project, there are many that fail to achieve their expected outcomes for the business. Even the most expert data science teams may only deploy a handful of models into production every year. This is partly because the mechanics of AI deployment can be complex, and there are still gaps in skills and tooling that can make it difficult for data science, IT operations and business teams to work together. But beyond the operational challenges, there are also much more profound issues of trust and transparency that businesses need to address before they can turn AI projects into true business advantage.

Enhancing trust

Knowledge workers must be able to trust AI and explain the results it produces before they use it confidently to augment decision-making across their business. If AI is a black box that simply takes in data and produces scores, there is no easy way for the business to judge whether those scores are a good guide to decision-making or not. Equally, the business will not be able to explain outcomes to customers, auditors or compliance teams.

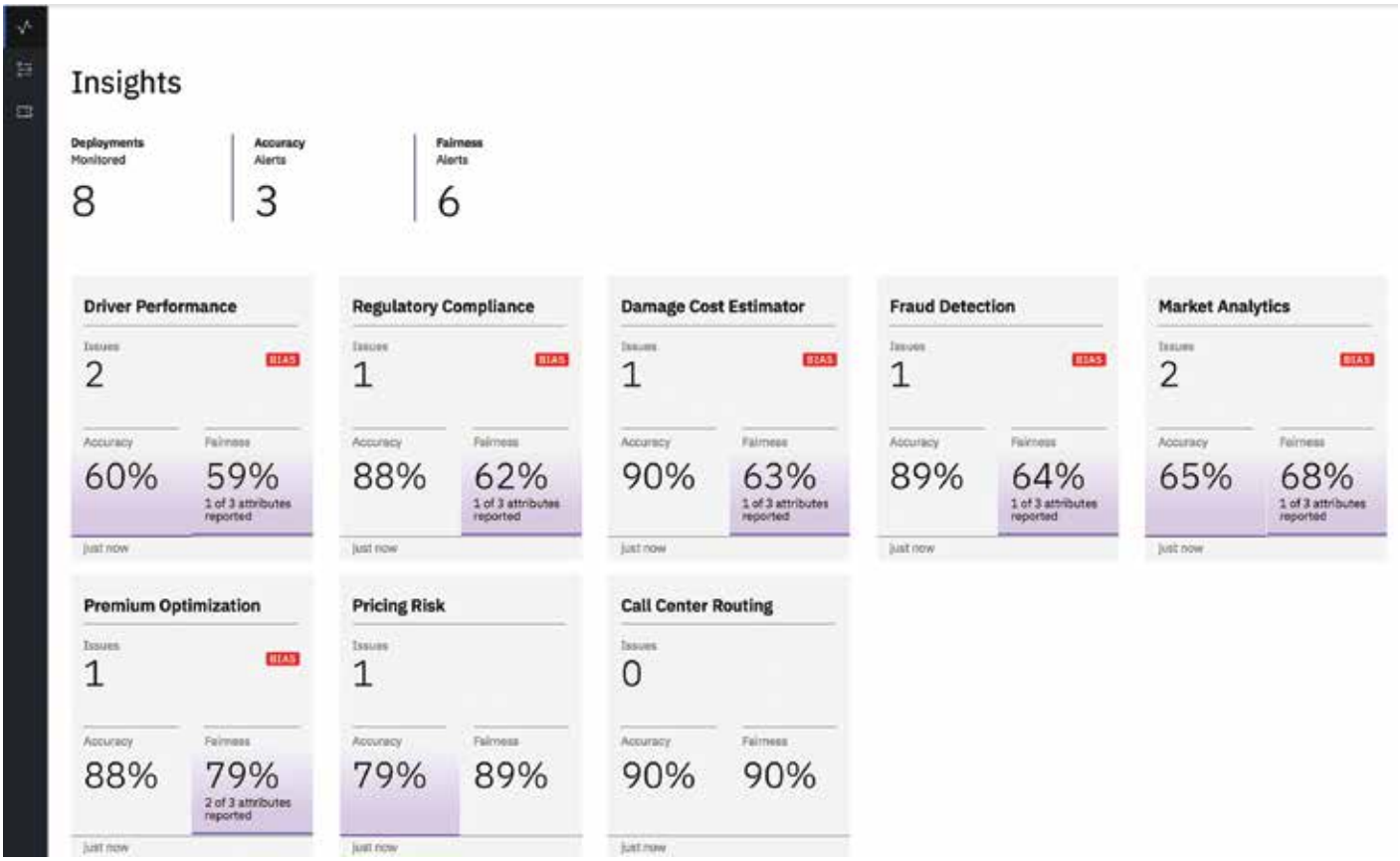

Figure A:
Track AI application health in real-time to assess accuracy, fairness, and performance

Today, many promising models never make it into production because businesses cannot afford to trust AI outcomes they do not fully understand. A business exposes itself to significant risk if it delegates responsibilities to an AI that does not fully align with its enterprise expectations and policies. For example, severe financial or reputational damage could result if a model unfairly discriminates against a particular group of customers because its training data did not represent a large enough sample of that population.

Ensuring fairness

Almost any AI model, no matter how carefully designed, is likely to exhibit a certain amount of bias. A model is only as good as the data on which it is trained, and because training datasets can never be 100 percent representative of real-world data, there is always a risk that a newly trained model may not perform well in production. Moreover, since most data domains are continuously evolving, model accuracy tends to drift over time.

The key is real-time visibility. If you can monitor the accuracy, performance and fairness of your AI models throughout their operational lifecycle, and provide analytics to help line-of-business users understand the reasoning behind the results, then you can overcome one of the most significant roadblocks on the AI journey.

Making decisions explainable

In many industries, regulatory scrutiny presents a significant barrier to AI adoption. Even if a company is satisfied that its models are fair and it can trust the results, regulators often demand a more rigorous approach.

For this reason, it is critical to ensure AI's input to any decision is fully explainable by keeping a complete track of the lineage of all the models, data, inputs and outputs of any AI-powered application. It should be possible to audit the lifecycle of every AI asset, from initial design and training and deployment, through to operation and retirement. For a given model, it should be possible to identify the team who built it and the datasets they used to train it, as well as the inputs it received in production and the outputs it produced.

Introducing IBM's new trust and transparency capabilities for AI on the IBM Cloud

IBM's new bias detection and mitigation, and explainability capabilities promote trust and transparency for AI by providing visibility into how it is used across an organization, augmenting the work of data scientists and application developers who are building, running and managing AI.

Not only are these capabilities optimized for models running in the IBM Cloud, but they also work with a wide variety of machine learning frameworks and AI build environments, including TensorFlow, SparkML, AWS SageMaker and Microsoft Azure Machine Learning.

How these capabilities work

Explain AI: This capability helps users across the business understand how AI reaches a decision to build trust in the technology. This capability automatically logs data that is processed by the model, enabling complete traceability of all decisions and predictions, and full data and model lineage. This logging data not only greatly improves auditability and compliance reporting, but it also supports powerful analytics. Users can query any business transaction and obtain an explanation of how the model arrived at its recommendation—in language that line-of-business users can easily understand.

Detect and mitigate bias at runtime: These capabilities help ensure a business's AI doesn't adopt or amplify any biases that would lead to unfair outcomes. Businesses are able to run a sophisticated set of diagnostic services to assess the accuracy and performance of the model. State-of-the-art anomaly and bias detection features, underpinned by innovations from IBM Research, help to identify harmful biases in both the data and the model. Bias checks can be performed both at build time and runtime to help ensure any issues are caught as early as possible. Then, to mitigate harmful bias, new data sets are recommended for use in model retraining.

Use case

IBM's new capabilities for AI make it possible for organizations to trust and explain their AI, helping to solve business problems and deliver value, while significantly mitigating risk. Here is an example of how the solution can help businesses across a range of industries:

Streamlining the loan approval processes

Getting a home loan approved is typically a 30-day process, and can take even longer during busy periods for lenders. The process depends on performing dozens of checks on each application to assess the risk of late payment or default.

AI models can use a wide range of data to help lenders expedite approval for low-risk candidates and identify high-risk applicants—helping to improve customer service, increase revenue and reduce the risk of losses. However, while these models often show excellent results in development, scaling them to support full loan approval processes in a reliable manner can be extremely challenging.

IBM's new bias detection and explainability capabilities enable companies to solve this problem by providing insights that help line-of-business users not only access the risk scores generated by the models, but also understand the logic behind the scores. If they know the models are accurate and are generating fair outcomes, they can make more confident decisions about whether to approve or reject a loan. Moreover, if a customer or regulator requests the reasoning behind a particular conclusion, the lender can easily explain how the model contributed to the decision.

AI you can trust

These new trust and transparency capabilities for AI give organizations full visibility into their models and provide for explainability of their AI, ensuring fair outcomes and granting business-process owners confidence in AI's ability to augment decision-making. With these capabilities, businesses can gain a deep understanding of the decisions their AI makes, providing them the confidence to accelerate and extend its use.

For more information

To learn more about these new capabilities for trust and transparency for AI on the IBM Cloud, contact your IBM representative or IBM Business Partner, or visit: ibm.com/watson/trust-transparency

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
September 2018

94019194USEN-00